**Empirical and Quantum-Mechanical Methods of ^{13}C Chemical Shifts Prediction:
Competitors or Collaborators?**

Short title:

Empirical and Quantum-Mechanical Methods of ^{13}C Shifts Prediction

Mikhail Elyashberg¹, Kirill Blinov¹, Yegor Smurnyy¹, Tatiana Churanova¹
and Antony Williams².

¹Advanced Chemistry Development, Moscow Department, Russian Federation, Moscow

² Royal Society of Chemistry, US Office, 904 Tamaras Circle, Wake Forest, NC-27587

Communicating Author: Antony J. Williams, 904 Tamaras Circle, Wake Forest, NC-
27587, Phone: 919 201 1516, Email: tony27587@gmail.com

Abstract

The accuracy of ^{13}C chemical shift prediction by both quantum-mechanical (QM) and empirical methods was compared using 205 structures for which experimental and QM-calculated chemical shifts were published in the literature. For these structures ^{13}C chemical shifts were calculated using both HOSE code and neural network (NN) algorithms developed within our laboratory. In total 2531 chemical shifts were analyzed and statistically processed. It has been shown that, in general, QM methods are capable of providing similar but nevertheless inferior accuracy relative to the empirical approaches, but quite frequently they give larger mean average error values. For the structural set examined in this work the following mean absolute errors (MAE) were found: MAE(HOSE)=1.58 ppm , MAE(NN)=1.91 ppm , MAE(QM)= 3.29 ppm. A

strategy of combined application of both the empirical and QM approaches is suggested. The strategy could provide a synergistic effect if the advantages intrinsic to each method are exploited.

Keywords

NMR, ^{13}C NMR, chemical shift prediction, GIAO, DFT, HOSE code, neural nets.

Introduction.

Different methods of ^{13}C NMR spectrum calculation have been developed over the years to provide a reliable choice for the most probable structural hypothesis, assist in the process of spectral signal assignment and to aid in the determination of stereochemistry for complex organic molecules. The first prediction algorithms were based on additive rules and referred to as an incremental method. They were intended for the empirical prediction of ^{13}C NMR chemical shifts and implemented in a series of programs (for example^[1-4]). The programs (for instance, ^[5-9]) utilizing a fragmental approach and HOSE codes ^[10] as well as efficient artificial neural net algorithms (NN) were developed ^[11, 12]. These algorithms are based on empirical methods, run *fully automatically* and require no user intervention. As the programs were required by expert systems for the purpose of computer-aided structure elucidation (CASE)^[13], they were implemented into the most advanced CASE systems ^[14-16].

Automated chemical shift prediction methods are under constant improvement ^[17-19]. Recently it has been shown ^[18] that programs based on NN algorithms and additive rules are capable of predicting ^{13}C chemical shifts for diverse classes of organic

molecules with a mean absolute error (MAE) value of 1.6-1.8 ppm and at a speed of 6000-10000 shifts per second. Programs utilizing HOSE codes ^[7, 9] provide similar or better accuracy. This approach also provides facilities which show all reference structures involved in a particular chemical shift calculation for a given atom. Visual analysis and comparison of atom environments in a reference structure and in the structure under investigation can be used to understand how the chemical shift was calculated. The shortcoming of these programs is that they are not very fast with the prediction speed varying between several seconds and tens of seconds depending on the size and complexity of a chemical structure.

The prediction of ¹³C chemical shifts using quantum-mechanical (QM) methods have been become the focus of many researchers and the GIAO approximation of the DFT approach has been increasingly applied to NMR spectral calculations. During the last decade, many publications devoted to the ¹³C chemical shift prediction of organic molecules using the QM approach were published. It is possible to distinguish the following goals of these works:

- Search for the most successful combinations of density functions and basis sets (calculation protocols) capable of providing a prediction of geometry and chemical shifts for sets of organic molecules characterized by structural diversity (for instance, ^[20-22]);
- Search for appropriate calculation protocols leading to acceptable predicted chemical shift values for a given compound or class of compounds (for instance, ^[23-25]);

- Detailed investigation of the structural and electronic properties for a single molecule or a series of selected molecules (for instance ^[26-28]);
- Selecting the most probable structural hypothesis in the process of molecular structure elucidation (for instance, ^[29-38]) and, once the genuine structure is determined, choosing its preferable stereochemical configuration.

There are a lot of examples demonstrating that successfully chosen calculation protocols lead to close coincidence between the predicted chemical and experimental shifts. It is rather common that the functions and basis sets selected for geometry optimization differ from those used for the chemical shift calculation which hampers guessing the best protocol. Attempts have been made to select an optimum protocol that fits for the purpose of ¹³C calculation for both rigid and flexible molecules. For instance, Cimino *et al* ^[20] tested about 50 protocols and concluded that the best prediction of the experimental ¹³C values is obtained at the mPW1PW91 level using the 6-31G(d,p) basis set both for the geometry optimization and chemical shift calculation.

Nevertheless the search for new approaches leading to improved calculation accuracy continues. Recently, for instance, Sorotti *et al* ^[39] suggested using for GIAO-based ¹³C chemical shift calculation a multi-standard method (MSTD). When the MSTD approach is employed, two reference compounds should be used: a) methanol – for prediction chemical shifts of sp³ hybridized carbon atoms and b) benzene – for sp and sp² hybridized ones. The authors concluded that the mPW1PW91/6-31G(d) protocol constituted a level of theory that provides maximal reliability and MAE values around 1.5 ppm at minimal computational cost when applying the MSTD approach. This approach looks attractive, and requires further investigation and testing.

Accessibility to programs performing QM calculations encouraged non-specialists in quantum chemistry to use them for the interpretation of different experimental data. Some authors ^[40] treat the GIAO chemical shift calculation as an almost routine method that can be easily utilized by organic chemists. However, the scattering of observed chemical shift MAE values found by different researchers is evidence that such generalities are not borne out in practice. Theoreticians developing QM-based methods of chemical shift calculations ^[41] note that “using to full advantage these (GIAO) interpretative potentialities requires perhaps a larger dose of theoretical experience”. Experienced researchers also comment that “since the quality of the results obtained depends on the functional and basis set used, their choice must be made wisely and with great attention”. We suppose that creation of an expert system capable of helping organic chemists to choose the appropriate protocol applied to a specific molecular structure could be useful.

