DOES QUANTUM CHEMISTRY HAVE A PLACE IN CHEMINFORMATICS?

TIMOTHY CLARK

Computer-Chemie-Centrum, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, 91052 Erlangen, Germany.
E-Mail: clark@chemie.uni-erlangen.de

ABSTRACT

The possible role of quantum mechanical (QM) techniques in cheminformatics is discussed. The advantages, disadvantages and capabilities of QM and its applicability to databases of thousands of molecules are discussed. The critical relationship between quantitative structure-property relationships (QSPRs) and the quality of the experimental data is discussed using aqueous solubility as an example. The use of QM-derived descriptors to investigate physical property space and to characterise compounds as drug-like or non-drug-like is illustrated. Finally, it is pointed out that not QM-calculations, but rather a knowledge of the molecular electron density is necessary for the examples shown, and a technique that can reproduce the electron density without QM-calculations is presented.

INTRODUCTION

Quantum mechanical calculations are not usually considered to be applicable to cheminformatics, although we have shown that semiempirical MO-calculations can be used on complete databases (1) and can play an important role in many cheminformatics applications (2,3). This article is intended to provide an overview of the applicability and capabilities of quantum mechanical techniques for cheminformatics and to discuss the relationships between data, descriptors and properties in quantitative structure-property relationships (QSPRs). Finally, an alternative technique for deriving the molecular electron density without quantum mechanics will be described.

Typically, cheminformatics applications use 2D- or very simple, classically derived 3D-descriptors for quantitative structure-activity relationships (QSARs) and QSPRs. As the border

between, for instance, QSAR and pharmacophore-based high-throughput virtual screening is very poorly defined, many cheminformatics tasks can be considered to be simply more traditional QSAR or QSPR applied to larger numbers of molecules. In this respect, the constant advances in hard- and software performance tend to make the border even less clear because larger datasets become manageable with every advance. We have shown (2,3) that complete databases of tens of thousand of compounds can be treated with economical quantum mechanical techniques and have discussed the advantages of detailed quantum mechanical descriptions of molecules for QSPR (2) and QSAR (3). What, however, has changed since in the two years since references 2 and 3 were written? Is quantum mechanics still a useful tool for cheminformatics? Will it displace more traditional techniques? Are there alternatives?

WHY USE QUANTUM MECHANICS?

The advantages of using semiempirical MO-calculations to calculate molecular descriptors have been described before (2,3) and will only be outlined briefly here. The resolution of the molecular electrostatic properties is generally higher (i.e. atoms are not treated isotropically) in quantum mechanical calculations. This generally results in better descriptions of the molecular electrostatic potential in important regions of the molecular surface (i.e. where bonding interactions occur). Furthermore, electronic properties such as polarisability, ionisation potentials, electron affinities, dipole and higher multipole moments etc. and descriptors derived from them often prove to be very useful descriptors, especially in QSPR-applications. An example of the use of such descriptors is given in our recent work on the hydrogen-bond acceptor strengths of nitrogen heterocycles (4). Note however, that many of the properties listed above can be obtained from the electron density, so that efficient methods for generating an accurate electron density without quantum mechanics are potentially of great interest for cheminformatics.

However, for the moment we should assess the reliability of the most computationally economical for of molecular orbital theory, the modern semiempirical techniques MNDO (5), AM1 (6), and PM3 (7) for calculating the properties listed above. These techniques are parameterised to reproduce experimental heats of formation, molecular structures, dipole moments and ionisation potentials (from Koopmans' theorem). These properties (especially the dipole moment) ensure that the electron densities calculated are generally quite accurate, so that the molecular electrostatics calculated by semiempirical techniques agree well with high level

ab initio data (8, 9). However, what about less directly parameterised properties like the polarisability, which is often thought to be very difficult to calculate and which requires very extensive basis sets at *ab initio* levels of theory? The solution, as for many properties in semiempirical theory, is to use a fast and effective level of theory to calculate the property in question and parameterise the method against experimental data. In this case, Rivail and his coworkers (10) published a very fast and simple variational technique for calculating the molecular electronic polarisability. However, this variational method is prone to systematic and element-specific errors, so that we (11) parameterised the element-specific integrals involved especially to reproduce the experimental data. This resulted in a fast and accurate method for calculation of the molecular electronic polarisability that is also amenable to partitioning into group, atomic or even orbital contributions (3, 12). The molecular electronic polarisability proves to be an important descriptor in most of our QSPR models and plays a very significant role in describing physical property space (13).

