Chemical Physics Letters 610–611 (2014) 163–166

Contents lists available at ScienceDirect

Chemical Physics Letters

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

Additive models for the molecular polarizability and volume

Shamus A. Blair, Ajit J. Thakkar*

Department of Chemistry, University of New Brunswick, Fredericton, New Brunswick E3B 5A3, Canada

ARTICLE INFO

Article history:
Received 26 June 2014
In final form 11 July 2014
Available online 18 July 2014

ABSTRACT

Additive models for molecular polarizabilities and volumes are created by fitting to data for 298 molecules. Tests on data for the 1641 organic molecules in the TABS database show that the best models have median absolute errors of 2.3% for the polarizability and 1.3% for the volume. Bonding lowers the volume in all 1641 molecules.

1. Introduction

Every chemistry textbook highlights the idea that molecular properties are determined principally by the properties of their constituent atoms, bonds, functional groups, and fragments. The foundations for this seminal idea were laid in the mid-19th century by Kopp [1]. He established that the molar volumes of organic liquids at their boiling points were close to additive functions of the molar volumes of their constituent elements. Soon thereafter, Gladstone and Dale [2] gathered experimental evidence to support nearly additive group contributions to molar refraction. Molar volumes are proportional to molecular volumes which in turn are closely associated with molecular polarizabilities as are molar refractions [3]. Thus, the additivity of molecular polarizabilities was already recognized well before the end of the 19th century.

An additive model expresses a molecular property as a weighted sum of transferable contributions from its constituent parts. Transferability is necessary to make predictions possible. Transferability and predictive power, even if limited, distinguish additive models from post hoc decomposition schemes. Since many molecules of biochemical interest remain outside the reach of contemporary experimental and theoretical techniques, additive models for molecular polarizabilities remain of interest to this day [4–12].

A hierarchical classification [12] helps bring order to the plethora of additive models in the literature. The lowest step in the hierarchy is a Level 1 (L1) model, also called a free-atom additive model (FAM), which can be expressed as:

$$\mathcal{P} \approx \mathcal{P}^{\mathrm{L1}} = \mathcal{P}^{\mathrm{FAM}} = \sum_i n_i \mathcal{P}_i^0 \tag{1}$$

in which $\mathcal{P}$ is a molecular property, n_i is the number of atoms of element i in the molecule, and $\mathcal{P}_i^0$ is the value of the pertinent property of a *free* atom of element i. Such a model is very approximate but has an important conceptual role. For example, discussions of the minimum polarizability principle may use an L1 model [13]. Since the properties of the free atoms are not known exactly, different L1 models can be defined by varying the source of the atomic property. For example, one may use experimental values in the gas- or solid-phase, or values computed using a particular quantum chemical method and basis set combination.

A significant improvement over an L1 model is obtained by moving up a rung in the hierarchy [12] to a Level 2 (L2) model, sometimes referred to as a dressed-atom additive model (DAM), which can be written as:

$$\mathcal{P} \approx \mathcal{P}^{\mathrm{L2}} = \mathcal{P}^{\mathrm{DAM}} = \sum_i n_i \mathcal{P}_i^{\mathrm{d}} \tag{2}$$

in which n_i is the number of atoms of element i in the molecule, and $\mathcal{P}_i^{\mathrm{d}}$ is the 'dressed' or effective property of an atom of element i in a molecule. A dressed atomic property incorporates the influence of a generic molecular environment. For example, Bosque and Sales [7] presented such a model, albeit with an additional non-physical constant term on the right hand side of Eq. (2).

The next step up is a Level 3 (L3) model which can be written as

$$\mathcal{P} \approx \mathcal{P}^{\mathrm{L3}} = \sum_i n_i \mathcal{P}_i^{\mathrm{t}} \tag{3}$$

in which the sum is over atom types, n_i is the number of atoms of type i in the molecule, and $\mathcal{P}_i^{\mathrm{t}}$ is the effective property of an atom of type i in a molecule. An L3 model reduces to an L2 model if the atomic number is the sole criterion used to distinguish atom types. One common way of creating atom types is to start with atoms of the elements and then subdivide them by hybridization scheme.

* Corresponding author.
E-mail address: ajit@unb.ca (A.J. Thakkar).

http://dx.doi.org/10.1016/j.cplett.2014.07.027
0009-2614/

For example, hydrogen and halogen atoms would form their own types but each carbon atom would be categorized as belonging to the sp^3, sp^2, or sp type. An early example of an additive model in which the separation into atom types is discussed explicitly in terms of hybridization is the work of Kang and John [14]. The use of group polarizabilities by Vogel [15] is a much earlier example of work equivalent to an L3 additive model. Miller [16] showed how Vogel's group polarizabilities could be factored into a set of atomic hybrid polarizabilities because Vogel's units coincided with atoms in the usual hybridization states. A different classification into atom types that is very nearly the same as using hybridization is categorization by coordination number. For example, carbon atoms are almost always tetra-, tri- or bi-coordinate; the corresponding atom types can be denoted C_4, C_3, and C_2, respectively. Penta- and mono-coordinate carbon atoms are rare and can be folded in with tetra- and bi-coordinate carbon atoms, respectively. One can have partial L3 schemes in which atoms of some elements are left at the dressed atom or L2 stage and others are subdivided into L3 types. Coordination number, rather than hybridization, is used in all the new L3 models reported in this work.

