ChemComm

COMMUNICATION

The Optimal DFT Approach in DP4 NMR Structure Analysis – Pushing the Limits of Relative Configuration Elucidation

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Kristaps Ermanis,[a] Kevin E. B. Parkes,[b] Tatiana Agback[b,c] and Jonathan M. Goodman*[d]

What computational methods should be used to achieve the most reliable result in computational structure elucidation? A study on the effect of quality and quantity of geometries on computational NMR structure elucidation performance is reported. Semi-empirical, HF and DFT methods were explored, and B3LYP optimized geometries in combination with mPW1PW91 shifts and M06-2X conformer energies was found to be best. The required number of conformers considered has also been investigated, as well as several methods for the reduction of this number. Clear guidelines for the best computational NMR structure elucidation methods for different levels of available computing power are provided.

Introduction

Determination of the structure of natural and synthetic organic molecules remains a challenge in many cases and methods for computational NMR prediction have become invaluable tools to facilitate this process.[1] The determination of relative configuration for complex natural and synthetic products is an especially difficult task, for which increment and machine-learning methods are less effective than DFT Gauge Invariant Atomic Orbital (GIAO) calculations. A key part of computational structure elucidation is deciding which of the candidate calculated spectra match the experimental data best. Mean absolute error, corrected mean absolute error and correlation coefficient can be used for this, but probabilistic CP3 and DP4 measures,[2,3,4] provide a clearer guide to assignment. These work by assigning probabilities to the NMR prediction errors for each diastereomer and comparing them to give the overall probability for a particular structure assignment. Modified DP4 models[5] and neural network analyses[6,7] have also been reported.

The key parts of computational NMR structure elucidation are (i) conformational search; (ii) DFT structure optimization (optional); (iii) DFT energy calculation; (iv) NMR shift calculation; (v) statistical decision on which set of computed NMR shifts fit the experimental data best. We have recently reported optimization of the DP4 statistical models and the DFT conditions for the computational elucidation of relative configuration of complex natural products and drug compounds.[8] The mPW1PW91 functional for shift calculation and M06-2X functional for conformer energy calculation, without complex statistical analyses, give particularly good results.

While we have shown from small studies that MMFF geometries are adequate for NMR prediction and no costly DFT optimization is necessary,[2,3] we decided to revisit this in a larger study. Specifically, we wondered if geometry optimization using methods cheaper than DFT, like semi-empirical or HF[9] methods, might be beneficial. We also investigated geometries optimized using two DFT methods – B3LYP[10] and M06-2X.[11] In addition, we have investigated the selection of the conformations required for these calculations: might a representative subset be as effective for structural elucidation as a complete list, whilst requiring substantially less computational time.

Results and Discussion

Study was conducted on a subset of the 25 compounds used in a previous study (Figure 1).[8,12] The set was focused on smaller structures to offset the computational cost involved in extensively testing geometry optimization procedures. The previously generated conformational searches were used as starting points for all geometry optimizations. In the case of MMFF geometries, no further optimization was done. In the case of PM7,[13] the geometry optimization was done using MOPAC.[14] PM6,[15] HF, B3LYP and M06-2X geometry optimizations were done using Gaussian.[16] Chemical shifts were calculated using GIAO[17] and the B3LYP,[10] mPW1PW91[18] or M06-2X[11] functionals. This also included a mixed DFT

[a] Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK. E-mail: ke291@cam.ac.uk
[b] Medivir AB, PO Box 1086, SE-141 22 Huddinge, Sweden
[c] Department of Molecular Sciences, Swedish University of Agricultural Sciences, PO Box 7015, SE-750 07 Uppsala, Sweden
[d] Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK. E-mail: jmg11@cam.ac.uk; Tel: +44 (0)1223 336434

† Electronic supplementary information (ESI) available: statistical model parameters, workflow performance data, per compound NMR spectra prediction and structure elucidation data, references for the sources of NMR data used in the study. DP4 automation tool PyDP4 is available at https://github.com/KristapsE/PyDP4/
All of the computational input and output files are available at https://doi.org/10.17863/CAM.32463

