QUANTUM CHEMINFORMATICS: AN OXYMORON?

TIMOTHY CLARK

Computer-Chemie-Centrum, Universität Erlangen-Nürnberg, Nägelsbachstraße 25, D-91052 Erlangen, Germany. Tel. +49-(0)9131-8522948; Fax +49-(0)9131-8526565.
E-mail: clark@chemie.uni-erlangen.de

Received: 13th June 2000/ Published 11th May 2001

ABSTRACT

The use of semiempirical MO-theory for complete databases is demonstrated using the example of the Maybridge Chemical Company Database (53,000 compounds). 3D-Descriptors derived from the quantum mechanical wavefunction are used to set up QSPR-models using neural nets as the interpolation technique. Techniques for cross-validation of such models and for calculating individual error estimates for each compound are discussed. The examples are illustrated for properties such as logP, the vapor pressure, aqueous solubility and boiling points. The multi-net method of estimating individual error bars appears to give a good approximation of error limits of ± one standard deviation for several datasets.

INTRODUCTION

Until recently, quantum mechanics calculations were thought of as CPU-intensive and only applicable to perhaps tens of moderately sized (typically under 100 atoms) molecules within a reasonable cost in computer resources. The often described phenomenal increase in the performance of computer hardware has, however, been accompanied by a similar increase in the efficiency of quantum mechanics software, so that, for instance the geometry optimization of ascorbic acid with MNDO, [1] which took about 40 minutes CPU-time on a Convex C1 superminicomputer at the end of 1983, now takes only 5 seconds on an average PC under Windows NT. This, and the fact that most cheminformatics applications are inherently massively parallel through the trivial parallelization of calculating one molecule per processor, make quantum mechanical techniques applicable to tens of thousands of compounds within a single day, as we were able to demonstrate a few years ago. [2] This article is intended to describe the use and applications of semiempirical molecular orbital techniques (exclusively AM1 [3] and PM3 [4]) to complete databases and for the prediction of physical properties. Such techniques are equally well suited to the estimation of biological activity, but this will be the subject of a second article. [5] This article will concentrate on the advantages of using quantum mechanical, rather than classical mechanical, methods and on the derivation of robust, reliable and accurate quantitative structure-property relationships (QSPRs) with individual error estimation for each.

WHY QUANTUM MECHANICS?

Classical mechanical (force field) techniques employ a simple mechanical model of the molecular system. It is therefore not surprising that they do not do as good a job of describing properties that can be derived from the electron density of the molecule such as the molecular

Published in “Chemical Data Analysis in the Large: The Challenge of the Automation Age”, Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/clark/clark.pdf

electrostatics, polarizability, ionization potential etc. as quantum mechanical techniques that treat the electrons explicitly. This is illustrated be the molecular electrostatic potentials shown for guanine in Figure 1. Figures 1(a) and 1(b) show the solvent-excluded surface [6] of guanine color coded according to the electrostatic potential at the surface. The color scale is the same for the two figures. Figure 1(a), however, shows the quantum mechanically calculated molecular electrostatic potential (MEP), whereas Figure 1(b) shows the MEP obtained from an atomic multipole model in which the partial atomic charges were fitted to the quantum mechanical MEP using the VESPA technique. [7] Thus, Figure 1(b) represents almost the best approximation to the quantum mechanical results obtainable from an atomic monopole model (not quite the best as VESPA fits to charges outside the molecular surface).

