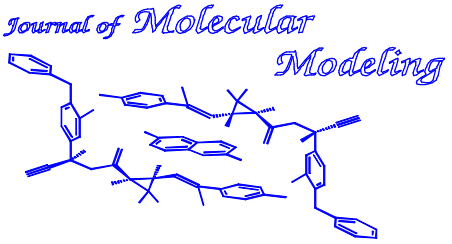
J. Mol. Model. **1995**, *1*, 22–35



A Combined Semiempirical MO/Neural Net Technique for Estimating ^{13}C Chemical Shifts

Timothy Clark,* Guntram Rauhut and Andreas Breindl

Computer-Chemie-Centrum des Instituts für Organische Chemie der Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstraße 25, D-91052 Erlangen, Germany. (clark@organik.uni-erlangen.de)

Received: 3 December 1994 / Accepted: 5 January 1995

Abstract

A back-propagation artificial neural net has been trained to estimate ^{13}C chemical shifts from the results of AM1 and PM3 semiempirical MO calculations. The input descriptors include the atom-centered monopole, dipole and quadrupole moments derived from the natural atomic orbital/point charge (NAO/PC) model, the four highest bond orders to the carbon atom being considered and the elements to which these bonds are made. The resulting net estimates the chemical shifts of a test set of 156 chemical shifts with a standard deviation of less than 7 ppm from the experimental values for AM1 and slightly more for PM3.

Keywords: ^{13}C Chemical Shift, Semiempirical MO, AM1, PM3, Neural Net

Introduction

Calculations of NMR chemical shifts using *ab initio* molecular orbital theory [1] have recently proven to be a powerful tool in the assignment of otherwise unknown structures [2]. However, the *ab initio* techniques used are still not suitable for rapid, everyday screening of large molecules on low-end hardware. Semiempirical MO-theory is several orders of magnitude faster than even low-level *ab initio* theory and can be used routinely for far larger molecules, but explicit GIAO-MNDO calculations of ^{13}C chemical shifts proved to be of limited accuracy and to need a reparametrization of the MNDO method [3]. In this work we consider an alternative approach that we have already used successfully to estimate esr hyperfine coupling constants [4]. Rather than trying to produce a physical model for the desired property, we use a simple back-propagation artificial neural net as a model-free device to derive the property in question from a series of related calculated properties. Thus, such a net can be trained to reproduce experimental esr coupling constants given a series of calculated spin densities and charges [4]. In a way, such methods represent the ultimate in "black box" technology as usually not even the programmer knows how the program arrives at the answer. Nevertheless, within the philosophical framework of semiempirical MO theory, techniques that relate the results of the calculations to commonly observed spectroscopic properties are of immense value to experimental chemists. We note here that the accuracy requirements for a theoretical method that allows complete assignment of a complicated spectrum are of the order of tenths of a ppm if very similar resonances are to be assigned uniquely, and that this sort of absolute accuracy is well beyond what we can expect from any existing calculational method, although cancellation of errors makes the situation much better for closely related carbon atoms. Our aim in this work is to produce a method fast enough that it can be applied routinely at the end of every semiempirical optimization and accurate enough to give a reasonable representation of the spectrum of the real compound. The theory should also be completely independent of any experimental data about the compound in question. This means that the entire calculational process, including

**To whom correspondence should be addressed*

determining the geometry, should be carried out using semiempirical theory.

The choice of descriptors (input data) for a net designed to estimate ^{13}C chemical shifts is wide. The most obvious choice are atomic charges of some sort, as used in Spiesecke-Schneider type relationships [5], although it is clear that such relationships apply only to planar π-systems and are not universally applicable. Quite generally, we expect that descriptors of the electron density around a given carbon atom will be needed to describe the diamagnetic contribution to the chemical shift [6] and that excitation energies or related quantities will be needed to describe the paramagnetic contribution. Rather than use excitation energies (which would need a configuration interaction calculation) directly, we chose to use the approach suggested by Karplus and Pople [7] in which the paramagnetic contribution is related to atomic charges and bond orders. Karplus and Pople assumed an average excitation energy, but we have also defined the atom to which the bond is made in order to allow the net to judge the magnitude of the various paramagnetic contributions.

Calculational Method

All calculations used the VAMP 5.5 program [8] on a Convex C-220/256, Hewlett-Packard 735 and Silicon Graphics Indigo workstations**.** Standard AM1 [9] and PM3 [10] parameters were used throughout. Geometries were optimized until the gradient norm was less than 0.4 kcal mol^{-1} $Å^{-1}$**.** Bond order calculations used the formalism proposed by Perkins and Stewart, [11] as also implemented in MOPAC and AMPAC.

Atom-Centered Multipoles

One of the most compact and useful ways of describing the electronic environment around an atom is a distributed multipole analysis [12] in which the multipoles are centered on the atoms themselves. We recently introduced the Natural Atomic Orbital/Point Charge (NAO-PC) model [13,14] for representing the electron density of a molecule in terms of an extended point charge model within semiempirical MO-techniques. This model, in which the electron density is represented by point charges situated at the centers of charge of the individual lobes of the natural atomic orbitals, is in effect a distributed multipole model and can be converted to an atom-centered multipole description very easily.

The calculation of the positions and sizes of the point charges has been described in detail elsewhere [13,14]. We note here, however, that very small charges far from the nucleus, which were neglected by our original cutoff procedure [13], play an important role in determining the molecular multipole moments exactly, and that the cutoff procedure was therefore abandoned for this work. Because only the nonhydrogen atoms are represented by an array of nine charges, we have limited ourselves to an atom-centered multipole analysis in which the hydrogens are represented as monopoles (which are identical to Coulson charges [15]) and the nonhydrogen atoms by a monopole, dipole and quadrupole. We also make the simple approximation that the multipoles derived for each atom are those resulting from the array of nine charges associated with the one-atom block of the density matrix for that atom. Using this assumption with the Born-Oppenheimer approximation and Buckingham's definition of the quadrupole moment [16] we obtain

$$Q_{kl} = \frac{1}{2}\left\{\sum_{\alpha} Z_{\alpha}\left(3C_{k.\alpha}C_{l,\alpha} - R_{\alpha}^{2}\delta_{kl}\right) + \left\langle \Psi_{el}\left|\sum_{i}\left(3C_{k,i}C_{l,i} - r_{i}^{2}\delta_{kl}\right)\right|\Psi_{el}\right\rangle\right\} \quad (1)$$

which can be reduced within the NAO-PC model *via*

(2)

$$\left\langle \Psi_{el}\left|\sum_{i}\left(3C_{k,j}C_{l,i} - r_{i}^{2}\delta_{kl}\right)\right|\Psi_{el}\right\rangle \cong \sum_{\alpha\neq H}\cdot\sum_{j=1}^{2n_{\alpha}} q_{i}\left(3C_{k,j}C_{l,j} - r_{j}^{2}\delta_{kl}\right) + \sum_{\alpha=H} q_{\alpha}\left(3C_{k,\alpha}C_{l,\alpha} - R_{\alpha}^{2}\delta_{kl}\right)$$

where Q_{kl} is the quadrupole moment of the Cartesian coordinates k and l, Z_{α} is the charge of atom α located at $\mathbf{R}_{\alpha}$ with the vector components $C_{k,\alpha}$ and $C_{l,\alpha}$. δ_{kl} is the Kronecker Delta and Ψ_{el} is the electronic part of the molecular wavefunction with the coordinate components $C_{k,i}$ and $C_{l,i}$ of the electron i at the point $\mathbf{r}_{i}$. q_{j} and q_{α} are the natural atomic orbital point charges. This simple approach enables us to calculate the multipole moments very quickly using highly vectorizable algorithms.

