JACS
ARTICLES

Published on Web 08/26/2010

Assigning Stereochemistry to Single Diastereoisomers by GIAO NMR Calculation: The DP4 Probability

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.

Received June 9, 2010; E-mail: j.m.goodman@ch.cam.ac.uk

Abstract: GIAO NMR shift calculation has been applied to the challenging task of reliably assigning stereochemistry with quantifiable confidence when only one set of experimental data are available. We have compared several approaches for assigning a probability to each candidate structure and have tested the ability of these methods to distinguish up to 64 possible diastereoisomers of 117 different molecules, using NMR shifts obtained in rapid and computationally inexpensive single-point calculations on molecular mechanics geometries without time-consuming *ab initio* geometry optimization. We show that a probability analysis based on the errors in each ^{13}C or ^{1}H shift is significantly more successful at making correct assignments with high confidence than are probabilities based on the correlation coefficient and mean absolute error parameters. Our new probability measure, which we have termed DP4, complements the probabilities obtained from our previously developed CP3 parameter, which applies to the case of assigning a pair of diastereoisomers when one has both experimental data sets. We illustrate the application of DP4 to assigning the stereochemistry or structure of 21 natural products that were originally misassigned in the literature or that required extensive synthesis of diastereoisomers to establish their stereochemistry.

Introduction

The *ab initio* calculation of NMR shifts is becoming an increasingly popular tool for the assignment of stereochemistry in organic and natural products chemistry. The technique was pioneered by Bifulco[1,2] and has played key roles in the stereostructure assignment or reassignment of several natural products, including hexacyclinol,[3,4] maitotoxin,[5] applidinones A−C,[6] jungianol,[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] hassananes,[16] ketopelenolides C and D,[17] 6*β*-hydroxyhyoscyamine,[18] dolichodial,[19] hypurticin,[20] santalol derivatives,[21] fusapyrones[22] and 9-*epi*-presilphiperfolan-1-ol.[23] NMR shift calculation has been used to determine or confirm the stereochemistry and/or structure of products obtained in synthetic chemistry: examples include a pair of bicyclic peroxides,[24] epoxides of carene,[25] and isohasubanan alkaloids.[26] The effect of using different levels of theory at various stages in the NMR shift calculation has

[1] Barone, G.; Gomez-Paloma, L.; Duca, D.; Silvestri, A.; Riccio, R.; Bifulco, G. *Chem.—Eur. J.* **2002**, *8*, 3233–3239.

[2] Barone, G.; Duca, D.; Silvestri, A.; Gomez-Paloma, L.; Riccio, R.; Bifulco, G. *Chem.—Eur. J.* **2002**, *8*, 3240–3245.

[3] Rychnovsky, S. D. *Org. Lett.* **2006**, *8*, 2895–2898.

[4] Saielli, G.; Bagno, A. *Org. Lett.* **2009**, *11*, 1409–1412.

[5] Nicolaou, K. C.; Frederick, M. O. *Angew. Chem., Int. Ed.* **2007**, *46*, 5278–5282.

[6] Aiello, A.; Fattorusso, E.; Luciano, P.; Mangoni, A.; Menna, M. *Eur. J. Org. Chem.* **2005**, 5024–5030.

[7] da Silva, G. V. J.; Neto, Á. C. *Tetrahedron* **2005**, *61*, 7763–7767.

[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.

[16] Yang, J.; Huang, S.-X.; Zhao, Q.-S. *J. Phys. Chem. A* **2008**, *112*, 12132–12139.

[17] Fattorusso, E.; Luciano, P.; Romano, A.; Taglialatela-Scafati, O.; Appendino, G.; Borriello, M.; Fattorusso, C. *J. Nat. Prod.* **2008**, *71*, 1988–1992.

[18] Muñoz, M. A.; Joseph-Nathan, P. *Magn. Reson. Chem.* **2009**, *47*, 578–584.

[19] Wang, B.; Dossey, A. T.; Walse, S. S.; Edison, A. S.; Merz, K. M., Jr. *J. Nat. Prod.* **2009**, *72*, 709–713.

[20] 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.

[21] Stappen, I.; Buchbauer, G.; Robien, W.; Wolchann, P. *Magn. Reson. Chem.* **2009**, *47*, 720–726.

[22] Honma, M.; Kudo, S.; Takada, N.; Tanaka, K.; Miura, T.; Hashimoto, M. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 709–712.

[23] (a) Pinto, S. C.; Leitão, G. G.; Bizzo, H. R.; Martinez, N.; Dellacassa, E.; dos Santos, F. M., Jr.; Costa, F. L. P.; de Amorim, M. B.; Leitão, S. G. *Tetrahedron Lett.* **2009**, *50*, 4785–4787. (b) Joseph-Nathan, P.; Leitão, S. G.; Pinto, S. C.; Leitão, G. G.; Bizzo, H. R.; Costa, F. L. P.; de Amorim, M. B.; Martinez, N.; Dellacassa, E.; Hernández-Barragán, A.; Pérez-Hernández, N. *Tetrahedron Lett.* **2010**, *51*, 1963–1965.

[24] Griesbeck, A. G.; Blunk, D.; El-Idreesy, T. T.; Raabe, A. *Angew. Chem., Int. Ed.* **2007**, *46*, 8883–8886.

[25] Koskowich, S. M.; Johnson, W. C.; Paley, R. S.; Rablen, P. R. *J. Org. Chem.* **2008**, *73*, 3492–3496.

[26] Nielsen, D. K.; Nielsen, L. L.; Jones, S. B.; Toll, L.; Asplund, M. C.; Castle, S. L. *J. Org. Chem.* **2009**, *74*, 1187–1199.

also been extensively investigated,[27] as has the use of multiple reference standards.[28] The area has been reviewed.[29]

Our approach to stereostructure assignment by NMR shift calculation is to calculate the shifts for the candidate structures (employing a Boltzmann weighted average of the shifts calculated for all low-energy conformers) and compare to the experimental data to decide which gives the best match. A key issue, however, is how best to quantify the fit for each possible match, since errors in both the calculated and experimental shifts mean that the agreement will not be perfect, even for a correct match.

We recently showed that, when one has two experimental spectra to be assigned to two possible structures, an effective approach is to compare the differences in calculated shifts between the two isomers with the corresponding differences in experimental shifts using our CP3 parameter.[30] CP3 has the dual advantage that systematic errors cancel out (because differences in shifts are being considered) and that, because of the way CP3 is calculated, the most weight is placed on the carbon or proton nuclei that have the greatest difference in experimental shift and so are most useful for stereostructure assignment. We also showed how, by using the CP3 parameter in conjunction with Bayes's theorem and a knowledge of the values of CP3 expected for a correct and incorrect assignment, a quantitative estimate of the probability that the assignment being made is correct can be obtained. This allows one not just to make a stereochemical assignment but also to provide an indication of the level of confidence that can be placed in it. CP3 usually gives correct assignments with a high level of confidence.[30]

The most significant limitation of CP3 is that it requires two sets of experimental data to be assigned to two possible structures. This situation is very common, especially in synthetic chemistry, where a stereoselective reaction may give a major and minor product and one wishes to know which is which. However, one is often faced with the problem of having only one set of experimental data to assign to one of several possible structures, and in these situations CP3 cannot be applied. For example, natural products are frequently isolated as single diastereoisomers, and assigning their relative stereochemistry can be a formidable task. In many cases it is necessary to resort to time-consuming and expensive total synthesis of potential diastereoisomers,[31] and even in structures apparently rigid enough for coupling constants and nuclear Overhauser effects (NOEs) to give stereochemical information, it is not uncommon for the originally proposed stereochemistry or structure to have to be revised following a total synthesis.[32] Many of the molecules that we consider later, such as neopeltolide, the aspergillides, tanarifuranonol, biyouyanagin A, and the tricholomalides, have had their stereochemistry revised from the original proposal following synthetic work.

We were therefore very interested in developing methodology for stereostructure assignment using NMR shift calculation that has the desirable features of CP3-based probabilities, namely the ability to make clear assignments with a high and quantifiable level of confidence, but that can be applied to the situation in which only one set of experimental data are available.

Computational Methods

All molecular mechanics calculations were performed using Macromodel[33] (Version 9.1, 9.5, or 9.7) interfaced to the Maestro[34] (Version 7.5, 8.0, or 9.0) program. All conformational searches used the Monte Carlo Multiple Minimum[35] (MCMM) method, the Systematic Pseudo Monte Carlo[36] (SPMC) method, or a 50:50 hybrid of MCMM and Low Mode[37] sampling, and the MMFF force field.[38] The searches were done in the gas phase, with a 50 kJ mol^{-1} upper energy limit and with the number of steps large enough to find all low-energy conformers at least 5−10 times.

Quantum mechanical calculations were carried out using Jaguar[39] (Version 6.5, 7.0, or 7.6). Test calculations showed that the different versions of the software give very similar results. For example, the mean absolute difference in calculated shift for aldol **2a** (see Figure 1) obtained by the standard procedure (see below) among the three versions was only 0.09 ppm for ^{13}C and 0.02 ppm for ^{1}H. As in our previous studies,[12,30,40] we employed the widely used B3LYP functional[41] and 6-31G(d,p) basis set[42] for all calculations. NMR shielding constant calculation used the GIAO method.[43]

In our previous investigations,[12,30,40] we showed 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, and we therefore continued to use this approach here.

Unless otherwise stated, 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 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

[27] (a) Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. J. *J. Chem. Phys.* **2006**, *104*, 5497–5509. (b) Forsyth, D. A.; Sebag, A. B. *J. Am. Chem. Soc.* **1997**, *119*, 9483–9494. (c) Giesen, D. J.; Zumbulyadis, N. *Phys. Chem. Chem. Phys.* **2002**, *4*, 5498–5507. (d) Cimino, P.; Gomez-Paloma, L.; Duca, D.; Riccio, R.; Bifulco, G. *Magn. Reson. Chem.* **2004**, *42*, S26–S33. (e) Tormena, C. F.; da Silva, G. V. *J. Chem. Phys. Lett.* **2004**, *398*, 466–470. (f) Bagno, A.; Rastrelli, F.; Saielli, G. *Chem.—Eur. J.* **2006**, *12*, 5514–5525. (g) Wu, A.; Zhang, Y.; Xu, X.; Yan, Y. *J. Comput. Chem.* **2007**, *28*, 2431–2442. (h) Wiitala, K. W.; Hoyle, T. R.; Cramer, C. J. *J. Chem. Theory Comput.* **2006**, *2*, 1085–1092. (i) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *112*, 6794–6799.

