The DP5 Probability, Quantification and Visualisation of Structural Uncertainty in Single Molecules

Alexander Howarth,[a] Jonathan M. Goodman*[a]

Whenever a new molecule is made, a chemist will justify the proposed structure by analysing the NMR spectra. The widely-used DP4 algorithm will choose the best match from a series of possibilities, but draws no conclusions from a single candidate structure. Here we present the DP5 probability, a step-change in the quantification of molecular uncertainty: given one structure and one ^{13}C NMR spectra, DP5 gives the probability of the structure being correct. We show the DP5 probability can rapidly differentiate between structure proposals indistinguishable by NMR to an expert chemist. We also show in a number of challenging examples the DP5 probability may prevent incorrect structures being published and later reassigned. DP5 will prove extremely valuable in fields such as discovery-driven automated chemical synthesis and drug development. Alongside the DP4-AI package, DP5 can help guide synthetic chemists when resolving the most subtle structural uncertainty. The DP5 system is available at https://github.com/Goodman-lab/DP5

Introduction

Molecular structure elucidation and verification are central problems in organic, synthetic and natural product chemistry. Due to richness of the structural information its spectra contain, NMR spectroscopy has cemented itself as the method chemists use to solve these problems. Due to the complex nature of NMR spectra and often subtle variation between similar molecules, interpretation of these spectra can sometimes present a significant challenge. As a result, incorrectly assigned structures remain pervasive in the literature.[1] Many of these incorrectly assigned are only discovered after costly and time consuming total syntheses are completed revealing a discrepancy between the experimental and literature NMR data.[2,3,4]

Over the last two decades, many computational tools have been developed to aid the assignment of NMR spectra and elucidation of molecular structures.[5–7] Comparing experimental NMR shifts with those calculated for a candidate structure using density functional theory (DFT) is now a well-established methodology and has been used to solve the structures of many molecules.[8–11] A powerful way of performing this analysis is to calculate the DP4 probability (and related metrics like DP4+ and J-DP4).[12–14] Unlike comparative metrics such as MAE and CMAE, the DP4 algorithm applies Bayes Theorem to calculate the probability that each candidate structure is the correct one. DP4 requires a list of possible structures as its input, and it assumes that one of these structures is correct. It is common for structures to be determined except for uncertainty in the details of their stereochemistry. DP4 has proved invaluable in the resolution of many such cases.[15–19] DP4 can also be used to resolve non-stereochemical uncertainty, provided that all of the acceptable possible structures can be enumerated. However, in cases where all the proposed candidate structures may be incorrect or only a single structure has been proposed, DP4 analysis cannot be applied. Until now in these very common situations chemists would have no quantitative way of assessing the probability of their proposed structure being correct given the NMR spectra.

To solve this problem, we present the DP5 probability, a new methodology and complete software package for quantifying uncertainty in molecular structures. Similar to the DP4 probability, the DP5 probability gives the probability that a candidate structure is correct. However, in contrast to DP4, DP5 calculates normalised stand-alone probabilities and hence, the user can propose one or many structures without having to assume any of their proposals are correct. As a result, DP5 can be used to answer different questions to DP4 and will prove valuable in situations where this type of analysis was previously impossible. The DP5 probability is calculated given only one-dimensional ^{13}C NMR data and utilises the same computational engine as the latest iteration of our DP4 software, DP4-AI. This program manages all NMR processing, assignment, DFT calculations and statistical modelling automatically. DP5 can also be used on a case-by-case basis utilising the graphical user interface (GUI).

This work represents a great leap forward in this field, in previous works systems have been developed using Neural Networks (NNs) to classify structure proposals as correct or incorrect.[20] These systems return binary metrics that cannot be interpreted as a probability, the statistical approach taken by