The general impression is that CPU-requirements for quantum mechanical calculations preclude them from being used for cheminformatics applications. This is not necessarily the case. In a first feasibility study (1), we were able to process the entire Maybridge database (about 53,000 compounds) on a 128-processor SGI Origin 2000 in half a day. However, computer performance has increased, and above all prices have decreased, since this study was performed, so that a Euro 1000 computer can process about 1,500 typical druglike compounds per day (full optimisations with AM1 or PM3).

Thus, there are relatively few real obstacles to using semiempirical MO-theory for cheminformatics. Do we, however, really need "better" descriptors, for instance for QSPR applications?

EXPERIMENTAL DATA AND QSPR MODELS

QSPR-models are derived by calculating descriptors for each molecule in the dataset and then using an interpolation technique (regression, neural net etc.) to relate the descriptors to the property. The quality of the model obtained is necessarily limited by the quality of the experimental data. Have we, however, already reached the limit of data-accuracy for some properties? Figure 1 shows results for an AM1/neural net model (14) for aqueous solubility based on a training set of solubilities for 559 compounds at 298K. This model gives a standard deviation between calculation and experiment of 0.51 log units, a mean unsigned error (MUE)

of 0.40, a maximum error of 1.67 and 35% of the predictions outside the calculated (15) ($\pm$ 1 standard deviation) error bars.
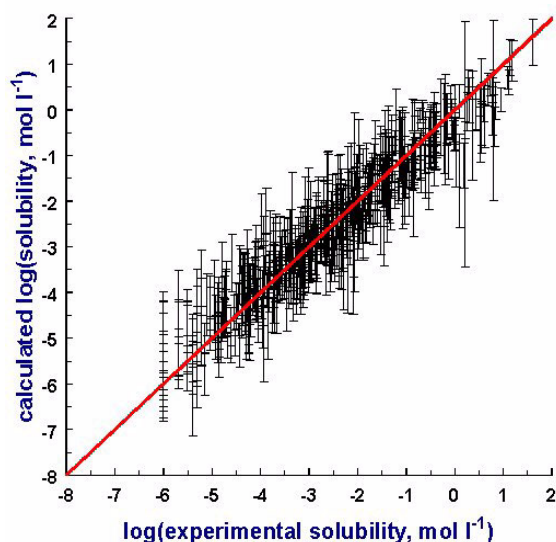

Figure 1. The performance of an artificial neural net QSPR-model[14] for aqueous solubility based on AM1 descriptors. The error bars shown are calculated according to the procedure outlined in reference 15.

These results are typical. A literature survey reveals that for 11 studies (some of which used related experimental data, not just calculated descriptors) using datasets between 399 and 1312 compounds, the six published standard deviations between calculation and experiment average to 0.60 log units (including one extreme outlier at 0.16) and the five published root mean square deviations (RMSD) average to 0.68. If we ignore the outlier, the standard deviations range from 0.57 to 0.79 and the RMSDs from 0.62 to 0.76. Thus, all published models perform very similarly, although they use very different types of descriptors and interpolation techniques. What is not different, however, are the data. How reliable are experimental solubility measurements? Yalkowsky and Banerjee[16] have outlined the experimental techniques for and difficulties encountered in measuring aqueous solubility. They also include a table of measured solubilities for some *"extremely hazardous substances"* (reference 16, Appendix C). As a crude estimate of the reliability of experimental data, Figure 2 shows the highest experimental value plotted against the lowest for the 18 compounds for which more than one measurement is listed.