[28] Sarotti, A. M.; Pellegrinet, S. C. *J. Org. Chem.* **2009**, *74*, 7254–7260.

[29] Bifulco, G.; Dambruoso, P.; Gomez-Paloma, L.; Riccio, R. *Chem. Rev.* **2007**, *107*, 3744–3779.

[30] Smith, S. G.; Goodman, J. M. *J. Org. Chem.* **2009**, *74*, 4597–4607.

[31] Walsh, L. M.; Goodman, J. M. *Chem. Commun.* **2003**, 2616–2617.

[32] (a) Nicolaou, K. C.; Snyder, S. A. *Angew. Chem., Int. Ed.* **2005**, *44*, 1012–1044. (b) Maier, M. E. *Nat. Prod. Rep.* **2009**, *16*, 1105–1124.

[33] 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.

[34] *Maestro*, Version 8.0; Schrödinger LCC: New York, 2007.

[35] Chang, G.; Guida, W. C.; Still, W. C. *J. Am. Chem. Soc.* **1989**, *111*, 4379–4386.

[36] Goodman, J. M.; Still, W. C. *J. Comput. Chem.* **1991**, *12*, 1110–1117.

[37] (a) Kolossváry, I.; Guida, W. C. *J. Am. Chem. Soc.* **1996**, *118*, 5011–5019. (b) Kolossváry, I.; Guida, W. C. *J. Comput. Chem.* **1999**, *20*, 1671–1684.

[38] (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. *J. Comput. Chem.* **1996**, *17*, 616–641. (f) Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 720–729. (g) Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 730–748.

[39] *Jaguar*, Version 7.0; Schrödinger LLC: New York, 2007.

[40] Smith, S. G.; Channon, J. A.; Paterson, I.; Goodman, J. M. *Tetrahedron* **2010**, *66*, 6437–6444.

[41] (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.

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

[43] (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.

compromise between computer time and the risk of missing important conformers (as judged by their subsequent *ab initio* energies) due to inaccurate ordering of the conformer energies by the MMFF force field. We have previously shown[12,30,40] that a 10 kJ mol^{-1} cutoff is generally sufficient to give good results, and we investigated this in detail for the stereopentads **18**, for which increasing the cutoff to 25 kJ mol^{-1} (at much greater computational cost) did not give a significant increase in accuracy.[40]

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

$$\sigma^x = \frac{\sum_i \sigma_i^x \exp(-E_i/RT)}{\sum_i \exp(-E_i/RT)} \tag{1}$$

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

$$\delta^x_{\mathrm{calc}} = \frac{\sigma^{\mathrm{o}} - \sigma^x}{1 - \sigma^{\mathrm{o}}/10^6} \tag{2}$$

where δ^x_{calc} is the calculated shift for nucleus x (in ppm), σ^x is the shielding constant for nucleus x from eq 1, 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

Molecules Studied. We initially studied the molecules in Figure 1, which are also listed with explanatory notes in Table 1. Unless otherwise stated, we considered all diastereoisomers generated by varying the stereocenters marked with an asterisk in Figure 1, but to save space each individual diastereoisomer is not drawn explicitly in Figure 1. A complete set of structures may be found in the Supporting Information. Molecules for which we have the calculated shifts but not the experimental data are shown with their labels in parentheses in Table 1.

The structures include the set of molecules that we studied in our previous investigation,[30] but because our current aim is to be able to assign one unknown compound to one of several possible structures, we have calculated the shifts in most cases for all remaining diastereoisomers of each compound in order to provide a greater number of decoys. In addition, we have incorporated the stereopentads from our recent paper[40] and a number of related molecules, and also some additional molecules whose stereochemistry has recently been reassigned. Shifts were calculated for all diastereoisomers of each molecule considered with the following exceptions (which were generally made to save computer time): For neopeltolide **7a** only the configurations of the stereocenters in the macrocyclic ring were varied, on the assumption that the relative stereochemistry of the six-membered ring can be reliably assigned by standard coupling constant and NOE analysis. For the same reasons, we considered only the diastereoisomers of nankakurine **15a** differing in configuration at the quaternary center (the center that has been subject to reassignment), and for biyouyanagin A **16a** we considered only the two originally proposed structures and the two alternative proposed structures (including the correct one). For the TBDPS-protected stereopentads **19** we considered only the three diastereoisomers for which we had experimental data. For

[44] Kaupp, M.; Malkina, O. L.; Malkin, V. G. *Chem. Phys. Lett.* **1997**, *265*, 55–59.

[45] Jervis, P. J.; Cox, L. R. *Beilstein J. Org. Chem.* **2007**, *3*, 6.

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

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

[48] Zampella, A.; D'Auria, M. V.; Gomez-Paloma, L.; Casapullo, A.; Minale, L.; Debitus, C.; Henin, Y. *J. Am. Chem. Soc.* **1996**, *118*, 6202–6209.

[49] Turk, J. A.; Visbal, G. S.; Lipton, M. A. *J. Org. Chem.* **2003**, *68*, 7841–7844.

[50] 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.

[51] Chen, P.; Wang, J.; Liu, K.; Li, C. *J. Org. Chem.* **2008**, *73*, 339–341.

[52] 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.

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

[54] Vintonyak, V. V.; Kunze, B.; Sasse, F.; Maier, M. E. *Chem.—Eur. J.* **2008**, *14*, 11132–11140.

[55] Kito, K.; Ookura, R.; Yoshida, S.; Namikoshi, M.; Ooi, T.; Kusumi, T. *Org. Lett.* **2008**, *10*, 225–228.

[56] Handea, S. M.; Uenishi, J. *Tetrahedron Lett.* **2009**, *50*, 189–192.

[57] Phommart, S.; Sutthivaiyakit, P.; Chimnoi, N.; Ruchirawat, S.; Sutthivaiyakit, S. *J. Nat. Prod.* **2005**, *68*, 927–930.

[58] Shiao, H.-Y.; Hsieh, H.-P.; Liao, C.-C. *Org. Lett.* **2008**, *10*, 449–452.

[59] Murphy, A. C.; Mitova, M. I.; Blunt, J. W.; Munro, M. H. G. *J. Nat. Prod.* **2008**, *71*, 806–809.

[60] Roulland, E. *Angew. Chem., Int. Ed.* **2008**, *47*, 3762–3765.

[61] Ko, C.; Feltenberger, J. B.; Ghosh, S. K.; Hsung, R. P. *Org. Lett.* **2008**, *10*, 1971–1974.

[62] Barrero, A. F.; Manzaneda, R, E. A.; Manzaneda, R, R. A. *Tetrahedron* **1990**, *46*, 8161–8168.

[63] Hirasawa, Y.; Morita, H.; Kobayashi, J. *Org. Lett.* **2004**, *6*, 3389–3391.

[64] Nilsson, B. L.; Overman, L. E.; Read de Alaniz, J.; Rohde, J. M. *J. Am. Chem. Soc.* **2008**, *130*, 11297–11299.

[65] Tanaka, N.; Okasaka, M.; Ishimaru, Y.; Takaishi, Y.; Sato, M.; Okamoto, M.; Oshikawa, T.; Ahmed, S. U.; Consentino, L. M.; Lee, K.-H. *Org. Lett.* **2005**, *7*, 2997–2999.

[66] Nicolaou, K. C.; Sarlah, D.; Shaw, D. M. *Angew. Chem., Int. Ed.* **2007**, *46*, 4708–4711.

[67] Appendino, G.; Tagliapietra, S.; Nano, G. M.; Jakupovic, J. *Phytochemistry* **1993**, *35*, 183–186.

[68] Channon, J. A.; Paterson, I. *Tetrahedron Lett.* **1992**, *33*, 797–800.

[69] Channon, J. A. Ph.D. thesis, University of Cambridge, 1992.

[70] Kang, B.; Mowat, J.; Pinter, T.; Britton, R. *Org. Lett.* **2009**, *11*, 1717–1720.

[71] Xu, R.-S.; Lu, Y.-J.; Chu, J.-H. *Tetrahedron* **1982**, *38*, 2667–2670.

[72] Sánchez-Izquierdo, F.; Blanco, P.; Busqué, F.; Alibés, R.; de March, P.; Figueredo, M.; Font, J.; Parella, P. *Org. Lett.* **2007**, *9*, 1769–1772.

[73] Geng, C.-A.; Jiang, Z.-Y.; Ma, Y.-B.; Luo, J.; Zhang, X.-M.; Wang, H.-L.; Shen, Y.; Zuo, A.-X.; Zhou, J.; Chen, J.-J. *Org. Lett.* **2009**, *18*, 4120–4123.

[74] Tsukamoto, S.; Macabalang, A. D.; Nakatani, K.; Obara, Y.; Nakahata, N.; Ohta, T. *J. Nat. Prod* **2003**, *66*, 1578–1581.

[75] Wang, Z.; Min, S.-J.; Danishefsky, S. J. *J. Am. Chem. Soc.* **2009**, *131*, 10848–10849.

[76] El-Naggar, M.; Capon, R. J. *J. Nat. Prod.* **2009**, *72*, 460–464.

[77] Grkovic, T.; Copp, B. R. *Tetrahedron* **2009**, *65*, 6335–6340.

[78] Evans, D. A.; Connell, B. T. *J. Am. Chem. Soc.* **2003**, *125*, 10899–10905.

[79] Taniguchi, T.; Shin'ichi, Y.; Ishibashi, H. *J. Org. Chem.* **2009**, *74*, 7592–7594.

[80] Morita, H.; Yoshinaga, M.; Kobayashi, J. *Tetrahedron* **2002**, *58*, 5489–5495.

[81] Fleury, E.; Lannou, M.-I.; Bistri, O.; Sautel, F.; Massiot, G.; Pancrazi, A.; Ardisson, J. *J. Org. Chem.* **2009**, *74*, 7034–7045.

