JOC Article

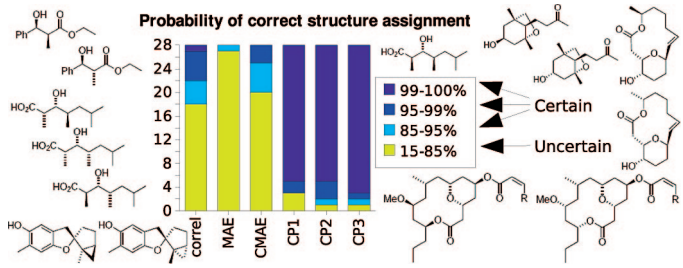
Assigning the Stereochemistry of Pairs of Diastereoisomers Using GIAO NMR Shift Calculation

Steven G. Smith and Jonathan M. Goodman*

Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, U.K.

j.m.goodman@cam.ac.uk

Received February 24, 2009



GIAO NMR chemical shifts have been calculated for a set of 28 pairs of diastereoisomers in order to test the ability of NMR shift calculation to distinguish between diastereomeric structures. We compare the performance of several different parameters for quantifying the agreement between calculated and experimental shifts from the point of view of assigning structures and introduce a new parameter, CP3, based on comparing differences in calculated shift with differences in experimental shift, which is significantly more successful at making correct structure assignments with high confidence than are the currently used parameters of the mean absolute error and the correlation coefficient. Using our new parameter in conjunction with Bayes' theorem, stereostructure assignments can be made with quantifiable confidence using shifts obtained in single point calculations on molecular mechanics geometries without computationally expensive ab initio geometry optimization.

Introduction

NMR spectroscopy is one of the most powerful tools for determining the structures of complex molecules such as natural products. Nevertheless, even with an extensive arsenal of 1D and 2D techniques, it is not uncommon for structures to be incompletely or incorrectly assigned.[1] Assignment of stereochemistry can be particularly challenging in conformationally flexible molecules where analysis of coupling constants and NOEs is not always reliable, and it is often necessary to resort to time-consuming total synthesis of potential diastereoisomers to find which of these matches the natural product.[2]

Recently there has been increasing interest in the use of ab initio NMR chemical shift prediction to aid structure assignment in difficult cases. The technique has been pioneered by Bifulco[3,4] and has since played key roles in the stereostructure assignment or reassignment of several natural products including hexacyclinol,[5] maitotoxin,[6] applidinones A−C,[7] gloriosaols A and B,[8] kadlongilactones D and F,[9] artarborol,[10] obtusallenes V−VII,[11] elatenyne,[12] spiroleucettadine,[13] samoquasine A,[14] mururin C,[15] ketopelenolides C and D,[16] dolichodial,[17] and hypurticin.[18] In

(1) Nicolaou, K. C.; Snyder, S. A. *Angew. Chem., Int. Ed.* **2005**, *44*, 1012–1044.
(2) Walsh, L. M.; Goodman, J. M. *Chem. Commun.* **2003**, 2616–2617.
(3) Barone, G.; Gomez-Paloma, L.; Duca, D.; Silvestri, A.; Riccio, R.; Bifulco, G. *Chem.—Eur. J.* **2002**, *8*, 3233–3239.
(4) Barone, G.; Duca, D.; Silvestri, A.; Gomez-Paloma, L.; Riccio, R.; Bifulco, G. *Chem.—Eur. J.* **2002**, *8*, 3240–3245.
(5) Rychnovsky, S. D. *Org. Lett.* **2006**, *8*, 2895–2898.
(6) Nicolaou, K. C.; Frederick, M. O. *Angew. Chem., Int. Ed.* **2007**, *46*, 5278–5282.
(7) Aiello, A.; Fattorusso, E.; Luciano, P.; Mangoni, A.; Menna, M. *Eur. J. Org. Chem.* **2005**, *2005*, 5024–5030.
(8) Bassarello, C.; Bifulco, G.; Montoro, P.; Skhirtladze, A.; Kemertelidze, E.; Pizza, C.; Piacente, S. *Tetrahedron* **2007**, *63*, 148–154.
(9) Pu, J.-X.; Huang, S.-X.; Ren, J.; Xiao, W.-L.; Li, L.-M.; Li, R.-T.; Li, L.-B.; Liao, T.-G.; Lou, L.-G.; Zhu, H.-J.; Sun, H.-D. *J. Nat. Prod.* **2007**, *70*, 1707–1711.
(10) Fattorusso, C.; Stendardo, E.; Appendino, G.; Fattorusso, E.; Luciano, P.; Romano, A.; Taglialatela-Scafati, O. *Org. Lett.* **2007**, *9*, 2377–2380.
(11) Braddock, D. C.; Rzepa, H. S. *J. Nat. Prod.* **2008**, *71*, 728–730.
(12) Smith, S. G.; Paton, R. S.; Burton, J. W.; Goodman, J. M. *J. Org. Chem.* **2008**, *73*, 4053–4062.
(13) White, K. N.; Amagata, T.; Oliver, A. G.; Tenney, K.; Wenzel, P. J.; Crews, P. *J. Org. Chem.* **2008**, *73*, 8719–8722.
(14) Timmons, C.; Wipf, P. *J. Org. Chem.* **2008**, *73*, 9168–9170.
(15) Hu, G.; Liu, K.; Williams, L. *J. Org. Lett.* **2008**, *10*, 5493–5496.

10.1021/jo900408d CCC: $40.75 © 2009 American Chemical Society
Published on Web 05/21/2009

synthetic chemistry, NMR shift calculations have been used in cases where a mixture of diastereoisomers is obtained in a reaction in order to identify the major and minor products; examples include a pair of bicyclic peroxides[19] and epoxides of carene.[20] The effect of using different levels of theory at various stages in the NMR shift calculation has also been extensively investigated.[21−29] The area has been reviewed.[30]

Our approach to stereostructure assignment by NMR shift prediction is to calculate the shifts for the potential structures (employing a Boltzmann weighted average of the shifts calculated for all low energy conformers), to compare to the available experimental data, and to decide which set of calculated data best matches which set of experimental data. The issue of how best to quantify the agreement between calculated and experimental data for the purposes of assigning the calculated data to structures is not a trivial one, because agreement will not be perfect in any example. Various approaches have been taken in the past.[3−20] The usual parameters used for the comparisons are

• **Correlation coefficient, *r*,** between calculated and experimental shifts. This has been used by Bifulco in his benchmarking studies[3,4] and also for studies on hexacyclinol,[5] aplidinones,[7] artarborol,[10] elatenyne,[12] dolichodial,[17] and carene epoxides.[20]

• **Mean absolute error, MAE**, calculated as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i}^{N} |\delta_{\mathrm{calc}} - \delta_{\mathrm{exp}}| \qquad (1)$$

was used in the obtusallene,[11] samoquasine A,[14] and dolichodial[17] studies.

• **Corrected mean absolute error, CMAE** (in some reports simply called MAE). In this case the data are empirically scaled according to

$$\delta_{\mathrm{scaled}} = \frac{\delta_{\mathrm{calc}} - \mathit{intercept}}{\mathit{slope}}$$

where *slope* and *intercept* are obtained from a plot of the calculated data against the experimental data to be assigned; the purpose of this approach is to remove systematic errors in the shift calculation. CMAE is then calculated according to

$$\mathrm{CMAE} = \frac{1}{N}\sum_{i}^{N} |\delta_{\mathrm{scaled}} - \delta_{\mathrm{exp}}| \qquad (2)$$

CMAE was used in Bifulco's benchmarking studies[3,4] and also for the studies on hexacyclinol,[5] aplidinones,[7] kadlongilactones,[9] artarborol,[10] elatenyne,[12] spiroleucettadine,[13] and ketopelenolides.[16]

The purpose of empirical scaling in the calculation of CMAE is to remove systematic errors in the shift calculation process. However, these systematic errors might be different for different types of carbon (for example, methyl groups vs aromatic carbons). An alternative approach to removing these errors is to compare the *differences* in calculated shift between the two possible diastereoisomers with the *differences* in experimental shift; cancellation of systematic errors should mean that these differences are reproduced more accurately than are the chemical shifts themselves.