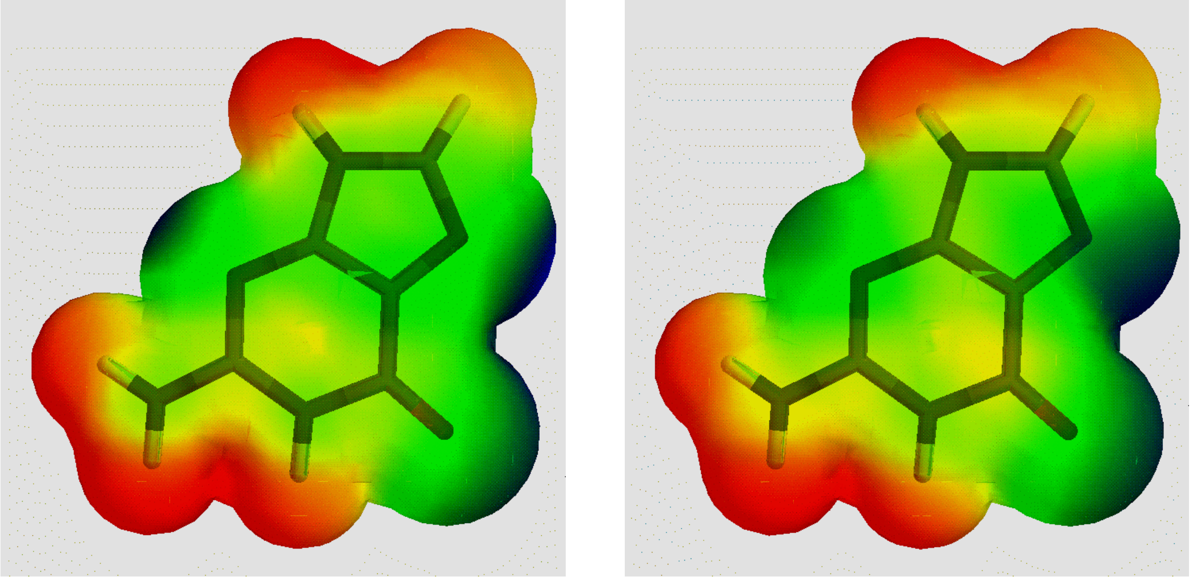

Figure 1: Color coded MEP-surface of guanine (red is positive, blue negative) calculated (left) using the NAO-PC technique [7] from the AM1 wavefunction and (right) using VESPA-derived [6] atomic monopoles.

Figure 2 shows the areas of the surface in which the difference between the two different MEPs is 10 kcal mol^{-1} or more. The surface is now color coded according to the difference in MEPs at the surface. Only the areas in which the absolute difference exceeds 10 kcal mol^{-1} are shown. Red indicates a positive difference and blue negative. The red circles indicate the nitrogen H-bond acceptor regions and the blue ellipse the H-bond acceptor region above the ring system.

The importance of the data illustrated by Figure 2 lies not in the magnitudes of the deviations, although these are significant, but in their positions, The largest concentrations of deviations between the two types of MEP lie at the two hydrogen-bond acceptor site on the ring nitrogens (marked by red

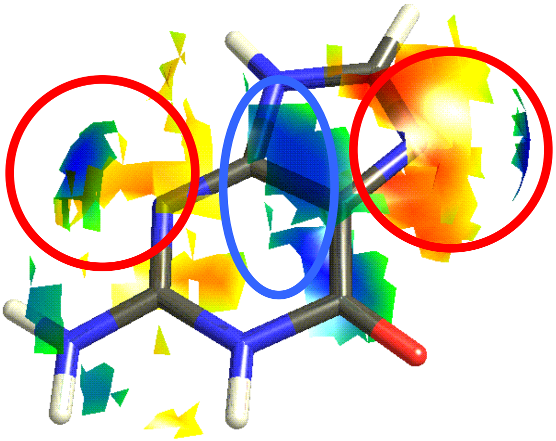

Figure 2: Difference {QM-monopole} of the two MEPs shown in Figure 1, again shown as a color-coded

circles) and at the H-bond acceptor site on the face of the ring system (marked by the blue ellipse). Thus, by projecting the quantum mechanical charge distribution onto an atomic monopole model we lose significant information exactly where it is important for intermolecular interactions.

Thus, we can expect that quantum mechanical methods should describe strong (electrostatic) intermolecular interactions better than atomic monopole based force field techniques. This is, however, not the only advantage of quantum mechanical techniques. Properties such as polarizability, ionization potentials, electron affinities, multipole moments etc. are readily available. Descriptors based on these properties can be expected to play a significant role in QSPRs designed to predict common physical properties.

The Molecular Polarizability

Apart from the often dominant and longrange electrostatic interactions, weak intermolecular forces (dispersion) play a major role in determining intermolecular interactions. [8] In order to treat these forces, which dominate for intermolecular interactions between nonpolar molecules, correctly, we need to be able to calculate the molecular electronic polarizability accurately. There are several types of calculational technique available for calculating the polarizability from the molecular wavefunction, but most are too unwieldy to be used routinely for applications on complete databases. Among these are the finite field perturbation method, [9] which, however, is compute-intensive and requires a large, flexible basis set in order to give good results, and the perturbational sum-over-states (SOS) technique. [10] The latter, however, requires a configuration interaction calculation in order to obtain the excited states and is therefore also very compute-intensive. The SOS-method does, however, have the advantage that it can give frequency-dependent polarizabilities.

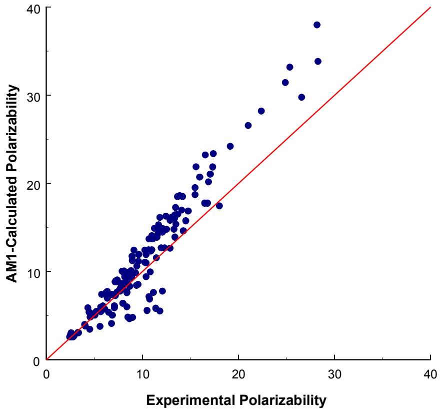

Figure 3: Calculated [11] and experimental molecular electronic polarizabilities (Å^3) using the original variational technique [10] with AM1.

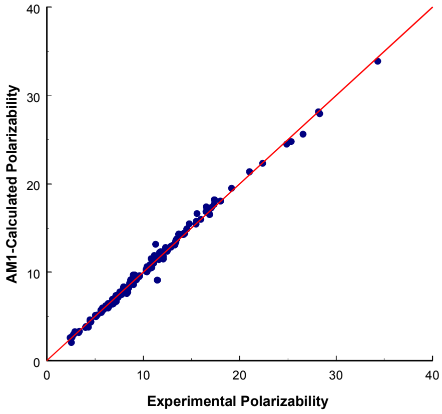

Figure 4: Comparison of calculated and experimental molecular electronic polarizabilities (Å^3) using the parametrized variational technique [11] with AM1.

