Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Molecular Graphics and Modelling 22 (2004) 519–525
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

# QSAR and QSPR based solely on surface properties?

Timothy Clark*

*Computer-Chemie-Centrum, Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, 91052 Erlangen, Germany*

Accepted 4 March 2004

Available online 5 May 2004

**Abstract**

The use of descriptors based on local properties calculated at the molecular surface for QSPR models is discussed. It is suggested that descriptors should be related to the physical theory of intermolecular interactions and the relationship between established surface-based descriptors and the fundamental types of intermolecular interaction is discussed. Descriptors based on local properties that do not encode the chemical constitution of the molecule directly are likely to provide less local QSPR models and favor scaffold hopping. The major disadvantage for surface-based descriptors is that they are difficult to interpret in the sense of relating predictions to the chemical composition of the molecule. This disadvantage must be alleviated by suitable model-interrogation techniques.


*Keywords:* QSPR; QSAR; Descriptors

## 1. Introduction

Quantitative structure–activity (QSAR) and structure–property (QSPR) relationships have a long and successful history, especially for predicting biological activity [1]. Perhaps the historically most successful approach to such studies is to use the so-called 2D-descriptors, which are based on the bonding topology of the molecules. A huge variety of such approaches, from topological descriptors [2] to "fingerprints" of various flavors [3] has been published. Such approaches have been remarkably successful for many applications and are still used extensively. Moving to three-dimensional descriptions of molecules brings extra geometrical information, which can be used to great advantage in rational drug design, for instance by interpreting the results of CoMFA analyses [4]. The major disadvantage of 3D-descriptors is, however, that the implicit treatment of multiple conformations inherent in 2D-descriptors is lost. Thus, either the active conformation must be known or an extensive conformational search is necessary. Even if the energetically accessible conformations are known, some technique must be used in the QSAR to select the correct conformation for each compound from the variety available. Thus, multi-conformational QSAR is usually limited to systems in which the bound conformation of the lead compound is known. The effect of conformation on QSPR studies is less extreme, but we can reasonably expect that different conformations of the same molecule should have different physical properties. In practice, techniques for estimating physical properties from 3D-descriptors usually use only one conformation per compound. This approach is generally justified because the quality of the experimental data does not allow resolution of conformational effects and because even quite significant conformational changes involving intramolecular hydrogen bonds only change, for instance, the predicted boiling point by about as much as the uncertainty in the predicted value [5].

However, even 3D-descriptors usually have a strong element of the molecular topology (atom or group counts, etc.). These elements are often regarded as being essential because they allow chemists to interpret the results in terms of the chemical structure and modifications that may improve the activity or a given physical property. One can, however, argue that these essentially 2D-elements in 3D-descriptor sets have little fundamental physical meaning and would not be necessary if the remaining true 3D-descriptors were able to describe the relevant properties of the molecule adequately. Fig. 1 shows three representations of the same molecule. The conventional structural drawing (Fig. 1(a)) conveys the most information about the chemical constitution of the molecule. Moving to the sticks representation (Fig. 1(b)) adds knowledge of the conformation without los-

* Tel.: +49-913-185-2948.
*E-mail address:* clark@chemie.uni-erlangen.de (T. Clark).