The results of quantum-mechanical ¹³C NMR shift predictions performed for organic molecules of different chemical compositions and different classes have been published in many articles. As far as we know the results have not yet been generalized and QM computational errors determined for a large enough structural set were not compared with those obtained from the empirical methods. It is worthy to note that the empirical methods of NMR shift prediction are either almost not mentioned at all in the articles devoted to QM-based computations of chemical shifts or the accuracy attained using QM approach is commented on without taking into account the latest achievements ^[7, 9] in the field of empirical methods.

Meanwhile, examples of the application of empirical methods for molecular structure elucidation and the determination of relative stereochemistry in parallel with QM methods have been considered ^[42-44]. The examples show that QM calculations, which are far more computationally expensive in comparison with empirical ones, are frequently used in such cases when empirical shift prediction allows one either to rapidly and reliably find the correct solution of a problem or suggest 1-3 structural hypotheses to be finally discerned by determining additional experimental data and theoretical considerations.

In this connection it would be worthy to cite the following quotation from Dirac's recollections ^[45] : “The engineering training which I received did teach me to tolerate approximations... If I had not had this engineering training, I should not have had any success with the kind of work that I did later on... Engineers were concerned only with getting equations which were useful for describing nature. They did not very much mind how the equations were obtained. Once they got them they proceeded to use them with their slide rules, and get results which were necessary for their work. And that led me of course to the view that this outlook was really the best outlook to have “. We suggest that Dirac's comment should be taken into account when choosing an appropriate method for ¹³C chemical shift prediction. It is quite probable that in many cases an “engineering outlook” represented by empirical methods can be successfully utilized without the additional work associated with the application of quantum-mechanical calculations. Speaking figuratively, it is possible to say that the empirical methods supply practicing chemists with a predictive tool that works automatically like an “engineering slide rule”.

The necessity of developing “engineering approaches” to improve the accuracy of

NMR chemical shift prediction was also recognized by theoretical chemists who suggested procedures for scaling non-empirically predicted chemical shifts ^[29] or scaling calculated isotropic tensors of magnetic shielding ^[46]. Aliev *et al* ^[47] suggested an universal equation for scaling ¹³C chemical shifts calculated with the GIAO B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d) protocol, which markedly reduces MAE values. Scaling procedures empirically take into account different effects (electron correlation, relativistic effects, interaction with solvent, etc.) influencing calculation accuracy. Reducing prediction errors is the main purpose of the scaling procedures. The MSTD approach mentioned above was also developed having in mind the same goal. One may say the non-empirical methods are indeed “semi-empirical” ones ^[40, 46]. The theoreticians conclude that “the choice of empirically scaled parameters could be mainly determined by an 'aesthetic drive', i.e. owing to the wish to consider apparently smaller values of the medium average error”^[20].

In our study, we made an attempt to compare the accuracy of ¹³C chemical shift prediction attained by QM and empirical methods for a large number of organic molecules. For this goal we extracted data from over 100 articles in the literature data associated with QM calculations published by different research groups over the last decade and compared the results with those obtained for the same structures using our HOSE code and ANN-based algorithmic approaches. We have been shown that, in general, QM methods are capable of providing the same accuracy as empirical approaches, but quite frequently they give larger MAE values, a situation that can be accounted for by the difficulties associated in selecting the appropriate calculation

protocols. A strategy for the combined application of both empirical and QM approaches is suggested.

Data selection and processing.

For our computational experiments we have found 205 structures for which both assigned experimental and QM-calculated ^{13}C chemical shifts were published in literature. Most of the data were obtained from the *Journal of Molecular Structure*, *Magnetic Resonance in Chemistry*, and other related journals. Only examples where the ^{13}C experimental spectra were of high quality were chosen for analysis. At the selection stage, we observed that some authors (for instance, ^[48]) used for the evaluation of QM methods experimental spectra which differed significantly from available reference spectra. In such cases we used the reference experimental ^{13}C NMR spectra which are present in the ACD/Labs database or in the Aldrich spectral atlas ^[49].

Figures 1 and 2 show the structure distribution as a function of the number of carbon atoms and molecular weight correspondingly. Almost 50% of the structures contained 10 or less carbon atoms and ~85% of the structures contained less than 20 carbon atoms. This distribution reflects the fact that QM chemical shift calculations were applied mostly to molecules of small and modest sizes. At the same time the figure demonstrates that QM chemical shift calculations are applicable to molecules with 20-30 carbon atoms, a common situation for natural products. Moreover, ^{13}C NMR prediction for a molecule of the size and complexity of Taxol has been reported recently ^[47]. Molecular masses can be evaluated from the plot shown in Figure 2.

Figure 1. Structure distribution as a function of the number of carbon atoms. The cumulative percentage is also displayed.

Figure 2. Structure distribution as a function of molecular weight. The cumulative percentage is also displayed.

All structures in the test set were input into ACD/Structure Elucidator software [15]. Carbon atoms were associated with both experimental and QM-calculated ^{13}C chemical shifts according to the assignment performed in corresponding articles. If the QM chemical shifts of a structure were computed using several different protocols, then the best approximation was chosen. In Structure Elucidator the structure set under test was included into a user database (UDB) where all results from the calculations could be stored. For all structures ^{13}C chemical shifts were calculated using ACD/CNMR Predictor [9] using all available algorithms: HOSE codes, NN and additive rules (increments, *Inc*). Before performing the HOSE based calculations the program checked whether a given structure was present in the ACD/Labs database (175,000 entries) employed for spectrum prediction. If a structure was detected in the database it was excluded from the spectrum prediction process. For each of the 205 structures the following values were estimated and stored in the user database relative to the HOSE, NN and QM methods of prediction:

- The experimental and predicted shifts for each individual carbon atom;
- The differences $\delta_{\text{exp}} - \delta_{\text{calc}}$ (with their signs) between the experimental and calculated chemical shifts for each carbon atom;
- Mean Absolute Error, MAE;
- Standard error (standard deviation, SD);
- Maximum absolute error (maximum deviation, d_{max})
- The regression parameters from linear regression (r , R^2 , SE, slope a , intersect b , etc.)