This type of approach has been used by Sun and co-workers in assigning the structure of kadlongilactone D.[9] Differences in experimental shift between kadlongilactones A and D were compared to differences in calculated shift between the known structure of kadlongilactone A and the proposed structure of kadlongilactone D. A method based on taking differences between shifts of chemically similar carbons in a particular molecule and comparing these differences to the corresponding calculated differences has also been recently used by Belostotskii to study the conformation of haouamine A[31] and by Rodriguez in a study of how diastereoisomers of epoxycholestanes can be distinguished by calculation of NMR parameters.[32]

In this study, we address the question of assigning two sets of experimental data to two possible structures. For this situation, we present a systematic investigation of structure assignment using each of the parameters listed above and show how an estimate of the level of confidence in the results can be obtained for each case. We also propose a new parameter, CP3, which is based on comparing differences in calculated shift with differences in experimental shift and which is significantly more successful than the above parameters at making correct stereostructure assignments with high levels of confidence.

Although it is often the case that one only has experimental data for one structure, for example, the data for a natural product that represents one out of many possible diastereoisomers, and that our new parameter cannot be applied in these situations, the situation considered here in which one has two sets of experimental data to assign to two possible diastereoisomers is also common. For example, a stereoselective reaction in synthetic chemistry will typically give a small amount of a minor diastereoisomer as well as the major one, and the ability to reliably confirm which of the major and minor product is which diastereoisomer (and hence whether the major isomer is the desired one) is a crucial first step in optimizing the reaction to give the maximum yield of the desired isomer that is then used for the next step of the synthesis. Of the examples considered in this study (see Figure 1), aldols **1**, tetrahydrofurans **11**

(16) Fattorusso, E.; Luciano, P.; Romano, A.; Taglialatela-Scafati, O.; Appendino, G.; Borriello, M.; Fattorusso, C. *J. Nat. Prod.* **2008**, *71*, 1988–1992.
(17) Wang, B.; Dossey, A. T.; Walse, S. S.; Edison, A. S.; Merz, K. M., Jr. *J. Nat. Prod* **2009**, *72*, 709–713.
(18) Mendoza-Espinoza, J. A.; López-Vallejo, F.; Fragoso-Serrano, M.; Pereda-Miranda, R.; Cerda-García-Rojas, C. M. *J. Nat. Prod.* **2009**, *72*, 700–708.
(19) Griesbeck, A. G.; Blunk, D.; El-Idreesy, T. T.; Raabe, A. *Angew. Chem., Int. Ed.* **2007**, *46*, 8883–8886.
(20) Koskowich, S. M.; Johnson, W. C.; Paley, R. S.; Rablen, P. R. *J. Org. Chem.* **2008**, *73*, 3492–3496.
(21) Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. J. *J. Chem. Phys.* **2006**, *104*, 5497–5509.
(22) Forsyth, D. A.; Sebag, A. B. *J. Am. Chem. Soc.* **1997**, *119*, 9483–9494.
(23) Giesen, D. J.; Zumbulyadis, N. *Phys. Chem. Chem. Phys.* **2002**, *4*, 5498–5507.
(24) Cimino, P.; Gomez-Paloma, L.; Duca, D.; Riccio, R.; Bifulco, G. *Magn. Reson. Chem.* **2004**, *42*, S26–S33.
(25) Tormena, C. F.; da Silva, G. V. *J. Chem. Phys. Lett.* **2004**, *398*, 466–470.
(26) Bagno, A.; Rastrelli, F.; Saielli, G. *Chem.—Eur. J.* **2006**, *12*, 5514–5525.
(27) Wu, A.; Zhang, Y.; Xu, X.; Yan, Y. *J. Comput. Chem.* **2007**, *28*, 2431–2442.
(28) Wiitala, K. W.; Hoyle, T. R.; Cramer, C. J. *J. Chem. Theory Comput.* **2006**, *2*, 1085–1092.
(29) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *112*, 6794–6799.
(30) Bifulco, G.; Dambruoso, P.; Gomez-Paloma, L.; Riccio, R. *Chem. Rev.* **2007**, *107*, 3744–3779.
(31) Belostotskii, A. M. *J. Org. Chem.* **2008**, *73*, 5723–5731.
(32) Poza, J. J.; Jiménez, C.; Rodríguez, J. *Eur. J. Org. Chem.* **2008**, *2008*, 3960–3969.

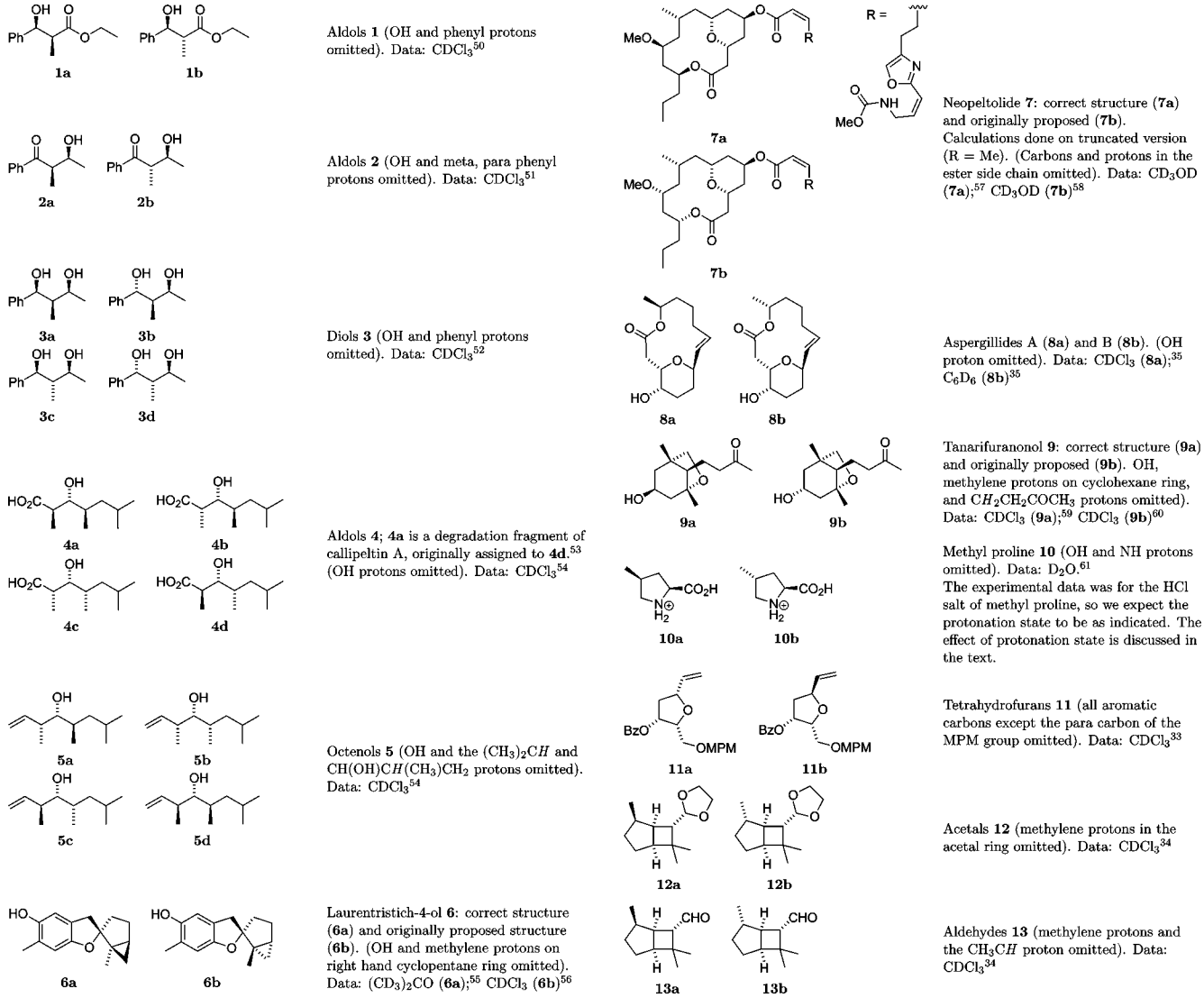


FIGURE 1. Set of structures used in this study. Protons for which the experimental data are unclear are omitted.

(intermediates in the synthesis of oocydin A[33]), and acetals **12**/aldehydes **13** (intermediates in the synthesis of raikovenal[34]) fall specifically into this category.