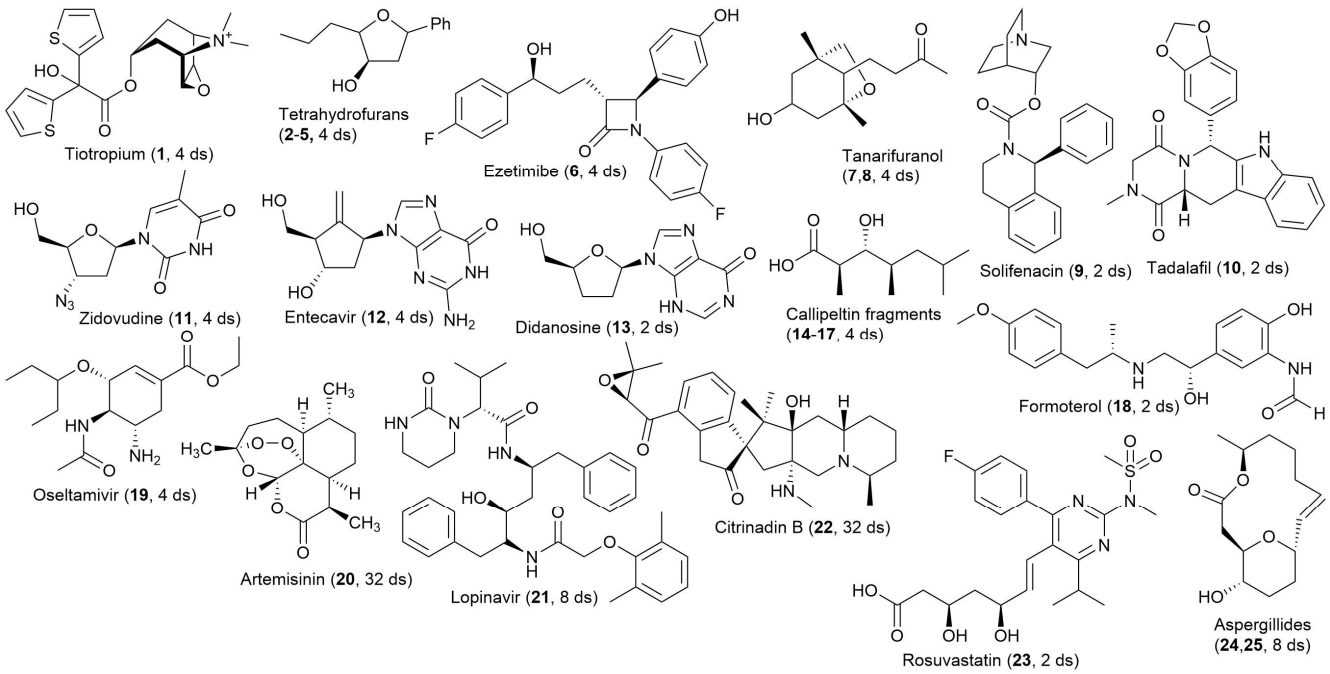

Figure 1 Compounds studied. Compound numbers for compounds with experimental data, and the total number of diastereomers considered shown in parentheses

calculation, were shifts were calculated using mPW1PW91, and the energies were calculated with M06-2X. The PCM solvent model was used for all DFT calculations. 6-311G* or 6-31G** basis sets were used for all DFT calculations (full details in SI). The statistical models used for DP4 probability calculation were generated for each workflow separately, and the statistical model parameters for the three best-performing workflows are provided in the SI.

In all cases both proton and carbon NMR shifts were calculated and the corresponding experimental data were used to test the NMR prediction and structure elucidation performance. We found that PM6 geometries gave larger carbon and proton chemical shift errors in most cases (Figure 2 for carbon data; proton data in SI. MAE is the Mean Absolute Error). This resulted in significantly reduced performance in structure elucidation (Figure 2), regardless of the computational conditions used for the shift calculation. PM7 geometries gave carbon chemical shift errors comparable to MMFF geometries, but the errors for proton spectra were still larger than when using MMFF. When combined with B3LYP NMR shifts, PM7 geometries gave better results than the corresponding MMFF workflow. However, when combined with mPW1PW91 or mixed mPW1PW91/M06-2X, PM7 still gave inferior results than in the MMFF case. All non-semi-empirical methods gave similar proton prediction accuracy and were not useful for method comparison (see SI).