For every structure plots showing the $\delta_{\text{calc}} = \delta_{\text{exp}}$ line (45-degree line) and linear regression lines for QM, HOSE and NN shift predictions were generated. Utilizing the UDB allows

us to access a routine which automatically produces electronic tables containing comprehensive statistical and descriptive information related both to each structure and to the full structural set. The obtained statistical data and plots were carefully analyzed.

RESULTS AND DISCUSSION.

Statistical comparison of methods.

The quantitative parameters characterizing the accuracy of the empirical and QM methods of ^{13}C NMR chemical shift prediction for the set of structures under examination are presented in Table 1.

Table 1.

The table shows that for the given test set of molecules the MAE value obtained for the HOSE-based prediction approach is less than half the value calculated when QM methods were utilized. MAE(NN) is less than MAE(QM) by a factor of 1.7. An analogue trend is observed for MAE(Inc) - the fastest method of chemical shift prediction based on additive rules^[17], while not the most accurate, also exceeds the QM methods in average precision.

Figures 3 and 4 show a plot of the MAE and maximal deviations d_{max} values found by the HOSE, NN and QM methods determined for every structure.

Figure 3. Mean absolute errors (MAE) calculated by QM, HOSE and ANN methods.

Figure 4. Maximum deviations (d_{max}) calculated by QM, HOSE and ANN methods.

Visual assessment allows us to conclude that the majority of MAE values calculated by all three methods are less than 4 ppm, while deviations exceeding 4 ppm were shown

mainly for the QM predictions. In this case the QM predictions also produce large deviations with values larger than those delivered by the empirical methods. The average values of the *maximum* deviations d_{max} are 4.75, 5.15 and 7.40 ppm for HOSE, ANN and QM approaches respectively.

Figure 5 shows a comparison of the errors associated with all prediction methods.

Figure 5. A comparison plot of the mean absolute errors established for HOSE, ANN and QM methods. The last black column means that the MAE(QM) exceeds 8 ppm for 25 structures.

The histogram shows that 60-70% of the MAE values provided by the empirical methods are less than 2 ppm and 90% –were less than 3 ppm. The corresponding percentages related to the QM methods are 45% and 60% respectively.

The results of a linear regression calculations performed for 2531 experimental and predicted ^{13}C chemical shifts are presented in Figures 6-8.

Figure 6. A linear regression plot showing the dependence of HOSE-based predicted chemical shifts versus experimental shifts. The linear regression equation:

$$\delta_{calc}=0.9991\delta_{exp}+0.0199, R^2=0.9975$$

Figure 7. A linear regression plot showing the dependence of NN-based predicted chemical shifts versus experimental shifts. The linear regression equation:

$$\delta_{calc}=0.9934\delta_{exp}+0.5916, R^2=0.9970$$

Figure 8. A linear regression plot showing the dependence of QM-based predicted chemical shifts versus experimental shifts. The linear regression equation: $\delta_{\text{calc}} = 0.9942\delta_{\text{exp}} + 1.0883$, $R^2 = 0.9906$

Comparison of the plots and statistical parameters calculated for the examined methods shows that all three models are characterized by acceptable quality. However, both visual inspection and comparison of the linear regression statistical terms shows that the quality gradually decreases in the following order: HOSE > NN > QM with the quantum-mechanical based predictions showing the poorest performance. The HOSE plot practically coincides with the 45°-grade line ($\delta_{\text{calc}} = \delta_{\text{exp}}$) and is almost coincident with the δ_{exp} axis zero point, while the QM plot is shifted up by 1 ppm, admittedly a small but notable difference. Larger scattering is observed in the QM plot in the interval 100-200 ppm indicating a decrease in the prediction accuracy. As mentioned earlier Aliev *et al* ^[47] suggested a universal equation $\delta_{\text{scal}} = 0.95\delta_{\text{calc}} + 0.3$ for scaling the ^{13}C chemical shifts calculated using a GIAO protocol B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d) (SHIFTS//GEOMETRY). The potential application of this equation to the >2500 chemical shifts calculated by *different* protocols to improve the average MAE value was

investigated. When scaling was applied the MAE *increased* from 3.29 ppm to 4.77 ppm and the error distribution shifted to the side of positive axis: the scaled chemical shifts in general were now underestimated (see Supporting materials, Figures 1S-3S) especially in the region 100-200 ppm. The suggested scaling equation may thus only be valid when a specific protocol is used.

The results were investigated in more detail specifically examining the calculated MAE values for the various hybridization states: CH₃, CH₂, CH and quaternary carbons. To extract statistical significance from the analyzed parameters atom types for which there were less than 50 representatives in the dataset were excluded from consideration. Following this process produced an atom set belonging only to cyclic structures (Table 2). This observation is accounted for by the fact that almost all compounds examined by QM chemical shift predictions were related to ring systems, mainly to natural products. The atom lists presented in Table 2 are ordered according to both the number of attached hydrogen atoms and the type of hybridization (the ordering also approximately corresponds to increasing chemical shifts) to ease investigation of patterns in the values obtained by QM and empirical methods.

Table 2.

Figure 9. A histogram of the mean absolute errors (MAE) associated with the corresponding ring carbon atoms in different hybridization states. The symbols C(ar) and CH(ar) denote atoms belonging to aromatic rings.

Figure 10. A scatterplot of the MAE values corresponding to different hybridization states of carbon atoms in cyclic structures. The symbols C(ar) and CH(ar) denote atoms belonging to aromatic rings.

The histogram presented in Figure 9 allows visual comparison of the MAE values associated with different atom types, while Figure 10 shows the corresponding scatter plots. It is evident that the accuracy associated with the *empirical* methods is essentially independent of the carbon atom type. This implies approximately equal reliability for the calculated shifts across the full chemical shift scale represented (0-200 ppm). In contrast, there is dependence between the MAE values and the atom types observed for QM-calculated points. A maximum MAE(QM) value of 5.18 ppm is observed for non-aromatic =C_q atoms which can be explained by the influence of substituents attached to quaternary sp²-hybridized carbons. Though it is also likely that the different number of

shifts for the non-aromatic and aromatic rings (188 for $=C_q$ and 405 for $=C(ar)$) leads to the observed difference. It has been noted ^[20] that the GIAO approximation of DFT based predictions frequently either overestimates or underestimates the predicted chemical shifts for sp^2 -hybridized carbon atoms depending on the calculation protocol used. This observation is in accord with the data presented here (Figures 9 and 10) for a large number of shifts (~ 1240). Figures 9 and 10 also clearly show that MAE(QM) values increase by a factor of 2 along the chosen plot order of CH3 to $=C_q$ carbon.