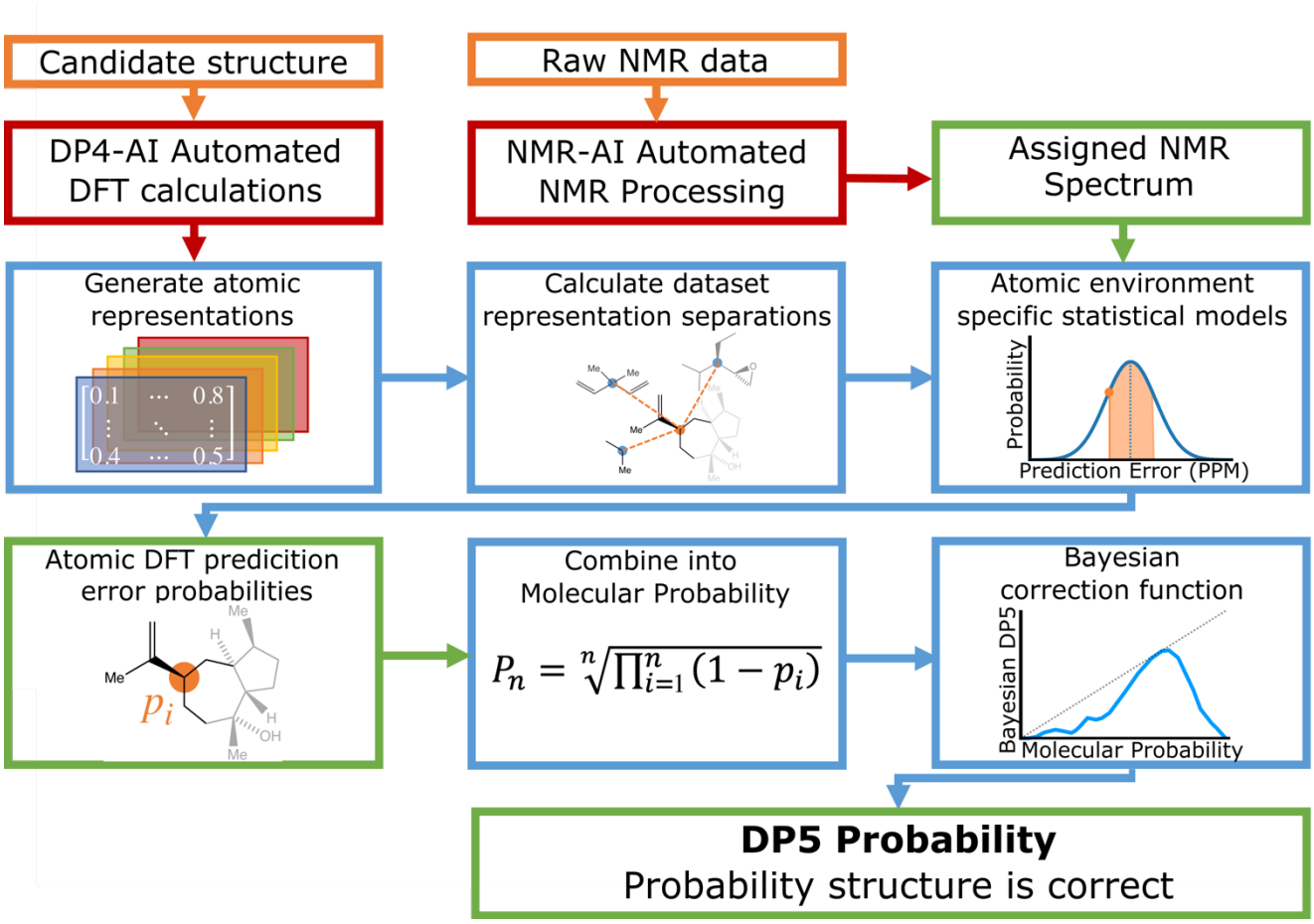

Figure 1. Schematic of the DP5 program. The required inputs from the user are a candidate structure and the raw ^{13}C NMR data (or a list NMR signals). The DP5 probability is built on top of the DP4-AI analysis.

DP5 solves the significantly more challenging problem of calculating a standalone normalised probability of a strcutre being correct. The NN systems are trained on vectors of metrics such as MAE and MSTD and hence repeat the analysis an expert chemist could perform. DP5 in contrast, takes into account the 3d geometry around each atomic environment to calculate the probability of observing the given NMR-DFT prediction error for each atom independently, this solves the well-known issue that DFT prediction errors vary in complex and non-linear ways with atomic environment. Finally, previous methods based on NNs are trained using a set of correct structure proposals and faked "incorrect" structure proposals. In contrast, the statistical approach taken by DP5 only utilises real data, avoiding utilising any fake datapoints and the subsequent effects of unbalanced training sets.

