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

described by PCM might provide a reasonable substitute for the non-specific interactions between molecules in a boiling liquid. This hypothesis proved to be valid, and in this paper we describe a model for normal boiling point prediction based on the use of PCM solvation energies. The model is easy to use, conceptually simple, and performs well for a wide variety of molecules, including 10 different elements (C, H, Br, Cl, F, N, O, P, S and Si).

2. Computational methods

2.1. Data sets

For this study, BP data for a training data set of 317 molecules and a test set of 74 molecules were obtained from the Reaxys database [21]. This database includes boiling points drawn from the primary chemical literature. A number of well-known errors are present in the data. First, some of the reported normal BP values are actually for boiling points at reduced pressure. Second, there is occasional confusion between units of °C and K. These two errors can usually be identified quite easily, since the former are generally large and the latter results in a fixed error of 273 K. Another potential source of error arises from BP values that were obtained under conditions where decomposition of the compound was evident; these values are generally associated with an appropriate note in the Reaxys database, and were avoided as far as possible. In order to minimise experimental errors, wherever practicable we chose BP's that had multiple but consistent independent entries in the database. This was not always possible, particularly for high boiling point compounds, and many of these are taken from single entries in the database. Nevertheless, we could detect no relationship between the number of independent values of the BP in the database and our final (observed – calculated) values. We found one example of an incorrect structural formula in the Reaxys database, as discussed below.

The training and test sets included hydrocarbons (aromatics, cyclic and acyclic aliphatics, alkenes and alkynes), alcohols, aldehydes, ketones, ethers, carboxylic acids, esters, amines, amides, halides (Cl, Br and F), spirocyclic compounds, heterocycles, sulfur compounds (thiols, thioethers, disulfides, sulfoxides and sulfones), phosphorus compounds (phosphines, phosphine oxides, phosponates and phosphates), silicon compounds (silanes and siloxanes) and 'exotics' (cyanides, isocyanides, isocyanates, thiocyanates, oximes, peroxides, carbonates, nitro compounds and nitrites). Many of the compounds contained more than one functional group. The molecules of the training set are given in Supplementary data (Table S1) and those of the test set are given in Table 1 and Scheme 1.

2.2. DFT calculations

Input geometries for DFT calculations were created using Hyperchem 7.51 [22] and subjected to initial energy minimization with the MM+ force field. In order to avoid symmetry-related saddle points, the Hyperchem models were subjected to small, arbitrary modifications of torsion angles to break any torsionally dependent symmetry before conversion into input coordinate files. Full geometry optimizations were then carried out within Gaussian03W or Gaussian09W [23], using the B3LYP functional and 6-31+G(d,p) basis set for all atoms. Output geometries were verified as true minima by the appropriate frequency calculations. PCM solvent corrections were then obtained for single point jobs using the optimized geometries and water as solvent. It should be noted that by default, Gaussian09W uses a different implementation of solvation calculations, so the SCRF = G03Defaults keyword is required to obtain comparable PCM solvent corrections when using Gaussian09W. Data were analysed using Microsoft Excel.

3. Results and discussion

3.1. Derivation of the general BP expression

For a boiling liquid, individual molecules will be tumbling rapidly and there will be no remaining structural order. In particular, any intermolecular hydrogen bonds that persist at lower temperatures should become transient only. Under these circumstances, the attractive forces that hold individual molecules within the bulk liquid should be averaged out, such that calculation of the implicit solvation energy should give an estimate of the strength of the intermolecular interactions. In this context, the implicit solvation energy ΔE_{solv} is defined as the difference between the gas phase energy and the energy of the molecule in solution, as calculated by PCM. Our preliminary calculations showed that, as expected, values of ΔE_{solv} for a given molecule in a range of different solvents are strongly correlated. We therefore chose water as the reference solvent, since solvation energies for water are generally larger than for other common solvents, and should therefore provide the widest range of calculated values.

If one considers a series of *n*-alkanes, then both the intermolecular forces in the pure liquid and the aqueous solvation energies are small, and the boiling points are determined primarily by the size of the molecules; indeed, it has long been known that the boiling points of *n*-alkanes can be predicted to within ~1 °C by Egloff's equation (1) [24];

$$\mathrm{BP} = 745.42\ \log(n + 4.4) - 689.4 \quad (1)$$

where *n* is the number of carbon atoms and in this and subsequent equations, the BP is given in °C. In this simplest case, no further analysis is needed. However, if one also considers branched alkanes, then it is clear that for a given molecular formula, the boiling point is lowered by branching. This is explained theoretically in terms of the reduced surface area, leading to decreased van der Waals interactions between molecules, and once again a relatively simple analysis is sufficient [1]. Building on these simple cases, a common approach is therefore to use a term derived from the molecular weight to calculate molecular bulk, and then add a correction for the degree of branching, such as the Wiener number [1] or globularity [15]. We decided to explore a different method, in which the molecular weight is not used, but rather Egloff's approach is generalised in order to calculate an effective molecular surface area using area coefficients for different types of atoms. In this way, the extent of branching can be built in by using different coefficients for carbon atoms with different levels of branching. For example, we used Eq. (2) below to fit the boiling points of a set of 60 acyclic alkanes;

$$\mathrm{BP} = 302.5\ \ln(SA) - 1511.5 \quad (2)$$

where *SA* is the effective molecular surface area, given by

$$SA = [(n^{CA} \times C^{A}) + (n^{CB} \times C^{B}) + (n^{CC} \times C^{C}) + n^{H} + a] \quad (3)$$