Results

The first test of any type of point charge or additive representation of the electronic part of the molecular wavefunction is the ability of the model to reproduce the dipole moment calculated directly [17] from the same wavefunction. We have used a set of 45 organic molecules to test the quality of the NAO-PC model in this respect. The results are shown in Figure 1. There is an excellent 1:1 linear correlation with a calculated correlation coefficient of 0.995 and a standard deviation between the two types of calculated dipole moment of 0.13 Debye. The slope of the best fit least squares line is 1.08 ± 0.02

The NAO-PC model is, however also well suited for the calculation of higher molecular moments. Although

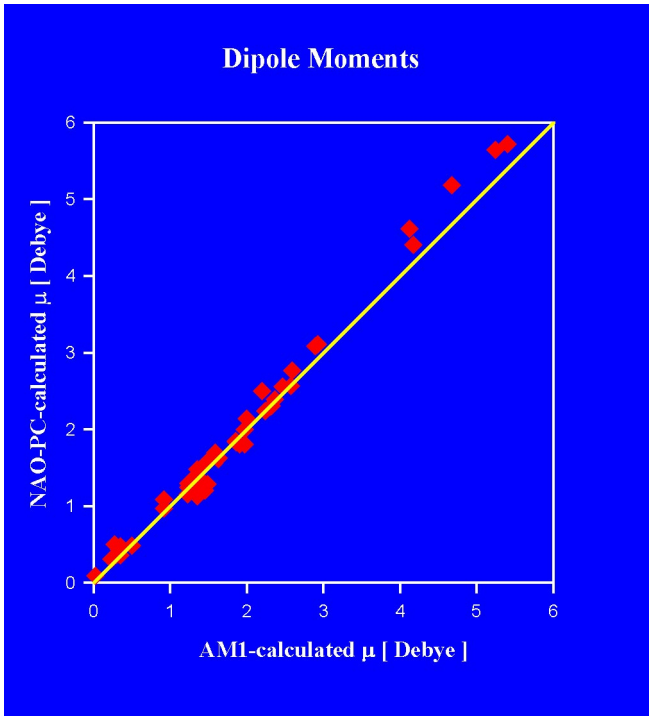


Figure 1: *Comparison between directly calculated AM1 dipole moments and those given by the NAO-PC model.*

experimental data are scarce, we tested the accuracy of AM1/NAO-PC calculated molecular quadrupole moments using some of the available data [18]. The results are compared with experiment in Table 1 and plotted against the experimental data in Figures 2a/2b. The agreement is surprisingly good and AM1 appears, at least for this very limited test set, to perform better than reasonable level *ab initio* calculations. This is possibly a direct consequence of the fact that AM1 uses Slater-type orbitals, which have far more pronounced tails than the Gaussian functions used in most *ab initio* work. Similarly to the small distant charges in the NAO-PC model, the distant low electron density regions of the wavefunction play a significant role in determining the magnitudes of the molecular electrostatic moments. PM3 has been shown [14] to perform similarly for molecular quadrupoles.

We therefore conclude that the NAO-PC method within the AM1 and PM3 frameworks provides a reliable description of the molecular electrostatics, as demonstrated for molecular electrostatic potentials [13,14], electrostatic fields [19] and molecular dipoles and quadrupoles. We therefore feel confident in using the atom-centered multipoles thus obtained as descriptors for an artificial neural net.

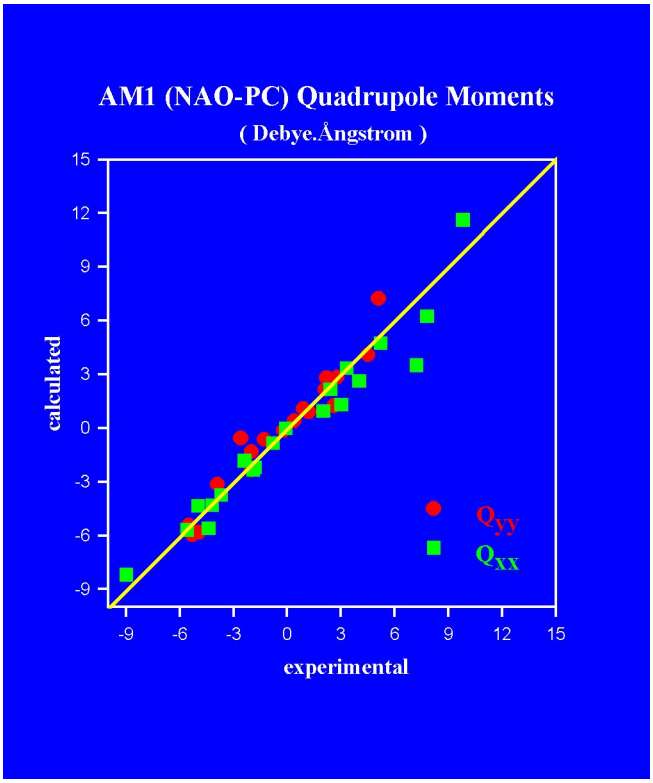


Figure 2a: *Comparison between experimental and AM1 (NAO-PC) calculated quadrupole moments. The Q_{zz}-values, which depend linearly on Q_{xx} and Q_{yy}, are not shown.*

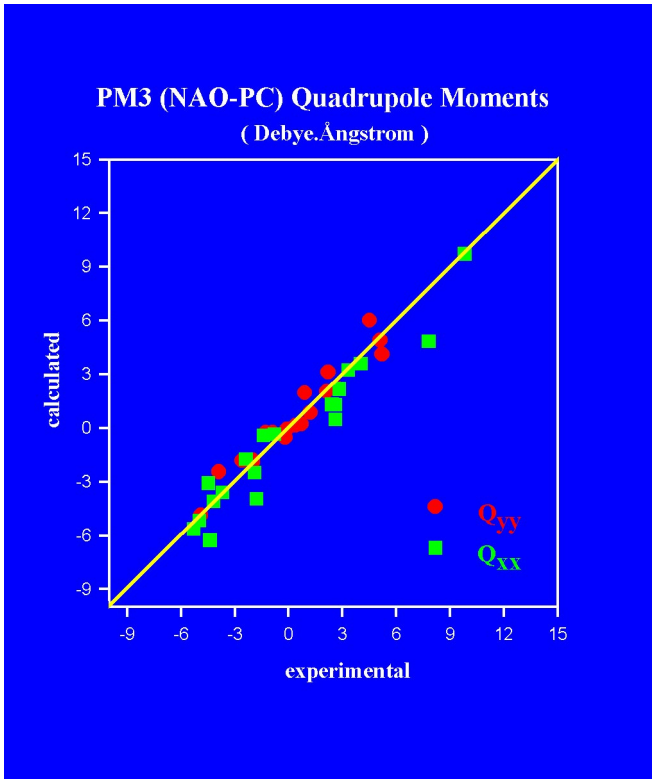


Figure 2b: *Comparison between experimental and PM3 (NAO-PC) calculated quadrupole moments. The Q_{zz}-values, which depend linearly on Q_{xx} and Q_{yy}, are not shown.*

The Back-Propagation Net and the Training Set

It is not appropriate to discuss the capabilities of back-propagation neural nets in detail here, but in the present context, the net can be considered to be a model-free device that attempts to derive the dependence of the target values (the chemical shifts) on the descriptors (input data). A simple back-propagation net achieves this by a steepest descent-type optimization procedure based on the generalized delta rule [20]. The weights and biases associated with the individual connections and nodes of the net are seeded to random values and then optimized in order to minimize the total RMS error between calculated and target values for all the chemical shift values in the training set. The inclusion of a hidden layer of nodes allows the net to learn functions such as the exclusive if and thus to react far more flexibly than a fitting procedure such as multivariate least squares.

As discussed in the introduction, a neural net for estimating ^{13}C chemical shifts must be given information about the electronic environment around the carbon atom in question and, at least in our case, about the number and types of bonds in which the carbon is involved. Furthermore, the descriptors must be rotationally invariant. We have therefore used individual standard multipole orientations derived from a diagonalization of the quadrupole tensor for each atom. The input for the net is thus rotationally invariant, but also includes information about the relative orientation of the dipole vector and quadrupole tensor components. The descriptors required for atomic multipoles up to quadrupole are thus the charge, the three dipole vector components in the orientation of the diagonalized quadrupole and the three nonzero values of the diagonalized quadrupole tensor.

In order to provide information that allows the net to deduce the paramagnetic contribution to the magnetic shielding, we followed the ideas proposed by Karplus and Pople [7] and included descriptors for the four highest bond orders to the carbon and about the four atoms to which these bonds are made. The atoms cannot be designated by their atomic numbers, as this would not describe the periodic properties of the elements. The four atoms were therefore each described using two descriptors, the numbers of the group and row in the periodic table.