The standard deviation between the two sets of experimental values is 0.79 log units, the MUE is 0.48, the RMSD 0.76 and the largest error 1.93. Even though a sample of only 18 compounds

cannot be considered reliable, the conclusion seems clear that QSPR-models cannot be very much better than this correlation.
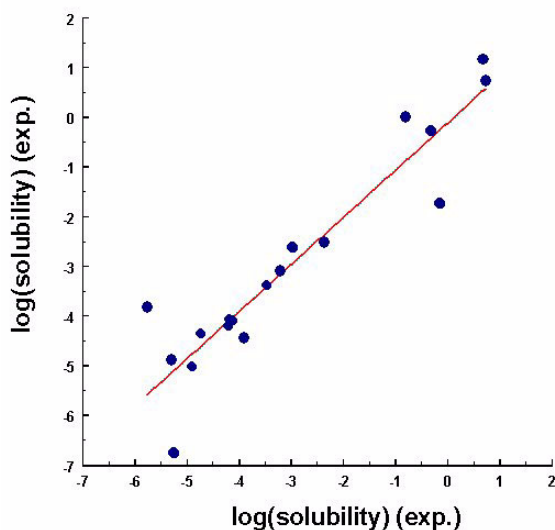

Figure 2. A plot of the highest experimental values for aqueous solubility against the lowest for 18 compounds taken from a table of *"extremely hazardous substances"* (reference 16, Appendix C).

We may even be in a situation in which our available descriptors could describe solubility very much better than they do if better experimental data were available (i.e. the models are limited by the data, not by the descriptors or the interpolation technique). Is it the worth developing new, "better" descriptors if the experimental data do not justify their use?

MAKING BETTER USE OF THE AVAILABLE DATA

It is not necessarily true that a QSPR model for a given property cannot be better than the dataset used to train for that property. Consider, for instance, molecular mechanics (force field) calculations for alkanes. Several high quality force fields give heats of formation for alkanes that are more reliable than experimentally measured values. This is possible because the force field (an unusual type of QSPR model) is not only trained to reproduce heats of formation, but also, for instance, structures, isomerisation energies etc. Thus, a very general QSPR model that is trained to reproduce several directly related properties can be more reliable than the experimental values for one or more of these properties. We recently tested (17) the applicability of a single QSPR model to more than one property for a simple example, vapour pressures. Note that we have not yet trained the model using data for more than one property, but have only tested the possibility that a single model can be used for several properties, in this

case vapour pressure, boiling point and heat of vapourisation. Our first QSPR model for vapour pressure (15) used only data measured at 298K taken from the Beilstein database. This limits the training/test dataset to 551 compounds. Other authors (18) have corrected some of their experimental data to 298K using the published temperature dependence. This is a legitimate way to extend the available data. It is, however, not necessarily the best because it does not allow use of data for compounds for which the temperature dependence of the vapour pressure is not known.

Our approach (17) is to include the temperature of the measurement as an additional descriptor in the QSPR model. This forces the interpolation technique (in this case a feed forward neural net) to learn the temperature dependence as part of the model. We can then gain extra information by interrogating the trained net about this temperature dependence. In this case, the boiling point at atmospheric pressure can be calculated by finding the temperature at which the vapour pressure reaches this value and the heat of vapourisation can be derived using the Clausius-Clapeyron equation. Figure 3 gives an idea of exactly now much more data is available in this case than for a model limited to 298K.
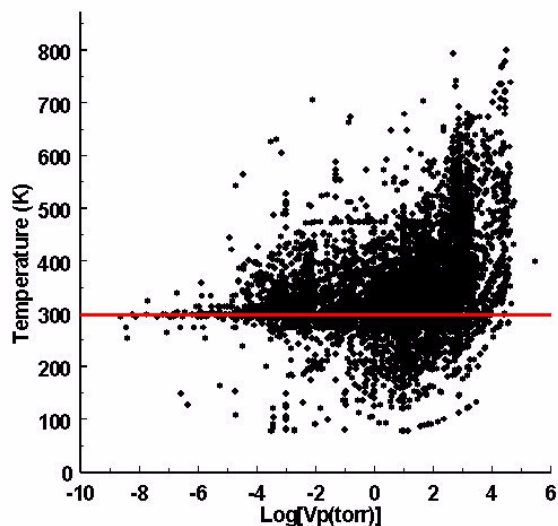

Figure 3. Distribution of the experimental data used to train a variable temperature vapour pressure model with respect to the vapour pressure itself and the temparature of the measurement. The horizontal line indicates the data available at 298 K.