Second, although natural products often occur as a single diastereoisomer, this is by no means always the case. The example of kadlongilactones A and D (which are diastereomeric at a single center)[9] has already been mentioned, and the aforementioned dolichodial study investigated the three diastereomeric natural products dolichodial, anisomorphal, and peruphasmal.[17] Aspergillides A and B, which are considered in this study (**8a** and **8b** in Figure 1), also fit into this category. The structure of these two diastereomeric natural products could be narrowed down to two possible diastereoisomers by NMR analysis; assigning which was which then required chemical degradation and derivatization.[35] There are many other examples of natural products that have been isolated as two or more diastereoisomers; two of the most recent examples include toosendanic acids A and B[36] and oxylipins from *Dracontium loretense*.[37]

Computational Methods

All molecular mechanics calculations were performed using Macromodel[38] (Version 9.1 or 9.5) interfaced to the Maestro[39] (Version 7.5 or 8.0) program. All conformational searches used the Monte Carlo Multiple Minimum[40] (MCMM) or Systematic Pseudo Monte Carlo[41] (SPMC) method and the MMFF force field.[42] The searches were done in the gas phase, with a 50 kJ mol^{-1} upper

(33) Roulland, E. *Angew. Chem., Int. Ed.* **2008**, *47*, 3762–3765.
(34) Ko, C.; Feltenberger, J. B.; Ghosh, S. K.; Hsung, R. P. *Org. Lett.* **2008**, *10*, 1971–1974.
(35) Kito, K.; Ookura, R.; Yoshida, S.; Namikoshi, M.; Ooi, T.; Kusumi, T. *Org. Lett.* **2008**, *10*, 225–228.
(36) Sang, Y. S.; Zhao, C. Y.; Lu, A. J.; Yin, X. J.; Min, Z. D.; Tan, R. X. *J. Nat. Prod.* 2009ASAP article, doi:, 10.1021/np800669c.
(37) Benavides, A.; Napolitano, A.; Bassarello, C.; Carbone, V.; Gazzerro, P.; Malfitano, A.; Saggese, P.; Bifulco, M.; Piacente, S.; Pizza, C. *J. Nat. Prod.* 2009ASAP article, doi:, 10.1021/np8006205.
(38) Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. *J. Comput. Chem.* **1990**, *11*, 440–467.
(39) *Maestro, Version 8.0*; Schrödinger, LLC: New York, NY, 2007.
(40) Chang, G.; Guida, W. C.; Still, W. C. *J. Am. Chem. Soc.* **1989**, *111*, 4379–4386.
(41) Goodman, J. M.; Still, W. C. *J. Comput. Chem.* **1991**, *12*, 1110–1117.

energy limit and with the number of steps large enough to find all conformers at least 5−10 times.

Quantum mechanical calculations were carried out using Jaguar[43] (Version 6.5 or 7.0). Test calculations showed that the newer version of the software gave essentially identical results (in terms of geometries, energies, and NMR shielding constants) to the older version. For example, the mean absolute difference in calculated shift for aldol **2a** (see Figure 1) obtained by the standard procedure (see below) in the older and newer versions was only 0.004 ppm for ^{13}C and 0.0003 ppm for ^{1}H.

As in our previous study,[12] we employed the widely used B3LYP functional[44] and 6-31G(d,p) basis set[45] for all calculations. NMR shielding constant calculation used the GIAO method.[46]

Our previous study suggested that single point ab initio calculations on MMFF geometries (i.e., with no computationally expensive ab initio geometry optimization) give good results for shift calculation. To further verify this point and also to test the need for a solvent model, we calculated the NMR shifts for aldols **1** (see Figure 1) with and without ab initio geometry optimization and/or a continuum dielectric solvent model. Solvent energies were calculated using the Poisson−Boltzmann solvent model as implemented by Jaguar.[47] We also carried out similar calculations using Gaussian 03 (Revision D.01),[48] interfaced to the GaussView (Version 4.1.2)[49] program; these calculations confirmed that the results obtained are not strongly dependent on the software package used (details in Supporting Information). We found that the computationally more expensive methods (involving ab initio geometry optimization and solvent models, which increased the computer time required by an order of magnitude) did not give much more accurate calculated shifts (see Supporting Information for details). Further, in the cases where there was a slight improvement (mainly just for the proton data) the agreement for the incorrect structure assignment was also better, so that no advantage in the ability to distinguish correct and incorrect structural assignments was gained. This investigation reinforces the conclusion in our previous study[12] that single point gas-phase calculations on MMFF geometries as outlined below offer the best balance between accuracy and calculation speed.

The following procedure was used for NMR shift calculation. First, a molecular mechanics conformational search was carried out using the MMFF force field (gas phase). Second, all identified conformers within 10 kJ mol^{-1} of the global minimum were subjected to single point ab initio calculations of energy and GIAO shielding constants at the B3LYP/6-31G(d,p) level (again in the gas phase). The choice of 10 kJ mol^{-1} as the cutoff was a compromise between computer time and the risk of missing important conformers (as judged by their subsequent ab initio energies) as a result of inaccurate ordering of the conformer energies by the MMFF force field. Our previous work suggested that, for the particular molecules studied, relative molecular mechanics energies often differed from the ab initio ones by several kJ mol^{-1} but that very few important conformers would have been missed using a cutoff of 10 kJ mol^{-1}: none of the 47 conformers with populations (at 298 K) >5% were missed by this procedure.[12] We assumed that the same would apply for the molecules in the present work, since for some of the larger molecules (such as neopeltolide **7**) a higher cutoff would have given a prohibitively large number of conformers. This assumption turns out to be justified by the fact that the resulting calculated shifts allow successful structure assignment, as we show later.

To calculate NMR shifts for a particular species, the shielding constants were first averaged over symmetry-related positions in each conformer and then subjected to Boltzmann averaging over the conformers i according to

$$\sigma^x = \frac{\sum_i \sigma_i^x \exp(-E_i/RT)}{\sum_i \exp(-E_i/RT)} \qquad (3)$$

where σ^x is the Boltzmann averaged shielding constant for nucleus x, σ_i^x is the shielding constant for nucleus x in conformer i, and E_i is the potential energy of conformer i (relative to the global minimum), obtained from the single point ab initio calculation. The temperature T was taken as 298 K.

Chemical shifts were then calculated according to

$$\delta_{\mathrm{calc}}^x = \frac{\sigma^{\mathrm{o}} - \sigma^x}{1 - \sigma^{\mathrm{o}}/10^6} \qquad (4)$$

where δ_{calc}^x is the calculated shift for nucleus x (in ppm), σ^x is the shielding constant for nucleus x from eq 3 and σ^{o} is the shielding constant for the carbon or proton nuclei in tetramethylsilane (TMS), which was obtained from a B3LYP/6-31G(d,p) calculation on TMS.

Results and Discussion

1. Molecules Studied. We considered each of the molecules shown in Figure 1. Experimental data is available for all of these molecules, which range from small synthetic structures to natural products and include examples of originally misassigned compounds. The molecules in Figure 1 represent 28 pairs of diastereoisomers, each of which provides a test of our methodology for assigning two data sets to two possible structures.

In cases where the experimental shifts were incompletely assigned to nuclei (common with the carbon shifts unless 2D

(42) (a) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490–519. (b) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 520–552. (c) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 553–586. (d) Halgren, T. A.; Nachbar, R. B. *J. Comput. Chem.* **1996**, *17*, 587–615. (e) Halgren, T. A.; Nachbar, R. B. *J. Comput. Chem.* **1996**, *17*, 587–615. (f) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 616–641. (g) Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 720–729. (h) Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 730–748.

(43) *Jaguar, Version 7.0*; Schrödinger, LLC: New York, NY, 2007.

(44) (a) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100. (b) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789. (c) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652. (d) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(45) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.

(46) (a) London, F. *J. Phys. Radium* **1937**, *8*, 397–409. (b) Ditchfield, R. *J. Chem. Phys.* **1972**, *56*, 5688–5691. (c) Wolinski, K.; Hinton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251–8260.

(47) (a) Tannor, D. J.; Marten, B.; Murphy, R.; Friesner, R. A.; Sitkoff, D.; Nicholls, A.; Honig, B.; Ringnalda, M.; Goddard, III, W. A. *J. Am. Chem. Soc.* **1994**, *116*, 11875–11882. (b) Marten, B.; Kim, K.; Cortis, C.; Friesner, R. A.; Murphy, R. B.; Ringnalda, M. N.; Sitkoff, D.; Honig, B. *J. Phys. Chem.* **1996**, *100*, 11775–11788.

(48) Frisch, M. J. et al. *Gaussian 03, Revision D.01*; Gaussian, Inc.: Wallingford, CT, 2004.