Experience with the above descriptor set showed larger than average deviations for cyclopropanes and other strained compounds, so that the smallest bond angle at the carbon

Table 1: *AM1-Calculated and Experimental Quadrupole Moments (Debye·Ångstrøm)*

Molecule	NAO-PC			Experimental [16]		
	Q_{xx}	Q_{yy}	Q_{zz}	Q_{xx}	Q_{yy}	Q_{zz}
1,3-Difluorobenzene	-4.34	-0.56	4.90	-5.0±0.9	-2.6±1.3	7.6±1.0
3-Methylfurane	2.16	-5.44	3.29	2.4±0.8	-5.5±1.1	3.1±0.7
Benzene	2.84	2.84	-5.68	2.8±1.4	2.8±1.4	-5.6±2.8
Fluorobenzene	-2.31	7.24	-4.93	1.9±0.8	5.1±1.0	-3.2±1.0
1,2-Difluoroethylene	-0.69	1.32	-0.63	-1.7±0.4	3.0±0.3	-1.3±0.5
CO2	-5.59	2.80	2.80	-4.4±0.2	2.2±0.2	2.2±0.2
Ammonia	-1.82	0.91	0.91	-2.4±0.1	1.2±0.1	1.2±0.1
Cyanogen fluoride	-4.31	2.15	2.15	-4.2±?	2.1±?	2.1±?
Difluoroformaldehyd	-3.74	-0.13	3.87	-3.7±0.7	-0.2±0.5	3.9±1.1
Fluoroacetylene	2.64	-1.32	-1.32	4.0±0.2	-2.0±0.2	-2.0±0.2
Chloroacetylene	6.26	-3.13	-3.13	7.8±?	-3.9±?	-3.9±?
Dicyanogen	-8.19	4.10	4.10	-9.0±?	4.5±?	4.5±?
Acetonitrile	-2.19	1.09	1.09	-1.8±1.2	0.9±1.2	0.9±1.2
Water	-0.01	1.25	-1.23	-0.1±0.1	2.6±0.1	-2.5±0.1
Formic acid	4.75	-5.95	1.20	5.2±0.4	-5.3±0.4	0.1±0.4
Dimethyl ether	3.35	-2.33	-1.03	3.3±0.6	-2.0±0.5	-1.3±1.0
Ethane	-0.82	0.41	0.41	-0.8±0.1	0.4±0.1	0.4±0.1
1,3-Pentadiyne	11.64	-5.82	-5.82	9.8±0.8	-4.9±0.8	-4.9±0.8

atom in question was also included as a descriptor. This gives a total of 20 descriptors, as shown in Table 2.

A variety of architectures were tested for a simple back-propagation neural net [20] before the final configuration of a three-layer net (one hidden layer) with 20 input descriptors, 14 neurons in the hidden layer and one output neuron (the chemical shift) was eventually selected. This net has a total of 309 variables (weights + biases) so that we used a training set of over 800 chemical shifts in order to ensure that the net was not learning shifts by heart, but rather deriving a more general set of rules. Factors larger than two between the number of data and the number of variables appears adequate for this purpose [21]. Using the same training set with a larger (20:18:1) net gave better performance for the training set, but not for the test set. Similarly a smaller (20:10:1) net performed less well for both the training and test sets. The training set consisted of 840 individual chemical shifts taken from 231 different molecules. The choice of molecules is critical. They must cover the entire range of compounds that the trained net is intended to be able to handle and the experimental chemical shifts should refer to a clearly defined state. This means that conformationally flexible molecules, or those for which tautomeric equilibria in solution are possible, should not be included. Similarly, because the net reacts non-linearly, chemically equivalent protons that are not symmetrically equivalent cannot be used in the training set. In this preliminary work, which is designed to test the viability of our approach, many types of functional groups have not been included. This neglect of certain types of functional group allows us to test the generality of the rules derived by the net by testing for „extrapolation“ compounds, which contain functional groups that the net does not know. We have also not included any ions in this initial training set for the same reason (see below). All chemical shift data for both the training set and the test set were taken from reference [22]. The test set covers the range of chemical shifts from -2.8 to 225 ppm. The nets were tested for overtraining periodically during training (i.e. the test set was calculated in order to make sure that its results were not depreciating as the net improved its performance for the training set). For both AM1 and PM3, the final stable net proved to perform best for both the training and the test sets.

Figures 3a/3b show the results obtained for the training set after about 400,000 training cycles. The net is stable at this stage, but because of the large data:variables ratio, it is unable to learn to recognise specific carbons and therefore generates a reliable set of rules for deriving the chemical shift from the descriptors. In this way we hope to avoid the common mistake in neural net applications that the net performs excellently for the training set, but has essentially no predictive power. Ideally, the net should perform as well for the test set as for the training set. The standard deviation between calculated and experimental chemical shifts is 3.9 ppm, with a maximum deviation of 17 ppm for AM1 and slightly worse (4.7 ppm standard deviation, -20 ppm maximum deviation) for PM3.

Table 2: *Input descriptors for the back-propagation neural net.*

Node #	Descriptor
1	Atomic monopole (charge)
2-4	Atomic dipole components
5-7	Atomic quadrupole components
8	Highest bond order, B1
9	Second highest bond order, B2
10	Third highest bond order, B3
11	Fourth highest bond order, B4
12	Element to which bond B1 is made (group of the periodic table)
13	Element to which bond B2 is made (group of the periodic table)
14	Element to which bond B3 is made (group of the periodic table)
15	Element to which bond B4 is made (group of the periodic table)
16	Element to which bond B1 is made (row of the periodic table)
17	Element to which bond B2 is made (row of the periodic table)
18	Element to which bond B3 is made (row of the periodic table)
19	Element to which bond B4 is made (row of the periodic table)
20	Smallest bond angle at the carbon in question

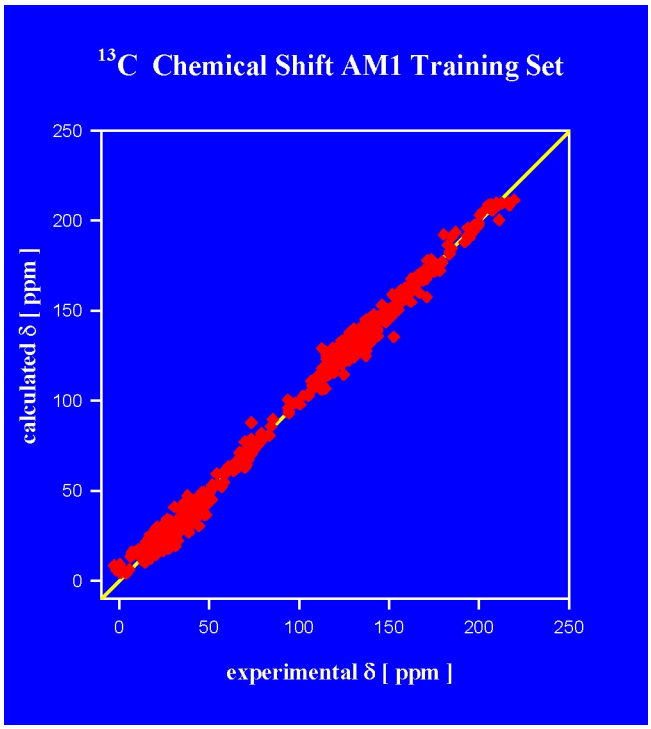


Figure 3a: *Comparison between experimental ^{13}C chemical shifts and those given by the 20:14:1 back-propagation net for the training set AM1.*

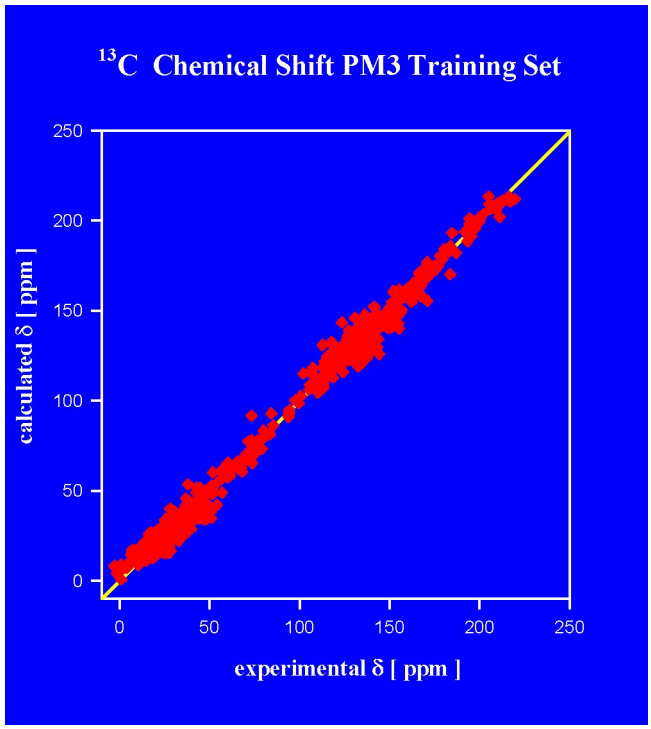


Figure 3b: *Comparison between experimental ^{13}C chemical shifts and those given by the 20:14:1 back-propagation net for the training set PM3.*