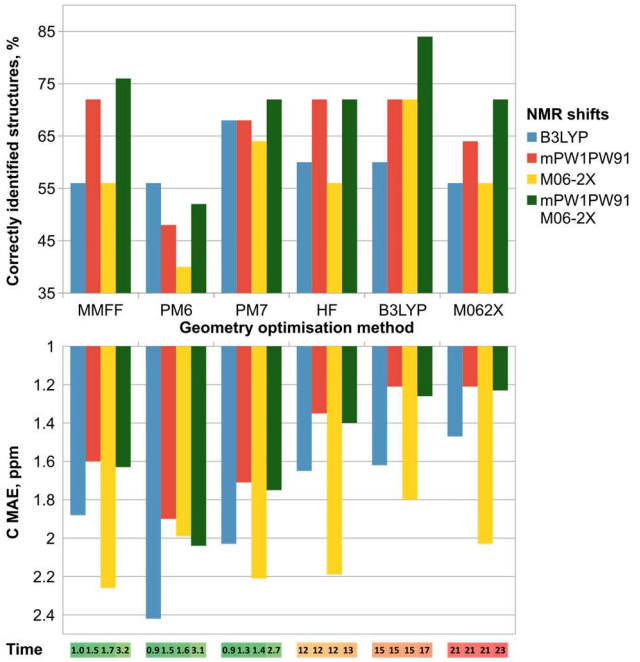

Figure 2 DP4 performance in relative configuration elucidation; Time relative to MMFF/B3LYP which is set to 1.0

Geometries optimized at HF level gave carbon NMR prediction errors that were about 0.2 ppm smaller than the ones arising from MMFF geometries. However, this did not give improvement in the identification of diastereomers, which was comparable to that given by MMFF geometries. Structure elucidation efficacy is not simply related to the MAE, as it depends on comparisons between similar molecules and not on absolute shifts.

B3LYP optimized geometries gave very accurate results, especially when shifts were calculated at mPW1PW91 level. In our dataset, a very low mean absolute error of 1.21 ppm was achieved for carbon NMR shifts. Despite the impressive accuracy, B3LYP optimized geometries in most cases gave similar performance in diastereomer identification to MMFF geometries. Over 50% improvement in performance was achieved only when B3LYP geometries were used with shifts calculated at mPW1PW91 level and the conformer energies were calculated at M06-2X level. M06-2X conformer energies appear to be crucial, since mPW1PW91 conformer energies in the same workflow gave similar results to MMFF. Finally, M06-

2X optimized geometries were tested. The carbon NMR prediction accuracy was very good and comparable to the one achieved from B3LYP. The diastereomer identification performance, however, was closer to the workflows using MMFF geometries.

In summary, B3LYP geometries combined with mPW1PW91 shifts and M06-2X energies give the best results in relative configuration elucidation. MMFF geometries with mPW1PW91 shifts and M06-2X energies also give good results. These workflows are quite different in the computational cost, however. While B3LYP geometries give the best results, this comes at a 16-fold increase in computational cost, when compared with the single-point B3LYP workflow. Therefore, we suggest using B3LYP optimized geometries when absolute best results are required and MMFF geometries when the computational cost is a concern.

Having explored the various levels of conformer geometry quality, we also wanted to explore the impact of the number of conformers considered on the efficacy of diastereomer identification. In the DP4 workflow, two parameters are used to control the number of conformers selected for DFT calculations from the conformational search: (i) MMFF energy threshold defines the maximum relative MMFF energy that a conformer can have to be selected for the DFT calculation. In the initial DP4 paper it was reported that 10 kJ/mol is a suitable value for this threshold, as no important conformers were missed. (ii) The second parameter is the maximum number of allowed conformers. When the number of conformers exceeds this limit after the removal of high-energy structures, the remaining structures are subjected to RMSD pruning. In this process, RMSD is calculated for all pairs of conformers. A low RMSD value indicates that the conformers are very similar and one of the pair can be removed without much effect on the NMR prediction accuracy. This is done starting with pairs with the lowest RMSD values and continuing until the total number of conformers no longer exceeds the limit.