[82] Crossman, J. S.; Perkins, M. V. *J. Org. Chem.* **2006**, *71*, 117–124.

[83] Watson, W. H.; Tairaa, Z.; Dominguezb, X. A.; Gonzalesb, H.; Aragon, R. *Tetrahedron Lett.* **1976**, *17*, 2501–2502.

[84] Majetich, G.; Grove, J. L. *Org. Lett.* **2009**, *11*, 2904–2907.

[85] Nieto, M.; García, E. E.; Giordano, O. S.; Tonn, C. E. *Phytochemistry* **2000**, *53*, 911–915.

[86] Yamaoka, Y.; Yamamoto, H. *J. Am. Chem. Soc.* **2010**, *132*, 5354–5356.

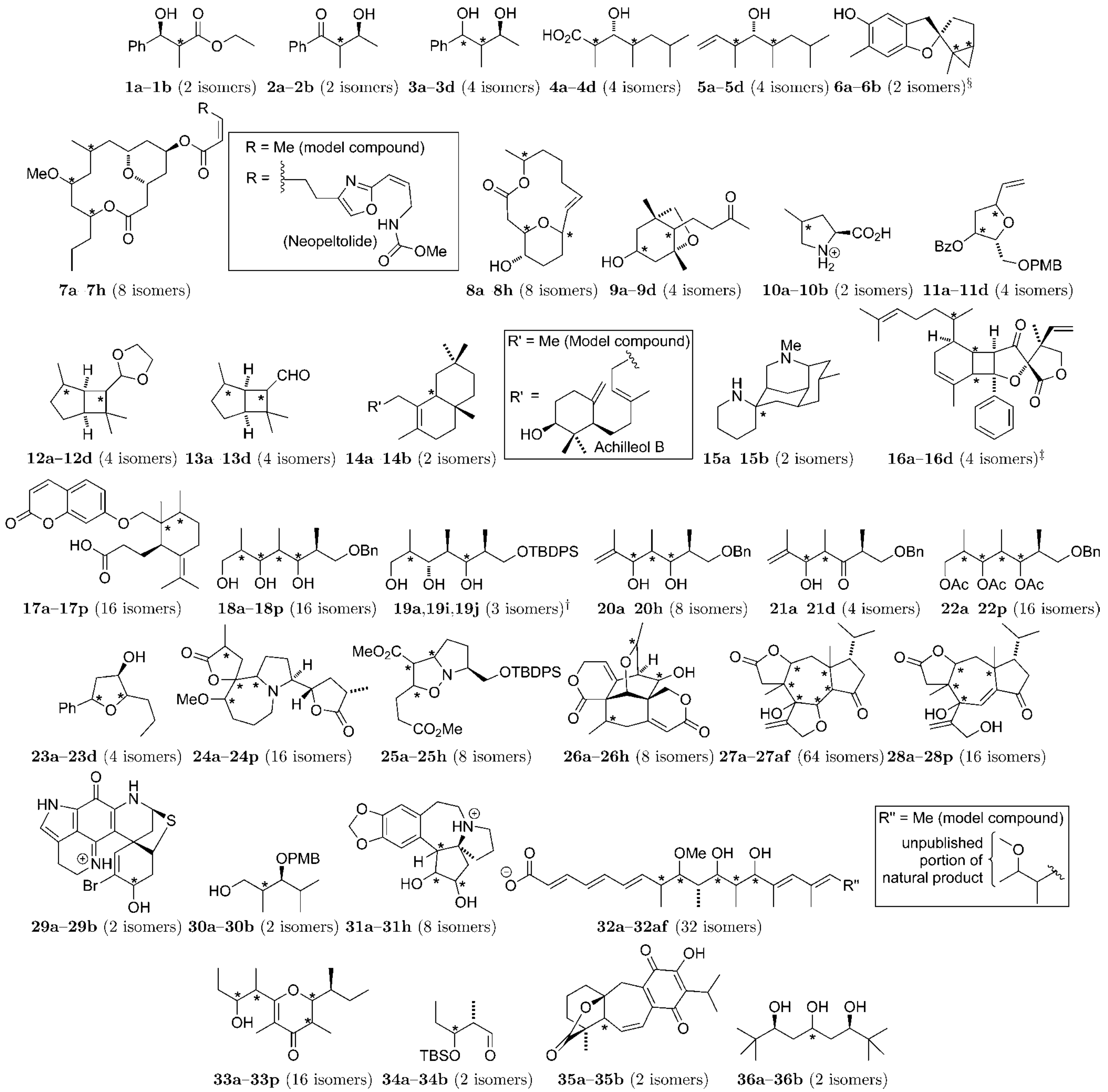


Figure 1. Molecules studied. Stereochemistry was varied at each of the carbons marked with an asterisk, and the structures of all the isomers are drawn with full stereochemistry in the Supporting Information. For explanatory notes see Table 1. Isomers for which we have calculated the shifts but do not have experimental data are indicated in Table 1. §3−5 ring system of **6** fixed as *cis*. ‡4−6 ring system of **16** fixed as *cis*. †Details of the three isomers of **19** considered are given in the Supporting Information.

stemonidine **24** we considered only the 16 diastereoisomers arising from varying the configuration at the four stereocenters where the stereochemistry has previously been misassigned.

In cases where the experimental shifts were incompletely assigned to nuclei (common with the carbon shifts unless 2D NMR data were available), any remaining assignment was done by sorting experimental and calculated shifts in order of size of chemical shift and pairing the resulting values. This step is required 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 were unclear, and the omitted nuclei are indicated in Table 1.

The flexible side chains of neopeltolide **7** and achilleol B **14** were truncated in order to reduce the number of conformers of the molecule and hence the computer time required, and Ardisson's polyketide, **32**, was also truncated in order to give a model compound that did not include the unpublished portion of the natural product. In each of these cases, nuclei close to the site of the truncation were also omitted, as detailed in Table 1.

NMR shift calculation of Br-substituted carbons is known to give poor results due to spin−orbit coupling effects,[44] so as in our previous studies[12] we did not include the Br-substituted carbons in dihydrodiscorhabdin A (**29**) in our analyses.

Table 1. Notes on Molecules Studied (Molecules for Which We Have Calculated the Shifts but Do Not Have Experimental Data Are Shown in Parentheses)

structures	notes
1a, 1b	Aldols **1**. Omit OH proton. Data: $CDCl_3$[45]
2a, 2b	Aldols **2**. Omit OH proton. Data: $CDCl_3$[46]
3a, 3b, 3c, 3d	Diols **3**. Omit OH protons. Data: $CDCl_3$[47]
4a, 4b, 4c, 4d	Aldols **4**; **4a** is a degradation fragment of callipeltin A, originally assigned to **4d**.[48] Omit OH proton. Data: $CDCl_3$[49]
5a, 5b, 5c, 5d	Octenols **5** Omit OH proton. Data: $CDCl_3$[49]
6a, 6b	Laurentristich-4-ol **6**: correct structure (**6a**) and originally proposed structure (**6b**). Omit OH proton. Data: $(CD_3)_2$ CO (**6a**);[50] $CDCl_3$ (**6b**)[51]
7a, 7b, (7c), 7d, (7e), (7f), (7g), (7h)	Neopeltolide **7**: correct structure (**7a**), originally proposed structure (**7b**), and selected other diastereoisomers (see text). Calculations done on truncated version (R = Me in Figure 1). Carbons and protons in the ester side chain omitted. Data: CD_3 OD (**7a**);[52] CD_3 OD (**7b**);[53] CD_3 OD (**7d**)[54]
8a, 8b, 8c, (8d), (8e), (8f), (8g), (8h)	Aspergillides A (**8a**) and B (**8b**), and other diastereoisomers. The originally proposed structures for aspergillides A and B were **8b** and **8c** respectively. Omit OH proton. Data: $CDCl_3$ (**8a**);[55] C_6 D_6 (**8b**);[55] $CDCl_3$ (**8c**)[56]
9a, 9b, (9c), (9d)	Tanarifuranonol **9**: correct structure (**9a**), originally proposed structure (**9b**), and other diastereoisomers. Omit OH proton. Data: $CDCl_3$ (**9a**);[57] $CDCl_3$ (**9b**)[58]
10a, 10b	Methyl proline **10** (OH and NH protons omitted). Data: D_2 O.[59] The experimental data were for the HCl salt of methyl proline, so as in our previous study[30] we used the protonation state indicated in Figure 1 in our calculations.
11a, 11b, (11c), (11d)	Tetrahydrofurans **11**. All aromatic carbons except the para carbon of the PMB group omitted due to unclear experimental data (details in Supporting Information). Data: $CDCl_3$[60]
12a, 12b, (12c), (12d)	Acetals **12**. Data: $CDCl_3$[61]
13a, 13b, (13c), (13d)	Aldehydes **13**. Data: $CDCl_3$[61]
14a, (14b)	Achilleol B: revised structure (**14a**) and originally proposed structure (**14b**). Calculations done on truncated version (R = Me in Figure 1). Carbons and protons in the truncated side chain omitted. Data: $CDCl_3$[62]
15a, 15b	Nankakurine A: correct structure (**15a**) and originally proposed structure (**15b**). NH proton omitted, and additional nuclei omitted for **15b** due to unclear experimental data (see Supporting Information for details). Data: CD_3 OD (**15a**),[63] CD_3 OD (**15b**)[64]
16a, 16b, (16c), (16d)	Biyouyanagin A: originally proposed structures (**16c** and **16d**) and second proposed structures (**16a** and **16b**), including correct structure **16a**. Data: $CDCl_3$ (**16a**),[65] $CDCl_3$ (**16b**)[66]
17a, (17b), (17c), (17d)	Galbanic acid: correct structure (**17a**), alternative structure (**17b**), and other diastereoisomers. Omit OH proton. Data: $CDCl_3$ (**17a**)[67]
(18a), 18b, 18c, 18d, 18e, 18f, 18g, 18h, (18i), (18j), 18k, 18l, 18 m, 18n, 18o, 18p	Stereopentads **18**.[68] Omit OH protons. Data: $CDCl_3$[40,69]
19a, 19i, 19j	TBDPS-protected stereopentads **19**.[68] Omit OH protons. Data: $CDCl_3$[40,69]
20a, 20b, 20c, 20d, 20e, 20f, (20g), 20h	Diols **20**. Omit OH protons. Data: ($CDCl_3$)[69]
21a, 21b, (21c), 21d	Aldols **21**. Omit OH proton. Data: $CDCl_3$[69]
(22a), 22b, 22c, 22d, 22e, 22f, 22g, 22h, (22i), (22j), 22k, 22l, 22 m, 22n, 22o, 22p	Stereopentad acetates **22**. Data: $CDCl_3$[69]
23a, 23b, 23c, 23d	Tetrahydrofurans **23**. Omit OH proton. Data: $CDCl_3$[70]
24a, (24b), 24c, (24d), (24e), (24f), (24g), (24h), 24i, (24j), (24k), (24l), (24 m), (24n), (24o), (24p)	Stemonidine (stemospironine) **24a**, originally proposed structures **24b** and **24c**, and other diastereoisomers. Data: $CDCl_3$ (**24a**,[71] **24c**,[72] **24i**.[72])
25a, 25b, (25c), (25d), (25e), (25f), (25g), (25h)	Stemonidine intermediate **25a** and other diastereoisomers. Data: $CDCl_3$.[72]
26a, 26b, (26c), (26d), (26e), (26f), (26g), (26h)	Swerilactone A (**26a**), swerilactone B (**26b**), and other diastereoisomers. Omit OH proton. Data: C_5 D_5 N.[73]
27a, (27b), (27c−27bl)	Tricholomalide A **27**: correct structure (**27b**), originally proposed structure (**27b**), and the 62 other diastereoisomers. We studied each of the 64 diastereoisomers in full, but to save space they are not all separately listed here; complete details may be found in the Supporting Information. Omit OH proton. Data: $CDCl_3$[74]
28a, 28b, (28c), (28d), (28e), (28f), (28g), (28h)	Tricholomalide B: correct structure (**28a**), originally proposed structure (**28b**), and other diastereoisomers. Omit OH protons. Data: $CDCl_3$ (**28a**),[74] $CDCl_3$ (**28b**)[75]
29a, 29b	Dihydrodiscorhabdin A **29**: correct structure (**29a**) and originally proposed structure (**29b**). Omit OH and NH protons, and the Br-substituted carbon (see text). The experimental data were for the TFA salt, so as for methyl proline **10** we used the protonation state indicated in our calculations. Data: CD_3 OD (**29a**),[76] DMSO-d_6 (**29b**)[77]
(30a), 30b	PMB ethers **30**. Omit OH proton. Data: $CDCl_3$[78]