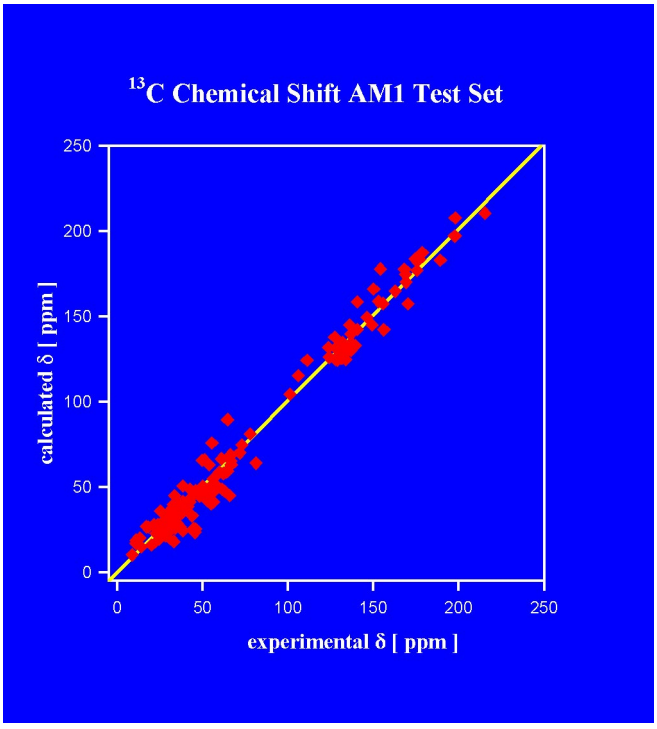


Figure 4a: *Comparison between experimental ^{13}C chemical shifts and those given by the 20:14:1 back-propagation net for the test set AM1 shown in Scheme 1.*

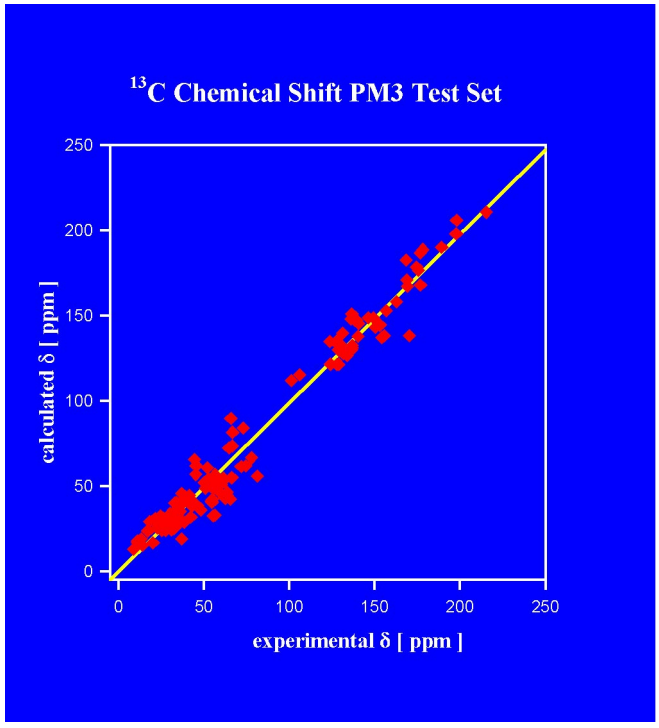
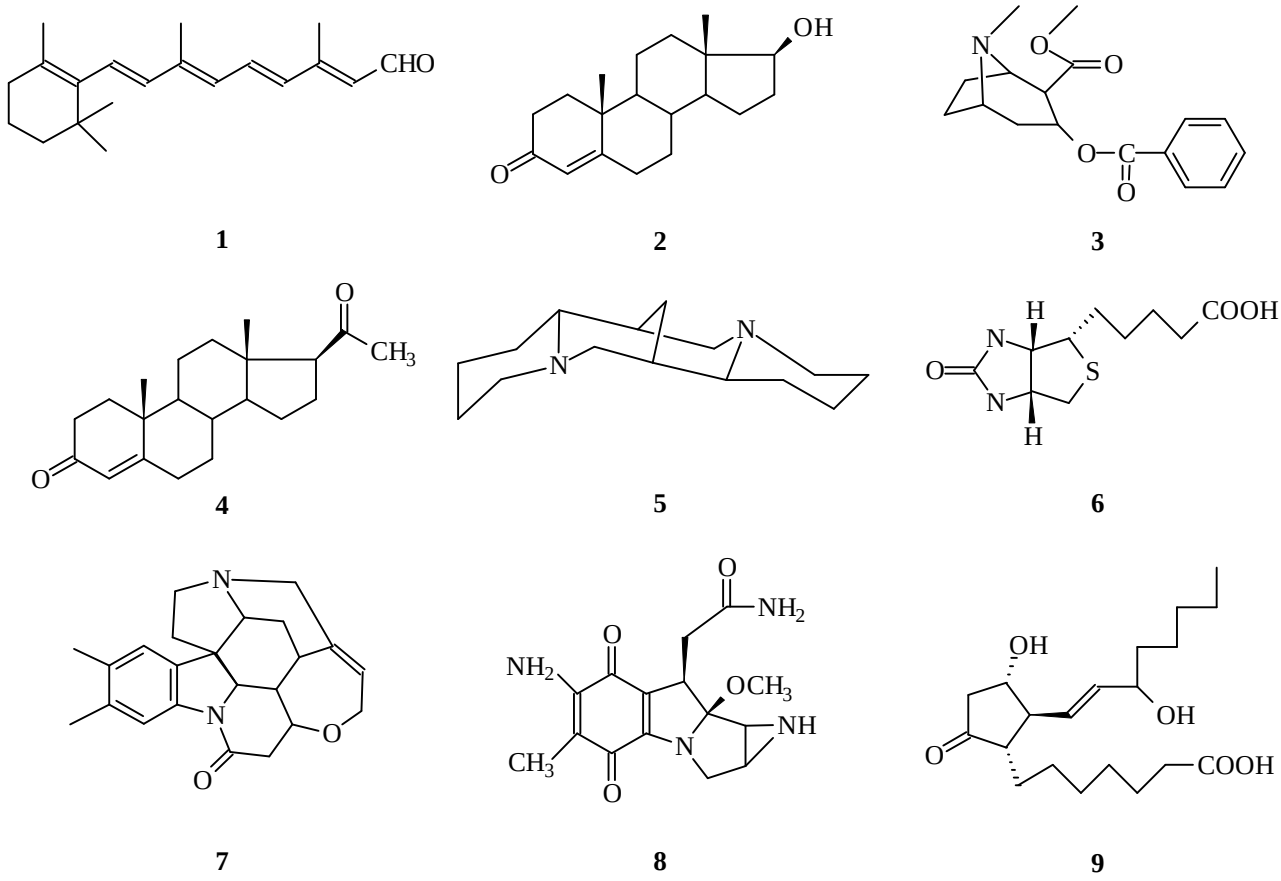


Figure 4b: *Comparison between experimental ^{13}C chemical shifts and those given by the 20:14:1 back-propagation net for the test set PM3 shown in Scheme 1.*



1 **2** **3**

4 **5** **6**

7 **8** **9**

Scheme 1: *The test set of molecules;* ***1****, trans-retinal;* ***2****, testosterone;* ***3****, cocaine;* ***4****, progesterone;* ***5****, sparteine;* ***6****, biotine;* ***7****, brucine;* ***8****, mitomycine and* ***9****, prostaglandine.*

The Test Set

The molecules used for the test set are shown in Scheme 1. We deliberately chose large molecules because they are usually treated well by semiempirical procedures and would require a major computational effort using *ab initio* methods. The functional groups contained in the test set are also represented in the training set, which, however, contains molecules whose AM1 or PM3 geometries may be significantly in error in order to test the entire geometry optimization/chemical shift estimation procedure for real examples.

The results obtained for the test sets are shown in Figures 4a/4b. The largest deviation for AM1 (19.6 ppm) is almost identical to that obtained for the training set and the standard deviation (6.9 ppm) is less than twice that given for the training set - both values that indicate that the net really has learnt to estimate chemical shifts on the basis of the descriptors. The PM3 net also gives results comparable to those obtained for the training set (31.8 ppm maximum deviation, 8.7 ppm standard deviation), but is generally not as accurate as the AM1 net. We therefore conclude that the neural net technology presented here provides a fast and economical method for estimating ^{13}C chemical shifts with the sort of accuracy given for the test set. Although GIAO/MNDO results have only been reported for a very limited set of hydrocarbons [3] we estimate that the neural net procedure achieves comparable accuracy to the direct calculation of the shieldings.