Instead of the data for 551 compounds, we now have 8,542 data points at temperatures between 76K and 800K for 2,349 different compounds. Applying the model to boiling points gives a standard deviation between calculated values and experiment of 28.6K and a MUE of 18.7K for our boiling point dataset, compared with values of 21.9K and 13.5K, respectively, for a model

(19) trained only to reproduce boiling points. A small sample of heats of vaporisation also showed a standard deviation between calculation and experiment of only 4.7 kcal mol^{-1} and a maximum error of 12.0 kcal mol^{-1}.

This is, however, only an indication of what is possible. We must learn to make the most possible use of all the <u>reliable</u> data available. Martin Hicks has pointed out (20) that there are different types of experimental data. Boiling points, for instance, are measured routinely, and probably not very carefully, for every liquid organic compound reported in the literature. These data are not a good basis for a boiling point model because they are incidental properties used to characterise the compound. Our experience suggests that in many cases it is not recorded that the boiling point was measured at reduced pressure, or even that the temperature scale used (Celsius or Kelvin) is reported wrongly. There are, however, certainly data measured in studies whose main aim was to determine boiling points accurately. These data can all be used, even those at reduced pressure, because they are simply another way of reporting an experimental vapour pressure. Similarly, heats of vaporisation are usually measured very carefully for very pure compounds and so should be as reliable as is possible for measurements on a difficult quantity. Heats of vaporisation cannot be used directly to train a neural net by back-propagation because they require the slope of the vapour pressure change with temperature. However, training feed forward neural nets by genetic algorithms, rather than direct back-propagation, is an established technique (21) that allows us to use derived properties to determine the error function to be minimised as well as those generated directly by the neural net. We are now investigating the effectiveness of such an approach in which linked physical properties are used to generate more general, and thus more reliable, QSPR models (22).

Another potential use of information-rich molecular descriptors is to map chemical compounds according to their physical properties, as has been demonstrated by Oprea *et al.* (23) in their *"Chemical GPS"* technique. We have used this approach to investigate the clustering of druglike compounds on such a map, primarily in order to distinguish drugs from nondrugs, but as a more general goal to be able to relate new compounds to known closely related ones.

PHYSICAL PROPERTIES, DESCRIPTORS AND COMPOUND MAPS

The idea of mapping compounds according to, for instance, their physical properties is that physically similar but possibly chemically diverse compounds should occur close to each other, and thus have similar ADME properties etc. Thus, rather than conventional QSPR being used

to predict individual properties, compounds would be compared with their known neighbours and assumed to behave similarly. So far so good, but how do we decide which descriptors (and how many) to use for the mapping? In order to be able to treat this question rationally, we should at least have some idea of the dimensionality and the descriptors appropriate for describing physical property space. Lipinksi (24) has described physical property space as being low dimensional (i.e. we only need a few descriptors to describe physical properties). We investigated (13) both the dimensionality and the nature of physical property space by calculating a range of descriptors known to be suitable for QSPR models for the entire Maybridge database plus a set of about 2,500 selected drugs. The principle components of the 26 descriptors that appear in many of our QSPR models were then calculated in order to characterise physical property space. The conclusions of this study are that 8-9 descriptors are enough to describe physical properties and that these can be loosely classified as shown in Table 1.

Table 1. Qualitative descriptions of the principle components of descriptors used to describe physical property space (13).

Principle component number	% variance explained	Main descriptors	Interpretation
1	23.3	Polarisability, molecular weight, surface area, globularity	Size, shape
2	18.5	Maximum MEP*, mean positive and negative MEPs, total variance (25)	Complementary electrostatic surface descriptors
3	9.1	Minimum MEP, mean negative MEP, balance parameter (25)	
4	7.6	Total MEP-derived charges on nitrogens (26), number of H-bond acceptors	Complementary hydrogen-bonding descriptors
5	5.4	Total MEP-derived charges on H and O (26), minimum MEP, number of aromatic rings	
6	5.4	Dipole moment, dipolar density (27)	Dipolar polarity
7-9	3.9 – 4.3	Total MEP-derived charges on different types of atoms	Chemical diversity

* MEP = molecular electrostatic potential at the solvent-excluded surface.