1093-3263/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2004.03.012

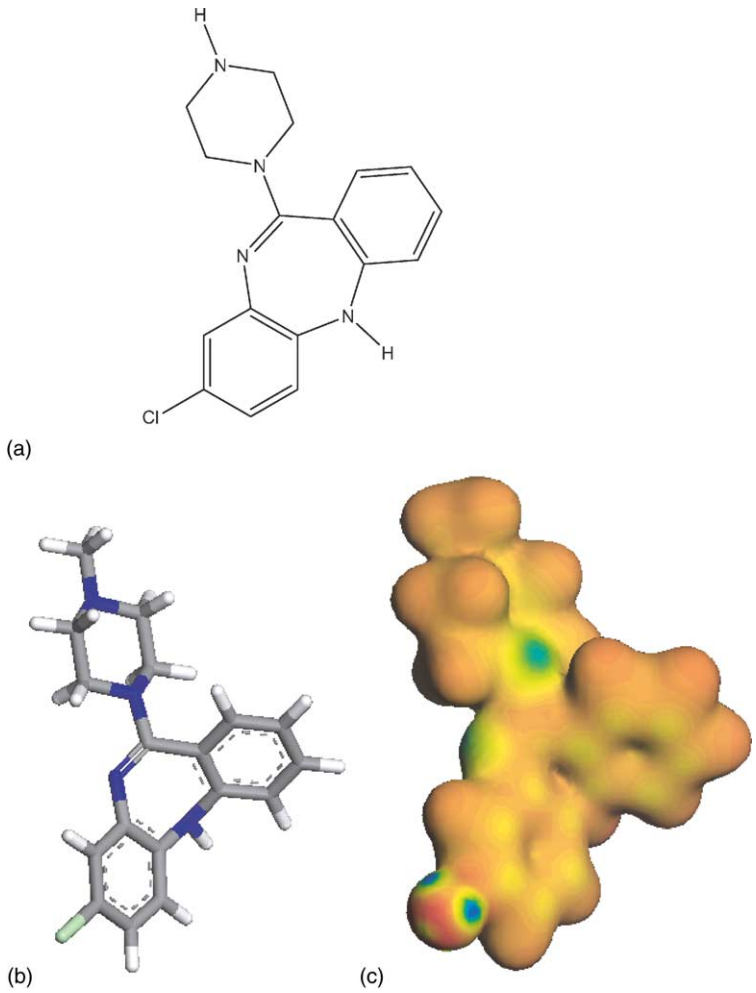


Fig. 1. (a) Conventional 2D structural drawing, (b) sticks representation and (c) molecular electrostatic potential mapped onto an isodensity surface all for the same molecule.

ing any information inherent in Fig. 1(a). However, many chemists would argue that the chemical constitution is most easily seen from Fig. 1(a). The color-coded isodensity surface (Fig. 1(c)) contains very little, if any, information about the chemical constitution. It is therefore of very little value to the average chemist. However, of these three formalisms for portraying the molecule, Fig. 1(c) contains by far the most information about the ability of the molecule interact with other identical or different molecules. The shape and extent of the isodensity surface is a very good approximation for the steric-repulsion surface of the molecule (see below). The strongest type of non-covalent intermolecular interaction, the purely electrostatic Coulomb term, is well represented by the molecular electrostatic potential (MEP) projected onto the surface. Thus, two of the major terms in intermolecular interaction potentials [6] are encoded in Fig. 1(c) at the expense of information about the chemical constitution of the molecule. Note that Fig. 1(c) was purposely drawn with an opaque surface in order to make the difference between molecular surfaces and molecular structural formulae clear. The purpose of this paper is to investigate the question as to whether we can encode all the information necessary to treat intermolecular interactions at the molecular surface. If so, we can treat almost all important QSAR and QSPR topics because both physical properties and biological activity are determined by intermolecular interactions. We therefore first consider intermolecular interactions in general and the properties needed to describe them.

## 2. Intermolecular interactions

The theory of intermolecular interactions is very well developed [6] so that we already know which molecular properties must be encoded at a molecular surface in order to describe the ability of the molecule to interact with another. These interactions are as follows.

### *2.1. Coulomb interactions*

Coulomb interactions are usually the strongest and most long range intermolecular interactions. They are usually

treated in classical mechanical models (force fields, etc.) by assigning fictitious and non-physical partial charges to individual atoms and calculating the interaction energy using Coulomb's law. The most usual color-coding seen on molecular surfaces is for the MEP, which describes the anisotropy of the molecular electrostatics well. The MEP at the molecular surface is an important descriptor for QSPR applications. Murray and Politzer [7] introduced a series of descriptors based on the statistics of the MEP-values at the triangulation points on a calculated molecular surface as powerful descriptors for physical properties. We [5,8–15] have used slight modifications of these descriptors in a series of QSPR models. Thus, Coulomb interactions have been treated systematically using surface-based descriptors.

### 2.2. van der Waals' interactions

van der Waals' interactions are usually considered to consist of two components, which we can consider separately here. The first is the steric repulsion. Both common van der Waals' potentials, Lennard-Jones [16] and Buckingham [17] have very steep repulsive segments, so that for most purposes it is adequate to define the position of the onset of the repulsive part of the curve. Thus, describing steric repulsion reduces to defining a molecular surface. The simplest approximation would be to use a van der Waals' surface, but these often contain deep narrow clefts that are inaccessible for other molecules. Thus, two different types of surface have become popular, isodensity surfaces for which the value of the electron density is chosen to approximate the steric extent of the molecule [18] and solvent-excluded surfaces (SES) based on van der Waals' radii and a fictitious spherical "solvent" molecule [19].