Table 1. Continued

structures	notes
31a, (31b), 31c, (31d), (31e), (31f), (31g), (31h)	Cephalezomines G and H: cephalezomine H revised structure **31a** and originally proposed structure **31b**, cephalezomine G revised structure **31c** and originally proposed structure **31d**, and other diastereoisomers. The experimental data have been reported to be for the protonated species,[79] so we used the indicated protonation state in our calculations. Omit OH and NH protons. Data: CD_3 OD[80]
32a, (32b), (32c)−(32af)	Ardisson's polyketide.[81] We studied each of the **32** diastereoisomers in full, but to save space they are not all separately listed here; complete details may be found in the Supporting Information. Only limited experimental data were available,[81] and only for the polypropionate portion (see the Supporting Information for details).
33a, 33b, 33c, 33d, 33e, 33f, 33g, 33h, (33i), (33j), (33k), (33l), (33 m), (33n), (33o), (33p)	Maurenone **33a**, and diastereoisomers. Omit OH proton. Data: $CDCl_3$[82]
34a, 34b	Aldehydes **34**. Data: $CDCl_3$[82]
35a, 35b	Icetexone **35a** and 5-*epi*-icetexone **35b**. The compound isolated by Watson[83] that was originally named icetexone and assigned structure **35a** has recently been renamed 5-*epi*-icetexone and assigned structure **35b**,[84] while the compound isolated by Tonn[85] which was originally named 5-*epi*-icetexone and assigned structure **35b** has now been renamed as icetexone and assigned structure **35a**.[84] Data: $CDCl_3$ (**35a**),[85] $CDCl_3$ (**35b**)[84]
36a, 36b	Triols **36**. Omit OH protons. Data: $CDCl_3$[86]

The DP4 Probability. If we have the experimental data for one molecule for which we do not know the stereochemistry, and we have calculated the shifts for all of the possible diastereoisomers, what is the best way to decide which set of calculated shifts provides the best fit to the experimental spectrum? Further, in order to give some indication of how certain we are about the conclusion, can we assign a numerical probability to each candidate structure?

The agreement between calculated and experimental data for a given assignment can be quantified using a parameter such as the correlation coefficient, mean absolute error (MAE), corrected mean absolute error (CMAE, which differs from MAE in employing empirically scaled calculated shifts[1]), or, in the case of assigning two spectra to two possible structures, our CP3 parameter.[30] We have also shown how the values of these parameters can be converted into a probability using Bayes's theorem[87] together with a knowledge of the values of these parameters expected for correct and incorrect assignments.[30] The DP4 probability, which we have developed for the task of assigning one experimental spectrum to one of many possible diastereoisomers, tackles the problem in a different way and gives a probability directly (i.e., there is no DP4 "parameter" as such, only a DP4 "probability").

DP4 is based on the following principle:

First, calculate the empirically scaled shifts[1] (δ_{scaled}) for each candidate structure and hence the error e between the scaled and experimental shifts (i.e., $e = \delta_{\text{scaled}} - \delta_{\text{exp}}$).

Then, assuming that the error for an atom in a correct structure obeys a t distribution with mean μ, standard deviation σ, and degrees of freedom ν, calculate the probability that each observed error is obtained. We initially tried using the corresponding normal distribution in place of the t distribution but found that the t distribution gave better results in terms of a reduced tendency to overstate the probability in extreme cases; this issue is discussed in more detail later. The use of Student's t distribution to model data with longer-than-normal tails has been suggested by Lange, Little, and Taylor.[89]

Next, assuming that the error in the shift of each atom in the molecule is an independent random variable, multiplying the probabilities just obtained gives the probability, for each candidate structure, of obtaining the particular set of errors observed for that structure. Finally, the resulting probabilities are converted to a set of probabilities that each candidate structure is the correct one using Bayes's theorem.

For our calculations we used values of μ, σ, and ν taken from an analysis of all the molecules in Table 1 for which we have experimental data; this gave a data set of 1717 ^{13}C shifts and 1794 ^{1}H shifts. The values we obtained were $\mu = 0$ (this is an automatic consequence of the empirical scaling process), $\sigma = 2.306$ ppm (^{13}C) or 0.185 ppm (^{1}H), and $\nu = 11.38$ (^{13}C) or 14.18 (^{1}H). The values of ν were obtained by fitting the data from all the molecules in Table 1 for which we have experimental shifts to a t distribution using the R statistical program.[90] Although in principle one might consider removing from the data set the particular molecules being studied in any given calculation, in order to avoid including the data for these molecules in the values of μ and σ, in practice this will make very little difference to the values of μ and σ because the data set of molecules is large. We previously showed for a data set of 28 pairs of molecules that excluding one pair does not change the expectation values and standard deviations by more than a few percent,[30] and the effect will be even smaller here because the data set is significantly larger: 117 molecules with experimental data, compared to the 32 molecules (giving a total of 28 pairs) in our previous study.

To facilitate the calculation of DP4, an applet is available from the authors at http://www.jmg.ch.cam.ac.uk/tools/nmr/DP4 for assigning one set of experimental data to one of many

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

(88) Hirasawa, Y.; Kobayashi, J.; Obara, Y.; Nakahata, N.; Kawahara, N.; Goda, Y.; Morita, H. *Heterocycles* **2006**, *68*, 2357–2364.

(89) Lange, K. L.; Little, R. J. A.; Taylor, J. M. G. *J. Am. Stat. Assoc.* **1989**, *84*, 881–896.

(90) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2010; ISBN 3-900051-07-0Specifically, we used the tFit method in the fBasics package (R package version 2110.79.

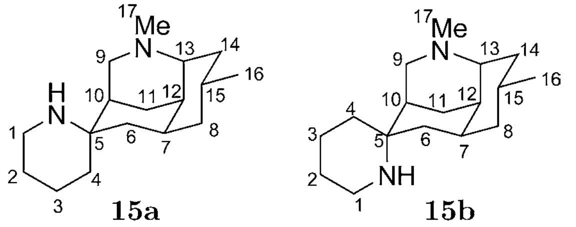


Figure 2. Nankakurine A: revised structure **15a**, originally proposed structure **15b**, and numbering system used.

possible diastereoisomers. Alternatively, DP4 may be computed "by hand" using the following equation (see the Supporting Information for a derivation):

$$P(i|\delta_1, \delta_2, ..., \delta_N) = \frac{\prod_{k=1}^{N} (1 - T^{\nu}(|(\delta^{i}_{\text{scaled},k} - \delta_{\text{exp},k}) - \mu|/\sigma))}{\sum_{j=1}^{m} [\prod_{k=1}^{N} (1 - T^{\nu}(|(\delta^{j}_{\text{scaled},k} - \delta_{\text{exp},k}) - \mu|/\sigma))]} \quad (3)$$

This equation gives the probability that candidate structure i (out of m possible candidates) is the correct one from the experimental shifts δ_1, δ_2, ..., δ_N. In this equation, T^{ν} is the cumulative t distribution function (i.e., $P(X < x)$) with ν degrees of freedom. $\delta_{\text{exp},k}$ is the experimental shift for nucleus k (where k runs from 1 to N), and $\delta^{j}_{\text{scaled},k}$ is the scaled calculated shift of nucleus k in candidate structure j; this is calculated according to $\delta^{j}_{\text{scaled},k} = (\delta^{j}_{\text{calc},k} - \text{intercept})/\text{slope}$, where intercept and slope are the intercept and slope of a plot of $\delta^{j}_{\text{calc},k}$ against $\delta_{\text{exp},k}$.