Here, n^{CA}, n^{CB} and n^{CC} are the numbers of (primary plus secondary), tertiary and quaternary carbons in the molecule, respectively, n^{H} is the number of hydrogens, and *a* is a constant. Regression analysis gave values of the area coefficients C^{A}, C^{B} and C^{C} of 17, 12 and 6, respectively, with $a = 70$, and an overall $R^2 = 0.995$. The average (obs. – calc.) was 5 °C, which is inferior to the accuracy of 1 °C achieved by Wiener for similar alkanes [1]; but the advantage of this method is that it can readily be generalised to virtually any type of molecule.

Table 1
Observed and calculated normal boiling points (°C) of the test set compounds.

#	Compound	Obs. BP	Calc. BP	#	Compound	Obs. BP	Calc. BP
B1	Propane	−43	−68	B38	$MeNHCOCH_2CHO$	178	184
B2	CF_2ClMe	−9.6	−16	B39	$(MeO)_3PO$	178	184
B3	Butane	−0.5	−9	B40	$CHBr_2CH_2Br$	188	193
B4	Me_2NH	7.4	2	B41	[a]	191	192
B5	MeN_3	22	19	B42	*o*-$C_6H_4(OH)Br$	195	196
B6	Cyclopentadiene	41	63	B43	*o*-$C_6H_4(OH)CHO$	196	183
B7	EtCHO	48	62	B44	$(CH_2OCFCl_2)_2$	200	204
B8	Oxetane	48	36	B45	[a]	203	162
B9	$(CHO)_2$	50	81	B46	[a]	218	222
B10	[a]	68	64	B47	$(n\text{-}Pr)_2CHCO_2H$	218	226
B11	MeN(F)CHO	77	75	B48	[a]	225	253
B12	$EtCO_2Me$	80	86	B49	$(CH_2=CH)_2SO_2$	236	254
B13	Cyclohexane	81	72	B50	*m*-$C_6H_4(OH)Br$	237	226
B14	$CF_3CFClCH_2Br$	82	115	B51	[a]	237	223
B15	$Me_2C(OMe)_2$	82	86	B52	*m*-$C_6H_4(OH)CHO$	240	238
B16	Et_3N	89	110	B53	[a]	241	227
B17	Et_2S	92	106	B54	$MeN(CH_2CH_2OH)_2$	243	242
B18	[a]	95	108	B55	[a]	270	266
B19	CCl_3CHO	98	114	B56	[a]	271	268
B20	CF_3SO_3Me	99	121	B57	[a]	275	295
B21	2-Bromofuran	102	109	B58	[a]	289	266
B22	MeC(O)SEt	116	125	B59	[a]	300	269
B23	*n*-$PrCO_2Et$	121	145	B60	[a]	308	333
B24	Thiophane	121	108	B61	1,3,5-$C_6H_3(NO_2)_3$	315	341
B25	$MeCS_2Et$	131	152	B62	[a]	315	304
B26	$H_2NCH_2CH_2PHMe$	132	139	B63	[a]	327	309
B27	$Me_3SiC{\equiv}CSiMe_3$	134	161	B64	Anthracene	328	345
B28	$Me(CH_2)_4C{\equiv}CMe$	138	147	B65	Fluorenone	342	344
B29	Me_3SiNCS	143	131	B66	[a]	344	329
B30	$MeCS_2Me$	143	125	B67	[a]	352	301
B31	Et_3SiCl	145	148	B68	Ph_3N	364	358
B32	Piperazine	147	150	B69	[a]	371	374
B33	[a]	151	144	B70	[a]	408	379
B34	[a]	154	155	B71	[a]	412	416
B35	*n*-$PrCO_2H$	163	164	B72	[a]	463	426
B36	[a]	175	183	B73	$Ph_3SiOSi(Me)Ph_2$	466	465
B37	$(Me_2N)_2CO$	177	216	B74	4,5-Benzopyrene	492	479

[a] See Scheme 1 for structures.

We now consider the use of solvation energies, ΔE_{solv}, for estimation of the boiling points of more polar molecules. Our initial analysis indicated that the best fit would be obtained using $\Delta E_{solv}/SA$, rather than ΔE_{solv} itself. This means that, for example, methanol receives a bigger solvent correction than *n*-butanol, even though these two molecules have similar values of ΔE_{solv}. The general BP expression then becomes