It was interesting to learn how the carbon atoms within the test set are distributed as a function of the differences between the experimental and calculated chemical shifts ($\delta_{exp} - \delta_{calc}$). The corresponding distribution plots computed for a deviation interval of ± 10 ppm with a summation step of 0.5 ppm are presented in Figure 11. The figure shows that the distribution corresponding to HOSE-based calculations is a near-normal distribution in nature and characterized by the sharpest peak. The error distribution for the NN approach is represented by a broad bell-shaped curve whose maximum is markedly shifted down relative to the maximum of the HOSE code distribution curve. The shape associated with the QM-distribution appears to be far from normal in nature. It has two additional maxima at ± 1 ppm and the negative wing abates markedly slower than the positive one. This observation confirms the fact that QM approach has a tendency to overestimate calculated chemical shifts when some frequently employed calculation protocols are used. ^[20]

Figure 11. The atom distributions with associated arithmetical differences between experimental and calculated chemical shifts ($\delta_{\text{exp}} - \delta_{\text{calc}}$).

Outliers and unusual structures.

It was interesting to consider the structures for which the ^{13}C chemical shift prediction by QM and/or empirical methods produced large MAE values. MAE values of close to 5 ppm are not rare cases for QM-based calculations (see Figure 5), and the structures for which $\text{MAE} > 5$ ppm was obtained at least by one of methods were examined. Typical structure-outliers with their corresponding MAE values and maximum errors d_{max} are presented in Table 1S (see Supporting materials). Analysis of the table shows that some large MAE values associated with the QM predictions relate to the presence of: halogen atoms, heteroatoms carrying unshared electron pairs and high molecular flexibility. The contributions from these factors have been discussed in many works devoted to QM chemical shift prediction (for instance, [20, 23, 50, 51]). Figures 12 and 13 show plots of the HOSE- and QM-calculated ^{13}C chemical shifts versus experimental shifts for all atoms included in the structures presented in Table 1S, 274 shifts in total.

Figure 12. A linear regression plot of HOSE-based predicted ^{13}C chemical shifts versus experimental shifts for atoms included in the structures listed in Table 1S.

Figure 13. A linear regression plot of QM-based predicted ^{13}C chemical shifts versus experiment shifts for atoms in structures listed in Table 1S.

=14.29, $d_{max}(NN)$ =17.12ppm, while the QM calculations predicted the ^{13}C NMR shifts more accurately giving $\text{MAE}(\text{QM}) = 3.92$ ppm. Using the facilities of ACD\CNMR Predictor to examine the calculation protocol we determined that the HOSE code algorithm failed to accurately predict the chemical shifts for two of the carbon atoms (those resonating at 179.5 and 113.8 ppm) because the data base has no reference structures containing the atoms with the necessary environments. Nevertheless, the program offered chemical shift values of 166.2 and 115. ppm corresponding to these atoms using as an approximation the NN algorithms.

The main application of chemical shift prediction is to confirm the correct structural hypothesis during the process of molecular structure elucidation. Therefore we investigated whether an empirical approach can be applicable to the identification of structure **1** in spite of the low prediction accuracy. The HMQC, HMBC and COSY data of structure **1** presented in the work ^[52] were input into the Structure Elucidator ^[15] software. The program automatically detected the presence of non-standard correlations (NSC) ^[53]. NSCs are HMBC and COSY correlations whose length exceeds 3 bonds. Because of the presence of these NSC so-called “fuzzy structure generation” ^[54] was initialized. Structure generation options were set which assume the presence of an unknown number, m , of NSCs having an unknown length in COSY and HMBC data. The following solution was found at a value of $m=5$: $k=10456 \rightarrow 5056 \rightarrow 2017$, $t_g=2$ m 58 s. In this representation k is number of structures that were generated (10,456), then stored after application of some filtering tools (5056) and finally saved after removal of duplicates (2017). The notation t_g indicates the CPU time consumed for the process of structure generation and filtering. According to our general CASE strategy ^[15, 55] the

final structures were then ranked by d_{NN} values, the average deviation between the neural net predicted chemical shifts and the experimental shifts. HOSE code based chemical shift predictions were then performed for the first 20 structures of the ranked file and then sorted based on increasing d_{HOSE} values. The first three structures ranked in ascending order of d_{HOSE} values are shown in Figure 14. As we see the suggested structure of daphnipaxinin was distinguished by the program to be the most probable. At the same time, automated ^{13}C NMR chemical shift assignment agreed with that suggested by the authors [30, 52]. The next two structures have slightly larger deviations and in addition they contain strained somewhat “exotic” fragments, which make them questionable.

Figure 14. The first three structures of the output file ordered in ascending order of d_{HOSE} values. The structure of daphnipaxinin is listed in first position.

The example shows that in spite of the unusual character of the structure and the large values of the deviations an “engineering approach” allows the program to correctly select this challenging structure from among 2000 candidate structures, though with very little preference on the closest members of an output file.

Bagno *et al* ^[30] also tested the method of QM-based ¹³C chemical shift prediction with other unusual structures which might seem challenging for empirical methods, namely strychnine, buletunone (**4**) and corianlactone (**5**).

We found that the empirical ¹³C NMR prediction for strychnine gave MAE(HOSE) = 0.61 ppm and MAE(NN) = 1.81 ppm, while the accuracy of the QM-based calculations performed by the authors ^[30] was characterized by MAE(QM) = 6 ppm. In respect to buletunone **4**, we have shown earlier ^[42] that application of Structure Elucidator allowed us to confidently identify this molecule from 2D NMR data with MAE(HOSE) and MAE(NN) equal to 0.63 and 1.99 ppm correspondingly (Bagno *et al* reported MAE(QM) = 5.3 ppm for this structure).