Calculation of DP4: Nankakurine A. As an example, we will demonstrate the calculation of DP4 for nankakurine A **15a** (Figure 2). This *Lycopodium* alkaloid was isolated in 2004 by Kobayashi and originally assigned to structure **15b**, with the configuration at the spiro center being determined by NOE data.[63] However, the configuration was reassigned by the same group to that of **15a** in 2006, following isolation of a related molecule,[88] and was confirmed as **15a** in 2008 when Overman synthesized both **15a** and **15b** and found **15a** to match the natural product.[64]

DP4 is designed for the situation in which one has experimental NMR data for one unknown compound that is to be assigned to one of two or more possible structures. For our example, we will attempt to assign the experimental data for the natural product (**15a**) to one of the two candidate structures **15a** and **15b**. We could consider other diastereoisomers as well, but for this example we will assume that the stereochemistry of the six-membered rings in the rigid cage structure can be reliably assigned by coupling constant and NOE analysis so that it is only the configuration of the spiro center that is in doubt.

The first step in making the assignment is to calculate the ^{13}C and ^{1}H shifts for each candidate structure (**15a** and **15b** in this case) as described in Computational Methods. The resulting shifts are tabulated in the columns labeled "^{13}C calcd" and "^{1}H calcd" in Table 2. These numbers can now be put into our Web applet, which will automatically calculate DP4, but here we demonstrate the process "by hand". If carrying out the calculation by hand, the next step is to match up the experimental shifts with the calculated ones and calculate the scaled calculated shifts. In the case of natural nankakurine A, the reported experimental ^{13}C data were all fully assigned to nuclei, so there is no ambiguity about which experimental shift goes with which calculated shift. Very often, however, one is faced with unassigned or partially assigned spectra. In the experimental ^{1}H data of nankakurine A, for instance, the diastereotopic protons were not assigned. Two shifts (1.53 and 1.58 ppm) were reported for the two protons on C2, and it is not clear which shift should be assigned to the *pro-R* hydrogen and which to the *pro-S*. In these situations it is necessary to assign the shifts by matching up in order with the calculated shifts. This means that the assignment given to a particular shift may not be the same in all of the candidate structures. For example, H2(*R*) in **15a** is calculated to have a bigger shift than H2(*S*), so when aligning the experimental data with those calculated for **15a**, the 1.58 ppm shift is assigned to H2(*R*) and the 1.53 ppm shift to H2(*S*) ("^{1}H expt" column in the bottom section of Table 2). However, in **15b** H2(*S*) is predicted to have the larger shift, so the assignment of the experimental shifts should be reversed. Only when the data are fully assigned experimentally (as is the case for the ^{13}C data of natural nankakurine A) are the assignments guaranteed to be the same for all candidate structures. We note in passing that having experimentally assigned data should improve the chances of a correct stereostructure assignment, since if the data are not assigned then the assignments can be swapped around for each alternative candidate structure so as to improve the match relative to that for the correct structure.

A further complication arises if experimental shifts are reported as a range. This issue does not arise in the experimental data for natural nankakurine A since all shifts were reported as single values, but had we been attempting to assign the experimental data corresponding to its isomer **15b**, we would have been faced with data such as "1.75−1.66 (m, 5H)". (Complete experimental data for **15b**, together with the results of assigning its structure by DP4, may be found in the Supporting Information, together with the corresponding data for all the other molecules in Figure 1.) Our approach to this situation is to replace the range by signals at the midpoint of the range (here five peaks at 1.705 ppm) in order to permit calculations on the shifts. This solution is not ideal because it may be that the five protons are spread throughout the range rather than all being at the center, but if the experimental spectra are too unclear for distinct environments within the range to be identified, we have no alternative apart from ignoring the data altogether.

Returning to the example in Table 2, the next step is to carry out empirical scaling[1] on the calculated shifts. This is done by plotting δ_{calc} vs δ_{exp} for the data being compared, obtaining the slope and intercept of the regression line, and using $\delta_{\text{scaled}} = (\delta_{\text{calc}} - \text{intercept})/\text{slope}$. For example, the scaled ^{1}H calculated shifts for **15a** ("scaled shifts", bottom of Table 2) are calculated using the values of slope and intercept obtained from a plot of the data in the "^{1}H calcd" column for **15a** (y-axis) against the experimental shifts in the "^{1}H expt" column (x-axis). Similarly, the scaled ^{1}H calculated shifts for **15b** are obtained using a similar plot, after the data in the "^{1}H expt" column have been swapped around as necessary to give the best match to the calculated data as discussed above.

Next, we calculate the error ($\delta_{\text{scaled}} - \delta_{\text{calc}}$) in each shift for the two possible assignments. For example, attempting to assign the ^{1}H experimental data to structure **15a** gives the errors in the "corrected error"column in the bottom section of Table 2, obtained by subtracting the calculated values from the experimental values, again after the shifts in the "^{1}H expt"

Table 2. DP4 Analysis of the Carbon and Proton Data for Natural Nankakurine A (**15a**)

	^{13}C calcd			scaled shifts		corrected error		probability	
position	15a	15b	^{13}C expt	15a	15b	15a	15b	15a	15b
1	40.56	40.97	41.0	40.62	40.38	−0.38	−0.62	0.44	0.40
2	27.05	28.56	26.3	24.75	24.49	−1.55	−1.81	0.26	0.22
3	22.15	22.53	20.9	18.93	18.65	−1.97	−2.25	0.21	0.18
4	34.93	37.92	34.6	33.71	33.46	−0.89	−1.14	0.35	0.32
5	58.27	58.34	56.1	56.91	56.70	0.81	0.60	0.37	0.40
6	41.39	39.27	40.0	39.54	39.30	−0.46	−0.70	0.42	0.38
7	36.06	34.78	34.5	33.60	33.35	−0.90	−1.15	0.35	0.31
8	41.01	41.11	41.9	41.59	41.35	−0.31	−0.55	0.45	0.41
9	56.48	57.31	58.5	59.50	59.29	1.00	0.79	0.34	0.37
10	40.18	41.75	37.4	36.73	36.49	−0.67	−0.91	0.39	0.35
11	33.74	32.35	32.5	31.44	31.19	−1.06	−1.31	0.33	0.29
12	39.09	39.35	36.9	36.19	35.95	−0.71	−0.95	0.38	0.34
13	63.03	63.21	65.1	66.62	66.43	1.52	1.33	0.26	0.29
14	39.54	39.5	40.0	39.54	39.30	−0.46	−0.70	0.42	0.38
15	24.46	24.43	22.0	20.11	19.84	−1.89	−2.16	0.22	0.18
16	24.25	24.33	23.0	21.19	20.92	−1.81	−2.08	0.22	0.19
17	41.07	41.39	43.4	43.21	42.97	−0.19	−0.43	0.47	0.43
product of probabilties								8.13×10^{-9}	2.09×10^{-9}
Bayes's theorem probability (%)								79.5	20.5

	^{1}H calcd			scaled shifts		corrected error		probability	
position	15a	15b	^{1}H expt	15a	15b	15a	15b	15a	15b
1(*R*)	2.79	2.66	2.82	2.79	2.70	−0.03	−0.12	0.44	0.26
1(*S*)	2.75	2.76	2.82	2.76	2.80	−0.06	−0.02	0.38	0.46
2(*R*)	1.39	1.16	1.58[a]	1.46	1.25	−0.12	−0.28	0.26	0.07
2(*S*)	1.32	1.28	1.53[a]	1.39	1.36	−0.14	−0.22	0.22	0.13
3(*R*)	1.42	1.57	1.57	1.48	1.64	−0.09	0.07	0.32	0.36
3(*S*)	1.47	1.52	1.57	1.53	1.59	−0.04	0.02	0.42	0.46
4(*R*)	1.56	1.60	1.66	1.61	1.67	−0.05	0.01	0.40	0.49
4(*S*)	1.68	1.37	1.66	1.73	1.45	0.07	−0.21	0.36	0.14
6(*R*)	2.52	2.58	2.29	2.54	2.63	0.25	0.34	0.10	0.05
6(*S*)	1.75	1.58	1.64	1.80	1.65	0.16	0.01	0.21	0.48
7	1.84	2.15	1.85	1.89	2.20	0.04	0.35	0.42	0.04
8(*R*)	1.49	1.46	1.49	1.55	1.53	0.06	0.04	0.37	0.41
8(*S*)	1.17	1.16	1.20	1.25	1.24	0.05	0.04	0.40	0.42
9(*R*)	3.07	2.87	3.00	3.06	2.90	0.06	−0.10	0.38	0.30
9(*S*)	1.89	1.97	2.14	1.93	2.03	−0.21	−0.11	0.14	0.28
10	1.73	1.46	1.81	1.78	1.54	−0.03	−0.27	0.43	0.08
11(*R*)	1.80	2.39	1.83	1.85	2.44	0.02	0.61	0.46	0.003
11(*S*)	1.39	1.15	1.53	1.46	1.23	−0.07	−0.30	0.35	0.06
12	1.47	1.50	1.53	1.53	1.57	0.00	0.04	0.50	0.42
13	1.88	1.88	2.03	1.92	1.94	−0.11	−0.09	0.29	0.32
14(*R*)	0.81	0.82	0.89	0.90	0.91	0.01	0.02	0.47	0.46
14(*S*)	1.97	1.95	2.02	2.01	2.01	−0.01	−0.01	0.47	0.47
15	2.26	2.21	1.95	2.29	2.26	0.34	0.31	0.04	0.06
16	0.80	0.79	0.85	0.89	0.88	0.04	0.03	0.42	0.44
17	1.94	1.93	2.12	1.98	1.99	−0.14	−0.13	0.23	0.25
product of probabilties								1.19×10^{-13}	3.85×10^{-19}
Bayes's theorem probability (%)								100.0	3.2×10^{-6}

[a] These signals should be swapped when comparing the data to those calculated for **15b** in order to give the best match to the calculated shifts. In this example these are the only two shifts that need to be swapped because the data are mostly experimentally assigned; in general, more substantial reordering of the shifts may be required. Combined ^{13}C and ^{1}H product of probabilities: **15a**, 9.66×10^{-22}; **15b**, 8.06×10^{-28}. Combined Bayes's theorem probability (%): **15a**, 100.0;**15b**, 8.3×10^{-7}.

column have been swapped around as necessary. The mean absolute value of these errors for ^{13}C or ^{1}H gives the CMAE parameter. However, in order to calculate DP4, each error is converted to a probability that such an error (or a more extreme one) is obtained. To do this we assume, as described previously, that the errors (for a correct assignment) follow a normal distribution with mean zero (this is an automatic consequence of the empirical scaling process) and standard deviation 2.306 ppm (^{13}C) or 0.185 ppm (^{1}H). These values were obtained from an analysis of all the molecules in Figure 1 for which we have experimental data. For example, when comparing the experimental data to structure **15a**, carbon C2 gives an error of −1.55 ppm (second row of the "probability" column for **15a** in the top section of Table 2). If the structure **15a** is right, then the probability of obtaining an error of −1.55 ppm or larger (i.e., more negative) is given by $1 - T^{11.38}(|-1.55 - 0|/2.306) = 0.26$ (part of the numerator of eq 3). Repeating the calculation for all the shifts gives the results in the "probability" columns of Table 2.