(49) Dennington, R., II; Keith, T.; Millam, J. *GaussView, Version 4.1.2*; Semichem, Inc.: Shawnee Mission, KS, 2007.

(50) Jervis, P. J.; Cox, L. R. *Beilstein J. Org. Chem* **2007**, *3*, 6.

(51) Ahmad, K.; Taneja, S. C.; Singh, A. P.; Anand, N.; Qurishi, M. A.; Koul, S.; Qazi, G. N. *Tetrahedron* **2007**, *63*, 445–450.

(52) Abate, A.; Brenna, E.; Fuganti, C.; Gatti, F. G.; Giovenzana, T.; Malpezzi, L.; Serra, S. *J. Org. Chem.* **2005**, *70*, 1281–1290.

(53) Zampella, A.; D L.; Casapullo, A.; Minale, L.; Debitus, C.; Henin, Y. *J. Am. Chem. Soc.* **1996**, *118*, 6202–6209.

(54) Turk, J. A.; Visbal, G. S.; Lipton, M. A. *J. Org. Chem.* **2003**, *68*, 7841–7844.

(55) Sun, J.; Shi, D.; Ma, M.; Li, S.; Wang, S.; Han, L.; Yang, Y.; Fan, X.; Shi, J.; He, L. *J. Nat. Prod.* **2005**, *68*, 915–919.

(56) Chen, P.; Wang, J.; Liu, K.; Li, C. *J. Org. Chem.* **2008**, *73*, 339–341.

(57) Wright, A. E.; Botelho, J. C.; Guzmán, E.; Harmody, D.; Linley, P.; McCarthy, P. J.; Pitts, T. P.; Pomponi, S. A.; Reed, J. K. *J. Nat. Prod.* **2007**, *70*, 412–416.

(58) Youngsaye, W.; Lowe, J. T.; Pohlki, F.; Ralifo, P.; Panek, J. S. *Angew. Chem., Int. Ed.* **2007**, *46*, 9211–9214.

(59) Phommart, S.; Sutthivaiyakit, P.; Chimnoi, N.; Ruchirawat, S.; Sutthivaiyakit, S. *J. Nat. Prod.* **2005**, *68*, 927–930.

(60) Shiao, H.-Y.; Hsieh, H.-P.; Liao, C.-C. *Org. Lett.* **2008**, *10*, 449–452.

(61) Murphy, A. C.; Mitova, M. I.; Blunt, J. W.; Munro, M. H. G. *J. Nat. Prod.* **2008**, *71*, 806–809.

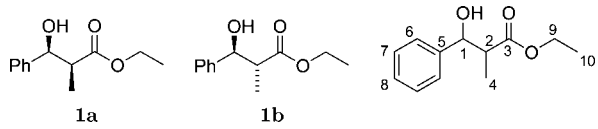


FIGURE 2. Aldols **1** and numbering system used.

NMR data was available), any remaining assignment was done by simply matching up the experimental shifts in order with the calculated shifts. This step is necessary because in order to calculate correlation coefficients, mean absolute errors, and other parameters it is necessary to know which experimental shift corresponds to which calculated shift. We ignored any nuclei for which the experimental data was unclear (for example proton shifts reported as ranges over several chemically distinct protons), and the omitted nuclei are indicated in Figure 1. The flexible side chain of neopeltolide **7** was truncated in order to reduce the number of conformers of the molecule and, hence, the computer time required. Nuclei close to the site of the truncation were also omitted as indicated in Figure 1. For methyl proline **10**, we considered three different protonation states (cationic, zwitterionic, and neutral) but focused on the cationic form, as we expect this to be the major species under the acidic conditions in which the experimental data were obtained.

Throughout this study we will use lower case letters (**a**, **b**) to refer to structures and the corresponding calculated data sets, and upper case letters (A, B) for the experimental data sets. If we have two experimental data sets, A and B, that we know correspond to the structures **a** and **b**, then there are two ways in which we can assign the structures: the right assignment, A = **a**, B = **b**, and the wrong assignment, A = **b**, B = **a**. What is the best way to distinguish them? We compare correlation coefficients, MAE, CMAE, and three comparison parameters that we develop: CP1, CP2 and CP3.

2. Example: Aldols 1. Our first test is aldols **1** (Figure 2). The two data sets can in fact be assigned using Heathcock's observation that the carbon of the α-methyl group (i.e., C4) in such compounds is generally more shielded in the *syn* isomer,[62] or Smonou's observation that the carbinol proton (H1) is more shielded in the *anti* isomer.[63] However, can they be distinguished using any of the methods for comparing calculated and experimental NMR shifts?

a. MAE, CMAE, and Correlation Coefficient. For the correct and incorrect assignments we can calculate values of MAE and CMAE using eqs 1 and 2 with the index *i* running over all carbon or hydrogen atoms in *both* molecules. The correlation coefficient for each assignment combination can be calculated in the same way. The results are given in Figure 3.

The "all data" graphs for MAE and CMAE show the geometric mean of the MAE or CMAE results for ^{13}C and ^{1}H; the normal arithmetic mean was not used since the values for ^{13}C are an order of magnitude larger than those for ^{1}H and so would have a disproportionately large effect on the average. In

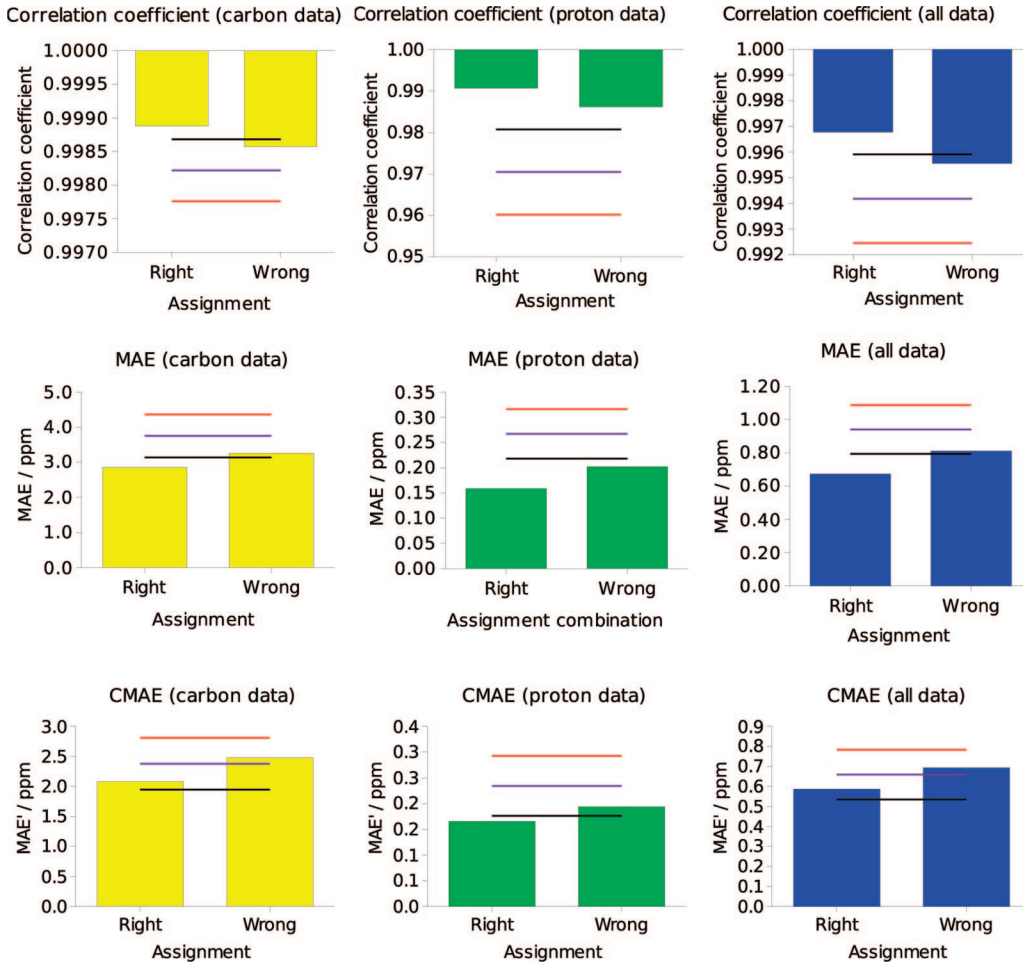


FIGURE 3. Assigning aldols **1** using the correlation coefficient, MAE, and CMAE after the proton and carbon signals have been matched. The horizontal lines represent the values of each parameter that are one, two, and three standard deviations away from the result expected for a correct structure.