A more computationally tractable technique that we have used for some years is the variational method developed by Rivail and his coworkers. [11] This technique requires only some multipole integrals and the density matrix and can therefore be appended to a normal SCF-calculation much like a population analysis and without increasing the time of the calculation significantly. Figure 3 shows the results of such calculations with AM1 for a test set of organic molecules.

The results show a systematic deviation for the larger molecules and a significant scatter for the smaller ones. The standard deviations between calculation and experiment for MNDO, AM1 and PM3 are 1.94, 2.99 and 4.44 Å^3, respectively.

Thus, although the original variational method fulfils the computational requirements for use in a cheminformatics application, it is not accurate enough. In order to remedy this situation, we developed a parameterized variational technique. [12]

If the multipole integrals, which are normally a function of the Slater exponents and ordinal numbers, are treated as variable parameters and the optimized for a set of 156 organic molecules, the results shown in Figure 4 are obtained for the independent test set of 83 organic molecules also shown in Figure 3.

The standard deviations between calculation and experiment for MNDO, AM1 and PM3 are now 0.78, 0.70 and 0.74 Å^3, respectively. Thus, the parameterized variational method offers a computationally economical and accurate method for determining molecular electronic polarizabilities. It also offers the advantage that, with certain restrictions, it can be partitioned into atomic polarizability tensors, which, although not physically measurable, are particularly useful for additive, atom-atom dispersion models.

AM1 Optimizations for a Complete Database

The computational software must fulfill two conditions for a semiempirical technique such as AM1 or PM3 to be applied to a database of perhaps hundreds of thousands of compounds. It must be fast and it must be extremely reliable. Perhaps surprisingly in the light of the introduction, speed is not really a problem. Database applications can use the full power of massively parallel architectures, or even of large compute clusters with relatively slow communication. This is of course because the computational effort per molecule is relatively large and data transfers relatively small and seldom. We reported [1] a benchmark application of AM1 to the Maybridge database [13] a few years ago. The computational protocol necessary to process a 2D-database like Maybridge is shown in Table 1.

Table 1: Processes, software and failure rates for processing the Maybridge database. [1]

Process	Software	# of failures
Data cleanup	**SDFClean [14]**	**211**
2D → 3D Conversion	**CORINA [15]**	**41**
AM1 optimization	**VAMP [16]**	**68**
Generate descriptors	**PROPGEN [17]**	**0**
Apply models	**PROPHET [18]**	**0**

The data cleanup process is necessary because, even if each structure were entered perfectly, the structures needed for quantum mechanical calculations are not necessarily those entered in databases. Ion pairs, for instance, may be entered as covalently bound structures, free base plus counterion, or in other less standard ways. Because generally the counterion is not considered in quantum mechanical calculations, it must be eliminated and the correct protonation site determined if the free base is entered. Finally, it is also necessary to check that the structures entered in the database make chemical sense. This process resulted in 211 compounds from Maybridge being marked for manual processing, mostly because the exact site of protonation was not absolutely clear. We note here that for many applications it may be preferable to calculate the free base, or even both the base and its conjugate acid.

The 2D to 3D conversion process has been discussed in detail before [19] and will therefore not be treated here. We used CORINA [15] for the Maybridge run, which resulted in only 41 failures.

The optimization of the molecular geometries with AM1 or PM3 is the most time-consuming step in the entire process. This was performed in parallel (one molecule per processor) on a 128-processor Silicon Graphics Origin 2000. At the time of the run, two processors were defective, giving a total number of processors used of 126. The details of this run have been published, but the essence is that the molecules in the database were optimized within 14 hours elapsed time with only 68 failures. [2] We have since repeated this run several times on distributed moderately parallel machines and on heterogeneous UNIX/Windows NT® clusters with excellent results. Using a Compaq-Alpha two-processor server, a Hewlett-Packard four-processor server and two Intel-based two-processor Windows-NT® machines, for instance, Maybridge can be processed in a weekend. [20]

The descriptors necessary to calculate physical properties can be calculated from the complete electrostatic information stored in the database in a relatively fast step (the most time-consuming task is to generate the potential-derived charges using the VESPA-technique [20]). Finally, the descriptors generated, which are added to the molecular description in the database, are used to calculate properties such as logP [21], the vapor pressure at 25° [22] or the aqueous solubility. [23]

What Factors are Important in QSPR-Models?

Figure 5 shows an overview of typical QSPR-techniques.