We explored the effect of the reduction of the admissible conformer MMFF energy threshold, and the reduction of the maximum allowed number of conformers, on diastereomer identification performance. Using the computational data set already generated in this study, the DP4 probability evaluations were repeated on data that had fewer and fewer conformers due to a stricter MMFF energy threshold or a lower maximum conformer number criteria. The results of this study for the original workflow and the three current best computational workflows are shown in Figure 3.

As expected, reduction of MMFF energy threshold caused a smooth increase in the carbon NMR MAE (Figure 3A). Even a seven-fold reduction in the number of conformers caused less than 10% increase in the MAE. The effect on the diastereomer identification rate is more dramatic (Figure 3B) and particularly pronounced in the better workflows. For the B3LYP/mPW1PW91/M06-2X workflow, the reduction of the threshold to 1 kJ/mol reduced the number of structures seven-fold, and doubled the number of incorrectly identified compounds. The likely reason for this sensitivity is that structure elucidation process compares NMR shift predictions

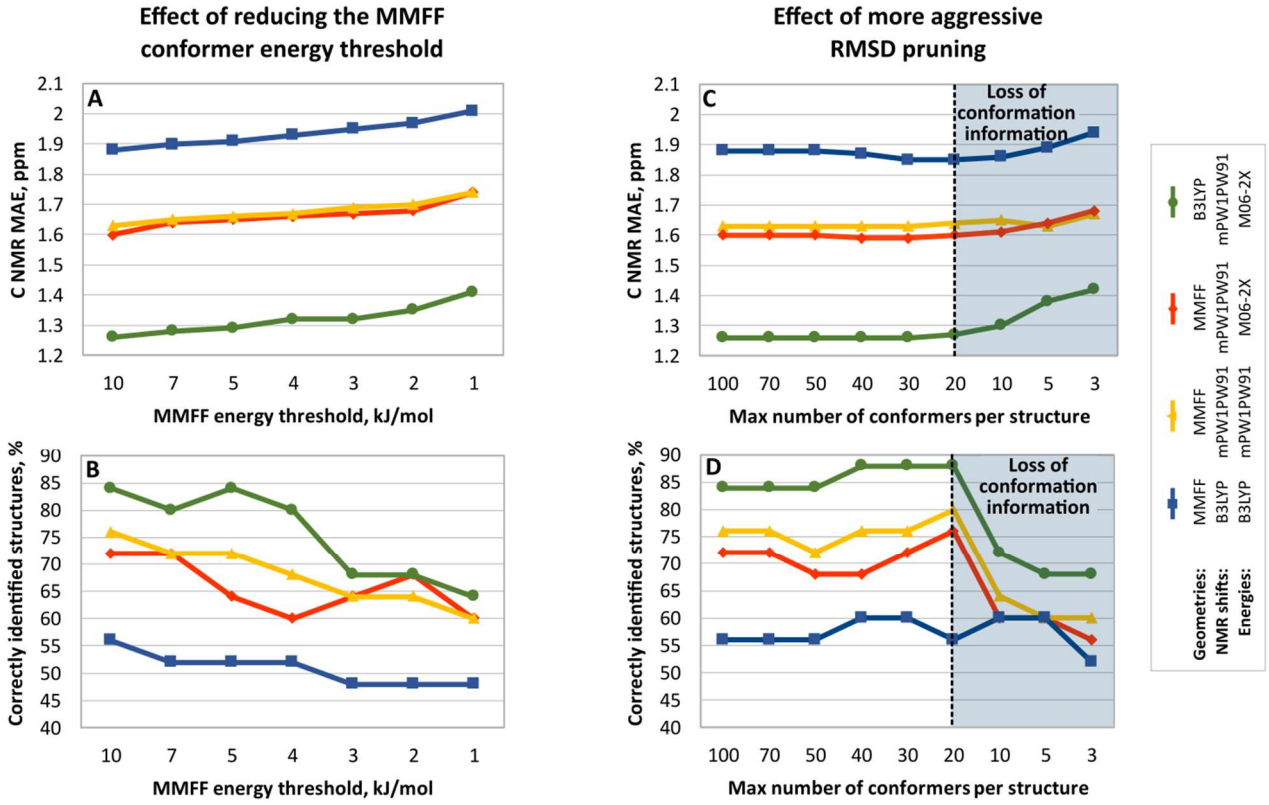

Figure 3 Exploring the effects of conformer number reduction on diastereomer identification performance (top, **lower is better**) and carbon NMR prediction accuracy (bottom, **higher is better**)

of several diastereomers among themselves, and so conformer pruning degrades shift prediction accuracy differently for each diastereomer and the balance between them is perturbed.

This re-emphasises that the correct elucidation of relative configuration is a harder problem than accurate prediction of NMR shifts, and there is no simple relationship between MAE and the efficacy of structural elucidation. Figure 3A shows the expected monotonous increase in MAE with decrease in energy threshold. Figure 3B is more complex, as the removal of higher energy structures may affect an important part of the structural comparison. Figures 3C and 3D show the different complexity as a result of removing structures which are geometrically similar but which may be low in energy. For example, in the B3LYP optimized geometry workflow even a small number of conformers will give much more accurate carbon NMR prediction than any other workflow, but the relative configuration elucidation performance will be essentially the same as for the MMFF geometry workflows.