$$BP = b\ \ln(SA) + c\left(\frac{\Delta E_{solv}}{SA}\right) - d \quad (4)$$

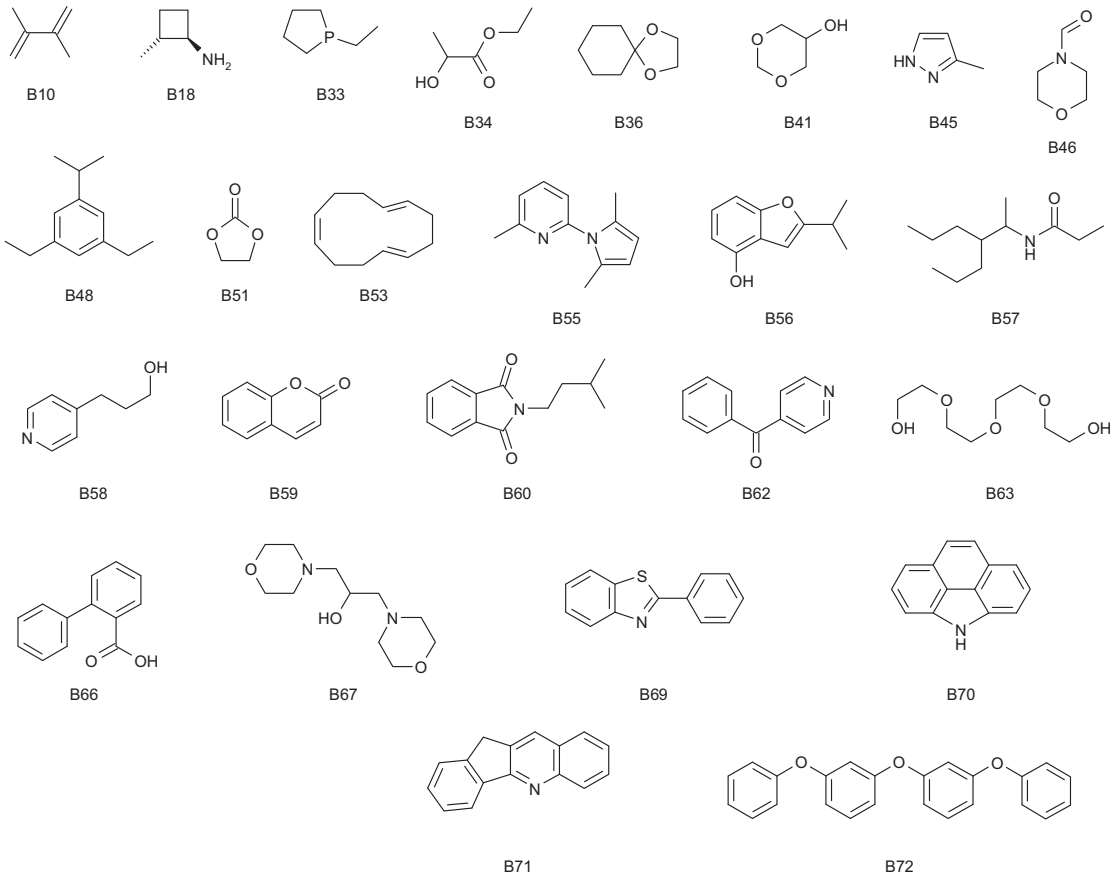
in which *b*, *c* and *d* are constants and ΔE_{solv} is given in atomic units. This expression proved satisfactory for most, but not all, classes of molecules; the exceptions are discussed below. For the calculation of *SA*, we considered a range of possible component atom types, using a process of trial and error to decide on the most suitable combination. These atom types and associated area coefficients are given in Table 2. In all, we used five types of carbon atom, including the three mentioned above, plus aromatic and spiro carbons. For most other elements, a single atom type was sufficient. However, for N, O and S, two atom types were required; one general type, and one with a much larger area coefficient. These special types were used for amide nitrogen (atom type N^A), a specific type of oxygen, defined below (atom type O^C), and S = O sulfur (atom type S^B). Thus cyclic sulfones require both special O and S types. The overall effect of these large area coefficients is to increase the calculated BP. Therefore, for highly polar molecules such as amides and sulfoxides, the observed BP's are higher than those calculated using the standard atomic area coefficients, even though their solvent corrections are typically large. It is worth noting that the area coefficient for silicon is smaller than might have been expected; this is probably a reflection of the fact that the silicon atoms are quaternary in many of the available molecules with known BP's, and therefore the organosilicon molecules in our dataset tend to be quite globular.

In terms of its conceptual meaning, *SA* is related to the total molecular surface area, although the use of larger coefficients for certain types of N, O and S atoms, as discussed above, means that *SA* is composite parameter that describes more than the surface area alone. There is no correlation between *SA* and the polar surface area (PSA), as described for example in the work of Bytheway et al. [6], which is derived from the molecular electron density but is determined primarily by the heteroatoms.

For most molecule types, use of Eq. (4) with *SA*, calculated using the general atom types given in Table 2, plus ΔE_{solv} as obtained from DFT calculations, was sufficient for calculation of the BP. However, a number of special cases emerged from our analysis, as now described.

3.2. Molecules containing fluorine

Fluorine is well known for both its distinctive electronic properties and exceptional effects on boiling points. For example, the BP's of known fluorobenzenes are all in the range of 82–95 °C, regardless of the number of fluorines. To allow for the anomalous behaviour of fluorine, we found it necessary to include different coefficients for the calculation of *SA* for the first and second terms in Eq. (4). Thus, the fluorine coefficient is taken as 1 for the first term of the equa-



Scheme 1.

tion, and 12 for the second; the effective surface areas calculated with the two parameters are distinguished as *SA* and *SA**, respectively. For non-fluorinated molecules, *SA* = *SA**. This was found to give satisfactory results for most fluorine-containing molecules, except when in combination with other exceptional atom types (see below). Allowing all of the atom types to have two different coefficients in this way gave only a marginal improvement to the overall R^2 value, which did not justify the increased number of variables.