The uncommon nature of the corianlactone structure **5** did not prevent us from solving this problem using empirical methods of ¹³C chemical shift prediction using the StrucEluc system. The 2D NMR data of this compound were taken from the original publication ^[56] and input into the Structure Elucidator software. The following results were obtained: k=83→72→65, t_g = 4.7 s. The three best structures in the ordered output file are shown in Figure 15.

Figure 15. The first three structures of the ordered output file resulting from the structure elucidation of the corianlactone molecule (**5**) using StrucEluc.

The structure of corianlactone was confidently identified with the aid of the StrucEluc software in combination with ACD/CNMR Predictor. As we demonstrated previously^[43] empirical methods of ^{13}C chemical shift prediction can also be used for selecting the preferable configurations from a full set of stereoisomers associated with a given molecular structure. StrucEluc generated all 256 stereoisomers of corianlactone and the most probable relative configuration, as shown by structure **5**, was determined using HOSE- and NN-based ^{13}C NMR spectrum prediction. Stereoisomer **5** was ranked as the most likely isomer with MAE(HOSE)=2.93ppm and MAE(NN)=3.89 ppm while the MAE(QM) value found for structure **5** using the GIAO approach was 5.3 ppm^[30].

In a separate study^[51] Bagno *et al* carried out QM ^{13}C chemical shift calculations for structure **6**. The MAE(QM) value = 6.83 ppm and the authors concluded that the QM approach allows ^{13}C NMR prediction for a polar, flexible molecule in aqueous solution with a high level of accuracy, comparable to that obtained for less complex systems.

6

The application of empirical methods to structure **6** led to the following results: MAE(HOSE)=1.15 ppm, MAE(NN)=1.75 ppm. Figure 16 shows the linear regression plots for all three methods, and the corresponding R^2 parameters are: $R^2(\text{HOSE})=0.997$, $R^2(\text{NN})=0.998$, $R^2(\text{QM})=0.996$

Figure 16. Linear regression plots for structure **6** generated from HOSE, NN and QM methods of ^{13}C chemical shift prediction. The solid line and black squares are related to QM prediction, the dotted line – to both HOSE and NN. The HOSE and NN predictions practically coincide with the 45-degree line ($\delta_{\text{calc}} = \delta_{\text{exp}}$).

Analysis of the data shows that the correlation coefficients are almost the same for all three methods of ^{13}C chemical shift prediction. The HOSE- and NN-plots are practically overlapped with the 45-degree line ($\delta_{\text{calc}} = \delta_{\text{exp}}$) while the intercept for the QM-calculated line is equal to 7.7 ppm (MAE(QM) equal to 6.83 ppm). The example shows that the R^2 value characterizes only the point scattering relative to the regression line but not the real accuracy of the chemical shift calculation which is more convincingly evaluated by the MAE or standard deviation values. It is known ^[57] that a very high value of R^2 can arise even though the relationship between the two variables is non-linear, so the fit of a model should never simply be judged from the R^2 value. Meanwhile, researchers frequently qualify the quality of prediction mainly from the R^2 value.

When the capabilities of different methods of chemical shift prediction are compared it is desirable to quantify the difference between the corresponding plots. The better a model ($\delta_{\text{calc}} = a\delta_{\text{exp}} + b$) then the closer the plot should be to the “reference” 45-degree grade line $\delta_{\text{calc}} = \delta_{\text{exp}}$. The two parameters characterizing the proximity of a given linear plot to the reference line are the intercept b and the angle γ between the reference line and the regression line. This angle can be calculated using the equation $\text{arctg}(\gamma) = (b-1)/(b+1)$. We suggest that the real difference between the *calculated* and *reference* values δ_{calc} and δ_{exp} may be represented more visually if, along with statistical parameters, the quality of prediction is additionally characterized by the angle γ .

As an example, the ^{13}C chemical shifts associated with structure **2** were successfully predicted using the QM approach accompanied by chemical shift scaling to give MAE(QM)=2.48 ppm ^[58]. Empirical methods gave large deviations: MAE(HOSE)=6.11

ppm, MAE(NN)=5.86 ppm. The linear regression plots associated with this structure are shown in Figure 17.

Figure 17. Linear regression plots for structure **2** generated using HOSE, NN and QM methods of ^{13}C chemical shift prediction. The solid line and black squares represent the QM prediction. The dotted line corresponds both to the HOSE and NN predictions. The QM predictions practically coincide with the 45-degree line ($\delta_{\text{calc}} = \delta_{\text{exp}}$).

The figure shows that the QM calculations are practically superimposed on the ($\delta_{\text{calc}} = \delta_{\text{exp}}$) line while the HOSE and NN plots can be characterized by the angle $\gamma(\text{HOSE})=\gamma(\text{NN})=-4^\circ$; both lines project angle of 41° relative to the δ_{exp} axis. It is evident that the signs of the deviations $d_{\text{exp}}^{\text{mod}} = (\delta_{\text{exp}} - \delta_{\text{model}})$ will be different at the scale segments situated before and after the point of line intersection and this may relate to model quality.

For structure **3** shift calculation using both empirical and QM methods ^[59] led to large MAE values of 6-8 ppm, which was associated with significant declinations from the 45°-degree line.

Synergistic interaction between empirical and non-empirical methods.

This work has shown that, in principle, both QM and empirical calculations can be performed with sufficient accuracy to solve practical problems in organic chemistry. Nevertheless, for the examined structural set the average accuracy of QM methods is 1.5-2 times lower than the accuracy of empirical methods (see Table 1). It is obvious that empirical methods possess the following merits: a) they are fully automatic; b) they are fast (prediction speed is thousands of shifts per second); c) they are quite accurate (MAE=1.5-1.8 ppm); d) there are no limitations imposed by molecule size. In regards to prediction speed, molecular size and level of automation QM approaches are inferior to empirical ones and these limitations, probably, are unlikely to be overcome in the near future. *Accuracy* is therefore the main criterion where QM methods have the potential to complement empirical methods and, in theory, maybe even surpass them.