The second component of intermolecular van der Waals' potentials is the dispersion term, which gives the weak van der Waals' minimum in the intermolecular potential. The dispersion energy has traditionally been treated using the London formula [20] in combination with the Slater–Kirkwood approximation [21], although there are many other variations. The London formula, however, relies on the electronic polarizability, which must therefore be encoded on the surface by a suitable local property.

### 2.3. Donor–acceptor interactions

Donor–acceptor interactions are the final type of intermolecular potential that must be described. Again, Murray and Politzer and their group have introduced a local property, the local ionization energy, that describes the donor ability of the molecule [22]. We [23] have recently extended this concept by introducing a local electron affinity, which can also be combined with the local ionization energy to give a local hardness. These local properties prove to be useful in predicting chemical reactivity, the ultimate consequence of donor–acceptor interactions. The exact nature of these local properties will be discussed below.

## 3. QSPR with surface-based properties

Abraham [24] has consistently emphasized the importance of sound physical principles in constructing his QSPR models. By far the majority of QSPR studies, however, have taken the pragmatic approach that if the model works it is sound. However, many, if not most, QSPR models produced in the last decade suffer from a pronounced locality. This means that they work well for the compound classes represented in the training and validation sets, but may fail catastrophically for other classes. This situation is not always easy to detect because for some key properties, such as aqueous solubility, the published experimental data is itself very local and moreover does not include important classes of compounds, for instance drugs. This is emphasized by the physical property map [25] shown in Fig. 2. The color-coded points mark the positions of compounds from the Aquasol database [26] whose aqueous solubility is known at 298 K. The contours indicate the areas in which drugs are found. There are essentially no training compounds within the drug areas, so that we cannot expect a solubility model based on the published data to treat drugs adequately.

A convincing hypothesis is that local models tend to result if the underlying physical properties are ignored. The 2D and to some extent the 3D-descriptors used for most models to date promote locality by relying on the presence of distinct features in the chemical constitution of the molecules. This leads to two dangers, that the features of a new compound are not adequately represented and that the properties attributed to a given feature may not always be transferable from molecule to molecule. Both of these disadvantages can lead to a local model. Although we can never ban the danger of locality from QSPR models with certainty, it is more likely that a model that uses exclusively descriptors that make no direct reference to the chemical constitution will be able to treat an unknown class of compounds. Surface-based descriptors are very promising in this respect but have often been used in the past in conjunction with constitution-dependent descriptors such as the number of aromatic rings or the sums of the MEP-derived charges on all atoms of a given element in the molecule [5,8–15]. We have been forced to use such inherently 2D-descriptors in past models because the surface-based descriptors constructed solely from the MEP only describe the Coulomb part of the intermolecular interactions, leaving element- or group-dependent descriptors to describe the rest.

Can we therefore use exclusively surface-based descriptors and whole-molecule properties (such as volume, surface area, molecular weight, etc.) for QSPR models? One way to answer this question is to investigate whether we can describe intermolecular interactions using local properties at the surface of the molecule, rather than the more