The interpretations of the individual PCs give an intuitive picture of the factors determining the physical properties of molecules. Most important are the size and shape, followed by two complementary descriptors that describe the higher multipole character of the electrostatics at

the surface of the molecules. Note that the dipole moment is not important in these two descriptors. Next, come two complementary descriptors that essentially describe the hydrogen-bond donor and acceptor properties (including aromatic ring acceptors) and then the simple dipolar polarity, which perhaps surprisingly only accounts for just over 5% of the variance described by all the descriptors. PCs 7-9 are essentially atom counts that describe chemical diversity. These descriptors probably occur in QSPR models to correct for systematic AM1 or PM3 errors for some elements.

Thus, one appropriate approach to mapping chemicals according to their physical properties would be to use the first six principle components described in Table 1 as descriptors and to ignore the "chemical diversity" PCs in order to train, for instance, a Kohonen net. According to our analysis, the resulting map should cluster compounds with similar physical properties. This work is still in progress.

Descriptors can, however, also be selected to map for a more limited goal, such as distinguishing drugs from nondrugs (13). We have used recursive partitioning (28) In order to select descriptors for such a mapping, thus sacrificing some of the unsupervised quality of the Kohonen net by using a supervised descriptor selection process. The resulting map can not only distinguish drugs from nondrugs with about the same efficiency as other published techniques (29), but also differentiate between, for instance, hormones and other drugs (13).

The above applications use descriptors that are predominantly derived from the electron density, in our examples calculated using semiempirical MO-theory. However, it would be more efficient to use classical methods to optimise the molecular geometries and then a non-quantum technique to approximate the electron density. We are currently developing such a technique on which to base future, faster QSPR methods.

A NON QUANTUM MECHANICAL APPROACH TO ELECTRON DENSITY

The principle of electronegativity equalisation (30) enjoyed some popularity 20 years ago and is the basis of the still popular Gasteiger-Marsili charges (31). However, all such models known to us calculate isotropic net atomic charges, rather than considering the inherent atomic anisotropies caused by the bonding situation. We are currently developing a procedure (32) that considers this atomic anisotropy by considering hybrid atomic orbitals and their interactions. Currently, we parameterise the model to reproduce the AM1 electron density, but high level *ab*

initio or density functional data could also be used. Figure 4 shows a flow chart of the calculations steps involved.

The input to the program is a simple Lewis structure, which may be derived from a force-field calculation (in which case the bonds, hybridisation etc. are fully defined) or a set of 3D-coordinates, from which a Lewis structure must be derived using bond-distance criteria.
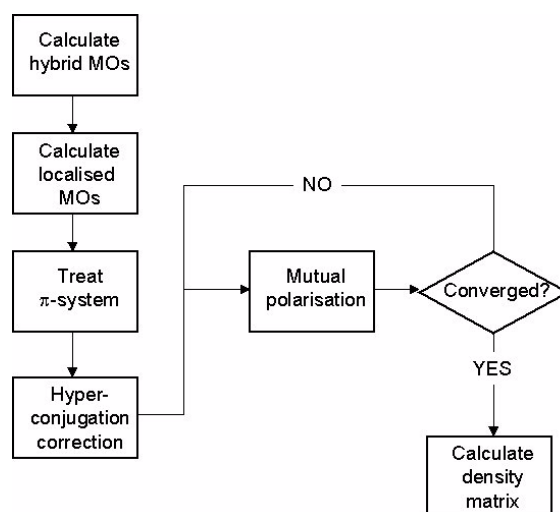

Figure 4. Flow chart of the classical procedure used to calculate electron densities (32).

The hybrid atomic orbitals are then calculated for each nonhydrogen atom using the procedure outlined previously for the hybrid orbital/point charge technique (33). The hybrid orbitals are then combined to bonding and antibonding localised molecular orbitals (LMOs) for the σ-framework and the lone pairs are identified. The remaining hybrid orbitals are considered components of the π-system, which is treated using a parameterised Hückel-like procedure with variable electronegativities. Negative hyperconjugation (donation from lone pairs into neighbouring σ^*-LMOs) (34) corrections are added specifically at this stage. The σ-system is then allowed to undergo a mutual polarisation step analogous to electronegativity equalisation, but based on the electrostatic potentials at the nuclei (35). This is an iterative procedure that usually converges within 4-10 cycles.

The initial parameterisation was restricted to compounds of the elements H, C, N and O and without π-systems. A training set of 52 representative compounds was calculated with AM1 and the geometries thus obtained used for the parameterisation. The error function was based on the one-atom (4×4) blocks of the AM1 density matrix. The error in the off-diagonal elements was weighted with a factor of 0.5 compared with the diagonal elements. The error function was minimised with either the simplex or the BFGS optimisation algorithms.