3.3. Flat molecules

During the course of our analysis, we discovered that large, flat molecules such as polycyclic aromatic hydrocarbons have anomalously high BP's compared to the calculated values provided by Eq. (4). Therefore, a correction was introduced for these cases, Eq. (5);

$$\mathrm{BP} = b\ \ln(SA) + c\left(\frac{\Delta E_{\mathrm{solv}}}{SA^*}\right) + C^{\mathrm{FLAT}} - d \tag{5}$$

where

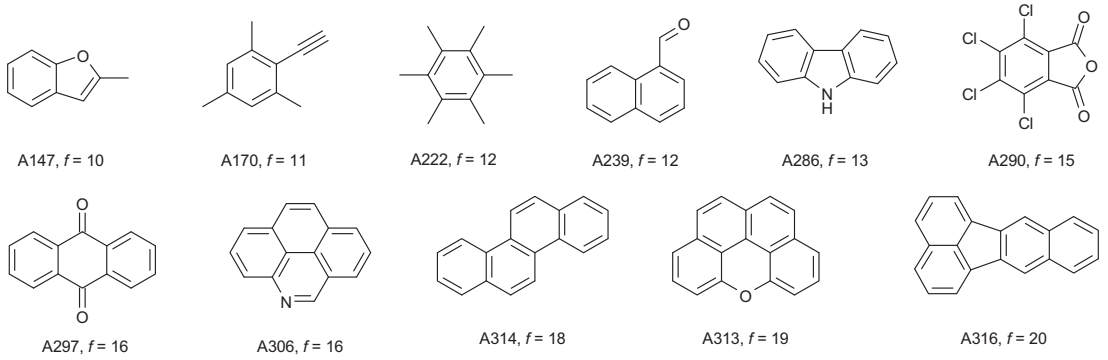
$$C^{\mathrm{FLAT}} = (f \times g) - h \tag{6}$$

Such that *f* is the number of co-planar heavy atoms (with a minimum of six), and *g* and *h* are constants. Here, a heavy atom is taken as any type of atom, except H or F. The planar molecules all contain at least one 6-membered aromatic ring, and can also include any number of the following groups; $-CH_3$, –Cl, –Br, C=O (where the C is part of the ring system), –OH, –SH, $-NH_2$, –C≡CH, –C≡CMe, or –N=C=S. Molecules with substituents such as –CH=O or $-NO_2$,

Table 2
Atom types and area coefficients used in the final model.[a]

Coefficient	Descriptor	Value	Coefficient	Descriptor	Value
C^A	1°, 2° carbon	16	O^A	General oxygen[b]	9
C^B	3° carbon	12	O^C	Cyclic X = O oxygen[c]	43
C^C	4° carbon	0.5	S^A	General sulfur[b]	37
C^F	Aromatic carbon	19	S^B	S = O type sulfur	91
C^G	Spiro carbon	13	P	Phosphorus	31
H	Hydrogen	1[d]	F	Fluorine	1 (12)[c]
N^A	Amide nitrogen	41	Cl	Chlorine	26
N^B	General nitrogen[b]	15	Br	Bromine	45
			Si	Silicon	5

[a] The values of the constants used in the final model, using atomic units for the solvation energy ΔE_{solv}, were as follows; $b = 212.1$; $c = -8.219 \times 10^5$; $d = 927.6$; $g = 10.79$; $h = 104.4$; $\Delta^{\mathrm{EN}} = -19$. Note that all the calculated PCM solvation energies were negative numbers.
[b] A general type atom is considered to be any atom other than the specific types listed.
[c] See text for explanation.
[d] Taken as fixed.



Scheme 2.

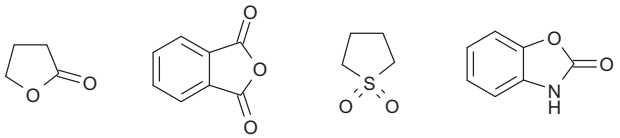
whose rotation could result in non-planarity, were considered to be flat provided that (1) no more than one of these substituents were present, and (2) the optimised geometry was planar. Some examples of flat molecules used in the training set are given in Scheme 2; in the test set, molecules B42, B43, B50, B52, B59, B64, B65, B69, B70, B71 and B74 were all classed as flat. Physically, it seems reasonable that a planar molecule might have higher BP than a comparable non-planar molecule, since the planar molecules would tend to stack in the bulk liquid. This could affect both the extent of van der Waals contact and the entropy-related cost of boiling. Nevertheless, we remain cautious in seeking a detailed physical interpretation for our model, bearing in mind that the corrections for the smaller flat molecules have negative sign.