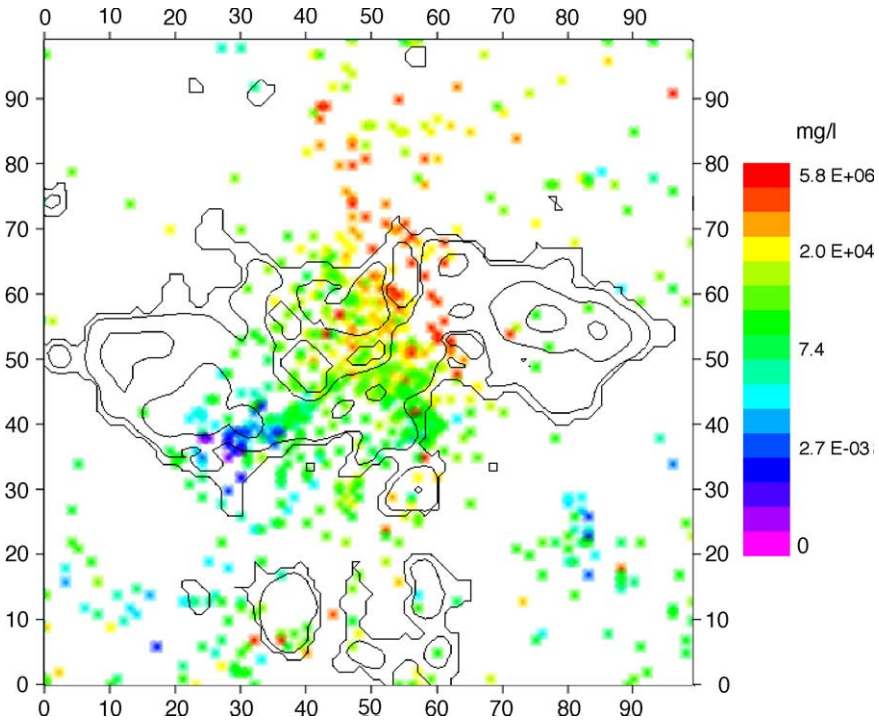


Fig. 2. 100 × 100 physical property map [25] of the occurrence of compounds with published aqueous solubilities compared with the areas in which drugs occur. The solubility data set is represented by crosses color coded according to their solubility and the drug areas are marked by contours and correspond to those reported in Ref. [26].

usual atom-based ansatz used for most force fields. It is often forgotten that atom-based force fields rely heavily on the atoms-in-molecules approximation, which has been popularized in another context by Bader [27]. If we adopt an extremist view, there are no “atoms-in-molecules”. We have traditionally used the transferability of atomic contributions to physical properties and structure to construct force fields, additive schemes for estimating properties, etc. The fact that these techniques work well confirms that atoms-in-molecules is a good approximation in most cases, but it is still an approximation. Introducing unique and physically sensible ways of determining the borders between individual atoms does not change the underlying approximation that atoms can be defined as individual entities within a molecule. Quantum mechanics defines a molecule as an electron density bound by the electrostatic effect (the external field) of the nuclei. The MEP-based descriptors introduced by Murray and Politzer [7] are calculated using only the effects of the electron density and the nuclei, even to the extent that the molecular surface is defined by the electron density. However, would these properties be adequate to construct an intermolecular force field to treat the Coulomb interactions?

## 4. Surface-based intermolecular Coulomb interactions

Imagine a molecule described as an irregular shaped surface, such as that shown in Fig. 1(c). One way to treat the electrostatics of this irregular body would be to use a multipole expansion [6]. However, multipole expansions often converge very slowly or not at all and the electrostatic center of, for instance, a U-shaped molecule may be outside the molecules itself. Another possibility that we have investigated recently [28] is to treat the effective charge of the molecule as being dependent on the orientation relative to the molecule itself. This simply means that, for instance, a second molecule approaching from above would “see” a different charge to one approaching from below. In order to set up such a model, we can describe the effective molecular charge in terms of spherical harmonics [29]. This ansatz has the disadvantages that it cannot describe surfaces that cross a line radiating from the center more than once but this proves not to be serious for molecules up to about 200 atoms. Fig. 3 shows that the errors introduced by such a model for benzoyl fluoride compared to our natural atomic orbital-point charge (NAO-PC) [30] electrostatic model are less than 0.1 kcal mol$^{-1}$ at points outside the van der Waals’ surface of the molecule. The exact form of the model is given in Eq. (1).

Thus, a feasible electrostatic model for an entire molecule is simply to fit the calculated MEP at points around the molecule to a function

$$\vec{V}_{\text{Coulomb}} = \tilde{q}\left(\frac{1}{r} + \frac{\tilde{a}}{r^2}\right) \quad (1)$$