The reduction in diastereomer identification with MMFF energy threshold reduction occurs even with small changes to the threshold. In contrast, increasingly strict limits for the number of conformers enforced by RMSD pruning is much more successful. A 20-conformer limit per structure has little effect on either structure identification (Figure 3D) or carbon NMR prediction accuracy (Figure 3C). Below this level there is a sharp drop in performance and increase in the MAE. This can be explained by the nature of RMSD pruning process. It strives to remove the most redundant conformers first, so the important conformational information is retained in the remaining conformers as far as possible. Once the remaining conformers cannot fully describe the conformational behaviour of the molecule the structure identification performance drops and NMR MAE rises. For the structures in our test set, 25 conformers per structure are sufficient for good results, but larger and more flexible molecules are likely to require more. Since the MAE rises at the same point that the identification performance drops, the MAE can be used as a guide to the number of conformers that are required for good results. The RMSD pruning process can order the structures and the chemical shift calculations run on the structures in the reverse order. As soon as the addition of new structures stops decreasing the MAE, the structural elucidation should be optimal.

RMSD pruning, therefore, appears to mitigate the increase in computational cost that comes with DFT geometry optimization rather effectively and is a much better approach than reducing the MMFF energy threshold. Even with this technique, DFT geometry optimisation remains much more expensive than the use of MMFF geometries in single-point DFT calculations. In situations when computational cost is a concern, the best approach is to do the NMR shift calculation on many lower-quality structures rather than a few higher-quality structures.

Conclusions

We have identified the following methods as optimal for DP4 NMR structure elucidation:

- Best method: B3LYP for geometry optimization, mPW1PW91 for shift calculation, M06-2X for conformer energies
- Cost-effective alternative: MMFF geometries, mPW1PW91 for shift calculation, M06-2X for conformer energies

Reducing the number of conformers with RMSD pruning effectively minimizes the computational cost of DFT geometry optimization. The number of conformers required for good results can be assessed by monitoring the changes to the MAE as more conformers are added in reverse-RMSD order. The process should be halted when the MAE reaches a minimum.

As previously, this study was greatly facilitated by the automated NMR calculation workflow PyDP4.[20] The latest version of the PyDP4 and additional scripts for custom statistical model generation can be obtained from the group website (http://www-jmg.ch.cam.ac.uk/tools/nmr), as well as from GitHub (https://github.com/KristapsE/PyDP4).

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors wish to thank Medivir, Leverhulme Trust and Isaac Newton Trust for the generous financial support.