The yellow boxes indicate the descriptors used to characterize the molecule. These may be atoms or groups, in which case the interpolation technique used (colored light blue) consists of a set of increments. Such atom- or group-additive methods assume that such increments are transferable and are best suited for properties where this is most likely to be true, such as heats of formation [24] or ^{13}C-chemical shifts. [25] There are a large variety of 2D-descriptors such as, for instance, the range of Kier and Hall indices, [26] although there are very many others. These indices are remarkably successful in treating a large number of properties. They have the advantage that they treat the molecular conformation, if at all, implicitly, so that there is no requirement to locate the most stable conformation or even perform a Boltzmann averaging over a number of conformations. 3D-descriptors, which will be used in the work described here, are derived from the molecule at a given geometry. They are often calculated from the electron density given by quantum mechanical calculations, but this must not be the case. Many descriptors, such as those introduced by Politzer and Murray, [27] describe a property such as the electrostatic potential at the molecular surface. 3D-descriptors are, however, conformationally dependent. This is in principle an advantage, but in practice practically always a disadvantage. This is because the search for the global conformational minimum or a representative set of stable conformations is an extremely compute-intensive task for molecules with a large number of rotatable bonds. Thus, many QSPR-models based on 3D-descriptors actually only use one conformation. This point will be discussed below. Table 2 shows

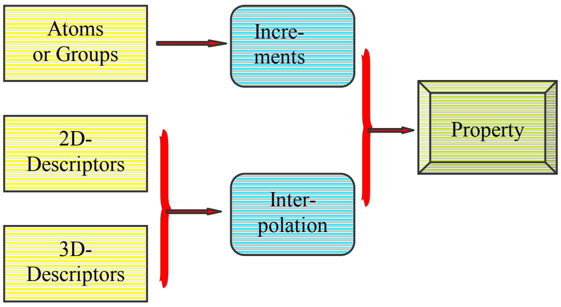

Figure 5: The typical features of QSPR models

the main characteristics of the different types of descriptors:

Table 2: The principal characteristics of different types of molecular descriptors.

Increments	2D-descriptors	3D-descriptors
Fast	fast	can be slow
not universally applicable	general	General
best for additive properties (heats of formation, chemical shifts)	good for many properties	good for properties involving intermolecular interactions
no conformational information	treats conformation implicitly (?)	conformationally dependent

The most traditional interpolation technique is a regression analysis in some form. Alternatives include nearest neighbor techniques, in which the property in question is estimated from those of the most similar known molecules, and artificial neural nets. When used carefully, the latter are extremely powerful but, like all interpolation techniques, they are open to misuse and can simulate a far better performance than they can actually deliver. This leads to a set of requirements for the interpolation used in a QSPR model:

The model should be well validated. This is typically done by some sort of cross-validation procedure in which the predictive ability of the technique, rather than its ability to reproduce known results, is assessed.

The second requirement is that the technique should be as robust as possible. This requirement is often translated as meaning that the model should give a small standard deviation from experimental values for a wide variety of compounds. I suggest, however, that the largest observed error is the most indicative variable for a the robustness of a QSPR-model. The largest likely error is a quantity that defines the reliability of the model for many experimentalists.

Leading from the requirement for robustness is the further desirable feature that the QSPR-model should be able to assess the likely reliability of its prediction *for each individual compound*. Clearly, the properties of s compound that is similar to many in the training set will be predicted more reliability than for one that lies outside its range. The ideal model should not only give its predicted value, but also its estimated error limits.

Quantum Mechanical/Neural Net QSPR-Models

We have in recent years developed a series of QSPR-models based on 3D-descriptors derived from semiempirical MO-calculations and using simple feedforward neural nets with one hidden layer as the extrapolation technique. The general scheme of such techniques is shown schematically in Figure 6.

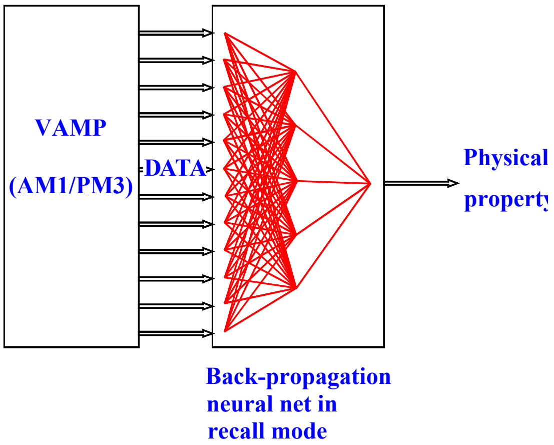

Figure 6: Schematic view of a typical QM/NN-QSPR-model.

However, such simple models do not usually satisfy the general conditions for a good QSPR-model given above. We must therefore address the questions of cross-validation and individual error estimates.