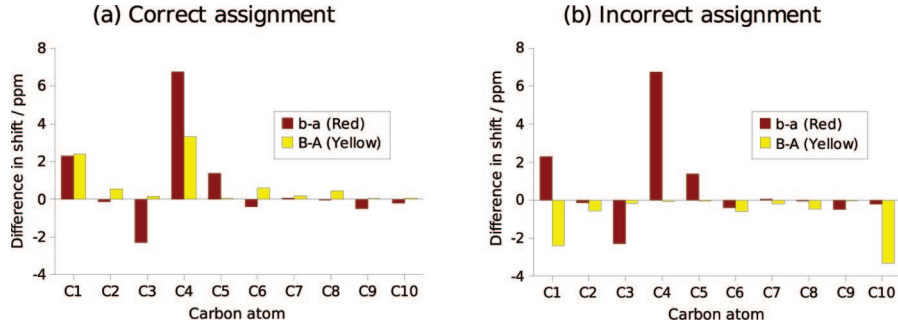


FIGURE 4. Differences in calculated and experimental shift for **1a** and **1b**. Good agreement is indicated by red and yellow pairs of bars pointing the same way.

a similar way the correlation coefficient for ^{1}H deviates from unity by roughly an order of magnitude more than that for ^{13}C, so the expression $r_{all} = 1 - \sqrt{(1 - r_C)(1 - r_H)}$ was used.

It is not obvious how large these parameters must be to be significantly good or significantly bad. For example, is $r = 0.99888$ substantially different to $r = 0.99857$? We address this issue by looking at the mean and variance of all parameters for the molecules under study (data in Supporting Information). We find, for example, that $\bar{r} = 0.99913$ for carbon for matching shifts (a correct assignment) and $\bar{r} = 0.99840$ for mismatched (an incorrect assignment). The difference of 0.00031 between $r = 0.99888$ and $r = 0.99857$ is fairly large in this context.

The horizontal lines on the graphs in Figure 3 represent the values of the correlation coefficient, MAE or CMAE that are one, two, and three standard deviations above (MAE, CMAE) or below (correlation coefficient) the value expected for a correct assignment. The expectation values and standard deviations were calculated from an analysis of the values of each parameter obtained for all of the pairs of diastereoisomers considered. Aldols **1** themselves were excluded from this analysis in order to avoid including the molecules under study in the data set of structures used to determine expectation values and standard deviations. In practice, this hardly affects the mean and standard deviation because the data set is large: we show later that including or excluding a particular pair of structures only changes the values by about 2%.

In each plot the value of the correlation coefficient, MAE, or CMAE for the correct assignment is within two standard deviations of the value expected for a correct structure assignment, but generally so too is the value for the incorrect assignment. Therefore, although in each case the correct assignment gives the best match, the incorrect combination cannot be ruled out with high confidence.

b. Comparison Parameters. Belostotskii[31] and Rodriguez[32] have recently pointed out that differences between the chemical shifts of similar carbons should be calculated more accurately than the shifts themselves because of cancellation of systematic errors. This can be useful for structure assignment.

Figure 4 plots the differences in calculated shift ($\delta_{\mathbf{b}} - \delta_{\mathbf{a}}$) and the differences in experimental shift ($\delta_B - \delta_A$) for both the correct and incorrect assignment combinations. For the correct assignment, the experimental difference at C1 and C4 is correctly reproduced by the calculated data. The other carbons show little difference in experimental shift and so are less useful for structure assignment. For the incorrect assignment, the agreement is much less good because the experimental difference at C1 now has the sign opposite to that calculated and the experimental shifts now show a big difference at C10 rather than at C4 as the calculations predict. This latter point is a consequence of the experimental resonances being incompletely assigned (see Figure 1); if they had been completely assigned then the yellow bars in Figure 4b would be a reflection in the x-axis of those in Figure 4a and the agreement would look even worse than it does here.

The information in each plot in Figure 4 can be combined into a single number by multiplying the red bars in each graph by the yellow bars and summing the products:

$$\sum_i (\delta^i_{\mathbf{a}} - \delta^i_{\mathbf{b}})(\delta^i_A - \delta^i_B)$$

A large positive value indicates good agreement (assignment likely to be correct), whereas a large negative value indicates poor agreement. Carbons for which there is very little difference in shift (for example C7) and which are, therefore, not useful for discriminating between structures, are automatically given a low weighting.

Dividing by the quantity $\Sigma_i(\delta^i_A - \delta^i_B)^2$, which represents the value of the above sum that would be obtained if all of the differences in experimental shift were reproduced perfectly by the calculation method, gives the new comparison parameter CP1:

$$\mathrm{CP1} = \frac{\sum_i \Delta_{exp}\Delta_{calc}}{\sum_i \Delta^2_{exp}} \qquad (5)$$

where Δ_{exp} and Δ_{calc} are the differences in the experimental and in the calculated shifts respectively. The value of CP1 is 1.54 for the correct assignment combination and −0.26 for the incorrect one.

If Δ_{calc} has the same sign but smaller magnitude than Δ_{exp}, then the effect is to reduce the value of CP1 (relative to the value obtained if all the shifts were perfectly reproduced by the calculations, i.e., $\Delta_{calc} = \Delta_{exp}$ for all nuclei). This is reasonable since the agreement is less good. On the other hand, if Δ_{calc} has the same sign but larger magnitude than Δ_{exp}, then the value of CP1 is increased, even though the agreement is also less good.

We correct for this problem in two ways and so generate two new comparison parameters; CP2 and CP3 correct the effect

(62) Heathcock, C. H.; Pirrung, M. C.; Sohn, J. E. *J. Org. Chem.* **1979**, *44*, 4924–4299.
(63) Kalaitzakis, D.; Smonou, I. *J. Org. Chem.* **2008**, *73*, 3919–3921.

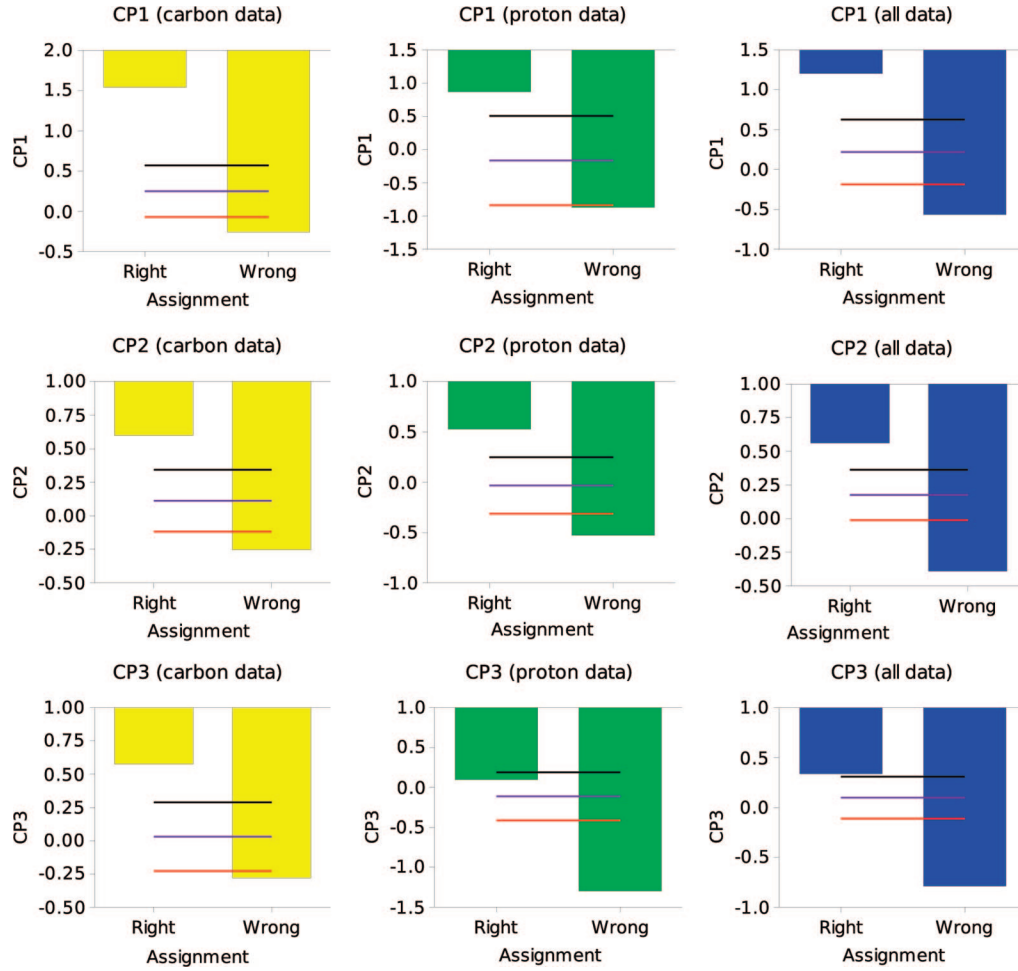


FIGURE 5. Assigning aldols **1** using the CP1−CP3 after the proton and carbon signals have been matched. The horizontal lines represent the values of each parameter that are one, two, and three standard deviations away from the result expected for a correct structure. In each case the correct assignment is picked out more clearly than in Figure 3.