A validation set of 25 cycloalkanes, ethers, amines, alcohols, sugars and steroids was used to test the resulting parameters. The preliminary results are shown in Table 2.

Table 2: Results obtained for the validation set of 25 compounds. "RMS" is the root mean square deviation between the target (AM1) value and that calculated by the non-quantum mechanical procedure.

Property	Numbers	RMS
Diagonal density matrix elements	1774	0.029
Off-diagonal one-atom density matrix elements	1986	0.029
Coulson atomic charges	781	0.037

The quality of the fit can also be expressed in properties that are more familiar. The steroid **1**, for instance, is part of the validation set.

1
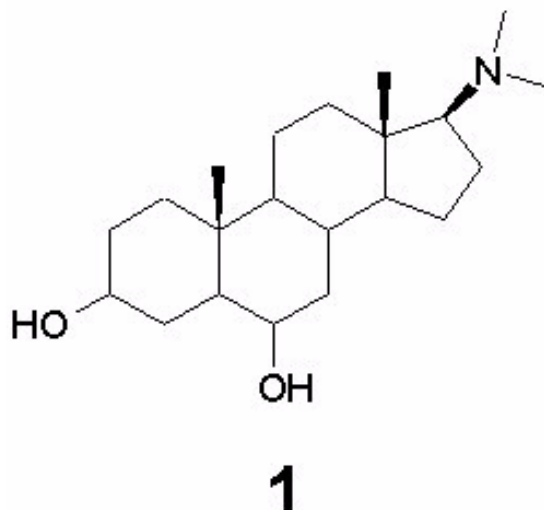
The parameterised procedure reproduces the AM1-calculated dipole moment with an error of 0.12 Debye, the root mean square of the Coulson net atomic charges is 0.027 and the root mean square deviation of the on-atom density matrix elements is 0.019. The molecular electrostatics of the molecule are thus well described by the new procedure, which requires only milliseconds of CPU-time. Figure 5 shows a plot of the molecular electrostatic potential at the solvent-excluded surface of the molecule. Only the areas with the largest deviation (below -5 kcal mol^{-1} and above 15 kcal mol^{-1}) are shown and the colour scale (blue to red) ranges from -9 kcal mol^{-1} to 19 kcal mol^{-1}.

The procedure outlined above would result in a vastly increased computational capacity for electron-density-based cheminformatics applications.
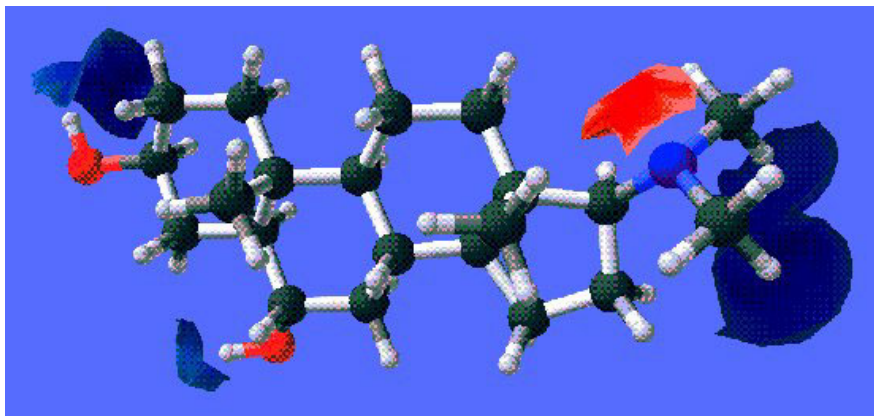
Figure 5. The molecular electrostatic potential at the solvent-excluded surface of **1**. Only the areas with the largest deviation (below –5 and above 15 kcal mol^{-1}) between the classical technique and the full AM1 calculation are shown. The colour scale (blue to red) ranges from –9 to 19 kcal mol^{-1}.