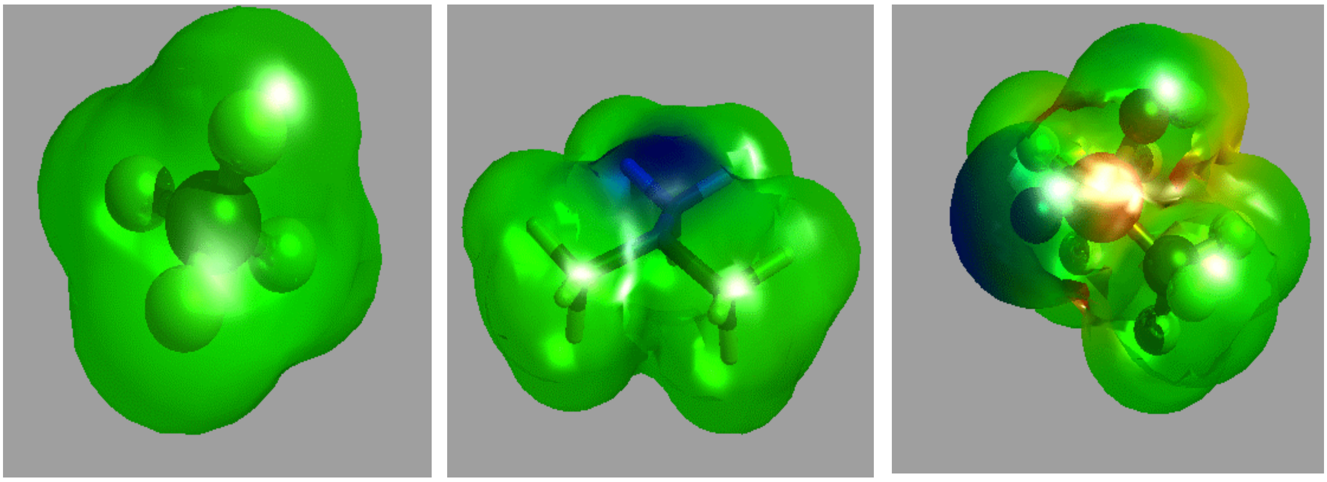

Figure 7: Molecular electrostatic potential surfaces for (from left to right) methane (total variance =5.4, balance parameter = 0.144), trimethylamine (total variance = 446.6, balance parameter = 0.009) and *bis*-(trifluoromethyl)phosphinic acid (total variance = 651.0, balance parameter = 0.246)

We [22] have approached cross-validation by dividing the dataset into about 10 equal, random portions and training 10 separate nets, each using one of the random portions as a test set. This results in 10 different nets, all of which use the same descriptors but which all have different test and training sets. The mean of the results of the 10 nets is used as the predicted value for the model as a whole and the results of the nets for which the compound in question was in the test set are used for cross-validation. In this way, cross-validated results are obtained for each compound in the dataset for a neural net in which it was a part of the test set.

The descriptors used for the QM/NN-models are often those introduced by Politzer and Murray for density functional calculations using the isodensity molecular surface. [27] We use semiempirical MO-theory with the NAO-PC model [28] for the molecular electrostatic potential at the solvent-excluded surface [6] of the molecule. Briefly, Politzer and Murray descriptors describe the statistics of the electrostatic potential distribution at the surface of the molecule. Figure 7 shows some illustrative examples. Methane is essentially nonpolar with very little variation of the electrostatic potential. This leads to a very low variance (5.4). Trimethylamine exhibits an area of negative potential due to the lone pair. This results in a higher variance (446.6) but, because there is no equivalent positive area, a very low balance parameter (0.009). The far more polar *bis*-(trifluoromethyl)phosphinic acid, with both positive and negative areas on the electrostatic potential surface, has an even higher total variance (651.0) and also a high balance parameter (0.246). Such descriptors were designed to describe the intermolecular electrostatic interactions. They have been used in all our QSPR models that estimate physical properties that depend on intermolecular forces. Table 3 shows the parameters used for our published logP model. [21]

These descriptors, of which the sums of the ESP-derived charges probably function as extended atom-counts, can all be linked to logP conceptually.

It is noteworthy that the molecular polarizability and the molecular volume, parameters that are generally very strongly correlated, are both necessary in order to generate a reliable model. Figure 8 shows the results obtained using the cross-validation technique described above.

Table 3: Descriptors used for logP. [21]

Descriptor	Definition
α	Molecular polarizability
μ	Dipole moment
A	Molecular surface area (SES)
V	Molecular volume
N_{sum}	Sum of ESP-derived charges on N-atoms
O_{sum}	Sum of ESP-derived charges on O-atoms
P_{sum}	Sum of ESP-derived charges on P-atoms
S_{sum}	Sum of ESP-derived charges on S-atoms
X_{sum}	Sum of ESP-derived charges on halogens
V_{max}	Maximum MEP at the SES
V_{min}	Minimum MEP at the SES
M_{+}	Mean positive MEP at the SES
M_{-}	Mean negative MEP at the SES
σ^{2}_{tot}	Total variance of the MEP
ν	Politzer/Murray balance parameter
G	Globularity [29]

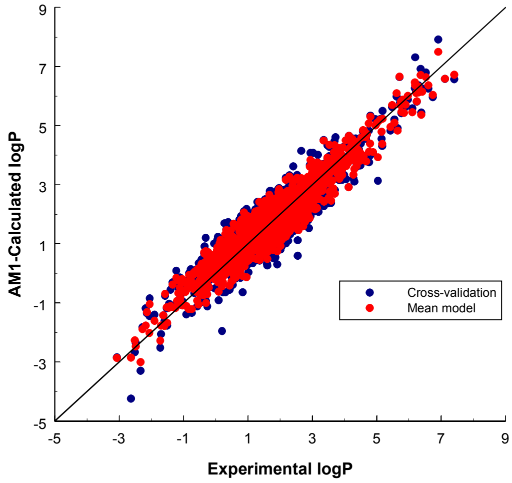

Figure 8: Mean and cross-validated results for the logP model. [21]

Table 4 gives the performance of the mean model and the cross-validation.