by returning the same result if Δ_{calc} is x times too large as if it were x times too small.

$$CP2 = \frac{\sum_i f_2(\Delta_{exp}, \Delta_{calc})}{\sum_i \Delta_{exp}^2} \quad \text{where}$$

$$f_2(\Delta_{exp}, \Delta_{calc}) = \begin{cases} \Delta_{exp}^3/\Delta_{calc} \text{ if } |\Delta_{calc}/\Delta_{exp}| > 1 \\ \Delta_{exp}\Delta_{calc} \text{ otherwise} \end{cases} \quad (6)$$

$$CP3 = \frac{\sum_i f_3(\Delta_{exp}, \Delta_{calc})}{\sum_i \Delta_{exp}^2} \quad \text{where}$$

$$f_3(\Delta_{exp}, \Delta_{calc}) = \begin{cases} \Delta_{exp}^3/\Delta_{calc} \text{ if } \Delta_{calc}/\Delta_{exp} > 1 \\ \Delta_{exp}\Delta_{calc} \text{ otherwise} \end{cases} \quad (7)$$

CP2 has the property of being between +1 (perfect agreement, i.e., $\Delta_{calc} = \Delta_{exp}$ for all nuclei) and −1 (perfect disagreement, i.e., the structures have been assigned the wrong way around and $\Delta_{calc} = -\Delta_{exp}$). CP3 differs from CP2 in that it only applies the correction if Δ_{calc} is larger in magnitude than Δ_{exp} *and* has the correct sign. This might be useful if the experimental data sets being assigned do not necessarily correspond to the two calculated data sets (perhaps because two sets of experimental data are being assigned to four possible diastereoisomers), in which case "perfect disagreement" is less meaningful as we would not necessarily expect $\Delta_{calc} = -\Delta_{exp}$ for perfectly calculated shifts and an incorrect assignment.

The values of CP1, CP2, and CP3 for the right (A = **a**, B = **b**) and wrong (A = **b**, B = **a**) assignments using carbon and proton data are plotted in Figure 5. The "all data" graphs show the arithmetic mean of the carbon and proton values, and the horizontal lines represent the values of each parameter that are one, two, and three standard deviations below that expected for a correct assignment. As before the expectation values and standard deviations were obtained from an analysis of the values of the parameter in question obtained for all pairs of experimental data apart from aldols **1**. Details may be found in Supporting Information. As before, leaving out data for each pair at a time made little difference.

All three parameters pick out the right assignment with much greater confidence than the MAE, CMAE, and correlation coefficient (Figure 3). This can be attributed to (a) the removal of systematic errors in the shift calculation for a particular carbon or proton by taking differences and (b) the fact that only those resonances showing a significant difference in shift between the two diastereoisomers make a significant contribution to the values of CP1−CP3. It is interesting to note that the resonances making the most contribution are C4 (and to a lesser extent

C1) among the carbons (Figure 4) and H1 among the protons; these are precisely the shifts proposed as being diagnostic by Heathcock[62] and Smonou.[63]

c. Confidence Levels and Probabilities. The horizontal lines drawn on Figure 3 and Figure 5 at one, two, and three standard deviations away from the value expected for a correct assignment give an indication of how confident one can be in the assignment.

It would be useful to be able to quantify the probability that the assignment made is the right one. Consider as an example CP3 for the carbon data. For a correct structure assignment using carbon data, CP3 has an expectation value of 0.546 and a standard deviation of 0.258 (these were the values used to draw the horizontal lines on Figure 5). For aldols **1**, the right assignment (i.e., A = **a**, B = **b**) gave a value of CP3 (using the carbon data) of 0.575, which is only 0.114 standard deviations away from the expectation value for a correct assignment. Assuming that CP3 (for a correct assignment) has a normal distribution, the probability of getting such a value, i.e. one that deviates from the expectation value by at least this amount, *if* the assignment being made is correct is 0.909. However this is *not* the same as the probability that the assignment is correct given that the value has been obtained. In terms of conditional probability and using Bayes' theorem, the latter is given by[64]

$$P(\mathrm{AC_1}|\mathrm{value}) = \frac{P(\mathrm{value}|\mathrm{AC_1}) \times P(\mathrm{AC_1})}{P(\mathrm{value})} \quad (8)$$

In this equation, $P(\mathrm{AC_1})$ is the probability, *in the absence of any of the NMR prediction evidence*, that our proposed assignment combination (A = **a**, B = **b**) is correct. Assuming that the NMR shift prediction is the only evidence available for structure assignment, without it we have nothing to say that assignment A = **a**, B = **b** is more likely than the alternative A = **b**, B = **a** and so we must assign each a probability of 0.5.

Second, $P(\mathrm{value})$ is the probability that we got the value of CP3 that we did regardless of which assignment combination is correct; it may be calculated as the sum of the probabilities that (i) $\mathrm{AC_1}$ is true and the value is obtained and (ii) $\mathrm{AC_1}$ is false (i.e., the alternative assignment, $\mathrm{AC_2}$, is true) and the value is obtained:

$$P(\mathrm{value}) = P(\mathrm{value}|\mathrm{AC_1}) \times P(\mathrm{AC_1}) + P(\mathrm{value}|\mathrm{AC_2}) \times P(\mathrm{AC_2}) \quad (9)$$

We already know the value of $P(\mathrm{value}|\mathrm{AC_1})$ (it is 0.909 in this case), and we have decided that in the absence of the NMR prediction data the probabilities $P(\mathrm{AC_1})$ and $P(\mathrm{AC_2})$ are equal, i.e. 0.5. However, we still need to know $P(\mathrm{value}|\mathrm{AC_2})$ before we can evaluate $P(\mathrm{value})$ using eq 9 and hence $P(\mathrm{AC_1}|\mathrm{value})$ using eq 8.

$P(\mathrm{value}|\mathrm{AC_2})$ represents the probability of getting our value of CP3 if our assumed assignment is wrong, and so we need some information about the distribution of CP3 values for wrong assignments. In the analysis of the values of each parameter expected for a correct assignment, the expectation values and standard deviations for an incorrect assignment (i.e., two structures assigned the wrong way round) were also obtained. This analysis gave for CP3 (carbon data) an expectation value of −0.495 and a standard deviation of 0.542. This means that if $\mathrm{AC_1}$ is wrong, i.e., A = **b** and B = **a**, we should expect to get a value of CP3 near −0.495, and the probability of getting the value that we did get, namely, 0.575, is 0.0483 (assuming that values of CP3 for wrong assignments are normally distributed). Now we can use eqs 8 and 9 to calculate the probability that assignment combination 1 is correct in the light of our value of CP3 as

$$P(\mathrm{AC_1}|\mathrm{value}) = \frac{0.909 \times 0.5}{0.909 \times 0.5 + 0.0483 \times 0.5} = 0.950$$

i.e., 95.0%.