Finally, multiplying together the values in each column and dividing by the sums of the products according to eq 3 gives the DP4 probabilities. For example, using the ^{13}C data alone, the structure **15a** is assigned a probability of 79.5%, which comes from multiplying the values in the "probability" column for **15a** in the top section of Table 2 and dividing by the sum of the product of "probability" column for **15a** and the product of the “probability” column for **15b**. To obtain the combined $^{13}C/^{1}H$ DP4 probability, the product of all the values in both “probability” columns for **15a** from the top and bottom sections of Table 2 is found and divided by the sum of the product of these columns and the product of the same columns for **15b**.

The final result, using both ^{13}C and ^{1}H data, is that candidate structure **15a** is assigned a probability of almost 100%, while the alternative structure **15b** is assigned a probability of about 10^{-7}. Assuming that our probabilities accurately reflect the certainty of the conclusion (this issue is addressed later), this corresponds to a near certainty that the structure of nankakurine A is **15a** and not **15b**. Thus, for nankakurine A, DP4 allows us to be confident in assigning the structure to **15a**. This will not always be the case, as it may be that the candidate structures are so similar that they cannot be distinguished with any significant confidence. Had the probabilities been 80% **15a** and 20% **15b** for instance, we could not place much confidence in the result because the structure with 20% probability would still have a reasonable chance of being correct: on average we should expect one in five assignments made on a probability of 80% to be wrong. If the probabilities had been 51% and 49%, we would have had to conclude that the two structures provide essentially an equally good match and no conclusion about which is most likely to be correct would be possible. DP4 therefore not only predicts a structure but provides an indication of how confident one can be in the conclusion in any particular case.

Inspection of the error and probability columns of Table 2 reveals why DP4 makes a clear assignment of nankakurine A to **15a** while, as will be seen later (Figure 4), the standard MAE and CMAE parameters do not. For nankakurine A it is the proton data that make the most decisive contribution to establishing **15a** as the best match: the product of ^{1}H probabilities for **15a** is 6 orders of magnitude greater than that for **15b**. Although other atoms play a role, this is in large part due to the particularly large difference (0.61 ppm) between the scaled calculated shift of H11(*R*) (the axial proton on C11) in **15b** and the corresponding experimental shift, compared to a much smaller difference for **15a**. A difference of this size would be very unlikely (probability 0.003) if structure **15b** were correct, and this feeds into the final conclusion that **15b** is unlikely to be the correct structure. By contrast, using the MAE and CMAE parameters, the effect of the large error in H11(*R*) for **15b** is diluted by the averaging with the errors in all the other nuclei.

It may be that the large difference in the calculated shift of H11(*R*) between **15a** and **15b** is related to the fact that this proton is close to the nitrogen attached to C5 in structure **15b** but not in **15a**. The same is true for H7, which is the second most diagnostic proton in Table 2. These differences are also seen in the experimental data: H11(*R*)has a significantly larger shift in **15b** (2.39 ppm) than in **15a** (1.83 ppm), in agreement with the calculated shifts, and the same is true to a lesser extent for H7.

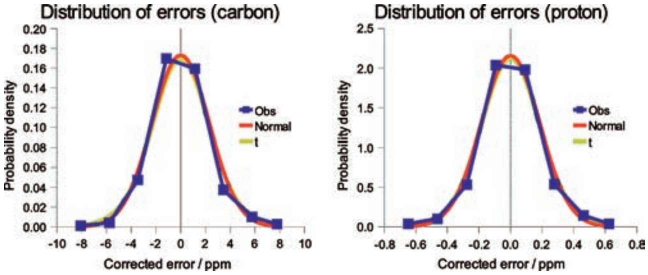


Figure 3. Testing whether the observed errors (Obs.) follow a normal or *t* distribution. The observed distribution was obtained by dividing the horizontal axis into blocks of one standard deviation’s width and counting the fraction of errors occurring in each block. This was then converted to a frequency density (by dividing by the width of the block) and a point plotted in the center of the block.

The results for nankakurine A are presented in Figure 4, along with those for the other natural products studied. We note that nankakurine A gives unusually low values of the correlation coefficient (0.9869 and 0.9869 for right (**15a**) and wrong (**15b**) assignments respectively, compared to an expectation value for a right assignment of 0.9989 with standard deviation 0.0011). We attribute this to nankakurine A possessing no sp^2 carbons and hence a relatively small range of shifts. For most molecules in the data set, plotting a graph of δ_{calc} vs δ_{exp} can be expected to give a cluster of points corresponding to the sp^3 carbons (and corresponding protons for the ^{1}H graph) and one for the sp^2 carbons; this will tend to give a high correlation coefficient in the same way that a graph with only two points gives a perfect line. Nankakurine A does not benefit from this effect and so gives an unusually low correlation coefficient compared to the expectation value which has been obtained by an analysis of all the molecules in the data set. We have observed similar results with other molecules, such as cephalezomine A **39a** (considered later), which has no sp^3 carbons and gives even lower values of the correlation coefficient. We believe that probabilities derived from the correlation coefficient should be treated with caution in such molecules because of this unusual behavior. We note that this behavior affects only the correlation coefficient, not the MAE and CMAE parameters or the DP4 probability.

Investigation of the Error Distribution. The numerical value of the DP4 probability can only be a rough guide to the probability of each candidate structure being correct due to the assumptions involved in the DP4 calculation. Specifically, we have assumed (i) that the corrected error ($\delta_{scaled} - \delta_{exp}$) in each shift is an independent random variable and (ii) that this error follows a *t* distribution.

To check the validity of assuming a *t* (or normal) distribution, we investigated the distribution of errors (i.e., of $\delta_{scaled} - \delta_{calc}$) for all the structures in Figure 1 for which we have experimental data. This gives a total of 1717 data points for ^{13}C and 1794 for ^{1}H. The results are shown in Figure 3, which compares the distribution of errors observed to that predicted by the fitted *t* and normal distributions. In each case the fitted distributions appear to be a reasonable fit. Additional justification comes from the success of the DP4 parameter in Figure 4 (below).

However, it is not possible to say for certain from these results whether the errors follow a *t* distribution, and if the model distribution is not followed exactly, then the calculated probabilities may not exactly reflect the level of confidence in the conclusion in all cases. If, for example, the tails of the distribution are fatter in the true distribution than in the fitted *t* distribution (which would not be obvious from Figure 3 but may be true for the proton data), DP4 would impose a greater penalty on anomalously large errors than it should, and this could lead to an overstatement of the probability in some cases. Nevertheless, even if this is the case, the numerical values of DP4 would still be a useful guide to the level of confidence in the conclusion, provided that the uncertainty in the level of confidence is kept in mind.

In fact, we found that in a few cases DP4 appears to overstate the probability, as indicated by a small number of incorrect assignments made with apparently very high confidence: aldol **21b** was incorrectly assigned as **21a** with 99.91% confidence, and cephalezomine H **31a** was incorrectly assigned as structure **31c** with 99.997% confidence. Although the nature of probability means that one must expect to be occasionally certain and wrong, the frequency with which this occurs should be in line with the apparent confidence in the result. For example, one should expect an assignment made with 99.9% confidence to be wrong only one time in 1000 (1/(1 − 0.999)), so in a data set of 117 molecules with experimental data the result for aldol **21b** and particularly cephalezomine H **31a** is unlikely (though not inconceivable) unless the probabilities are being overstated. Accurate determination of the shape of the probability distribution in the region of the low-probability tails would require significantly more data, so we will continue to use the *t* distribution for the present study. We note that, using the normal distribution, cephalezomine H **31a** and **21b** are assigned with 99.99999997% and 99.95% confidence, respectively, suggesting that the normal distribution has a greater tendency to overstate the probability in extreme cases. This is most likely due to the *t* distribution having slightly thicker tails than the normal distribution and giving a better fit to the experimental distribution. On the other hand, the *t* distribution looks very similar to the normal distribution in Figure 3 and for most molecules gives very similar probabilities. This is confirmed by a comparison of the results of DP4 for each molecule using the two distributions; these probabilities may be found in the Supporting Information. If it is desirable to use the normal distribution in place of Student's *t* distribution, T^{ν} should be replaced by the standard cumulative normal distribution function in eq 3, and our Web applet includes an option for using the normal distribution in place of the *t* distribution. However, we recommend using the *t* distribution for the reasons just discussed, and the *t* distribution has been used to generate all the graphs in this paper.

Application of DP4 to Assigning the Stereochemistry of Natural Products. Figure 4 shows the results of using DP4 to assign the stereochemistry of all the natural products in Figure 1. Results for the remaining molecules in Figure 1 may be found in the Supporting Information.