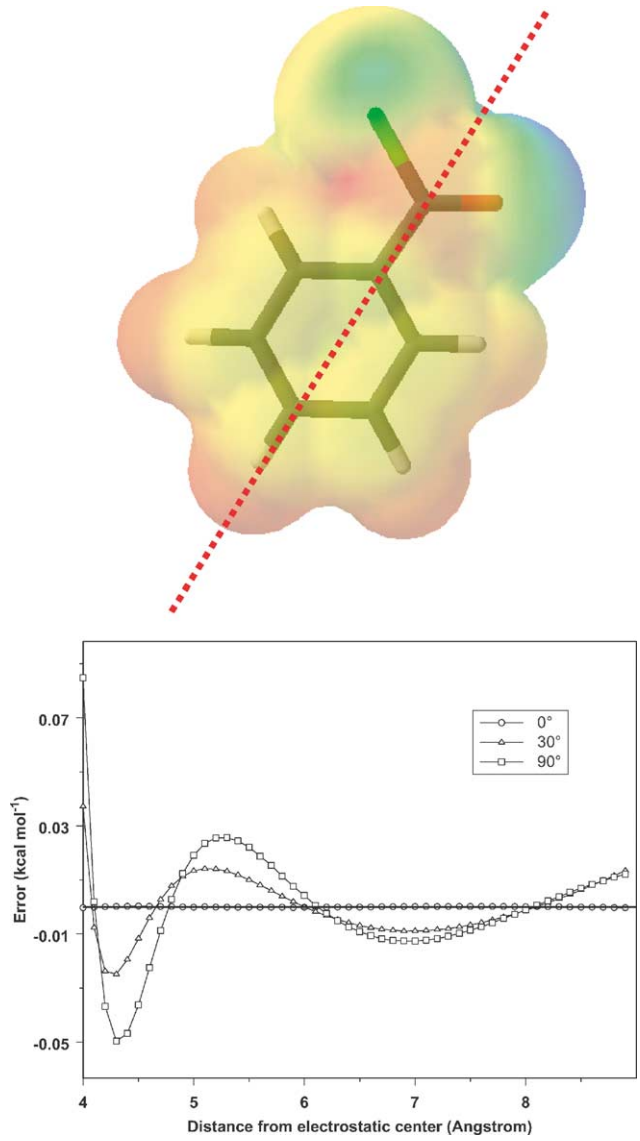


Fig. 3. Errors in the molecular electrostatic potential (kcal mol$^{-1}$) on radial vectors from the electrostatic center of benzoyl fluoride in a plane containing red dotted line and perpendicular to the ring plane. The errors plotted are the difference between the MEP calculated with the full NAO-PC model [29] and that calculated with a simple $q(1/r + a/r^2)$ technique, as described in the text. Results are shown for radial vectors at angles of 0°, 30° and 90° to the ring plane.

where $\vec{V}_{\text{Coulomb}}$ is the MEP at the point being considered and $\tilde{q}$ is an effective charge dependent on the orientation of the point relative to the molecular axes and $\tilde{a}$ is a fitted constant that is also described by spherical harmonics. Most importantly, such a model need not use the atoms-in-molecules concept. Although this model does not depend specifically on the MEP at the molecular surface, there is clearly a strong correlation between the effective charge of the molecule in a given direction and the MEP at the surface on the relevant radial vector, so that we can conclude that surface-based electrostatics can indeed describe intermolecular Coulomb interactions.

## 5. Donor/acceptor and van der Waals' interactions

The above discussion treats only the Coulomb interaction. Donor/acceptor interactions are also orientation-dependent and can be treated using a similar ansatz to Eq. (1). We need, however, to define local properties that describe the donor and acceptor properties of the molecule in a given direction. Sjoberg et al. [22] introduced the local ionization energy as a measure of the electron-donating capacity of a molecule relative to an acceptor at a given point in space and we [23] extended this concept to the local electron affinity, which is the acceptor equivalent. These properties can also be described in terms of spherical harmonics in conjunction with a distance-dependent overlap term to calculate donor–acceptor interactions. Thus, surface-based descriptors calculated from the local ionization potential and electron affinity are able to describe intermolecular donor/acceptor interactions [31].

Dispersion (weak) interactions are often defined using the London formula [20] in combination with the Slater–Kirkwood approximation [21]. In its anisotropic form, the London formula is inherently orientation-dependent, but greater orientational resolution would be required for an intermolecular model. We [23] have recently defined the local polarizability within NDDO-based semiempirical molecular orbital theory using our parameterized [32] and extended [33] version of the variational technique introduced by Rivail [34,35]. This local property can be used analogously to the MEP or the local ionization energy and electron affinity for calculating dispersion interactions. Note, however, that the partitioning of the molecular polarizability is arbitrary [32] and based on an atoms-in-molecules concept, so that the current version is not entirely consistent with our aims.