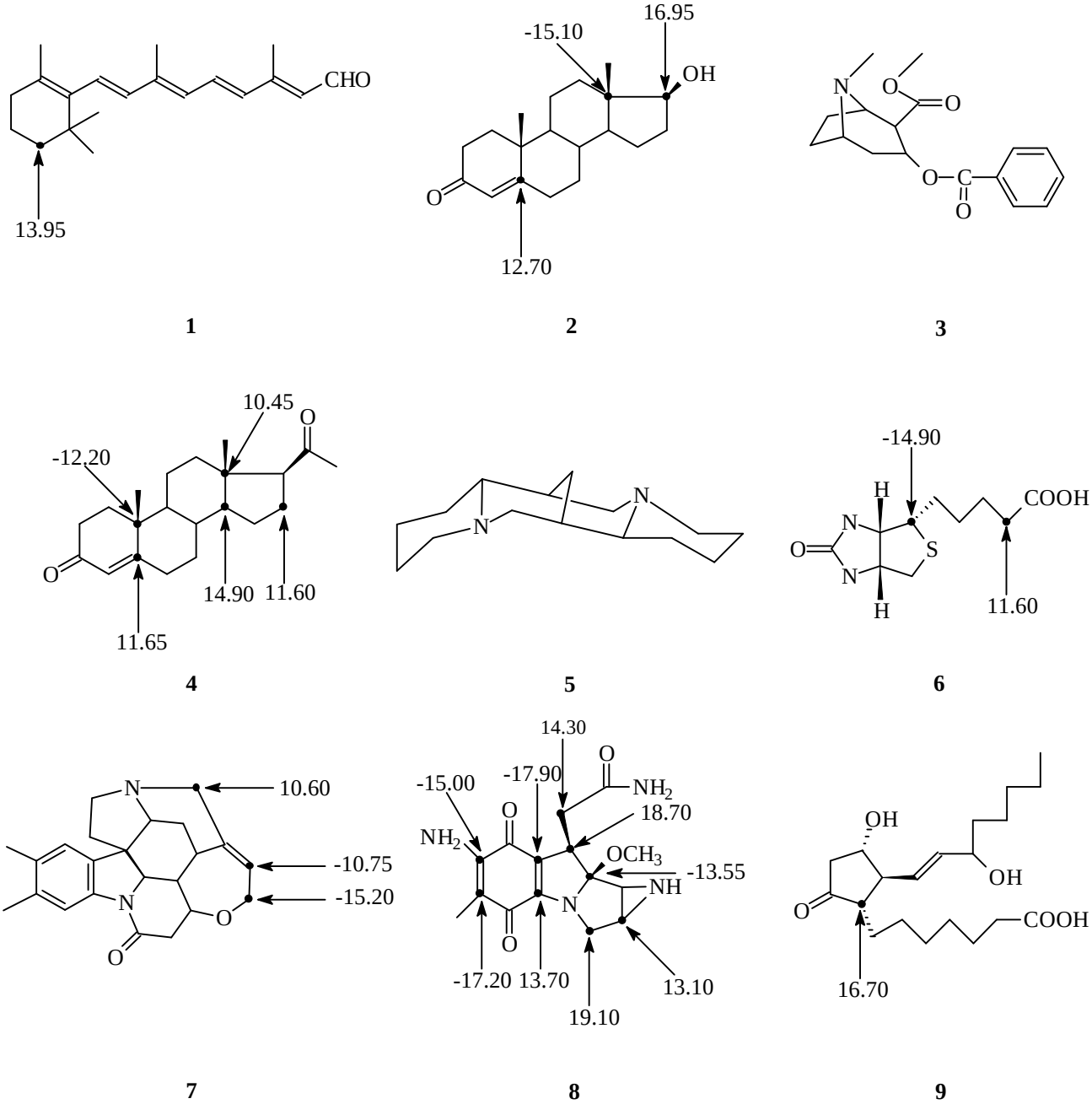
Discussion

This work has established the viability and the limitations in accuracy of the type of approach outlined here. The accuracy obtained is not sufficient to, for instance, assign the individual olefinic carbons in *trans*-retinal, but is closer to that which can be expected from good additive schemes. Schemes 2a and 2b show all the carbon atoms in the test set for which the difference between calculated and experimental chemical shift is larger than ±10 ppm for AM1 and PM3, respectively. For cocaine **3** and sparteine **5**, all carbon shifts are predicted within 10 ppm. For *trans*-retinal **1**, and the prostaglandin **9**, one carbon shift (for a ring carbon in **1** and the tertiary carbon á to the carbonyl in **9** shows an error of 14-17 ppm. The performance for the bridgehead C-atoms for the two steroids testosterone **2** and progesterone **4** is less impressive. This may reflect a lack of highly substituted ring systems in the training set or the

known tendency of AM1 [23] to make rings too flat. The latter suggests that systematic AM1 geometry errors may influence the results adversely. Brucine **7** shows some large errors for three of the peripheral carbons (one olefinic and two allylic), but is otherwise treated well. Mytomycine **8** is treated far less well than the other test molecules. This is probably a result of the unusual amino-substituted quinone system that was not present in the training set. Generally, the results indicate both an adequate overall performance and some systematic weaknesses that can probably eliminated by retraining with a suitably extended training set.

For PM3, only the chemical shifts in sparteine are all within the 10 ppm limit. Generally, fully substituted olefinic carbons show large errors, as do quaternary carbons and those α to sulfur. The performance of PM3 for the test set is inferior to that of AM1, although it apears to be able to treat the quinone system of **8** better than AM1.

Despite these errors, we feel that this sort of procedure represents a significant step forward in semiempirical



1 **2** **3**

4 **5** **6**

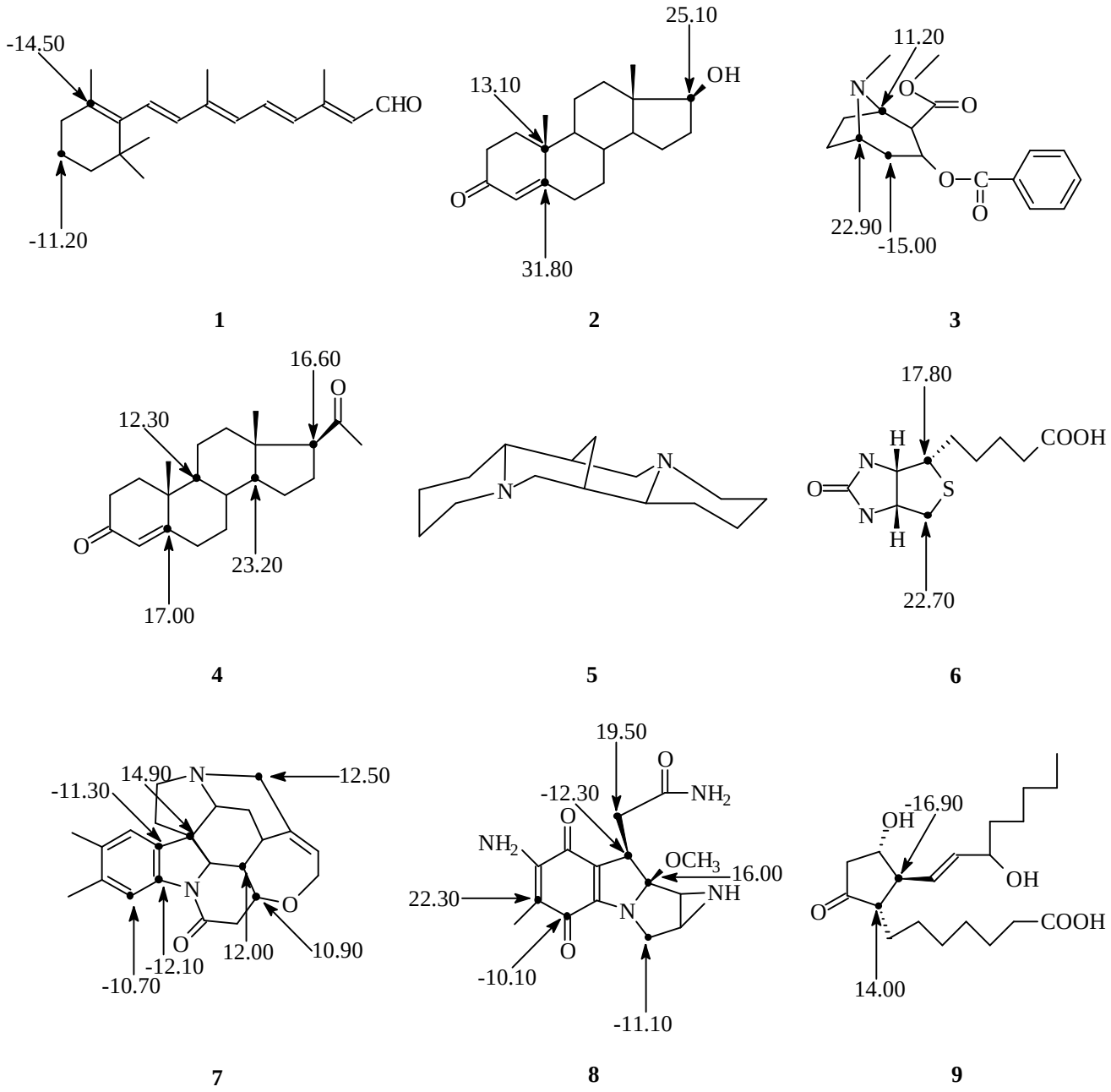
7 **8** **9**

Scheme 2a: *AM1 test set with the location of the Carbon atoms with a calculated error > ±10 ppm.*

techniques. The chemical shift calculation is so fast that there is no need to make it an option within the VAMP program. Every AM1 or PM3 calculation on neutral, closed-shell carbon-containing molecules in the development (6.0) version [24] gives the estimated ^{13}C chemical shifts as an extension of the population analysis. The most important feature of this procedure is that it establishes a direct link between the calculations and spectroscopic data that are almost always available. The question as to how the net actually derives its result is, however of interest and was investigated by systematic variation of the descriptors in order to determine the net's reaction.