The system was developed and rigorously tested utilising a dataset of 5140 organic molecules from NMRShiftDB originally selected for NMR prediction using machine learning by Paton *et al.* (see supporting information section 6).[21,22] To demonstrate the performance of the DP5 probability in even more challenging situations, the system was also evaluated using 13 case studies of molecular structures that have undergone reassignments in the literature and addition 42 challenging relative stereochemistry elucidation examples.

DP5 represents an exciting leap forward in quantifying molecular uncertainty. This system will prove valuable in fields requiring high throughput molecular structure elucidation such as automated chemical synthesis, but also in traditional organic chemistry as a tool to aid and guide expert chemists in their development of complex syntheses. DP5 has been made possible following recent advances in molecular machine learning techniques and increased data availability.[23–28]

Computational Methods

DFT calculations for the structure reassignment and stereochemistry elucidation examples were performed using the method developed in previous works.[29–31] All molecular mechanics calculations were performed using MacroModel (Version 9.9)[32]. All conformational searches were performed in the gas phase utilizing the MMFF force field[33–38] and a mixture of Low Mode following and Monte Carlo search algorithms.[39,40] The step count for MacroModel was set so that all low energy conformers were found at least 5 times. Quantum mechanical calculations were carried out using Gaussian09[41]. NMR shielding constants were found using the GIAO method.[42–44] The functional mPW1PW91[45] was chosen with the 6-311G(d)[46,47] basis set for NMR shift prediction as this has been shown to be optimal for DP4 calculation. For molecules containing iodine,

the basis set def2-SVP[48,49] was chosen. All DFT calculations were performed using the implicit PCM solvent model.[50] The molecular geometries were also optimized at the DFT level of theory, this was performed using the B3LYP functional[51,52] with the 6-31G(d) basis set. Finally, single-point energies were separately calculated using M06-2X[53] functional and def2-TZVP[48,49] basis set.

The calculations were managed by the DP4-AI[31] Python script written in Python 3.7. DP4-AI is available from http://www-jmg.ch.cam.ac.uk/tools/nmr/ and GitHub https://github.com/Goodman-lab/.

DFT optimised geometries and NMR shift calculations for the molecules from NMRShiftDB were obtained from the training data of the GNN NMR shift prediction software CASCADE.[22] A single conformer of each of these molecules was optimised utilising the M062X functional and def2-TZVP basis set and NMR shift calculations performed using in 6-311g(d) basis set and mPW1PW91 functional.

Calculation of FCHL atomic representations, l2 distances and gaussian kernel transformations were performed using the python package qml.[54]

Program Description

A schematic of the DP5 program is displayed in Figure 1. Structure inputs can be made as any combination of, .sdf files, SMILES, SMARTS or InChIs. ^{13}C NMR data can be input as raw data (for automated analysis) or as a list of peaks from a user analysis.

DP5 calculates NMR shifts for the atoms in populated conformers of the candidate structure utilising the highly optimised and well established methods within DP4-AI (see supporting information section 2.1).[29–31]

Raw NMR data interpretation is handled by a part of DP4-AI called NMR-AI.[31] This system was developed to remove the requirement for the user to process and assign NMR spectra and has been demonstrated to complete this task to at least the same high standard as an expert chemist.

Once the geometries of populated conformers have been calculated, the probabilities of the observed DFT-NMR prediction errors for each atom in that conformer can be found. To do this a probability density function (PDF) describing the DFT-NMR prediction error distribution for that atomic environment is required (see supporting information section S2.2). This PDF is found empirically by performing a Kernel Density Estimation (KDE) on a dataset of 63542 known prediction errors calculated for the DFT optimised geometries of 5140 molecules from NMRShiftDB. This dataset was originally developed for training machine learning models for NMR shift prediction, the generality and near chemical accuracy achieved by these models has been taken as justification for using this dataset in this similar task (details regarding this dataset can be found in the original publication).[22] It is well known that the expected magnitude and variance of DFT prediction errors for different functionals show strong complex, nonlinear dependencies on atomic environment.[55,56] This process takes this into account by weighting the contribution to the error PDF for the test atom of each atomic environment in the database by its similarity to the test environment. The similarity of these atomic environments is calculated by finding the Euclidian distance between a vector representation of the test atomic environment and those in the training set. These distances are converted into covariances utilising a gaussian kernel (equation (1); see supporting information section S2.4). By setting the pre-exponential scaling factor to one, these covariances can be interpreted as a measure of the similarity. The resulting PDF is integrated by equation (2) (see supporting information section S2.5) to yield a prediction error probability for the test atom.