Steric repulsion can also be treated similarly, although in this case it would probably be sufficient to define an effective distance from the molecule that is dependent on the orientation. Meyer [36] has treated molecular shape in detail and describes many relevant concepts. We still need, however, an orientation-dependent description of the molecular shape that is currently still missing from our set of molecular descriptors [31]. Currently, the very simple globularity [37] is used to describe the deviation of the molecule from spherical.

## 6. Conclusions

The above discussion is intended to extend the concept of surface-based descriptors so that it can be used to describe all intermolecular interactions. In order to do this, we have considered a novel force-field-like approach based on the established theory of intermolecular interactions. The essential requirement is that we define local properties that can be calculated at the molecular surface in order to derive statistical descriptors such as those introduced by Murray and Politzer [7]. These descriptors are well established for

Coulomb interactions, which often suffice as these are by far the strongest intermolecular forces for polar molecules. The local ionization energy [22] and the local electron affinity [23] serve analogously to describe donor/acceptor interactions. For non-polar molecules, the local polarizability [23] becomes particularly important for describing physical properties such as the boiling point [31] because it can be related to dispersion interactions. We still need to define a local property related to the molecular shape in order to be able to describe intermolecular steric repulsion.

The purpose of such descriptors is to remove 2D-like information about the molecular constitution from the descriptor set used for QSPR and later QSAR models. This change has three potential advantages that remain to be demonstrated:

- The models based on descriptors that do not use information about the molecular constitution should be more general because they are based on the physics on intermolecular interactions. Thus, we can hope that such models will not suffer from the locality often observed for current models.
- Surface-based descriptors are able to describe properties contained in many 2D-like descriptors (such as sums of MEP-derived charges on different elements, hydrogen bond donor and acceptor counts, etc.). This means that we can use less descriptors [31], which often results in a more robust model.
- For QSAR applications, scaffold hopping becomes inherently more likely if the molecular constitution is not used for the descriptors. This enhances the chances of finding analogous active compounds that are not related to the lead in their chemical constitution.

A further advantage of the techniques discussed above is that they can be used for surface-integral models for physical properties, such as those introduced by Brickmann and coworkers [37]. The combination, for instance, of a traditional descriptor-based QSPR with a surface-integral model for the same property would provide a powerful reality check for predicted properties of unknown compounds.

However, the traditional argument against descriptors that cannot easily be interpreted remains. Individual surface properties can be visualized and interpreted [23] but descriptors based on them are more difficult to understand [22]. Thus, until we have convincing evidence that surface-based descriptors give us better QSPR and QSAR models, the advantage of interpretability favors traditional descriptors. However, surface-based descriptors can be related directly to the theory of intermolecular interactions, so that they are interpretable in a different, less chemical way. Nevertheless, should surface-based descriptors prove to be superior to the conventional approach, we will have to resort to interrogation techniques ("*what if I replaced this group by* $CF_3$?") with the necessary software.