Each graph in Figure 4 shows the DP4 probabilities assigned to each candidate structure, and these are compared to the probabilities obtained using the correlation coefficient, MAE, and CMAE parameters. In each graph the probability assigned to the correct structure is represented by the white section of each bar, so an ideal parameter would have as much of the bar shown in white as possible. Unlike DP4, the correlation coefficient, MAE, and CMAE parameters do not yield probabilities directly; instead probabilities were calculated using Bayes's theorem together with a knowledge of the values of the parameters expected for a correct and incorrect assignment (see the Supporting Information for full details). Specifically, we used eq 4 (which is written in terms of a generalized parameter *p* that could be the correlation coefficient *r*, MAE, or CMAE) and the values of μ and σ obtained from an analysis of all the molecules in Figure 1 for which we have experimental data.

$$P(i|p_1, p_2, ..., p_m) = \frac{(1 - \Phi(|p_i - \mu_r|/\sigma_r)) \prod_{j \neq i} (1 - \Phi(|p_j - \mu_w|\sigma_w))}{\sum_{k=1}^{m} [(1 - \Phi(|p_k - \mu_r|/\sigma_r)) \prod_{j \neq k} (1 - \Phi(|p_j - \mu_w|/\sigma_w))]} \quad (4)$$

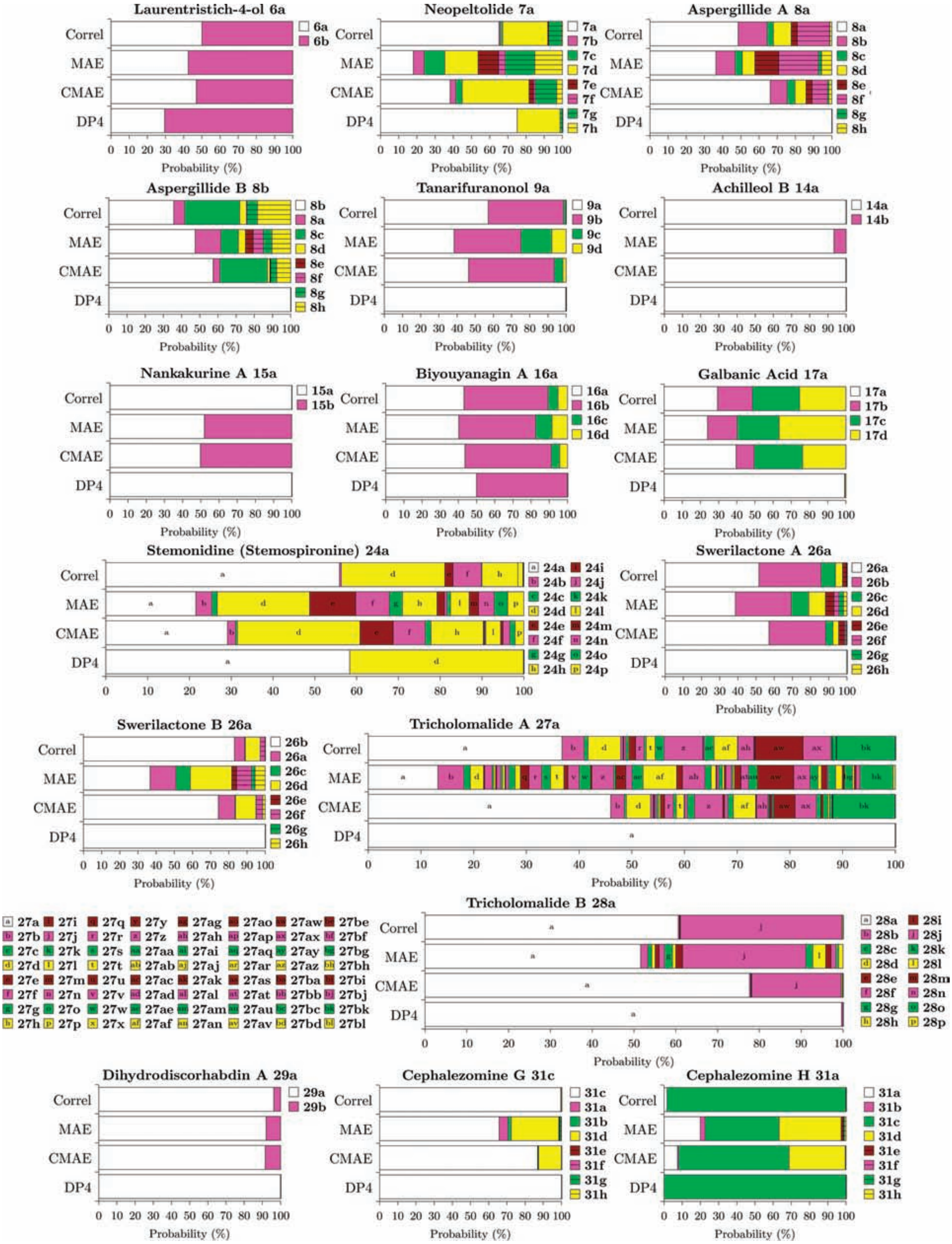
As in our previous study[30] we used the "combined" values of *r*, MAE, and CMAE, in which the quantities (1 − *r*), MAE, and CMAE for ^{13}C and ^{1}H are combined using the geometric mean. μ_r and μ_w are the expectation values of the parameter in question for a right and wrong assignment, respectively, and σ_r and σ_w are the standard deviations in these values. $\Phi(x)$ is the standard cumulative normal distribution function; we did not use the *t* distribution for these parameters because they do not show the same tendency to overstate the probabilities that DP4 with the normal distribution does; indeed they suffer most often from the opposite problem of not giving a clear assignment when DP4 correctly identifies the right structure with high confidence.

Figure 4 shows that DP4 gives excellent results, with a clear and correct prediction made with nearly 100% confidence in almost all cases, as indicated by all or most of the DP4 bar being white in each case. Even if the probability assigned to the correct structure is not 100%, the correct isomer is almost always picked out as the most likely candidate with good confidence (for example, the case of neopeltolide **7** and Ardisson's polyketide **32**) or is one of two identified candidates (biyouyanagin A **16**, stemonidine **24**, maurenone **33**, and laurentristich-4-ol **6**). By contrast, the other parameters (correlation coefficient, MAE, and CMAE) often give inconclusive results, with significant probability being assigned to several candidate structures.

Of the natural products studied, only for cephalezomine H **31a** does DP4 make an incorrect assignment with significant confidence, and in this case the other parameters also give the wrong answer. The correct structure is, however, the second most likely candidate out of the eight possibilities. We note that the nature of probability means that we should not expect to be right in every single case. For example, if we make 10 assignments with 90% confidence (so that if the our probabilities are accurate, each assignment has a 10% chance of being wrong), we would expect on average for one of these 10 assignments to be incorrect. However, as already mentioned, the certainty with which DP4 gives the wrong answer for cephalezomine H suggests that DP4 may be overstating the confidence in this case, most likely due to the assumptions involved in the DP4 calculation that have already been discussed.

For systems with a large number of possible structures with very similar spectra, assignment is particularly challenging. For tricholomalide A **27**, Ardisson's polyketide **32**, and maurenone **33**, DP4 gives the highest probability to the correct structure, outperforming analyses using correlation coefficients, MAE, and CMAE. Maurenone **33** is of particular interest, because DP4 has 40% confidence in the correct structure and 90% confidence in its top three structures, whereas all the other methods give very similar probabilities to a large number of structures, suggesting that there is insufficient information in the spectra to distinguish the structures effectively.

In a few cases, the conclusion that it is impossible to assign the structure on the basis of the data available is probably the correct analysis. Stereopentads **18** and stereopentad acetates **22** are series of very similar molecules for which many spectra are available.[40] In about half of these difficult cases, DP4 had quite high confidence in an incorrect assignment, whereas the analyses with the correlation coefficient, MAE, and CMAE suggest that confident assignments should not be made (details are available in the Supporting Information). **18** and **22** are small and flexible molecules for which solvent effects may be particularly important, and this may be the origin of DP4's overconfidence in this case. For two of the similar cases where DP4 outperforms the other methods, tricholomalide A **27** and maurenone **33**, the molecules are rather less flexible. Ardisson's polyketide **32**, however, has similar flexibility to the stereopentad 18, and yet DP4 is very effective.



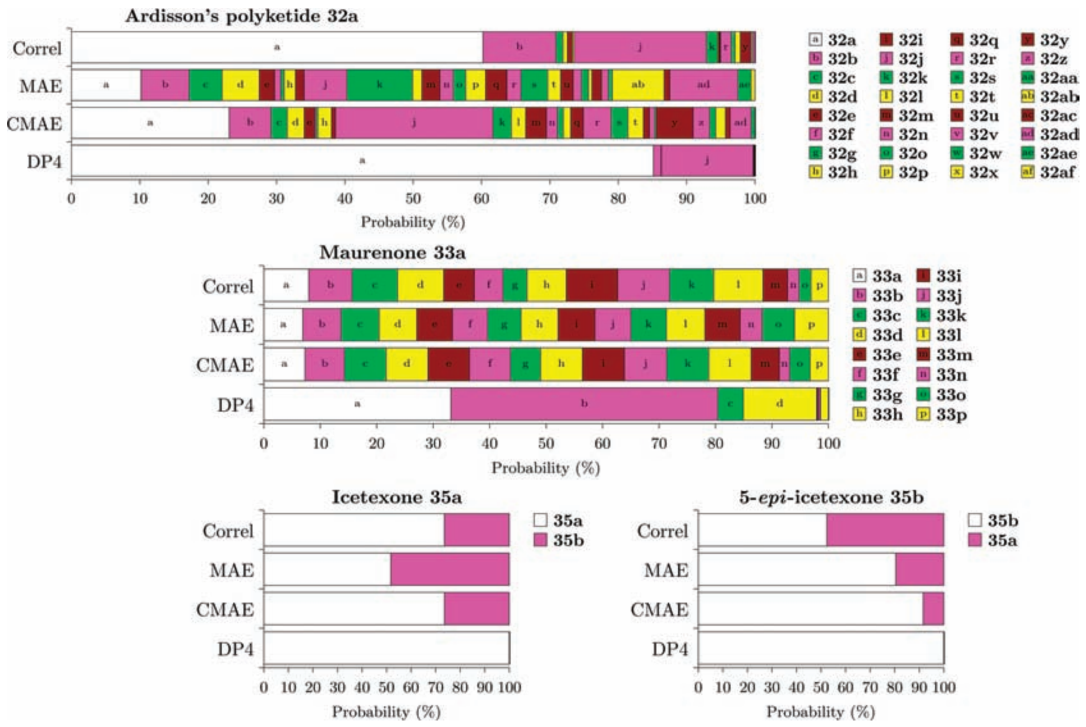


Figure 4. Assigning natural products using DP4. The graphs show the DP4 probabilities for each candidate structure, together with the probabilities that are assigned by an analysis based on the values of the correlation coefficient, MAE and CMAE. The probability for the correct structure is shown in white in each case, so the best results are where most or all of the horizontal bar is white (i.e., the correct structure assigned a probability of near 100% and the other candidate structures assigned a low probability). The DP4 approach is much more successful than are the approaches based on the correlation coefficient, MAE, and CMAE.