Dependence of the Estimated Shifts on the Descriptors

The dependence of the predicted chemical shifts was investigated by systematically changing the input descriptors for a benzene carbon and observing the behavior of the net. Figure 5a shows the dependence of the chemical shift estimated



Scheme 2b: *PM3 test set with the location of the Carbon atoms with a calculated error > ±10 ppm.*

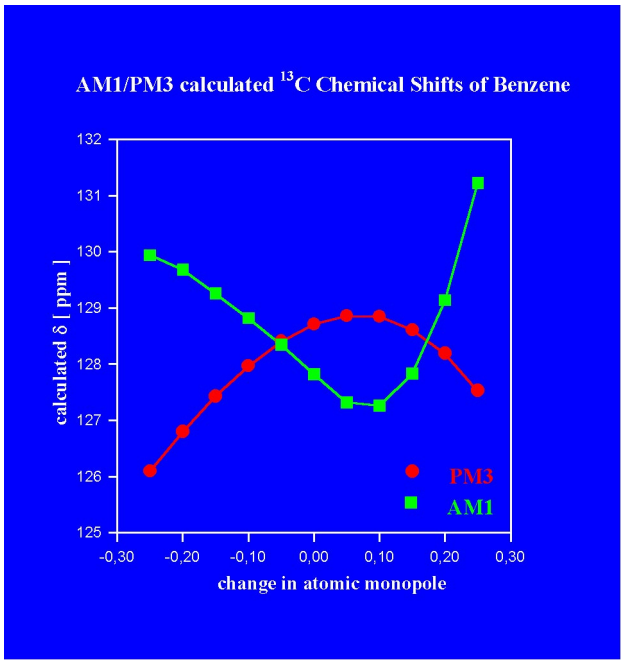


Figure 5a: *Dependence of the AM1/PM3 predicted chemical shift of a benzene carbon on the atomic charge. Charges are given in electronic charges. The "Change in parameter" axis indicates that the numerical value of the parameter was changed by the given amount.*

by the AM1 and PM3 nets on the atomic multipole. The effect is generally small (2-3 ppm for a change of 0.3 in the monopole) and the two nets give different trends. AM1 shows a minimum in the estimated chemical shift around +0.07 and PM3 a minimum. These results suggest that the monopole is not playing a major role in determining the chemical shift for this type of carbon.

The effect of the three dipole components on the estimated shifts for the two nets is shown in Figures 5b/5c. Component 1 (in-plane, roughly C-H direction) causes an increase in the estimated chemical shift as its value is increased in both cases, but this effect is far stronger for the PM3 net than for AM1. The second component (in-plane, perpendicular to the first) shows the same effect. The out-of plane component 3 has a negative slope in both cases, but falls of far more sharply for AM1 than for PM3. Changes in these parameters have, however, little physical meaning for a benzene carbon.

The effect of the three quadrupole components is shown in figures 5d/5e. The two plots are completely different. All three components give a monotonic increase in the calculated chemical shift as they are increased, but for AM1 component 3 (out-of plane) shows the opposite trend. This suggests that the effect of increasing π-electron density is opposite in the two nets. This unexpected result probably reflects linear dependencies between descriptors that make interpretation difficult.

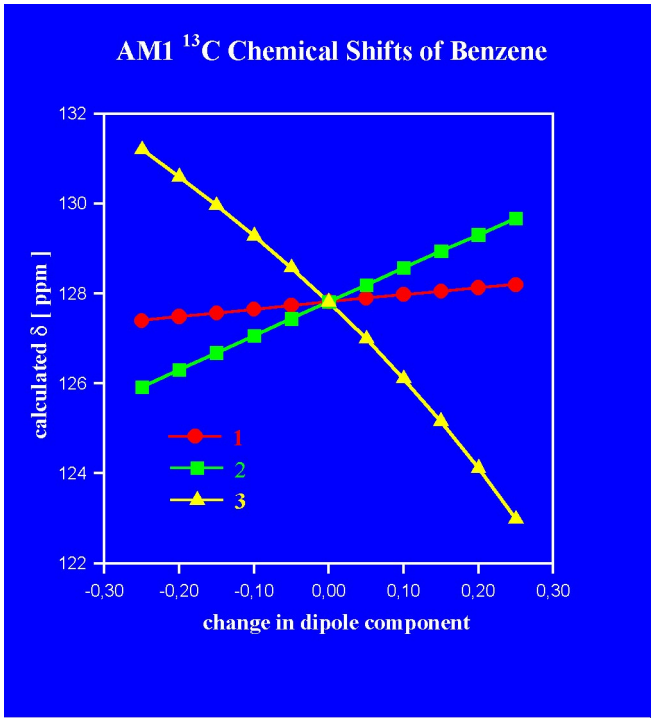


Figure 5b: *Dependence of the AM1 predicted chemical shift of a benzene carbon on the dipole components. Dipoles are given in Debyes. The "Change in parameter" axis indicates that the numerical value of the parameters dipole were changed by the given amount.*

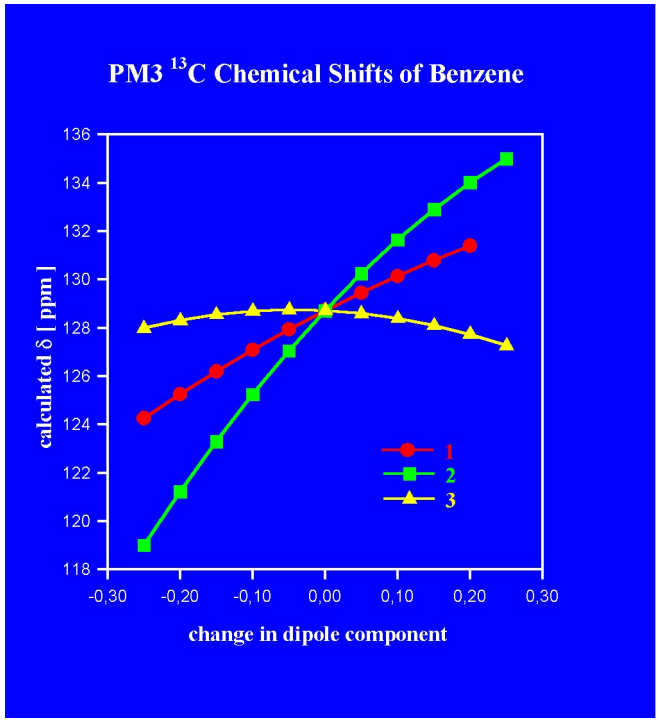


Figure 5c: *Dependence of the PM3 predicted chemical shift of a benzene carbon on the dipole components. Dipoles are given in Debyes. The "Change in parameter" axis indicates that the numerical value of the parameters dipole were changed by the given amount.*

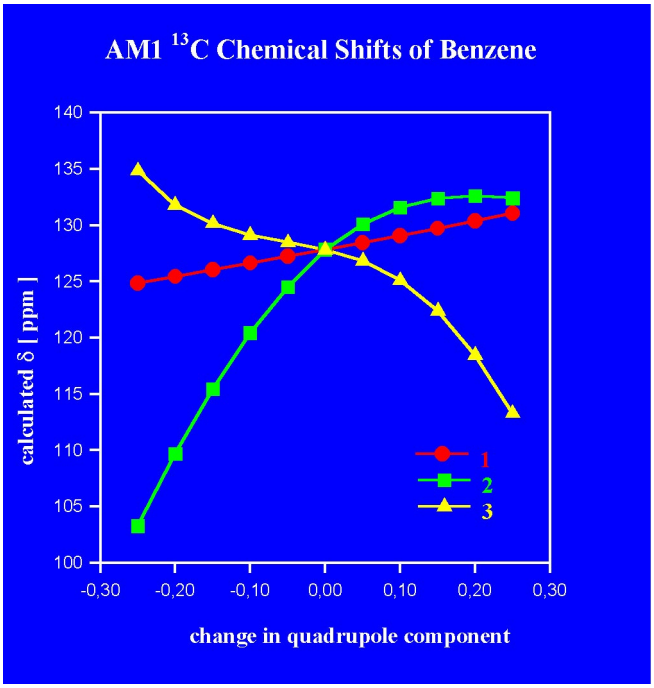


Figure 5d: *Dependence of the AM1 predicted chemical shift of a benzene carbon on the quadrupole components. Quadrupoles are given in Debye.Ångstrøm. The "Change in parameter" axis indicates that the numerical value of the parameters were changed by the given amount.*