3.4. Anomalous cases

Although Eq. (5) is generally applicable, we found a few types of molecule that require special treatment. The first type is exemplified by propylene carbonate, for which Eq. (5) gave a BP of 170 °C, far lower than the observed value of 242 °C. To correct this anomaly, we introduced the atom type O^C; this is defined as any carbonyl O for which the carbonyl C is part of a 4- or 5-membered ring, and one or more of the adjacent atoms in the ring is a heteroatom; or any sulfone O for which the sulfone S is part of a 4- or 5-membered ring. This atom type is used in conjunction with another parameter, w. For cyclic carbonyls, w is taken as the number of heteroatoms that neighbour the carbonyl group(s), divided by the number of carbonyl groups. For cyclic sulfones, $w = 1$. Examples of such molecules, taken from the training set, are given in Scheme 3. The multiplier for the coefficient O^C is then $(w \times n^{OC})$, where n^{OC} is the number of O^C type oxygens in the molecule.

A second class of anomalous molecules are amides, sulfoxides and sulfones, when the amide N or S=O group(s) are attached to one or more electronegative groups, such as O, N, F or CF_3. Predicted BP's for these compounds were generally too high. Therefore, an empirical correction C^{SO} was used, according to Eq. (7);

$$C^{SO} = (n^{EN} \times \Delta^{EN}) \tag{7}$$



Scheme 3.

where n^{EN} is the number of electronegative atoms; e.g. for –OR, $–NR_2$ or –X, $n^{EN} = 1$ per group, for $–CF_3$, $n^{EN} = 3$ per group, up to a maximum of 3; and the variable $\Delta^{EN} = -19$.

Considering these empirical corrections, both types of anomaly seem to be associated with X=O type double bonds (X=C or S) in specific environments, resulting in distinctive electronic effects. Further research into these effects is hampered by the fact that the number of well-characterised examples of each type of molecule is rather limited. It is perhaps not surprising that these molecules are difficult cases, since molecules such as fluorinated amides and sulfoxides provide an exacting test for the quantum calculations. Omitting the molecules discussed in this section (and their associated parameters w and n^{EN}, plus variables O^C and Δ^{EN}) made no significant difference to the overall R^2 value.

The most general form of our boiling point expression is then given by Eq. (8);

$$BP = b\ \ln(SA) + c\left(\frac{\Delta E_{solv}}{SA^*}\right) + C^{FLAT} + C^{SO} - d \tag{8}$$

3.5. Data analysis and scope of the model

Having identified the required variables, we carried out a regression analysis of the training set, in order to optimise their values. For this purpose, we partitioned the observed BP (BP^{OBS}) into three contributions, according to Eqs. (9)–(11);

$$BP^{SA} = BP^{OBS} - \left(c\left(\frac{\Delta E_{solv}}{SA^*}\right) + C^{FLAT} + C^{SO}\right) \tag{9}$$

$$BP^{SOLV} = BP^{OBS} - (b\ \ln(SA) + C^{FLAT} + C^{SO} - d) \tag{10}$$

$$BP^{FLAT} = BP^{OBS} - \left(b\ \ln(SA) + c\left(\frac{\Delta E_{solv}}{SA^*}\right) + C^{SO} - d\right) \tag{11}$$

The graph of BP^{SA} *versus* ln (SA) is shown in Fig. 1; the straight line has $R^2 = 0.984$ and its slope and intercept provide the constants b and d, respectively. Similarly, Fig. 2 shows a plot of BP^{SOLV} *versus* $(\Delta E_{solv}/SA^*)$; the straight line was constrained to pass through the origin and has $R^2 = 0.900$. The slope of this graph gives the constant c. It should be noted that most of the data points in Fig. 2 are clustered at the lower end of the range, i.e. there are few molecules with large values of $\Delta E_{solv}/SA^*$ in our data set. This is undesirable but also unavoidable, since the number of such molecules with experimentally determined boiling points is very small.

Finally, a plot of BP^{FLAT} *versus* the number of co-planar heavy atoms f is shown in Fig. 3. This gives a straight line with $R^2 = 0.929$, and the slope and intercept provide the constants g and h, respectively. These three graphs were used to fit all of the variables by regression analysis. The values of the constants are given in the footnotes to Table 2. In all, our model uses a total of 23 independent variables, fitted with the training set of 317 molecules. The overall fit of observed *versus* calculated BP's for both the training and test

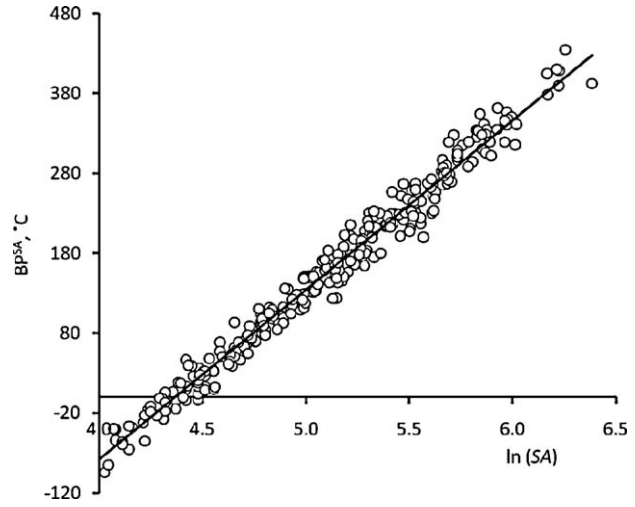


Fig. 1. Plot of BP^{SA} (i.e. observed boiling point, corrected for contributions from solvation and flatness) *versus* ln (*SA*), for which $R^2 = 0.984$.