It is computationally very efficient, so that large databases or even complete enzymes can be treated easily. The electron density is also polarisable so that, for instance, the method could be used to calculate the electrostatics of the classical part of a hybrid QM/MM calculation without major inconsistencies in the electrostatic treatment of the classical and the quantum mechanical parts. Similarly, classically derived electron densities can be used as economical initial guess densities for MO-calculations. In this case, they have the advantage that no matrix diagonalisation is necessary, making the technique eminently suitable for parallel computers. A fast method for calculating accurate electron densities for proteins would also help the refinement of their X-ray structures.

CONCLUSIONS

Quantum mechanical methods, especially semiempirical MO-theory, can be used for cheminformatics applications. Advances in computer hardware have made semiempirical MO geometry optimisations on databases of 50-100,000 compounds commonplace on economical compute clusters. However, if we are to use the additional information provided by quantum mechanics relative to classical techniques, we must adopt a new paradigm for our QSPR and QSAR models, which are now often limited not by the descriptors, but rather by the quality of the training data. More general, physically rational models are needed that relate several physical properties to each other in order to eliminate biases or weaknesses in the training data for any one property. It is, for instance, unlikely that a dataset of heats of vapourisation will suffer from the same systematic problems as one for boiling points. High quality, possibly

quantum mechanical descriptors will be needed should such compound QSPR models prove successful.

Many of our current descriptors, however, only require the electron density, not a wavefunction. An extension of the well known electronegativity equalisation technique to the calculations of a detailed electron density may prove to offer the ideal compromise between the detail offered by quantum mechanical calculations and the computational efficiency of classical methods. Our initial model has demonstrated the viability of such techniques. It promises to be of very general use wherever a fast, relatively accurate calculation of the electron density is required. As the algorithm is inherently parallel, it can be used for very large systems and may even be suitable for use in a polarisable force field.

ACKNOWLEDGEMENTS

REFERENCES

[1]   Beck, B., Horn, A., Carpenter, J. E., Clark, T. (1998). *J. Chem. Inf. Comput. Sci*. **38**:1214.

[2]   *Quantum Cheminformatics: An Oxymoron?,* (Part 1) Published in *"Chemical Data Analysis in the Large: The Challenge of the Automation Age"*, M. G. Hicks (Ed.), *Proceedings of the Beilstein-Institut Workshop*, May 22nd - 26th, 2000, Bozen, Italy: http://www.beilstein-institut.de/bozen2000/proceedings

[3]   *Quantum Cheminformatics: An Oxymoron?,* (Part 2) T. Clark (2001). In *Rational Approaches to Drug Design*, H.-D. Höltje & W. Sippl (Eds), Prous Science, Barcelona.

[4]   Hennemann, M. & Clark, T. (2002). *J. Mol.Model*. **8**:95-101.

[5]   Dewar, M. J. S. & Thiel, W. (1977). *J. Am. Chem. Soc.* **99**:4899, :4907; Thiel, W. (1998). *Encyclopedia of Computational Chemistry*, P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer,III and P. R. Schreiner (Eds.), Wiley, Chichester, 1599.

[6]   Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., Stewart, J. J. P. (1985). *J. Am. Chem. Soc.* **107**:3902; Holder, A. J. (1998). *Encyclopedia of Computational Chemistry*, P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, and P. R. Schreiner (Eds), Wiley, Chichester, 8.

Clark, T.

[7] Stewart, J. J. P. (1989). *J. Comput. Chem.* **10**:209, :221; Stewart, J. J. P. (1998). *Encyclopedia of Computational Chemistry*, P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer,III and P. R. Schreiner (Eds), Wiley, Chichester, 2080.

[8] Rauhut, G. & Clark, T. (1993). *J. Comput. Chem.* **14**:503.

[9] Beck, B., Rauhut, G., Clark, T. (1994). *J. Comput. Chem*. **15**:1064.

[10] Rinaldi, D. & Rivail, J.-L. (1974). *Theoret. Chim. Acta*, **32**:57, :243; Rivail, J.-L. & Carter, A. (1978). *Mol. Phys*. **36**:1085.

[11] Schürer, G, Gedeck, P., Gottschalk, M., Clark, T. (1999). *Int. J. Quant. Chem.* **75**:17.

[12] Martin, B., Gedeck, P., Clark, T. (2000). *Int. J. Quant. Chem.* **77**:473.

[13] Brüstle, M., Beck, B., Schindler, T., King, W., Mitchell, T., Clark, T. (2002). *J. Med. Chem.,* in press.