Traditional models on the fourth rung of the hierarchy [12] use information about bonds and functional groups in addition to atom types. Writing the molecular polarizability as a sum of bond polarizabilities dates back to the early mid-20th century work of von Steiger [17] and Smyth [18]. A Level 4 (L4) model can predict different properties for different structural isomers but there are ambiguities about the most accurate way to do this [10]. There is a different way to think about progressing from atom types to bond-additive models. At the L3 stage, one effectively uses the valence state and the number of nearest (bonded) neighbors to create atom types. Taking the identity of the nearest neighbors into account leads to L4 models. For example, a tricoordinate carbon atom type C_3 can be further subdivided into subtypes by the number of C_3 atoms it is bonded to: $C_{3,0}$, $C_{3,1}$, $C_{3,2}$, and $C_{3,3}$. These types are illustrated by the following simple examples. The carbon atom in the carboxylic group of acetic acid is of type $C_{3,0}$, the terminal carbon atoms in butadiene are of type $C_{3,1}$, the interior carbon atoms in butadiene and the carbon atoms in benzene are of type $C_{3,2}$, and the two carbon atoms shared by the edge-fused six-membered rings in naphthalene are of type $C_{3,3}$. For example, Zhokhova et al. [9] enhanced the L2-like model of Bosque and Sales [7] by adding a correction term which is proportional to the number of $C_{3,3}$ carbon atom types.

Models at the next level (L5) of the hierarchy [12] incorporate network effects, that is they allow for distant atoms that alter the environment. Levels higher than that require geometrical information, such as Cartesian coordinates, about the molecules they are being applied to.

The purpose of this work is to parameterize and test additive models for molecular polarizabilities and volumes using a large, consistent, and balanced set of data. The limits of accuracy that can be achieved by using L2, L3, and partial L4 additive models are assessed. Section 2 details the database chosen and the fitting procedures. Additive models for polarizabilities and volumes are presented and discussed in Sections 3 and 4, respectively. A few concluding remarks follow in Section 5. Atomic units are used throughout.

2. Methodological detail

A reliable comparison of different additive models requires a moderately large, consistent, balanced, and reasonably accurate set of data [19]. We used the TABS database [20] containing structures of 1641 molecules of organic, biochemical, and pharmaceutical interest. The molecules contain as many as 34 atoms including at least one C atom and may have one or more H, N, O, F, S, Cl, and Br atoms. The TABS database contains at least 25 molecules representing each of 24 functional categories. These structures were all obtained [20] by computations with the B3LYP hybrid density functional [21–24] and the aug-cc-pVTZ basis set [25,26]. Static dipole polarizabilities computed at the same level are available for the TABS database [27,19]. Use of a range-separated functional would probably lead to more accurate polarizabilities. However, we think that using a model chemistry [28] (a consistent method and basis set) for all properties is essential to avoid conflating different types of errors. The molecular volume, defined [29,30] as the volume contained within the 0.001 a_0^{-3} isodensity surface of the electron density, has also been calculated [31,19] at the same B3LYP/aug-cc-pVTZ level for all the molecules in the TABS database.

It is prudent to use only a subset of TABS for fitting purposes so that the rest of the database provides an independent test. The training set should ideally be less than a quarter of the full set. An algorithm was devised to build a training set by selecting molecules at random from the full set subject to the constraint that each of the desired atom types occurs at least $N_{\min}$ times. If the full set does not contain $N_{\min}$ instances of a desired atom type, then all the available instances are put into the training set. This procedure was applied, using $N_{\min} = 35$, to produce an initial training set containing all the atom types required for a selected partial L4 additive model. This initial training set had 274 molecules. Then, for all other partial L4 models of interest, the same procedure was used to determine which molecules needed to be added to the initial training set to fulfill the constraint on minimum numbers of each atom-type. All the extra molecules so identified were added to the initial training set to produce the final training set containing 298 molecules.

Atom-L3-type counts for TABS and its training subset are listed in Table 1. Observe that TABS does not have as many hypervalent S atoms as could be desired. Although less than 10% of the 1641 molecules in TABS contain Br atoms, the total number of Br atoms is probably large enough to ensure stable fits. The same training set was used for all models reported here. Using a uniform training set for all models and properties facilitates comparisons. The TABS ID numbers of the molecules in the training set are listed elsewhere [19].

A rule-of-thumb that was mentioned recently [32], but whose origin has not been traced, suggests that no more than $\sqrt{N_t}$ parameters can be determined reliably if the training set contains N_t points. This rule suggests that no more than $\sqrt{298} \approx 17$ parameters can be determined reliably by fits to the chosen training set. This is only one more parameter than needed to create a complete L3 additive model for the molecules in the TABS database.

Table 1
Atom-type counts for the TABS database and training set, and training fraction.

Element	TABS	Training set	Fraction
H	10 489	1745	0.17
C_4	3325	514	0.15
C_3	3439	639	0.19
C_2	583	153	0.26
N_3	555	93	0.17
N_2	489	92	0.19
N_1	196	42	0.21
O_2	663	85	0.13
O_1	662	161	0.24
F	388	50	0.13
S_4	34	34	1.00
S_3	25	25	1.00
S_2	287	54	0.19
S_1	83	37	0.45
Cl	305	53	0.17
Br	128	41	0.32