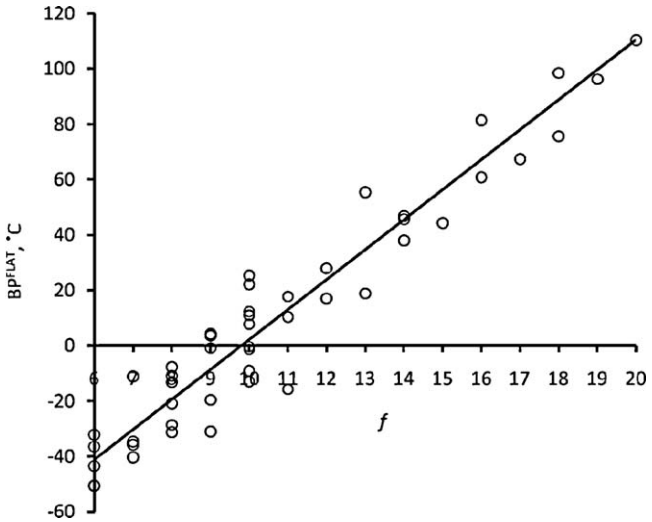


Fig. 3. Plot of BP^{FLAT} (i.e. observed boiling point, corrected for contributions from the surface area *SA* and solvation) *versus* the number of co-planar heavy atoms *f*, for which $R^2 = 0.929$.

data sets is shown in Fig. 4, and the RMS error for the training data set was 16.1 °C.

It is instructive to consider how the various terms in Eq. (8) contribute to the BP's of individual molecules. The value ranges for the first three terms of Eq. (8), namely the surface area, solvation and flatness terms, are 654, 295 and 149 °C, respectively.

For acyclic alkanes, the contribution from solvation is small (<10 °C) and the BP's are essentially determined by the first term of Eq. (8). However, other hydrocarbons have larger solvation energies, for example allene (A3), but-2-yne (A20) and benzene (A53) all have solvation energy corrections of ~40 °C. At the other extreme, polar groups such as alcohols are associated with large solvent corrections; the largest value of 299 °C is for glycerol (A237). The largest flat molecule in the training set is benzo[k]fluoranthene, $C_{20}H_{12}$ (A316), for which $C^{FLAT} = 112$ °C.

In terms of overall scope, the training set included molecules with a range of BP's from −50 (ketene, A1) to 495 °C (perylene, A317). The model performed poorly for very small molecules such as methane and HCN, perhaps because the properties of such compounds are determined by the precise balance of effects that are more averaged out for larger molecules. The upper limit of ~500 °C is a consequence of both limited experimental data (few compounds are stable at such high temperatures) and the computational requirements of DFT calculations on large molecules.

Having optimised our model, we applied it to a test set of 74 molecules, including the full range of types used in the training set (Table 1 and Scheme 1). A plot of BP^{OBS} *versus* BP^{CALC} for this data set gave a straight line with $R^2 = 0.979$, and the RMS error was 18.0 °C. Hence, the test set gave similar results to the training set, and our model seems to be generally applicable for the prediction of BP's of a wide range of organic compounds.

3.6. Internal hydrogen bonding

Our model assumes that there are no explicit intermolecular interactions between molecules in the liquid phase at the boiling point. However, the possibility of specific intramolecular hydrogen bonds must also be considered. This could serve to reduce the degree of intermolecular attraction, and so lower the BP. With this in mind, we revisited the case of 2- and 3-hydroxybenzaldehyde (compounds B43 and B52, respectively in our test set). These two

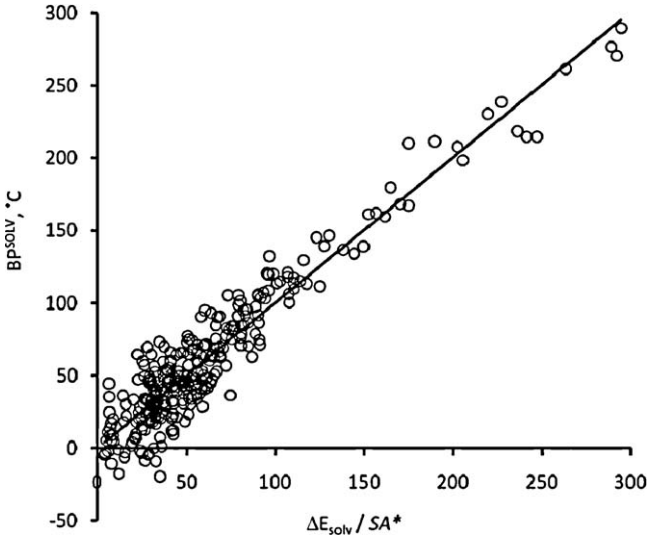


Fig. 2. Plot of BP^{SOLV} (i.e. observed boiling point, corrected for contributions from the surface area *SA* and flatness) *versus* ($\Delta E_{solv}/SA^*$), for which $R^2 = 0.900$.