However, we have more information that we have not yet used. We have CP3 = 0.575 for one assignment *and* −0.280 for the other (see Figure 5). Using all this information, we can write, analogously to eq 8:

$$P(\mathrm{AC_1}|\mathrm{R_1\ and\ R_2}) = \frac{P(\mathrm{R_1\ and\ R_2}|\mathrm{AC_1}) \times P(\mathrm{AC_1})}{P(\mathrm{R_1\ and\ R_2})} \quad (10)$$

where $\mathrm{R_1}$ is the result obtained for CP3 assuming $\mathrm{AC_1}$ (i.e., 0.575) and $\mathrm{R_2}$ is the result assuming $\mathrm{AC_2}$ (i.e., − 0.280). Assuming that $\mathrm{R_1}$ and $\mathrm{R_2}$ are independent variables and using eq 9 to expand the denominator, eq 10 can be rewritten as

$$\begin{aligned} &P(\mathrm{AC_1}|\mathrm{R_1\ and\ R_2}) \\ &= \frac{P(\mathrm{R_1}|\mathrm{AC_1})P(\mathrm{R_2}|\mathrm{AC_1})P(AC_1)}{P(\mathrm{R_1\ and\ R_2})} \\ &= \frac{P(\mathrm{R_1}|\mathrm{AC_1})P(\mathrm{R_2}|\mathrm{AC_1})P(\mathrm{AC_1})}{P(\mathrm{R_1\ and\ R_2}|\mathrm{AC_1})P(\mathrm{AC_1}) + P(\mathrm{R_1\ and\ R_2}|\mathrm{AC_2})P(\mathrm{AC_2})} \\ &= \frac{P(\mathrm{R_1}|\mathrm{AC_1})P(\mathrm{R_2}|\mathrm{AC_1})P(\mathrm{AC_1})}{P(\mathrm{R_1}|\mathrm{AC_1})P(\mathrm{R_2}|\mathrm{AC_1})P(\mathrm{AC_1}) + P(\mathrm{R_1}|\mathrm{AC_2})P(\mathrm{R_2}|\mathrm{AC_2})P(\mathrm{AC_2})} \end{aligned} \quad (11)$$

Putting in the numbers for assigning aldols **1** using CP3 we find

$$\begin{aligned} &P(\mathrm{AC_1}|\mathrm{R_1\ and\ R_2}) \\ &= \frac{0.909 \times 0.692 \times 0.5}{0.909 \times 0.692 \times 0.5 + 0.0483 \times 0.00137 \times 0.5} \\ &= 99.9895\% \end{aligned}$$

corresponding to a very high probability that the combination (i.e., A = **a**, B = **b**) is correct. (The probability that the alternative assignment is correct is 0.0109%) In view of the CP3 graph for the carbon data in Figure 5 this is not too surprising, given that the bar for the right assignment is around the value expected for a correct assignment while that for the wrong assignment is more than three standard deviations below the value expected for a correct assignment near the value expected for an incorrect assignment. Thus, Bayes' theorem *increases* the certainty of our conclusion.

The above calculations can be repeated for the other parameters using carbon, proton, and the combined data.

It is also possible to use extended versions of eqs 10 and 11 to incorporate additional information into the calculation. For example, rather than just using $\mathrm{R_1}$ and $\mathrm{R_2}$ one could use $\mathrm{R_1}$−$\mathrm{R_4}$, with $\mathrm{R_1}$ and $\mathrm{R_2}$ being the two results using the carbon data (as before) and $\mathrm{R_3}$ and $\mathrm{R_4}$ those with the proton data. In fact, this approach to incorporating both carbon and proton data was slightly less successful (in terms of number of assignments made

(64) Riley, K. F.; Hobson, M. P.; Bence, S. J. *Mathematical Methods for Physics and Engineering*, 3rd ed.; Cambridge University Press: New York, 2006.

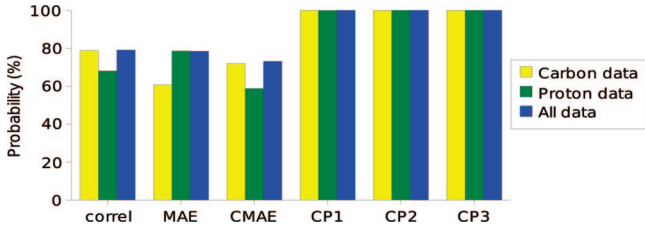


FIGURE 6. Probabilities of the assignment A = **a**, B = **b** being correct as calculated by each of the parameters.

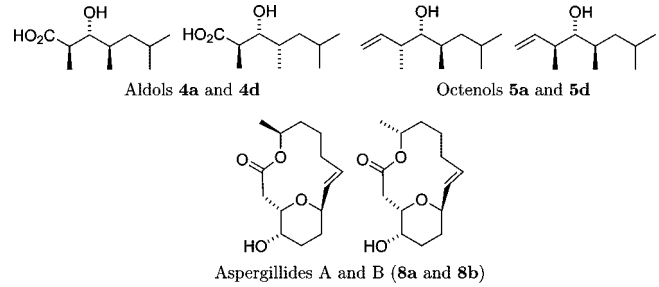


FIGURE 7. Pairs of structures that are difficult to distinguish between.

correctly) than the operationally simpler method that we have already been using of taking the geometric or arithmetic mean of the parameter values for carbon and proton data (right-hand graphs in Figures 3 and 5). We therefore do not consider this approach further.

Figure 6 plots the quantities $P(AC_1|R_1$ and $R_2)$ obtained with each parameter, reflecting the confidence with which each assigns the data as A = **a**, B = **b** rather than the other way around. Values near 100% are good as they indicate that the data was correctly assigned with a high level of confidence; in Figure 6 such values are typically observed for each of the comparison parameters. Values near zero are undesirable as they indicate that data was assigned *incorrectly* with a high level of confidence; no such values are observed in this case. Values around 50% indicate that a firm conclusion was not possible.

We should note that the probabilities calculated can only be rough guides due to the assumptions made in calculating them. Specifically, we have assumed that (i) the value of each parameter for both a correct and an incorrect assignment is normally distributed, (ii) the expectation values and standard deviations for these distributions can be approximated by those obtained from an analysis of the data set of molecules studied here, and (iii) where probabilities are combined the parameters in question are independent random variables.

We note that, from Figure 6, best results are obtained using the comparison parameters CP1−CP3.

3. Testing the Approach across a Range of Molecules. The calculations illustrated for aldols **1** were repeated for all pairs of molecules in Figure 1. The assignments that each parameter made correctly (using the equivalent of Figures 3 and 5) are shown in Table 1.

All of the analyses got most of the assignments correct, and CP2 and CP3 for the combined data assigned every pair correctly. The success rates are quite impressive, particularly for the larger, more flexible structures (such as neopeltolide **7**). Having experimental data recorded in polar solvents (neopeltolide **7**, CD_3OD; methyl proline **10**, D_2O; laurentristich-4-ol **6a**, $(CD_3)_2CO$) does not seem to cause a particular problem despite all calculations being done in the gas phase. The more challenging structures include pairs **4a** and **4d**, **5a** and **5d**, and **8a** & **8b** (Figure 7). The results for methyl proline **10** in Table 1 are for the cationic form (Figure 1); using the zwitterionic form instead gave poor results, which we ascribe to the fact that our calculations were done in the gas phase. The neutral form of **10** gave similar results to the cationic form (identical for CP1−CP3, similar for MAE, and slightly less good for CMAE).

The percentage of assignments made correctly by each parameter is plotted in Figure 8. The best results are obtained with CP2 and CP3, and in all cases the combined values give better results than either ^{13}C or ^{1}H alone.

However, as seen when considering aldols **1**, the level of confidence with which the assignments are made is also important. The same probability calculations as illustrated for aldols **1** were carried out for all pairs of structures in Figure 1; in each case the expectation values and standard deviations used

TABLE 1. Correct Assignments Made by Each of the Parameters

		1a,1b	2a,2b	3a,3b	3a,3c	3a,3d	3b,3c	3b,3d	3c,3d	4a,4b	4a,4c	4a,4d	4b,4c	4b,4d	4c,4d	5a,5b	5a,5c	5a,5d	5b,5c	5b,5d	5c,5d	6a,6b	7a,7b	8a,8b	9a,9b	10a,10b	11a,11b	12a,12b	13a,13b	Total correct
Carbon data	correl	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	25
	MAE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				✓	✓	✓	✓		✓	✓	✓	✓				✓	✓	✓	21
	CMAE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓	✓	25
	CP1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	24
	CP2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	26
	CP3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	27
Proton data	correl	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	26
	MAE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓			✓	✓	✓	✓	✓	✓		✓	✓	✓	24
	CMAE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	25
	CP1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	26
	CP2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓		✓	✓	✓	✓	✓	✓	✓	26
	CP3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	27
All data	correl	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	28
	MAE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓		✓	✓	✓	24
	CMAE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	26
	CP1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	27
	CP2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	28
	CP3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	28

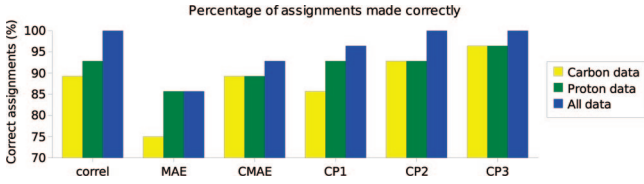


FIGURE 8. Parameter success rates.