The above model appears to be robust as the cross-validation results are comparable to those of the mean of the ten nets. It does not yet, however, give error estimates for individual compounds.

In order to be able to assess individual errors, we [22] calculated the standard deviations of the 10 net predictions for each compound. In principle, the larger the disagreement among the 10 nets, the less reliable should be the predicted value. If now the absolute difference between the calculated (mean model) and experimental value for each compound is divided by the standard deviation of the 10 net predictions for that compound, we obtain the histogram shown in Figure 9.

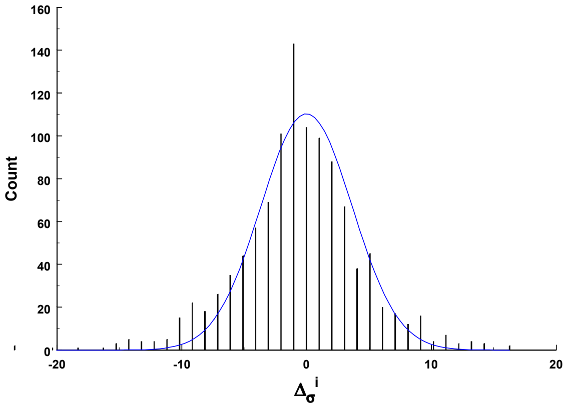

Figure 9: Histogram of the experimental errors in units of the standard deviations of the predictions of the 10 nets for the logP model. [21, 22]

Table 4: Analysis of the mean model and the cross-validation results for the logP model.

Parameter	Mean model	Cross-validation
Std. dev	0.47	0.56
Max. error	1.21	2.15
r^2	0.91	0.87
slope	1.01	0.97
intersect	0.01	0.06

The mean absolute value of the deviation in units of the individual standard deviation for each compound is 3.58. We therefore suggest that an intuitively reasonable error estimate for each compound is simply the product of the standard deviation of the net predictions times this mean deviation for the training dataset. [22] If we calculate the error bars in this way for the logP model, we obtain the data shown in Figure 10.

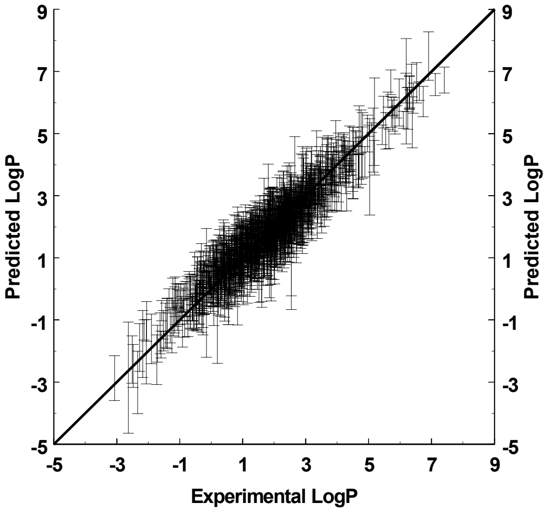

Figure 10: Performance of the logP model with error bars. [21, 22]

This results in 408 compounds (37%) with errors outside the error bars, which corresponds fairly closely to an error estimate of ± one standard deviation. Two questions remain. Is this behavior general for all models and how appropriate are the error bars for completely unseen data?

In order to answer the latter question, we investigated the dataset of nucleotides published bay ACD-labs. [30] These data are not only outside our dataset, but also apply to a class of models explicitly excluded from our data because of the ambiguity of the exact form of the compounds in different media. The results obtained are shown in Figure 11.

Table 5: Performance of three QM/NN-QSPR models.

	Aqueous solubilty	Vapor pressure	Boiling point
Reference	[23]	[22]	[31]
Units	Log (solubility)	Log (vapor pressure)	°C
Number of compounds	559	551	6,000
Std. dev.	0.51	0.29	16.5
mean unsigned error	0.40	0.22	11.8
maximum error	1.67	1.00	-119
r²	0.90	0.94	0.96
slope	1.03	1.01	1.01
intersect	0.08	-0.01	-4.6
mean Δ	2.11	2.98	2.15
compounds outside the error bar	201 (35%)	199 (36%)	2244 (37%)

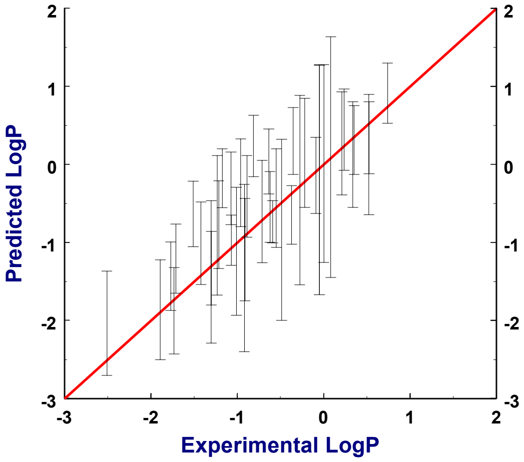

Figure 11: LogP results obtained for the nucleotide dataset. [22, 30]

In this case only 8 compounds (20%) are outside the error bars. This, however, is an anomalous result probably caused by the very low diversity of the dataset, as will be seen in the following examples. Table 5 shows the statistics of the results obtained for three further models, aqueous solubility [23], vapor pressure at 25° [22] and boiling points at atmospheric pressure. [31] In all cases, the error estimates given by the multi-net technique described above are close to those expected from error bars of ± one standard deviation, confirming the hypothesis that the multi-net technique as described gives reliable error estimates.