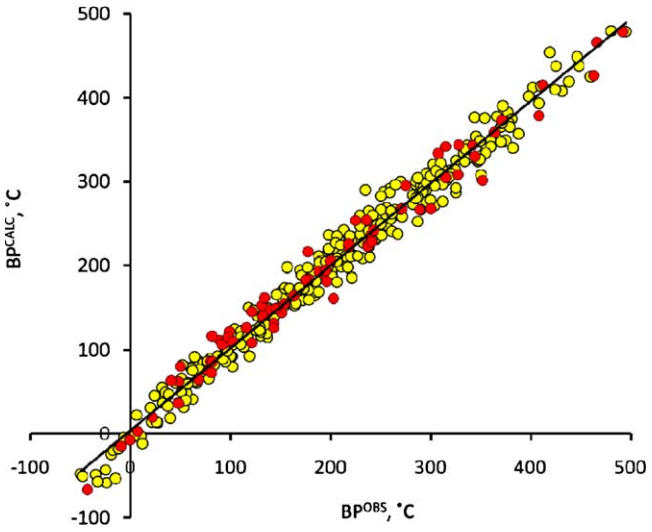


Fig. 4. Plot of observed *versus* calculated BP's for the training set (yellow points) and test set (red points). $R^2 = 0.980$ for the training set and 0.979 for the test set.

compounds boil at 196 and 240 °C, respectively, and this large difference was attributed by Stanton [14] to the possibility of formation of an internal hydrogen bond in 2-hydroxybenzaldehyde. Our calculations support this view; the solvation energies for the 2- and 3-isomers were −32.1 and −56.5 kJ mol^{-1}, respectively, leading to calculated BP's of 183 and 238 °C, respectively, in good agreement with experiment. A further example is provided by 2-, 3- and 4-bromophenol (B42, B50 and A190, respectively), for which the experimental BP's are 195, 237 and 238 °C, respectively. The 2-isomer shows an internal hydrogen bond, giving a solvation energy of −26.6 kJ mol^{-1} and a calculated BP of 196°C, whilst the 3- and 4- isomers have solvation energies of −43.1 and −44.1 kJ mol^{-1}, respectively, leading to calculated BP's of 226 and 229 °C, respectively. It appears that more flexible molecules can also display internal hydrogen bonds. For example, ethyl lactate (B34) can adopt a conformation with an internal hydrogen bond between the –OH and C=O groups, which is calculated to be 22.4 kJ mol^{-1} more stable than the more linear form. This internal hydrogen bond again reduces the magnitude of the solvation energy, from −46.7 to −30.4 kJ mol^{-1}; resulting in a reduction in the calculated BP from 202 to 155 °C, compared to the experimental value of 154 °C. Another interesting example is provided by *N*-methyl pyruvamide, MeNHC(O)C(O)Me, B38, BP 178 °C, which was incorrectly given in the Reaxys database as the isomeric *N*-methyl-3-oxopropanamide, MeNHC(O)CH_2CHO. We verified that the former is the correct isomer from the original paper [25]. The hydrogen bond between the amide NH and ketonic C=O reduces the magnitude of the solvation energy from −55.1 to −36.0 kJ mol^{-1}, thereby lowering the calculated BP from 232 to 184 °C, in agreement with the experimental value. Hence, internal hydrogen bonding can cause a reduction in BP of up to 50 °C.

4. Conclusion

We have developed a model for the prediction of boiling points based on solvation energies, using DFT calculations with PCM solvent corrections. The model is easy to use and gives reasonably accurate results for a very wide range of organic molecules. It also provides a simple conceptual framework, in that the three main terms used in the calculations (effective surface area, strength of intermolecular interactions, and planarity) are readily understood. The experimentally observed reduction of BP by intramolecular hydrogen bonding is correctly predicted by our model. To the best of our knowledge, this is the first time that the flatness of large planar molecules has been associated with elevation of their boiling points. The use of our method as a cross-check in cases where the literature data appears doubtful, and indeed to uncover errors in chemical databases, represents a useful application. It is possible that further improvements in the methodology for calculation of solvation energies will allow us to further refine our model, and this is the focus of our ongoing research in this area.

Acknowledgement

We thank Mr Thomas Durrant (University of York) for help in inputting the data.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2011.06.010.