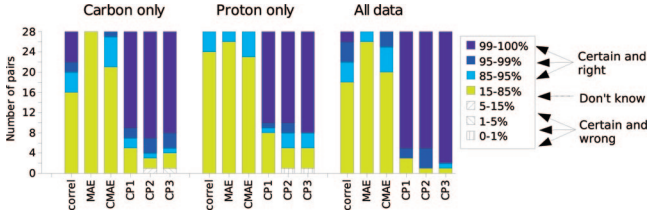


FIGURE 9. Distribution of probabilities for each parameter.

were obtained from an analysis of all the data points that did not include the species in question. (The set of expectation values and standard deviations when all data points are included may be found in Supporting Information).

The results are summarized in Figure 9, which shows the distribution of probabilities for each parameter. For example, the correlation coefficient with the carbon data correctly assigned six pairs with over 99% confidence, two pairs with between 95% and 99%, and six pairs with 85−95%, with the remaining 14 pairs not clearly assigned.

A good parameter has most pairs in the dark blue section (correctly assigned with high confidence) and none in the hatched sections, since these correspond to misleading incorrect assignments made with apparently good confidence. On this basis the CP3 parameter gives the best results, and we therefore recommend the use of this parameter for structure assignment in place of the more commonly used correlation coefficient, MAE, and CMAE. To facilitate this, we have produced a web applet for assigning two sets of experimental data to two sets of calculated data using CP3 and the probability approach; this applet may be found at http://www-jmg.ch.cam.ac.uk/tools/nmr/. Alternatively, probabilities may be computed "by hand" using eq 10 and the data in Supporting Information.

4. Testing the Robustness of the Process. Unassigned Spectra. As discussed in Section 1, the experimental data was incompletely assigned in almost all cases, i.e., not all of the resonances were assigned to a specific carbon or proton nucleus, and any remaining assignment was done by simply matching up the experimental shifts in order to the calculated shifts to which they could potentially correspond. If only 1D spectra are available, then while it may be possible to assign some of the resonances based on intensities, multiplicities, and coupling constants, many of the protons and particularly carbons will be unassigned or only partially assigned (for example it may be possible to identify a resonance as one of several methyl groups). With 2D data most of the carbon and proton resonances can typically be assigned, but even then it may not be possible to distinguish diastereotopic protons.

To investigate how important it is for the experimental data to be as fully assigned as possible (and hence for 2D NMR data to be obtained), the above calculations were repeated assuming *no assignment* in all cases. The results are shown in Figure 10 (equivalent to Figure 9).

Although Figure 10 does show slightly less of the dark blue regions than when the data was partially assigned (Figure 9),

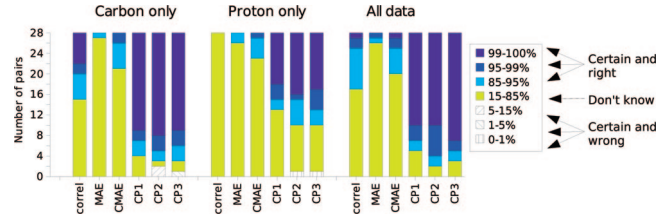


FIGURE 10. Distribution of probabilities using experimental data not assigned to nuclei.

the effect is relatively small. Indeed, the probabilities summarized in Figure 10 deviated from those calculated with the partially assigned data by an average of only 4%. Therefore, although it may in some cases be advantageous to have experimental data as fully assigned as possible, it is by no means essential. From Figure 10 the CP3 parameter is still the most successful at making correct assignments with good confidence.

Robustness of the Expectation Values and Standard Deviations. The expectation values and standard deviations used to calculate the probabilities came from an analysis of the results for each of the pairs in Figure 1 that did not include the pair in question. One may ask whether this data set of 28 pairs is large and diverse enough to give values that apply to molecules outside of the data set. To test its robustness, we examined the expectation values and standard deviations obtained from removing each pair in turn from the data set. If the values are strongly dependent on the exact composition of the data set (indicating that the data set is too small), one would expect there to be significant differences between the values obtained using all 28 pairs and the sets of values obtained using each of the 28 possible combinations of 27 pairs. In fact the changes were small: the mean unsigned percentage change (across all the parameters) when each pair in turn was removed was only 1.7% for the expectation values and 1.8% for the standard deviations.

We also carried out a similar experiment in which we removed all 378 ($^{28}C_2$) possible combinations of two pairs. Once again the changes were small: the 378 expectation values for each parameter differed by an average of only 2.5% (mean unsigned percentage change) from the expectation values using all 28 pairs, and the corresponding value for the standard deviations was only 2.8%.

These small changes should have a correspondingly small effect on the probabilities that the expectation values and standard deviations are used to calculate. To confirm this, we recalculated the probabilities for each pair of molecules using each of the 28 or 378 sets of expectation values and standard deviations obtained by removing from the data set all possible combinations of one or two pairs plus, where necessary, the pair of molecules in question. The mean unsigned changes in the probabilities were indeed small: only 0.4% for when one pair was removed from the data set and 0.6% when two pairs were removed.

These results indicate that the current data set is robust to small changes in its composition. Therefore, although work is currently ongoing to expand and diversify the data set still further, this study gives confidence that the current data set is large enough to provide meaningful results.

To confirm that the results are not strongly dependent on the software package used and to investigate whether the results are sufficiently similar that the expectation values and standard deviations obtained from the Jaguar calculations are transferable to calculations using Gaussian and vice versa, we recalculated

the shifts for all of the molecules in Figure 1 using the Gaussian program. The mean absolute difference in calculated shift between Gaussian and Jaguar was 0.30 ppm for ^{13}C and 0.027 ppm for ^{1}H, and the average percentage change in the expectation values and standard deviations across all the parameters were only 2.1% and 2.4%, respectively. These are comparable to the changes that were observed when one or two pairs of structures were removed from the data set; this suggests that separate sets of expectation values and standard deviations for Jaguar and Gaussian are not required, since the differences between them are comparable to the small changes that one would get simply by choosing a slightly different data set of molecules for the data set. In fact, using the Jaguar expectation values and standard deviations (the Jaguar data set) when the shifts have been calculated using Gaussian changed the probabilities by an average of only 1.7% from those calculated using the Gaussian data set. Similarly, using the Gaussian data set with shifts calculated using Jaguar changed the probabilities by an average of only 0.6% from those calculated using the Jaguar data set.

Conclusions

Based on this study, we recommend the following approaches for structure assignment of a pair of diastereoisomers by NMR shift prediction:

• Calculate shifts using single point ab initio calculations on molecular mechanics geometries obtained from a conformational search with the MMFF force field. Good results can be obtained without expensive ab initio optimizations and solvent models.

• Use CP3 rather than the correlation coefficient, MAE, or CMAE for analyzing the data. Combining ^{13}C and H values is more successful than using either individually.

• Calculate probabilities using Bayes' theorem to give an indication of the level of confidence in the conclusion. This can be done using our applet at http://www-jmg.ch.cam.ac.uk/tools/nmr/ or "by hand" using eq 10 and the data in Supporting Information.

We have shown here that NMR shift calculation, combined with analysis using CP3, is an effective way to assign two experimental spectra to two possible structures. Work is underway to extend this methodology to situations in which one only has data for one unknown diastereoisomer (rather than two) and also to apply it to a greater number and range of examples. Calculation of probabilities using other nuclei is likely to work well but would, however, require development of a database to obtain the necessary expectation values and standard deviations. Finally, although we have only considered ^{1}H and ^{13}C NMR in diastereomeric structures, we note that the methodology can also be used for structural isomers and should also be applicable to other nuclei, including in inorganic systems, since to calculate CP3 one only has to have calculated and experimental data for two isomers.

Acknowledgment. We thank the University of Cambridge (S.G.S.) and Unilever for financial support. We acknowledge the use of the CamGrid service in carrying out this work and Dr. Charlotte Bolton for the IT support.

Supporting Information Available: Complete ref 48, complete details of calculated and experimental shifts, expectation values, and standard deviations for correct and incorrect structure assignments, and results of including ab initio geometry optimization and a solvent model in the calculation. This material is available free of charge via the Internet at http://pubs.acs.org.

JO900408D