References

1 For reviews in the area see: (a) M. W. Lodewyk, M. R. Siebert, D. J. Tantillo, *Chem. Rev.*, 2012, **112**, 1839; (b) D. J. Tantillo, *Nat. Prod. Rep.*, 2013, **30**, 1079
2 S. G. Smith, J. M. Goodman *J. Org. Chem.*, 2009, **74**, 4597.
3 S. G. Smith, J. M. Goodman *J. Am. Chem. Soc.*, 2010, **132**, 12946
4 Some examples of DP4 use in natural product structure elucidation and confirmation: (a) K. M. Snyder, J. Sikorska, T. Ye, L. Fang, W. Su, R. G. Carter, K. L. McPhail, P. H.-Y. Cheong *Org. Biomol. Chem.*, 2016, **14**, 5826 (b) L.-B. Dong, X.-D. Wu, X. Shi, Z.-J. Zhang, J. Yang, Q.-S. Zhao *Org. Lett.* 2016, **18**, 4498 (c) T. P. Wyche, J. S. Piotrowski, Y. Hou, D. Braun, R. Deshpande, S. McIlwain, I. M. Ong, C. L. Myers, I. A. Guzei, W. M. Wrestler, D. R. Andes, T. S. Bugni *Angew. Chem., Int. Ed.*, 2014, **126**, 11767; (d) I. Paterson, S. M. Dalby, J. C. Roberts, G. I. Naylor, E. A. Guzmán, R. Isbrucker, T. P. Pitts, P. Linley, D. Divlianska, J. K. Reed, A. E. Wright, *Angew. Chem., Int. Ed.*, 2011, **50**, 3219;
5 N. Grimblat, M. M. Zanardi, A. M. Sarotti, *J. Org. Chem.*, 2015, **80**, 12526.
6 A. M. Sarotti *Org. Biomol. Chem.* 2013, **11**, 4847
7 M. M. Zanardi, A. M. Sarotti *J. Org. Chem.*, 2015, **80**, 9371
8 K. Ermanis, K. E. B. Parkes, T. Agback, J. M. Goodman, *Org. Biomol. Chem.* 2017, **15**, 8998-9007
9 (a) R. Ditchfield *Mol. Phys.* 1974, **27**, 789; (b) K. Wolinski, J. F. Hinton, P. Pulay, *J. Am. Chem. Soc.* 1990, **112**, 8251; (c) M. Haser, R. Ahlrichs, H. Baron, P. Weis, H. Horn *Theor. Chim. Acta* 1992, **83**, 455.
10 (a) A. D. Becke, *Phys. Rev. A* 1988, **38**, 3098. (b) C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* 1988, **37**, 785. (c) A. D. Becke, *J.*

Chem. Phys. 1993, **98**, 5648. (d) P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch, *J. Phys. Chem.* 1994, **98**, 11623
11 Y. Zhao, D. Truhlar, *Theor. Chem. Acc.* 2008, **120**, 215.
12 Full references for the sources of all compounds spectral data can be found in the Supplementary Information
13 J.J.P. Stewart *J. Mol. Model.* 2013, **19**, 1.
14 MOPAC2016, James J. P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA, HTTP://OpenMOPAC.net (2016)
15 J.J.P. Stewart *J. Mol. Model.* 2007, **13**, 1173
16 Gaussian 09, Revision D.01, Gaussian, Inc., Wallingford CT, 2009, see full reference in the SI.
17 (a) F. London, *J. Phys. Radium* 1937, **8**, 397. (b) R. Ditchfield, *J. Chem. Phys.* 1972, **56**, 5688. (c) K. Wolinski, J. F. Hinton, P. Pulay, *J. Am. Chem. Soc.* 1990, **112**, 8251.
18 C. Adamo, V. Barone, *J. Chem. Phys.* 1998; **108**, 664
19 B. Mennucci and J. Tomasi, *J. Chem. Phys.*, 1997, **106**, 5151.
20 K. Ermanis, K. E. B. Parkes, T. Agback, J. M. Goodman, *Org. Biomol. Chem.* 2016, **14**, 3943