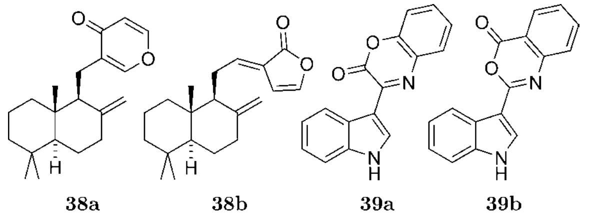


Figure 5. Ottensinin, correct structure (**38a**) and originally proposed structure (**38b**), and cephalandole A, correct structure (**39b**) and originally proposed structure (**39b**).

Structural Isomers: Ottensinin 38a and Cephalandole A 39a. We also tested the applicability of DP4 to structural (rather than stereo) isomers. We consider ottensinin, originally assigned as **38b**[91] but later shown to be **38a**,[92] and cephalandole A, originally assigned to **39b**[93] but subsequently found to be **39a** (Figure 5).[94] The results are shown in Figure 6.

Refining Stereostructural Assignment. In most cases, DP4 assigns the correct structure with high confidence. In the few cases it does not, it generally prioritizes the trial structures effectively. For example, in Ardisson's polyketide **32** the correct structure is only 85% certain, but this is dramatically better than the 3% confidence of a random selection from 32 possibilities. We will now address the question of how much better we can do (in terms of the probability assigned to the correct structure) with the DP4 approach compared to if we have not done any NMR shift calculations. Specifically we determined, for each of the molecules in Figure 1, whether each parameter gave the correct structure a higher or lower probability than one would get from guessing at random. To quantify the improvement in the assigned probability over guessing at random, we define the following "improvement factor":

$$\begin{aligned}\text{improvement factor}\\ &= \frac{\text{probability assigned to correct structure by parameter}}{\text{probability assigned by guessing at random}}\\ &= \frac{\text{probability assigned to correct structure by parameter}}{1/N}\end{aligned} \tag{5}$$

where N is the number of possible candidate structures.

We calculated the improvement factor for each of the molecules in Figure 1 using each of the parameters. The distribution of improvement factors for each parameter are shown in Figure 7. The bars on the left in each graph are for when a correct structure is being assigned, and the bars on the right are for when one is attempting to assign a wrong structure.

To see how these plots are generated, consider as an example assigning aspergillide B using the correlation coefficient, the probabilities for which were shown in Figure 4. There are eight candidate structures, so by guessing at random one would assign each a probability of being right of 12.5%. The correlation coefficient, however, assigns the correct structure a probability of 35.7% (Figure 4) which is a 2.86-fold increase on 12.5%. Thus, aspergillide B is one of the structures represented by the 1.25−4 bar in the left-

[91] Akiyama, K.; Kikuzaki, H.; Aoki, T.; Okuda, A.; Lajis, N. H.; Nakatani, N. *J. Nat. Prod.* **2006**, *11*, 1637–1640.

[92] Boukouvalas, J.; Wang, J.-X. *Org. Lett.* **2008**, *10*, 3397–3399.

[93] Wu, P.-L.; Hsu, Y.-L.; Jao, C.-W. *J. Nat. Prod* **2006**, *69*, 1467–1470.

[94] Mason, J. J.; Bergman, J.; Janosik, T. *J. Nat. Prod.* **2008**, *71*, 1447–1450.

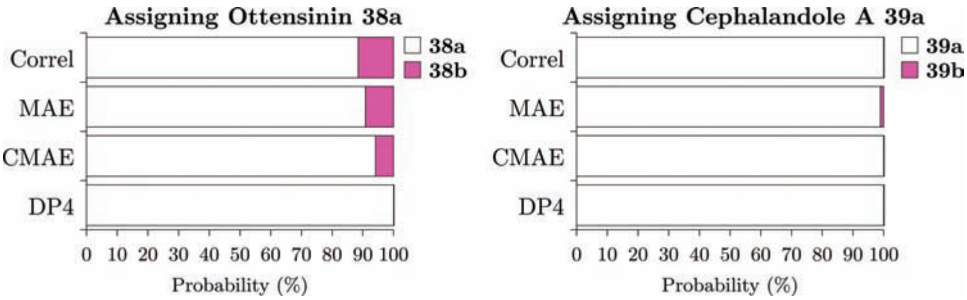


Figure 6. Assigning ottenisin and cephalandole A using the same approaches as used in Figure 4. As in Figure 4, bars that are entirely or mostly white are good, as they indicate a correct assignment made with high confidence. Both the DP4 probabilities and the probabilities based on the correlation coefficient, MAE, and CMAE give good results for these molecules, but the DP4 probabilities are the most successful.

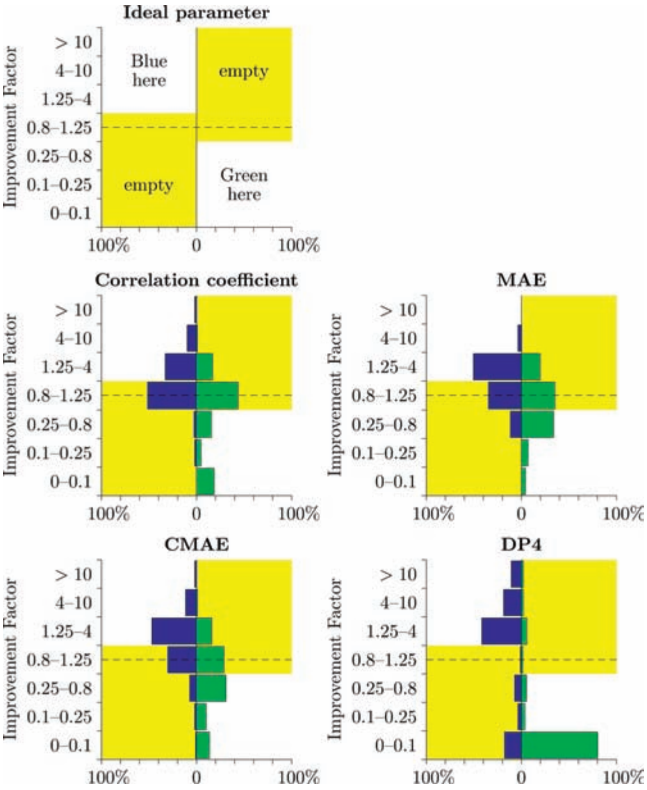


Figure 7. Distribution of improvement factors across all the structures. The bars on the left are for when a correct structure is being assigned, and the bars on the right are for when one is attempting to assign a wrong structure. An ideal result would have bars in only the top left-hand corner (structures assigned a high probability when the assignment being made is right) and the bottom right-hand corner (structures assigned a low probability if the assignment being made is wrong). The DP4 probabilities are more successful in achieving this distribution than are probabilities based on the correlation coefficient, MAE, and CMAE parameters.

hand half of the correlation coefficient graph in Figure 7. Structure **8d**, on the other hand, is assigned a probability of 3.7% (Figure 4) by the correlation coefficient when assigning aspergillide B, indicating that the correlation coefficient has correctly identified this wrong structure as being less likely than if one were guessing at random. 3.7/12.5 = 0.30, so attempting to wrongly assign aspergillide B to **8d** contributes to the 0.25−0.8 bar in the right-hand half of the correlation coefficient graph in Figure 7. Repeating this process for all of the correct assignments and all possible incorrect assignments, and counting up the number of cases in each bar, then allows Figure 7 to be generated. The bars on the left represent a total of 113 structures and the bars on the right 928 possible incorrect assignment combinations. The heights of the bars have been normalized so that the total areas under the bars on the left and on the right are the same.

Since a factor of 1.0 in Figure 7 is equivalent to guessing at random, an ideal parameter will have the bars on the left-hand half the graph skewed toward the top of the graph (i.e., correct structures assigned an enhanced probability) and the bars on the right-hand side skewed toward the bottom of the graph (incorrect structures assigned a reduced probability). The ideal distribution is indicated by the top graph in Figure 7. Each of the parameters does in fact show this skew to some extent. However, the difference in skew between right and wrong structures is relatively small for the correlation coefficient, MAE, and CMAE parameters, which indicates that these parameters often do not assign correct structures a very high probability of close to 100% or incorrect structures a very low probability. This is borne out in Figure 4. The skew is much greater for DP4, which accounts for the greater success of this parameter in correctly assigning structures with high confidence in Figure 4. We also note that DP4 is much less likely to give an uncertain conclusion than the other parameters: the range 0.8−1.25 is virtually empty in the DP4 graph in Figure 7 but is considerably populated (for both correct and incorrect assignments) in the case of the other parameters.

Conclusions

We have shown that NMR shift calculation in conjunction with our new DP4 probability is a powerful tool for assigning stereochemistry in the challenging case of only having one set of experimental data. DP4 is much more successful at making correct assignments with high confidence than are probabilities based on the mean absolute error and correlation coefficient. It is also much more efficient at assigning an enhanced probability (compared to that from guessing at random) to the correct structure. In a few cases where there are a large number of structures with very similar spectra, DP4 can be overconfident in its assignments. DP4 deals efficiently with both stereochemical and structural assignment, and we recommend its use when one has experimental data for only a single compound to assign.

Acknowledgment. We thank the University of Cambridge (S.G.S.) and Unilever for financial support. We are very grateful to Dr. Julia Channon and Prof. Ian Paterson (University of Cambridge) for providing the experimental data for stereopentads **18** and **19**, diols **20**, aldols **21**, and stereopentad acetates **22**. We thank Dr. Charlotte Bolton for the IT support.

Supporting Information Available: Derivations of eqs 3 and 4, complete details of calculated and experimental shifts, detailed results of assigning each of the molecules in Figure 1 using each of the parameters, expectation values and standard deviations for all of the parameters, and calculated geometries and energies for all the molecules studied. This material is available free of charge via the Internet at http://pubs.acs.org.

JA105035R