Empirical methods are known to suffer from at least one principal drawback: if the database created for HOSE prediction or the training set for the neural net algorithm do not contain specific atoms representing the atom environments existing in the molecule under investigation, then the empirical methods can fail to predict the chemical shift of such atoms with sufficient accuracy. In these situations QM methods can compensate for the lack of representative data. However, the problem of accuracy should be solved to allow QM methods to be considered as a real analytical tool. We believe that current advances in QM, HOSE and NN ^{13}C NMR chemical shift prediction allow for the creation of an efficient strategy for jointly utilizing both empirical and non-empirical methods to solve actual analytical problems.

The most important task requiring the application of chemical shift prediction is that of complete structure elucidation, including stereochemistry. Empirical methods have been successfully used in this field for many years. Considering the growing capabilities of non-empirical approaches it is possible to suggest the following strategy for a combined approach using both methods and, in theory, deliver a synergistic effect.

Recently ^[42] we demonstrated the advantages of a systematic approach to forming and verifying structural hypotheses. According to this approach, the most efficient strategy consists of applying the Structure Elucidator expert system for automatic generation of all (without exclusion) conceivable structural hypotheses with their subsequent verification using ¹³C NMR spectrum prediction. Experience accumulated over the last decade shows ^[60] that, in the overwhelming majority of cases, empirical methods allow the successful sorting of structures using MAE(HOSE) values and determination of the most probable structure. The most probable structure is that which satisfies all constraints imposed by both the 1D and 2D NMR spectra and has the minimal MAE(HOSE) value. Generally speaking this structure fully satisfies the partial axiomatic theory formulated regarding the given spectrum-structural problem ^[42]. If the MAE(NN) value is also minimal for the preferred structure this is considered as additional support for the selection made. We have observed ^[60] that if the difference between the average HOSE deviations $\Delta = d(2) - d(1)$ found for the second and first structures in the ordered structural file is >1 ppm then the selected structure is, as a rule, the correct one. Otherwise, the selected structure should be confirmed with additional data, both experimental and/or theoretical, including the application of chemical common sense.

For instance, in the case of daphnipaxinin, the difference in deviation values between the preferred and second structure is very modest: $\Delta=d(2) - d(1) = 0.13$ ppm. The identification of the appropriate structure would require additional experimentation (for instance, NOESY or ROESY data) or alternatively QM-based chemical shift calculation could be helpful. The size of the molecule can be an insurmountable hindrance for QM calculations. For instance, when we input into the StrucEluc software the 1D and 2D NMR (HSQC, HMQC, COSY) data for the recently published ^[61] molecule, belizeanolide (C₈₁H₃₂O₂₀), the following solution was obtained: $k=938044 \rightarrow 7845 \rightarrow 3926$, $t_g=3$ h 9 m.

Figure 18. The first three structures of the ordered output file resulting from the structure elucidation of belizeanolide molecule.

The three best structures identified by the program from nearly 4000 hypothetical molecules are shown in Figure 18. The correct structure was placed in third position. The difference in deviations $d(3) - d(1)$ is very small - 0.08 ppm. Here the QM ¹³C chemical shift calculation is unlikely to be helpful due to the large size of the molecule. In such a

situation only additional experimental data, chemical knowledge and chemical common sense can help solve the problem.

If questionable structures ranked first contain some fragment which seems “exotic” in nature, then it is possible to perform a preliminary search of this fragment in the database used for ^{13}C chemical shift prediction. Once it is identified that such a fragment is not contained within the database then a QM calculation could be applied to a rationally selected fragment from the molecule and could be used to deliver reliable chemical shifts which could then be merged in an appropriate fashion with the shifts which were calculated by HOSE and NN methods for the rest of the molecule. Of course, the shifts would be tagged appropriately to label their underlying prediction algorithm. This approach could also be used when the calculation protocol facility of the HOSE-based shift predictor informs the user that it is impossible to predict the chemical shifts for some atoms due to absence of related structures in the database. There are already publications where fragmental QM chemical shift calculations were utilized to select or confirm a structural hypothesis [35, 62].

It should be underlined that the rank-ordered StrucEluc output file contains structures for which all experimental NMR chemical shifts are already assigned in accordance with their 2D NMR correlations. This circumstance significantly simplifies application of the QM ^{13}C chemical shift prediction for selection of the “best” structure: the first several structures for which the QM calculations would be employed can be ranked in ascending order of MAE(QM) values as is commonly the case when HOSE and NN prediction approaches are used. An example demonstrating how the fast NN chemical shift prediction accompanied with bar-graph based spectrum comparison

allowed avoiding QM calculations was presented previously^[42]. In this case the correct structure was easily distinguished visually without utilizing any chemical shift assignment.

Since the shielding of nuclei resonating in a magnetic field crucially depends on their 3D coordinates, the calculation of the most probable stereo-configuration of a molecule followed by NMR chemical shift prediction is a conventional procedure for molecular stereochemistry determination. Nevertheless empirical methods of ¹³C chemical shift calculation have been shown^[43] to be useful for preliminary filtering of the full set of stereoisomers conceivable for a given chemical structure, as well as for determining the relative stereochemistry of comparatively rigid molecules by geometry optimization guided by spatial constraints produced on the basis of NOESY correlations^[63]. Since the time required for empirical NMR spectral prediction is negligibly small in comparison with that required for QM calculations it would be useful to empirically detect a set of the most probable stereoisomers prior to comprehensive QM-based investigations. A restricted set of several selected stereoconfigurations could be used as initial approximations necessary for the purpose of geometry optimization and theoretically resulting in reduced computational costs.

We hope that as QM methods for NMR spectrum prediction are improved and the choice of the appropriate calculation protocol becomes a user-independent procedure, these methods will be more readily available for solving different spectrum-structural problems. A reasonable combination of QM and empirical approaches should provide a synergistic effect and will make both approaches more powerful and amenable to be used for practical purposes.

Computational Details.

All calculations were performed using ACD/NMR predictor Version 12.00. A personal computer equipped with a 2.8 GHz Intel processor and 2Gb of RAM and running the Windows XP operating system was used. All computer programs are an integral part of the Structure Elucidator expert system. ^{13}C NMR chemical shift calculations require no intervention from the chemist and are performed fully automatically.