[14] Beck, B. & Clark, T., manuscript in preparation.

[15] Beck, B., Breindl, A., Clark, T. (2000). *J. Chem. Inf. Comput. Sci*. **40**:1046.

[16] Yalkowsky, S. H. & Banerjee, S. (1992). *Aqueous Solubility*, Marcel Dekker, New York.

[17] Chalk, A. J., Beck, B., Clark, T. (2001). *J. Chem. Inf. Comput. Sci*. **41**:1053.

[18] McClelland, H. E. & Jurs, P. C. (2000). *J. Chem. Inf. Comput. Sci*. **50**:967.

[19] Chalk, A. J., Beck, B., Clark, T. (2001). *J. Chem. Inf. Comput. Sci*. **41**:457.

[20] M. Hicks, personal communication and comment at the Bozen Workshop (2002).

[21] Montana, D. J. & Davis, L. D. (1989). In *Proceedings of the International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Francisco; Mitchell, M. (1998). *An Introduction to genetic Algorithms*, The MIT Press, Cambridge, MA.

[22] Brüstle, M. & Clark, T., unpublished.

[23] Oprea, T. O. & Gottfries, J (2001). *ChemGPS: A Chemical Space Navigation Tool*. In *Rational Approaches to Drug Design: 13th European Symposium on QSAR*, H.-D. Höltje and W. Sippl (Eds), Prous Science, Barcelona, p. 437; Oprea, T. I. & Gottfries, J. (2001). *J. Comb. Chem*. **3**:157.

[24] Lipinski, C. A., Lombardo, F., Dominy, B. W., Feeney, P. J. (1997). *Avd. Drug. Delivery Rev.* **23**:3.

[25] Murray, J. S. & Politzer, P. (1998). *J. Mol. Struct. (Theochem)* **425**:107; Murray, J. S., Lane, P., Brinck, T., Paulsen, K., Grince, M. E., Politzer, P. (1993). *J. Phys. Chem.* **97**:9369.

[26] Beck, B., Clark, T., Glen, R. C. (1995). *J. Mol. Model.* **1**:176.

[27] Cronce, D. T., Famini, G. R., DeSoto, J. A., Wilson, L. Y. (1998). *J. Chem. Soc., Perkin Trans*. **2**:1293.

[28] Zhang, H. & Singer, B. (1999). *Recursive Partitioning in the Health Sciences*, Springer Verlag, Telos; Hawkins, D. M. http://www.stat.umn.edu/users/FIRM/.

[29] Sadowski, J. & Kubinyi, H. A. (1998). *J. Med. Chem.* **41**:3325; Wagener, M. & van Geersestein, V. J. (2000). *J. Chem. Inf. Comput. Sci.* **40**:280; Ajay; Walters, W. P. & Murcko, M. A. (1998). *J. Med. Chem.* **41**:3314.

[30] Sanderson, R. T. (1974). *Educ. Chem.* **11**:80.

[31] Gasteiger, J. & Marsili, M. (1978). *Tetrahedron Lett.* **34**:3181; Gasteiger, J. & Marsili, M. (1980). *Tetrahedron* **36**:3219; Marsilim, M. & Gasteiger, J. (1981). *Stud. Phys. Theor. Chem.* **16**:56; Marsili, M. & Gasteiger, J. (1981). *Croat. Chem. Acta* **53**:601.

[32] Horn, A. C. & Clark, T., unpublished.

[33] Gedeck, P., Schindler, T., Alex, A., Clark, T. (2000). *J. Mol. Model.* **6**:452.

[34] Schleyer, P. v. R. & Kos, A. J. (1983). *Tetrahedron* **39**:1141.

[35] Politzer, P. & Parr, R. G. (1974). *J. Chem. Phys.* **61**:4258; Politzer, P., Daiker, K. C., Trefonas, P., III. (1979). *J. Chem. Phys.* **70**:4400; Politzer, P. (1980). *Isr. J. Chem.* **19**:224; Politzer, P. & Sjoeberg, P. (1983). *J. Chem. Phys.* **78**:7008; Politzer, P. & Levy, M. (1987). *J. Chem. Phys.* **87**:5044; Erratum (1988). *J. Chem. Phys.* **89**:2590.