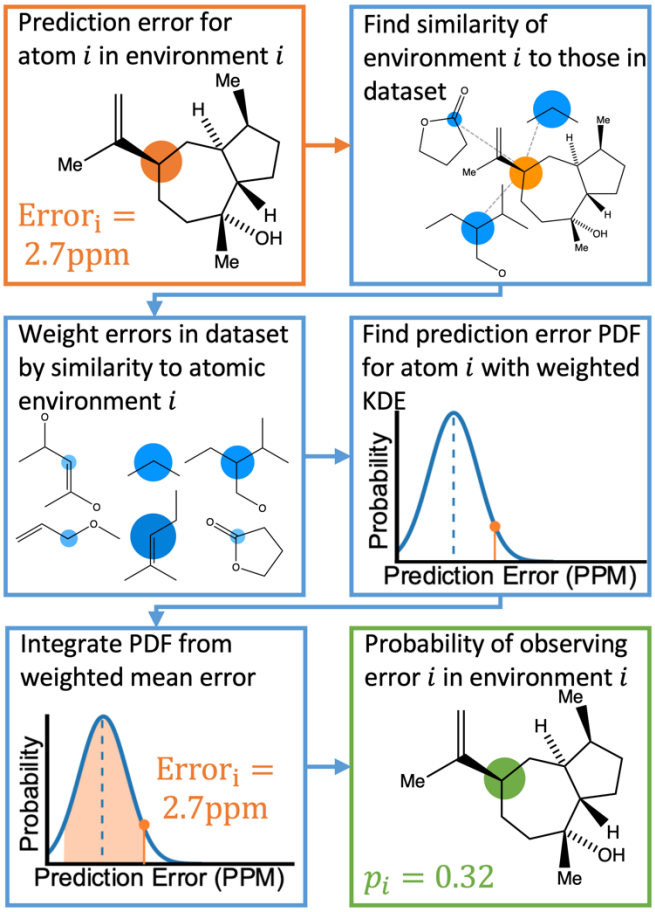

Figure 2. Schematic diagram of how the probability of observing a DFT-NMR prediction error for an atom in a given environment is calculated as described in the text.

$$\text{similarity}_{ij} = \text{A}\exp\left(-\frac{\left\|X_i - X_j\right\|_2^2}{2\sigma^2}\right) \quad (1)$$

The similarity between the FCHL representation of atom i, X_i and that of atom j in the test set X_j is calculated using a gaussian kernel.

$$\text{p}_i = \int_{-|\bar{\Delta}-\Delta_i|}^{+|\bar{\Delta}-\Delta_i|} pdf_i(\Delta)\, d\Delta \quad (2)$$

A prediction error probability for atom i is calculated by integrating the bespoke prediction error function generated for that atom, where Δ_i is the (internally scaled) prediction error for atom i and $\bar{\Delta}$ corresponds to the mean absolute prediction error for the training set.