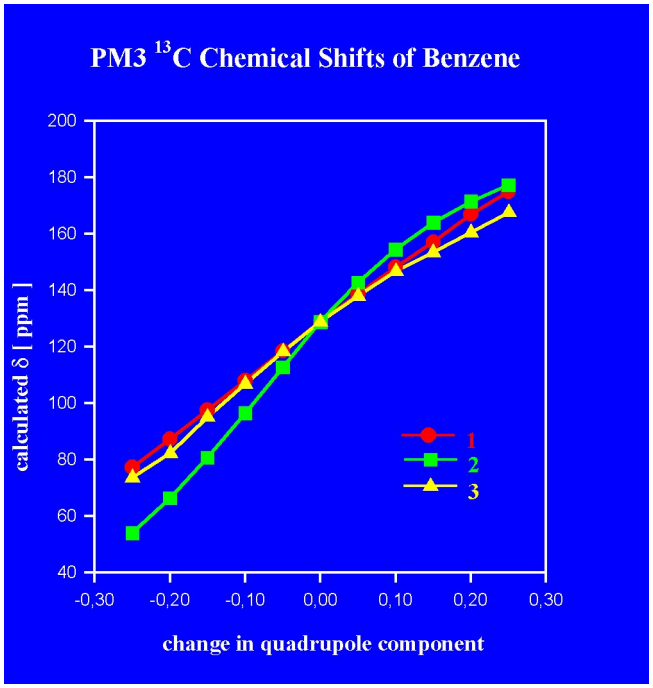


Figure 5e: *Dependence of the PM3 predicted chemical shift of a benzene carbon on the quadrupole components. Quadrupoles are given in Debye.Ångstrøm. The "Change in parameter" axis indicates that the numerical value of the parameters were changed by the given amount.*

The effect of changing the aromatic bond order descriptor is shown in Figure 5f. Increasing bond order leads to a higher chemical shift for both nets, but the effect is much larger for AM1. The AM1 curve also shows a maximum at about +0.15.

The above results shed little light on the internal dependencies of the two nets. What, however, is clear is that the two apparently similar nets are reaching their estimates by significantly different paths. This may mean that the two nets have trained to two nonequivalent local minima, but retraining the PM3 net starting from the AM1 weights suggests that this is not the case.

Predictive power of the Net

The test set shown in Scheme 1 only probes the ability of the net to interpolate within the known range of data because it does not contain any groups that are not also in the training set. A far more general test of the predictive power of the net (and whether it really has established general rules for estimating the chemical shift) is to use a second test set containing chemical entities that are not represented in the training set. Schemes 3, 4 and 5 show such molecules with the calculated and experimental chemical shifts. We have included charged species, sulfones, pyridine-N-oxide *etc*. No such species appear in the test set.

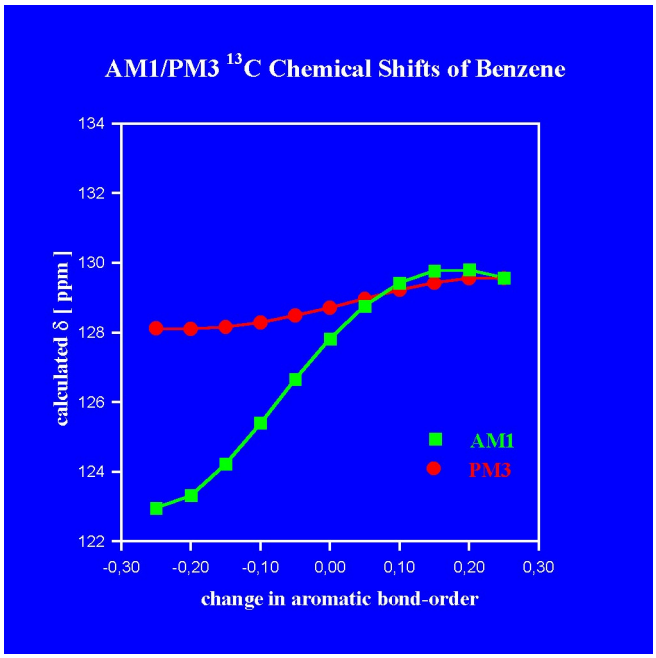
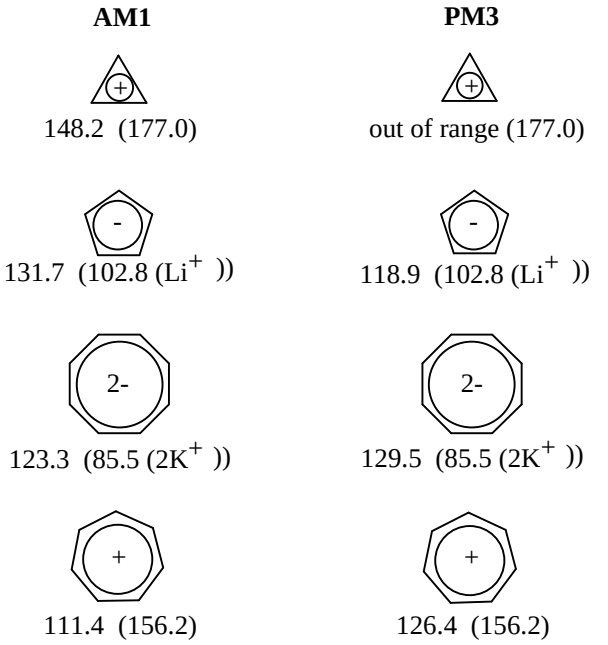


Figure 5f: *Dependence of the AM1/PM3 predicted chemical shift of a benzene carbon on the C-C-bond order. Bond orders correspond roughly to the customary single, double, triple bond convention. The "Change in parameter" axis indicates that the numerical value of the bond order was changed by the given amount.*



Scheme 3: *Calculated and experimental (in parentheses) chemical shifts for anulene ions.*

The results for the annulene ions (cyclopropenium, tropylium, cyclopentadienyl anion and cyclooctatetraene dianion, Scheme 3) show very systematic errors. For AM1, simply adding 25 ppm per charge to these results would give moderate agreement between the net and experiment, so that the addition of one extra descriptor (the total molecular charge) would give a generally applicable net. This simple correction would also work moderately well for PM3 with the notable exception of the cyclopropenium ion. It is remarkable that a species such as the cyclooctatetraene dianion, which cannot exist in the gas phase [25], should give results that obey this simple description even roughly.

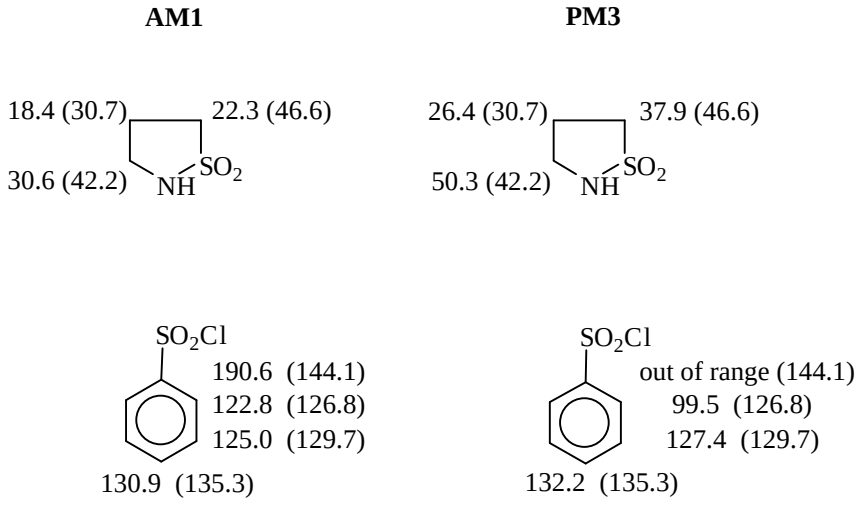
The sulfones and the sulfonamide (Scheme 4) also give interesting results. The estimated chemical shifts for the sulfonamide are acceptable with PM3, but the carbon α- to the sulfone group deviates by 24 ppm using AM1. There are large deviations for the α-carbons in the sulfonyl chloride (50-60 ppm), but AM1 does well for the remaining carbons. The agreement for the *meta*-carbon is also poor with PM3.

Finally, pyridine-N-oxide and the methyl-pyridinium ion (Scheme 5) were used to test the net's ability to deal with unusual inductive effects in aromatic systems. The N-oxide shows large (66 and 33 ppm for AM1 and PM3, respectively) deviation for the *ortho*-carbons. These, however, have descriptors outside the range of those used to train the net. This situation can be detected by testing that all descriptors are within the normalization range. For out-of-range carbon atoms, a diagnostic message is printed. AM1 treats the other two carbons well, whereas PM3 has a 19 ppm error for the *para*-position.