References

[1] H. Wiener, Structural determination of paraffin boiling points , J. Am. Chem. Soc. 69 (1947) 17–20.
[2] R.D. Cramer, BC(DEF) parameters. 1. The intrinsic dimensionality of intermolecular interactions in the liquid state , J. Am. Chem. Soc. 102 (1980) 1837–1849.
[3] R.D. Cramer, BC(DEF) parameters. 2. An empirical structure-based scheme for the prediction of some physical properties , J. Am. Chem. Soc. 102 (1980) 1849–1859.
[4] P.L.A. Popelier, U.A. Chaudry, P.J. Smith, Quantum topological molecular similarity. Part 5. Further development with an application to the toxicity of polychlorinated dibenzo-p-dioxins (PCDDs) , J. Chem. Soc., Perkin Trans. 2 (2002) 1231–1237.
[5] P.L.A. Popelier, P.J. Smith, QSAR models based on quantum topological molecular similarity , Eur. J. Med. Chem. 4 (1) (2006) 862–873.
[6] I. Bytheway, M.G. Darley, P.L.A. Popelier, The calculation of polar surface area from first principles: an application of quantum chemical topology to drug design , ChemMedChem 3 (2008) 445–453.
[7] R. Gilson, M.C. Durrant, Estimation of the pKa values of water ligands in transition metal complexes using density functional theory with polarized continuum model solvent corrections , Dalton Trans. (2009) 10223–10230.
[8] A.R. Katritzky, L. Mu, V.S. Lobanov, M. Karelson, Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organics and a test set of 9 simple inorganics , J. Phys. Chem. 100 (1996) 10400–10407.
[9] L.M. Egolf, M.D. Wessel, P.C. Jurs, Prediction of boiling points and critical temperatures of industrially important organic compounds from molecular structure , J. Chem. Inf. Comput. Sci. 34 (1994) 947–956.
[10] A.R. Katritzky, V.S. Lobanov, M. Karelson, Normal boiling points for organic compounds: correlation and prediction by a quantitative structure–property relationship , J. Chem. Inf. Comput. Sci. 38 (1998) 28–41.
[11] M.D. Wessel, P.C. Jurs, Prediction of normal boiling points for a diverse set of industrially important organic compounds from molecular structure , J. Chem. Inf. Comput. Sci. 35 (1995) 841–850.
[12] E.S. Goll, P.C. Jurs, Prediction of the normal boiling points of organic compounds from molecular structures with a computational neural network model , J. Chem. Inf. Comput. Sci. 39 (1999) 974–983.
[13] D. Sola, A. Ferri, M. Banchero, L. Manna, S. Sicardi, QSPR Prediction of N-boiling point and critical properties of organic compounds and comparison with a group-contribution method , Fluid Phase Equilib. 263 (2008) 33–42.
[14] D.T. Stanton, Development of a quantitative structure–property relationship model for estimating normal boiling points of small multifunctional organic molecules , J. Chem. Inf. Comput. Sci. 40 (2000) 81–90.
[15] A.J. Chalk, B. Beck, T. Clark, A quantum mechanical/neural net model for boiling points with error estimation , J. Chem. Inf. Comput. Sci. 41 (2001) 457–462.
[16] M.P. González, A.A. Toropov, P.R. Duchowicz, E.A. Castro, QSPR calculation of normal boiling points of organic molecules based on the use of correlation weighting of atomic orbitals with extended connectivity of zero- and first-order graphs of atomic orbitals , Molecules 9 (2004) 1019–1033.
[17] P.R. Duchowicz, E.A. Castro, F.M. Fernández, M.P. Gonzalez, A new search algorithm for QSPR/QSAR theories: normal boiling points of some organic molecules , Chem. Phys. Lett. 412 (2005) 376–380.
[18] S.H. Kumar, A comparative QSPR study of alkanes with the help of computational chemistry , Bull. Korean Chem. Soc. 29 (2008) 67–76.
[19] J.-T. Chen, H.-L. Liu, F.-Y. Wang, H.-X. Yu, D.-L. Li, QSPR study on the boiling points of some oxygen- and sulfur-containing organic compounds , Chin. J. Struct. Chem. 28 (2009) 1561–1568.
[20] K. Cossi, G. Scalmani, N. Rega, V. Barone, New developments in the polarizable continuum model for quantum mechanical and classical calculations on molecules in solution , J. Chem. Phys. 117 (2002) 43–54.
[21] http://www.info.reaxys.com, Data represented here used with kind permission of the copyright owner Elsevier Properties SA, Copyright 2010©. Elsevier Properties SA, All rights reserved. Authorized use only. Reaxys® is a registered trademark owned and protected by Elsevier Properties SA and used under license.
[22] Hyperchem release 7.51 for Windows, Hypercube Inc., 2002.
[23] (a) M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J.A. Montgomery Jr., T. Vreven, K.N. Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J.E. Knox, H.P. Hratchian, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Baboul, S. Clifford, J. Cioslowski, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez, J.A. Pople, Gaussian 03W (Revision C.02) , Gaussian, Inc., Wallingford CT, 2004;
(b) M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H.P. Hratchian, A.F. Izmaylov, J. Bloino, G. Zheng, J.L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J.A. Montgomery Jr., J.E. Peralta, F. Ogliaro, M. Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J.M. Millam, M. Klene, J.E. Knox, J.B. Cross, V. Bakken, C. Adamo,

J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, R.L. Martin, K. Morokuma, V.G. Zakrzewski, G.A. Voth, P. Salvador, J.J. Dannenberg, S. Dapprich, A.D. Daniels, Ö. Farkas, J.B. Foresman, J.V. Ortiz, J. Cioslowski, D.J. Fox, Gaussian 09W (Revision A.1) , Gaussian, Inc., Wallingford, CT, 2009.

[24] G. Egloff, J. Sherman, R.B. Dull, Boiling point relationships among aliphatic hydrocarbons , J. Phys. Chem. 44 (1940) 730–745.

[25] P.M. Pojer, I.D. Rae, Reactions of methylamine and aniline with methyl pyruvate, Aust. J. Chem. 23 (1970) 413–418.