Thus, the purely empirical technique of error estimation appears to give reliable results for a variety of QSPR-models and can help to point to compounds for which the neural nets are attempting to extrapolate outside the range of their training sets.

The Effect of Conformational Changes

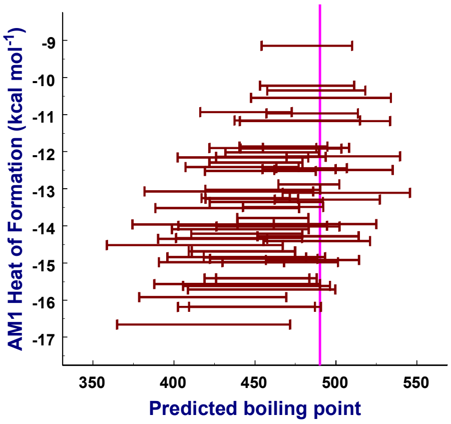

Figure 12: Calculated boiling points for different conformations of bis-(2-aminoethyl)amine plotted against the heat of formation of the individual conformers.

The above models used only one molecular conformation per molecule – that obtained from the Corina-calculated structure after AM1-optimization with VAMP. In principle, models based on 3D-descriptors such as these should be able to describe conformational effects on the property. However, the available data, most of which is for flexible compounds, does not provide us with the necessary experimental resolution to be able to produce a conformationally dependent model. We thus rely on the standard computational protocol to provide us with reasonable conformations. How does this affect the results, however? In order to investigate this effect, we [31] calculated all the minimum energy conformations of *bis*-(2-aminoethyl)amine using the systematic torsional search facility in VAMP. The boiling point model was then applied to each of these conformations, some of which, for instance, contain internal hydrogen-bonds. The results are shown in Figure 12.

In general, the fluctuations in the calculated boiling point are of the same order as the error estimate. The Boltzmann-averaged calculated boiling point is 444±36°, compared with an experimental value of 480°. We therefore feel justified in using the present single conformation approach.

Summary and Conclusions

The techniques described here have demonstrated the applicability of quantum mechanical techniques to cheminformatics. Surprisingly for some, the CPU-requirements are not the major disadvantage of such techniques, but rather the lack of reliable and consistent experimental data and, to some extent, the limitations of current semiempirical methods. For some properties such as aqueous solubility, the published experimental data is too sparse and too noisy to produce a first class QSPR-model. In any case, the available data do not usually allow us to produce a conformationally-dependent model, although normal boiling points may be an exception to this rule. Modern techniques allow us to store essentially the entire electrostatic and polarizability information about a molecule as well as a host of other quantum mechanically derived parameters, so that an amazingly complete description of the molecules is now available form databases of this type.

Just as the work reported here was impossible at the time of the first Beilstein Workshop (1988), so will the techniques described here be superseded in ten years time? A prime requirement is a semiempirical MO-method that does not suffer the weaknesses of the current techniques for heavy atoms, hydrogen bonds, branching errors and weak interactions. We are currently developing such a technique, which should then provide an even better description of the molecules. However, the “magic limit” of about ±0.5 log units mean error for QSPR-models of physical properties is only likely to be lifted when large (10^3-10^4) numbers of consistent and accurate datapoints become available.

ACKNOWLEDGEMENTS

I thank my postdocs and students, Dr. Bernd Beck, Dr. Andrew Chalk, Dr. Andreas Breindl, Dr. Peter Gedeck, Gudrun Schürer, Maik Gottschalk, Bodo Martin, Matthias Hennemann and Michael Will for their constant support and constructive contributions to this work, which was supported by the Deutsche Forschungsgemeinschaft and Oxford Molecular.

LITERATURE AND NOTES

[1] Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.*, 1977, **99**, 4899; 4907; Thiel, W., *MNDO*, in *Encyclopedia of Computational Chemistry*, Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F. III; Schreiner, P. R. (Eds), Wiley, Chichester, 1998, **3**, 1599.

[2] Beck, B.; Horn, A.; Carpenter, J. E.; Clark, T. *J. Chem. Inf. Comput. Sci*. 1998, **38**, 1214.

[3] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P.; *J. Am. Chem. Soc*., 1985, **107**, 3902; Holder, A. J.; *AM1* in *Encyclopedia of Computational Chemistry*, Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F. III; Schreiner, P. R. (Eds), Wiley, Chichester, 1998, **1**, 8.