The pyridinium salt shows moderate agreement with experiment except for the *meta*-carbon for both methods. AM1 gives a 30 ppm error for the *ortho* and *para* positions, but PM3 does much better. Clearly, the uniform charge correction suggested for the annulene ions above would not work in this case, but a net with one extra descriptor should be able to deal with this example.

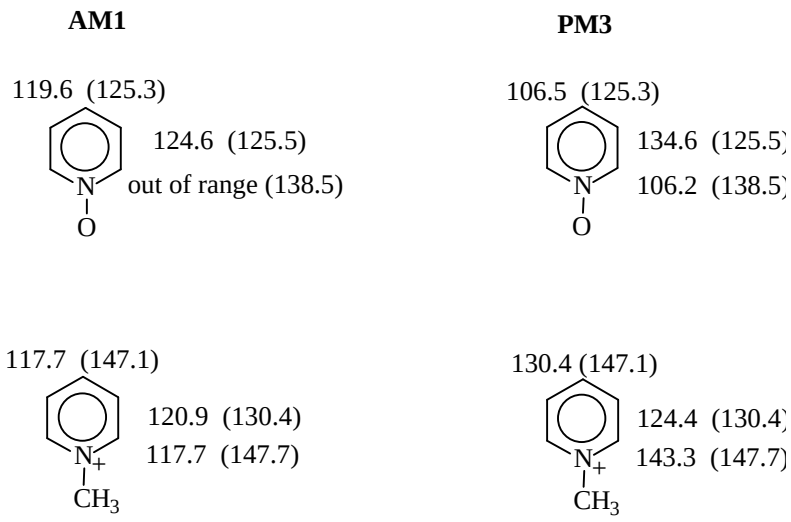
Summary and Conclusions

The technique presented here combines semiempirical MO-theory with an artificial neural net to provide estimates of ^{13}C chemical shifts that are accurate enough to be of general use to the non-specialist, but not for detailed assignment of spectra, which would need almost two orders of magnitude lower errors. Many of the failures can be traced to systematic



Scheme 4: *Calculated and experimental (in parentheses) chemical shifts for sulfones.*

Scheme 5: *Calculated and experimental (in parentheses) chemical shifts for pyridine derivatives.*



AM1 and PM3 errors, suggesting that the basic neural net methodology itself is sound. This conclusion is also reinforced by the net's estimates for the "extrapolation" compounds shown in Schemes 3, 4 and 5. In these cases the net must be using some sort of generally applicable rule that also works moderately well for new compound classes. Nevertheless, it is clearly possible that the net will give errors of up to 30 - 35 ppm even for descriptors within the range of its training set. Generally, however, the standard deviation of the estimated chemical shifts from the experimental values is 6-10 ppm. Paradoxically, we would have distrusted our results if they had been significantly better than this because direct calculation of the magnetic shielding within AM1 or PM3 cannot be expected to be very much more accurate than the net.

The present net is limited to neutral molecules and has not been trained for all substituents. We believe, however, that we have established that the general calculational technique is useful and that a net that includes a molecular charge descriptor with a more universal training set should be able to deal with the exceptions shown above with about the same accuracy that the current net attains for the test set of molecules. The major advantage of ^{13}C chemical shifts is that experimental data are plentiful, so that we can now embark on the training of a more universal net, which will eventually be included in the VAMP program.

Acknowledgements

This work was supported by Oxford Molecular Ltd. and the Deutsche Forschungsgemeinschaft. G. R. thanks the Studienstiftung des Deutschen Volkes for the award of a Fellowship.

References and Notes

1. See, for example, Ditchfield, R. *J. Chem. Phys.* **1972**, *56*, 5688; Kutzelnigg, W. *Isr. J. Chem.* **1980**, *19*, 193; Schindler, M.; Kutzelnigg, W. *J. Chem. Phys.* **1982**, *76*, 1919; Wolinski, K.; Hinton, J. F.; Pulay, P. *J. Am. Chem. Soc.*, **1990**, *112*, 8251.
2. See, for example, Bühl, M.; van Eikema Hommes, N. J. R.; Schleyer, P. v. R.; Fleischer, U.; Kutzelnigg, W. *J. Am. Chem. Soc*. **1991**, *113*, 2459 and references therein.
3. Wei-Xiong, W.; Xiao-Zeng, Y.; An-Bang, D. *Acta Chimica Sinica* **1987**, *44*, 195; *Acta Chimica Sinica (Series B)* **1988**, *31*, 1048; see also Garber, A.R.; Ellis, P. D.; Seidman, K.; Schade, K. *J. Magn. Res*. **1979**, *34*, 1.
4. Nelsen, S. F.; Clark, T. manuscript in preparation
5. Spiesecke, H.; Schneider, W. G. *Tetrahedron Lett*. **1961**, 468; *J. Chem. Phys.* **1961**, *35*, 731.
6. Saika, A.; Slichter, C. P. *J. Chem. Phys.* **1954**, *22*, 26; Pople, J. A. *Disc. Far. Soc*. **1963**, *24*, 7; *Mol. Phys.* **1963**, *7*, 301; Lamb, W. E. *Phys. Rev.* **1941**, *60*, 817.
7. Karplus, M.; Pople, J. A. *J. Chem. Phys.* **1963**, *38*, 2803.
8. Rauhut, G.; Chandrasekhar, J.; Alex, A.; Beck, B.; Sauer, W.; Clark, T.; VAMP 5.5, available from Oxford Molecular, The Magdalen Centre, Oxford Science Park, Sandford on Thames, Oxford OX4 4GA, England.
9. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
10. Stewart, J. J. P. *J. Comput. Chem.* **1989**, 10, 209; 221; Stewart, J. J. P. *J. Comput. Chem.* **1991**, 12, 320.
11. Armstrong, D. R.; Perkins, P. G.; Stewart, J. J. P. *J. Chem. Soc. Dalton Trans.* **1973**, 838.
12. See, for instance, Stone, A. J. *Chem. Phys. Lett.* **1981**, *83*, 233; Stone, A. J.; Alderton, M. *Mol. Phys.* **1985**, *114*, 359; Stone, A. J.; Price, S. L. *J. Phys. Chem.* **1988**, *92*, 3325.

13. Rauhut, G.; Clark, T. *J. Comput. Chem*. **1993**, *14*, 503.
14. Beck, B. Clark, T. *J. Comput. Chem*. **1994**, *15*, 1064.
15. Coulson, C. A.; Longuet-Higgins, H. C. *Proc. Roy. Soc. A* **1937**, *93*, 39.
16. Buckingham, A. D. *Quart. Rev.* **1959**, *13*, 189.
17. Pople. J. A.; Beveridge, D. L. *Approximate Molecular Orbital Theory*, McGraw-Hill, New York, 1982.
18. Gierke, T. D.; Tigelaar, H. L.; Flygare, W. H. *J. Am. Chem. Soc.* **1972**, *94*, 330; Flygare, W. H.; Benson, R. C. *Mol. Phys.* **1971**, *20*, 225; Coonan, M. H.; Craven, I. E.; Hesling, M. R.;Ritchie, G. L. D.; Spackman, M. A. *J. Phys. Chem.* **1992**, *96*, 7301; Czieslik, W.; Wiese, J.; Sutter, D. H. *Z. Naturforsch.* **1976**, *31a*, 1210; Brobjer, J. T.; Murrell, J. N. *J. Chem. Soc., Faraday Trans. II* **1982**, *78*, 1853.
19. Blaney, F. E.; Edge, C. M.; Phippen, R. W. *Abstracts of the 12th Annual Conference of the Molecular Graphics Society*, Interlaken, Switzerland, 1993.
20. See, for instance, Pao, Y.-H. *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley, Reading, Massachusetts, 1989; for general applications of artificial neural nets in chemistry, see e.g. *Neural Networks*, Lacy, M. E. (Ed.) *Tetrahedron Computer Methodology, Symposia in Print No. 2* **1990**, *3*, 119-245.
21. Manallack, D. T.; Livingstone, D. J. *Med. Chem. Res.* **1992**, *2*, 181.
22. Kalinowski, H.-O.; Berger, S.; Braun, S. *Carbon-13 NMR Spectroscopy*, Wiley, Chichester, 1988.
23. Stewart, J. J. P., *J. Comput. Chem*. **1989** *10*, 221.
24. Rauhut, G.; Chandrasekhar, J.; Alex, A.; Beck, B, Sauer, W.; Hutter, M.; Clark, T.; VAMP 6.0, to be released by Oxford Molecular Limited, The Magdalen Centre, Oxford Science Park, Sandford on Thames, Oxford OX4 4GA, England.
25. Schleyer, P. v. R.; Wilhelm, D.; Clark, T. *J. Organomet. Chem.* **1985,** *281*, C1.