Conclusions

We have compared the accuracy of ^{13}C chemical shift prediction achieved by both quantum-mechanical (QM) and empirical methods. To achieve this goal we extracted from the literature data associated with QM calculations published by different research groups during the last decade and compared the results with those obtained for the same structures using HOSE code and neural network algorithms developed within our laboratory. In totally 2531 chemical shifts associated with 205 molecules were analyzed. It has been shown that, in general, QM methods are capable of providing similar but inferior accuracy to the empirical approaches, but quite frequently they give larger mean average error values. This is accounted for mainly with difficulties in selecting the appropriate calculation protocols and difficulties arising from molecular flexibility. The data show that the average accuracy of the QM methods is 1.5-2 times lower than the accuracy shown by the empirical methods. For the structural set examined in this work the following mean absolute errors were found: $\text{MAE}(\text{HOSE})=1.58$ ppm, $\text{MAE}(\text{NN})=1.91$ ppm, $\text{MAE}(\text{QM})= 3.29$ ppm.

A strategy of combined application of both the empirical and QM approaches is suggested. The strategy could provide a synergistic effect if the advantages intrinsic to each method are exploited. The suggested strategy requires verification on a diverse data set and our group welcomes cooperation with theoreticians interested in such a study. We have >300 problems, all related to natural products, for which structure elucidation from 1D and 2D NMR spectra has been performed using the StrucEluc system and using empirical methods for selection of the most probable structure. These data could provide an interesting dataset for further informative computational experiments.

References

- [1] J.-T. Clerc, H. A. Sommerauer. *Anal. Chim. Acta* **1977**, 95, 33.
- [2] Fürst A., E. Pretsch. *Anal. Chim. Acta* **1990**, 229, 17.
- [3] E. Pretsch, A. Fürst, M. Badertscher, R. Burgin, M. E. Munk. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 291.
- [4] R. B. Schaller, M. E. Munk, E. Pretsch. *J. Chem. Inf. Model.* **1996**, 36, 239.
- [5] H. Kalchhauser, W. Robien. *J. Chem. Inf. Comput. Sci.* **1985**, 25, 103.
- [6] W. Robien. *Nachr. Chem. Tech. Lab.* **1998**, 46, 74.
- [7] Modgraph, http://www.Modgraph.Co.Uk/product_nmr.Htm.
- [8] Upstream Solutions GMBH.
- [9] Advanced Chemistry Development. ACD/NMR Predictors. Prediction suite includes ¹H, ¹³C, ¹⁵N, ¹⁹F, ³¹P NMR prediction. .
- [10] W. Bremser. *Anal.Chim. Act. Comp. Techn. Optimiz.* **1978**, 2, 355.
- [11] J. Meiler, R. Meusinger, M. Will. *J. Chem. Inf. Comp. Sci.* **2000**, 40, 1169.
- [12] J. Meiler, W. Maier, M. Will, R. Meusinger. *J. Magn. Reson.* **2002**, 157, 242.

- [13] M. E. Elyashberg, A. J. Williams, G. E. Martin. *Prog. NMR Spectrosc.* **2008**, 53, 1.
- [14] M. E. Munk. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 997.
- [15] M. E. Elyashberg, K. A. Blinov, S. G. Molodtsov, A. J. Williams, G. E. Martin. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 771.
- [16] M. E. Elyashberg, K. A. Blinov, S. G. Molodtsov, Y. D. Smurnyy, A. J. Williams, T. S. Churanova. Computer-assisted methods for molecular structure elucidation: Realizing a spectroscopist's dream. *J. Cheminform.*, vol. 1:3, 2009.
- [17] Y. D. Smurnyy, K. A. Blinov, T. S. Churanova, M. E. Elyashberg, A. J. Williams. *J. Chem. Inf. Model.* **2008**, 48, 128.
- [18] K. A. Blinov, Y. D. Smurnyy, M. E. Elyashberg, T. S. Churanova, M. Kvasha, C. Steinbeck, B. E. Lefebvre, A. J. Williams. *J. Chem. Inf. Model.* **2008**, 48, 550.
- [19] K. A. Blinov, Y. D. Smurnyy, T. S. Churanova, M. E. Elyashberg, A. J. Williams. *Chemometr. Intell. Lab. Syst.* **2009**, 97, 91.
- [20] P. Cimino, L. Gomez-Paloma, D. Duca, R. Riccio, G. Bifulco. *Magn. Reson. Chem.* **2004**, 42, S26.
- [21] A. Balandina, A. Kalinin, V. Mamedov, B. Figadere, S. Latypov. *Magn. Reson. Chem.* **2005**, 43, 816.
- [22] N. J. R. Eikema Hommes, T. Clark. *J. Mol. Model.* **2005**, 11, 175.
- [23] A. R. Katritzky, N. G. Akhmedov, J. Doskocz, C. D. Hall, R. G. Akhmedova, S. Majumder. *Magn. Reson. Chem.* **2007**, 45, 5.
- [24] W. Migda, B. Rys. *Magn. Reson. Chem.* **2004**, 42, 459.
- [25] K. W. Wiitala, C. J. Cramer, T. R. Hoye. *Magn. Reson. Chem.* **2007**, 45, 819.