[4] Stewart, J. J. P.; *J. Comput. Chem*., 1989, **10**, 209; 221; Stewart, J. J. P. *PM3*, in *Encyclopedia of Computational Chemistry*, Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F. III; Schreiner, P. R. (Eds), Wiley, Chichester, 1998, **3**, 2080.

[5] Clark, T. *QSAR 2000; proceedings of the 13th European Symposium on QSAR*, to be published.

[6] Pascal-Uhuir, J. L.; Silla, E.; Tuňon, I., *J. Comput. Chem.* 1994, **15**, 1127.

[7] Beck, B.; Clark, T.; Glen, R. C.; *J. Comput. Chem.*, 1997, **18**, 744.

[8] Stone, A., *The Theory of Intermolecular Forces*; Vol. 32 of *International Series of Monographs in Chemistry*; Oxford University Press, Oxford, 1996.

[9] Kurtz, H. A.; Stewart, J. J. P.; Dieter, K. M., *J. Comput. Chem*. 1990, **11**, 82; Cardelino, B. H.; Moore, C. E.; Stickel, R. E., *J. Phys. Chem.* 1991, **95**, 8645.

[10] Docherty, V. J.; Pugh, D.; Morley, J. O. *J. Chem. Soc. Faraday Trans. 2*, 1985, **81**, 1179; Zamini-Khamiri, O.; Hameka, H. F. *J. Chem. Phys.*, 1979, **71**, 1607.

[11] D. Rinaldi and J.-L. Rivail, *Theoretica Chmica Acta* 1974, **32**, 243; J.-L. Rivail and A. Carter, *Mol. Phys.* 1978, **36**, 1085.

[12] G. Schürer, P. Gedeck, M. Gottschalk and T. Clark, *Int. J. Quant. Chem*. 1999, **75**, 17.

[13] Maybridge Chemicals Company Ltd., Trevillet, Tintagel, Cornwall PL34 OHW, England.

[14] Beck, B. Oxford Molecular, 1999.

[15] Sadowski, J.; Gasteiger, J. *Corina v. 1.8*, Oxford Molecular, Medawar Centre, Oxford Science Park, Oxford, OX4 4GA, UK.

[16] Clark, T.; Alex, A.; Beck, B.; Chandrasekhar, J.; Gedeck, P.; Horn, A.; Hutter, M.; Martin, B.; Rauhut, G.; Sauer, W.; Schindler, T.; Steinke, T. *Vamp 7.0*, Oxford Molecular, The Medawar Centre, Oxford Science Park, Sandford-on-Thames, Oxford OX4 4GA, United Kingdom, 1999/2000.

[17] Beck, B.; Burkhardt, F.; Clark, T., *Propgen 1.0*, Oxford Molecular, The Medawar Centre, Oxford Science Park, Sandford-on-Thames, Oxford OX4 4GA, United Kingdom, 1999.

[18] Beck, B.; Burkhardt, F.; Clark, T., *Prophet 1.0*, Oxford Molecular, The Medawar Centre, Oxford Science Park, Sandford-on-Thames, Oxford OX4 4GA, United Kingdom, 2000.

[19] Sadowski, J.; Gasteiger, J. *Chem. Rev.* 1993, **93**, 2567.

[20] Beck, B. unpublished results

[21] Breindl, A.; Beck, B.; Clark, T.; Glen, R. C. *J. Mol. Model* 1997, **3**, 142.

[22] Beck, B.; Chalk, A.; Clark, T. *J. Chem. Inf. Comp. Sci*., in the press.

[23] Beck, B.; Clark, T. unpublished.

[24] Benson, S. W. *Thermochemical Kinetics* 2nd ed.; Wiley: New York, 1976; Clark, T.; McKervey, M. A. Saturated Hydrocarbons in *Comprehensive Organic Chemistry*, Barton; D. H. R. and Ollis, W. D. Eds.; Pergamon Press: Oxford, 1979, Volume 1, Chapter 2, 37-120.

[25] Kalinowski, H.-O.; Berger, S.; Braun, S. *Carbon-13 NMR Spectroscopy*, Wiley, Chichester, 1988.

[26] Hall, L. H.; Kier, L. B. *Reviews in Computational Chemistry*, Lipkowitz, K. B; Boyd, D. B. (Eds), VCH, Weinheim, 1990, p. 367; Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1976; *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press, Wiley, Letchworth, England, 1986.

[27] Murray, J. S.; Lane, P.; Brinck, T.; Grice, M. E.; Politzer, P. *J. Phys. Chem.* 1993, **97**, 9369

[28] Rauhut, G.; Clark, T. *J. Comput. Chem*., 1993, **14**, 503; Beck, B.; Rauhut, G.; Clark, T. *J. Comput. Chem.*, 1994, **15**, 1064.

[29] Meyer, A. Y. *Chem. Soc. Rev*. 1986, **15**, 449.

[30] http://www.acdlabs.com

[31] Chalk, A.; Beck, B.; Clark, T. *J. Chem. Inf. Comput. Sci*., 2000, **40**, 1046.