## References


[1] C. Hansch, A. Leo, Exploring QSAR: Fundamentals and Applications in Chemistry and Biology, ACS Professional Reference Book, American Chemical Society, Washington, DC, 1995.
[2] L.B. Kier, L.H. Hall, Molecular Structure Description: The Electrotopological State, Academic Press, San Diego, CA, 1999.
[3] A.C. Good, J.S. Mason, S.D. Pickett, Meth. Principles Med. Chem. 10 (2000) 131–159.
[4] R.D. Cramer III, D.E. Patterson, J.D. Bunce, Prog. Clin. Res. 291 (1989) 161–165.
[5] A.J. Chalk, B. Beck, T. Clark, J. Chem. Inf. Comput. Sci. 41 (2001) 457.
[6] A.J. Stone, The Theory of Intermolecular Interactions, Clarendon Press, Oxford, 1996.
[7] J.S. Murray, P.A. Politzer, J. Mol. Struct. (Theochem.) 425 (1998) 107–114;
J.S. Murray, P. Lane, T. Brinck, K. Paulsen, M.E. Grice, P.A. Politzer, J. Phys. Chem. 97 (1993) 9369–9373.
[8] T. Clark, in: M.G. Hicks (Ed.), Chemical Data Analysis in the Large: The Challenge of the Automation Age, Proceedings of the Beilstein-Institut Workshop, Bozen, Italy, May 22–26, 2000, Logos Verlag, Berlin, 2000, pp. 93–104.
[9] T. Clark, in: H.-D. Höltje, W. Sippl (Eds.), Rational Approaches to Drug Design, Prous Science, Barcelona, 2001.
[10] T. Clark, in: M.G. Hicks, C. Kettner (Eds.), Molecular Informatics: Confronting Complexity, Proceedings of the Beilstein-Institut Workshop, Bozen, Italy, May 13–16, 2002, Frankfurt am Main, July 2003. http://www.beilstein-institut.de/bozen2002/proceedings/clark/clark.pdf.
[11] T. Clark, A. Breindl, G. Rauhut, J. Mol. Model. 1 (1995) 22.
[12] A. Breindl, B. Beck, T. Clark, R.C. Glen, J. Mol. Model. 3 (1997) 142.
[13] B. Beck, A. Breindl, T. Clark, J. Chem. Inf. Comput. Sci. 40 (2000) 1046–1051.
[14] A.J. Chalk, B. Beck, T. Clark, J. Chem. Inf. Comput. Sci. 41 (2001) 1053.
[15] M. Hennemann, T. Clark, J. Mol. Model. 8 (2002) 95–101.
[16] J.E. Lennard-Jones, Proc. R. Soc. 43 (1931) 461.
[17] R.A. Buckingham, Proc. R. Soc. A 168 (1938) 264.
[18] G.D. Purvis III, J. Comput.-Aid. Mol. Des. 5 (1991) 55–80.
[19] J.L. Pascual-Ahuir, E. Silla, I. Tuñon, J. Comput. Chem. 15 (1994) 1127–1138.
[20] F. London, Trans. Faraday Soc. 33 (1937) 8–26.
[21] J.C. Slater, J.G. Kirkwood, Phys. Rev. 37 (1931) 682–697.
[22] P. Sjoberg, J.S. Murray, T. Brinck, P.A. Politzer, Can. J. Chem. 68 (1990) 1440–1443;
P.A. Politzer, J.S. Murray, M.E. Grice, T. Brinck, S. Ranganathan, J. Chem. Phys. 95 (1991) 6699–6704;
P.A. Politzer, J.S. Murray, M.C. Concha, Int. J. Quant. Chem. 88 (2002) 19–27.
[23] B. Ehresmann, B. Martin, A.H.C. Horn, T. Clark, J. Mol. Model. 9 (2003) 342–347.
[24] M.H. Abraham, H.S. Chadha, Org. React. (Tartu) 30 (1996) 13–20.
[25] M. Brüstle, B. Beck, T. Schindler, W. King, T. Mitchell, T. Clark, J. Med. Chem. 45 (2002) 3345;
M. Brüstle, T. Clark, in preparation.
[26] S.H. Yalkowsky, Y. He, Handbook of Aqueous Solubility Data, CRC Press, Cleveland, OH, 2003.
[27] R.F.W. Bader, Atoms in Molecules: A Quantum Theory, Oxford University Press, Oxford, 1994.
[28] J.-H. Lin, T. Clark, in preparation.
[29] H. Groemer, Geometric Applications of Fourier Series and Spherical Harmonics, Cambridge University Press, Cambridge, 1996.
[30] G. Rauhut, T. Clark, J. Comput. Chem. 14 (1993) 503;
B. Beck, G. Rauhut, T. Clark, J. Comput. Chem. 15 (1994) 1064.

[31] B. Ehresmann, A. Alex, M. de Groot, T. Clark, J. Chem. Inf. Comput. Sci. 43 (2004) 658–668.
[32] G. Schürer, P. Gedeck, M. Gottschalk, T. Clark, Int. J. Quant. Chem. 75 (1999) 17.
[33] B. Martin, P. Gedeck, T. Clark, Int. J. Quant. Chem. 77 (2000) 473–497.
[34] D. Rinaldi, J.-L. Rivail, Theoret. Chim. Acta 32 (1974) 243–251; D. Rinaldi, J.-L. Rivail, Theoret. Chim. Acta 32 (1973) 57–70.
[35] P.G. Mezey, Shape in Chemistry, VCH, New York, 1993.
[36] A.Y. Meyer, Chem. Soc. Rev. 15 (1986) 449–475.
[37] R. Jaeger, S.M. Kast, J. Brickmann, J. Chem. Inf. Comput. Sci. 43 (2003) 237–247.