- [26] R. Infante-Castillo, S. P. Hernandez-Rivera. *J. Mol. Struct.* **2009**, 917, 158.
- [27] M. Karabacak, A. Coruh, M. Kurt. *J. Mol. Struct.* **2008**, 892, 125.
- [28] M. Karabacak, M. Cinar, A. Coruh, M. Kurt. *J. Mol. Struct.* **2009**, 919, 26.
- [29] G. Barone, L. Gomez-Paloma, D. Duca, A. Silvestri, R. Riccio, G. Bifulco. *Chemistry* **2002**, 8, 3233.
- [30] A. Bagno, F. Rastrelli, G. Saielli. *Chemistry* **2006**, 12, 5514.
- [31] A. Balandina, D. Saifina, V. Mamedov, S. Latypov. *J. Mol. Struct.* **2006**, 791, 77.
- [32] A. A. Balandina, V. A. Mamedov, E. A. Khafizova, S. K. Latypov. *Russ. Chem. Bull.* **2006**, 55, 2256.
- [33] P. Wipf, A. D. Kerekes. *Journal of Natural Products* **2003**, 66, 716.
- [34] K. N. White, T. Amagata, A. G. Oliver, K. Tenney, P. J. Wenzel, P. Crews. *J. Org. Chem.* **2008**, 73, 8719.
- [35] T. A. Johnson, T. Amagata, A. G. Oliver, K. Tenney, F. A. Valeriote, P. Crews. *J. Org. Chem.* **2008**, 73, 7255.
- [36] C. Fattorusso, E. Stendardo, G. Appendino, E. Fattorusso, P. Luciano, A. Romano, O. Taglialatela-Scafati. *Org. Lett.* **2007**, 9, 2377.
- [37] E. Fattorusso, P. Luciano, A. Romano, O. Taglialatela-Scafati, G. Appendino, M. Borriello, E. Fattorusso. *J. Nat. Prod.* **2008**, 71, 1988.
- [38] S. D. Rychnovsky. *Org. Lett.* **2006**, 8, 2895.
- [39] A. M. Sarotti, S. C. Pellegrinet. *J. Org. Chem.* **2009**, ASAP.
- [40] C. A. Franca, R. P. Diez, A. H. Jubert. *J. Mol. Struct. THEOCHEM* **2008**, 856, 1.
- [41] V. Barone, P. Cimino, O. Crescenzi, M. Pavone. *J. Mol. Struct.* **2007**, 811, 323.

- [42] M. E. Elyashberg, K. Blinov, A. W. Williams. *Magn. Reson. Chem.* **2009**, 47, 371.
- [43] M. E. Elyashberg, K. Blinov, A. W. Williams. *Magn. Reson. Chem.* **2009**, 47, 333.
- [44] I. Stappen, G. Buchbauer, W. Robien, P. Wolschann. *Magn. Reson. Chem.* **2009**, 47, 720.
- [45] P. A. M. Dirac. *History of twenties century physics: Proceedings of the international school of physics "enrico fermi". Course LVII*. Academic Press: London, 1977.
- [46] D. B. Chesnut. *Chem. Phys. Lett.* **2003**, 380, 251.
- [47] A. E. Aliev, D. Courtier-Murias, S. Zhou. *Mol. Struct. THEOCHEM* **2009**, 893, 1.
- [48] R. Infante-Castillo, L. A. Rivera-Montalvo, S. P. Hernandez-Rivera. *J. Mol. Struct.* **2008**, 887, 10.
- [49] C. J. Pouchert, J. Behnke. *Aldrich library of ¹³C and ¹H FT-NMR spectra* 1993.
- [50] K. Dybiec, A. Gryff-Keller. *Magn. Reson. Chem.* **2009**, 47, 63.
- [51] A. Bagno, F. Rastrelli, G. Saielli. *Magn. Reson. Chem.* **2008**, 46, 518.
- [52] S.-P. Yang, J.-M. Yue, . *Org.Lett.* **2004**, 6, 1401.
- [53] S. G. Molodtsov, M. E. Elyashberg , K. A. Blinov, A. J. Williams, G. M. Martin, B. Lefebvre. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 1737.
- [54] M. E. Elyashberg, K. A. Blinov, S. G. Molodtsov, A. J. Williams, G. E. Martin. *J. Chem. Inf. Model.* **2007**, 47, 1053.

- [55] K. A. Blinov, D. Carlson, M. E. Elyashberg, G. E. Martin, E. R. Martirosian, S. G. Molodtsov, A. J. Williams. *J. Magn Reson. Chem.* **2003**, *41*, 359.
- [56] Y.-H. Shen, S.-H. Li, R.-T. Li, Q.-B. Han, Q.-S. Zhao, L. Liang, H.-D. Sun, Y. Lu, P. Cao, Q.-T. Zheng. *Org. Lett.* **2004**, *6* (10), 1593.
- [57]
- [http://www.babylon.com/definition/Multiple_regression_correlation_coefficient_\(R2\)/English](http://www.babylon.com/definition/Multiple_regression_correlation_coefficient_(R2)/English).
- [58] M. Szafran, P. Barczynski, A. Komasa, Z. Dega-Szafran. *J. Mol. Struc.* **2008**, *887*, 20.
- [59] O. Tsikouris, T. Bartl, J. Tousek, L. N., T. Tite, P. Marakos, N. Pouli, E. Mikros, R. Marek. *Magn. Reson. Chem.* **2008**, *46*, 643.
- [60] M. E. Elyashberg, K. A. Blinov, A. J. Williams, S. G. Molodtsov, G. E. Martin. *J. Chem. Inf. Model.* **2006**, *46*, 1643.
- [61] J. G. Napolitano, M. Norte, J. M. Padron, J. J. Fernandez, A. H. Daranas. *Angew. Chem. Int. Ed.* **2009**, *48*, 796.
- [62] D. Sanz, R. M. Claramunt, A. Saini, V. Kumar, R. Aggarwal, S. P. Singh, I. Alkorta, J. Elguero. *Magn. Reson. Chem.* **2007**, *45*, 513.
- [63] Y. D. Smurnyy, M. E. Elyashberg, K. A. Blinov, B. Lefebvre, G. E. Martin, A. J. Williams. *Tetrahedron* **2005**, *61/42*, 9980.

Tables

Table 1. Average statistical parameters calculated for the test set of molecules^a.

Method	MAE, ppm	SD, ppm	d(max), ppm
HOSE	1.58	2.55	18.9
NN	1.91	2.79	21.7

Inc	2.15	3.12	22.2
QM	3.29	4.98	28.3

^aThe total number of chemical shifts was 2531. MAE is calculated by summation of absolute errors found for each carbon atom divided by the total number of shifts.

Table 2. The mean absolute errors (MAE) corresponding to the ring carbon atoms in different hybridization states. The symbols C(ar) and CH(ar) denote atoms belonging to aromatic rings.

	sp ³				sp ²			
	CH ₃	CH ₂	CH	C _q	=CH	CH(ar)	C(ar)	C _q
Count ^a	273	459	278	99	59	586	405	188
HOSE	1.51	1.46	1.97	1.34	1.90	1.20	2.05	1.79
NN	1.61	1.79	2.40	1.87	2.61	1.51	2.20	2.46
QM	2.35	1.66	2.61	2.65	2.91	3.64	4.72	5.18

^aTotal number of shifts used is 2347 out of a total of 2531.