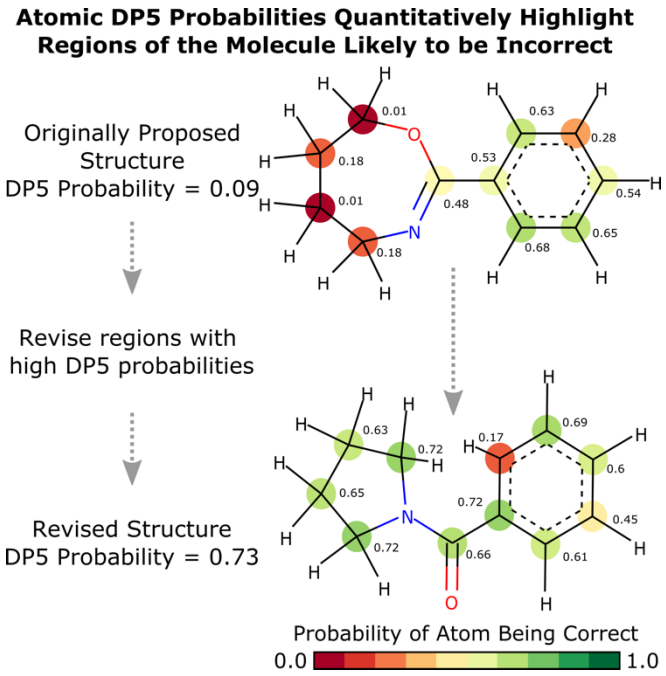

Figure 3. The GUI accompanying DP5 pictorially overlays atomic DP5 probabilities onto the molecular structure. This clearly displays regions of the structure that are expected to be correct and conversely regions that may require revision. This functionality will help chemistry assess and revise structure proposals. This structure revision example has been taken from a real-world case study of an incorrectly assigned molecule in the literature (see Results)

This process is then repeated for each atom in each conformer of the proposed structure. Once atomic probabilities have been calculated for each atom in each conformer, these values are Boltzmann weighted to produce overall atomic probabilities for the structure. This process is summarised in Figure 2.

The atomic representation used by DP5 was investigated in great detail. In recent years many representations have been developed for applications in molecular machine learning, such as the coulomb matrix,[57] bag-of-bonds,[58] aSLATM[59] and FCHL.[27] Kernel ridge regression (KRR) utilising the FCHL atomic representation have been shown to predict NMR shielding constants with near chemical accuracy (also tested in this work see supporting information section S3.4.1).[60] These works demonstrate that FCHL contains the information required to accurately encode atomic environments. Due to the similarity of these tasks, the FCHL representation has been chosen for use in the DP5 probability calculation (see supporting information section S2.3).

A particular challenge in the development of the DP5 probability was determining an equation to combine individual atomic probabilities to yield probability for the whole structure. If any single atom is given too much influence, a molecular probability of one or zero will usually be assigned, whilst if there is too much smoothing over individual atomic probabilities, the resulting molecular probabilities will not show enough useful variation. A number of formulae were tested during this study (see supporting information section S2.7). Overall equation (3) was found to yield useful variation in molecular probabilities, whilst combining the atomic probabilities in a mathematically meaningful way. The inclusion of the geometric mean in equation (3) was found to be necessary to prevent single atoms with a very high or low probability having too much influence on the final result.

The last stage in the calculation scales the molecular probability using a Bayesian correction function to yield the final DP5 probability (see supporting information section S2.8). This empirical stage of the process ensures the DP5 probability assigned matches the probability of the structure being correct as closely as possible. This empirical correction function was found by first calculating a PDF for the molecular probabilities assigned to the 5140 NMRShiftDB molecules. By finding all the possible pairs of spectra and structures in this dataset with the same number of carbon atoms, a PDF of the molecular probabilities of incorrect spectra-structure pairs was also generated. In this instance, each pair was assigned a weight to ensure the mean absolute DFT-NMR prediction error distribution of these incorrect pairs matched that of the correct structure-spectra pairs (see supporting information section S3.2). Given any proposed structure must be either correct or incorrect, by applying Bayes Theorem the DP5 probability is defined by equation (4).

$$\mathrm{P_n} = \sqrt[n]{\prod_{i=1}^{n}(1-p_i)} \tag{3}$$

Atomic DP5 probabilities are combined to form the molecular probability P_n by equation 3, where n is the number of atoms in the molecule and p_i is the DFT-NMR prediction error probability for atom i

$$\mathrm{DP5} = \frac{\mathrm{P(correct|P_n)}}{\mathrm{P(incorrect|P_n) + P(correct|P_n)}} \tag{4}$$

The DP5 probability is calculated by applying Bayes Theorem to the molecular probability calculated in equation 3, where $\mathrm{P(correct|P_n)}$ gives the probability of a molecule being correct given its calculated molecular probability $\mathrm{P_n}$.

Calculation of the DP5 probability has been integrated into the well-established DP4-AI workflow.[31] All the required calculations and analysis of NMR data can be performed automatically with no user input required. DP5 can hence be integrated into pre-existing automatic reaction/characterisation workflows. DP5 analysis can also be performed on single molecules with the GUI. This GUI can be used to launch calculations, analyse NMR assignments made by NMR-AI and also to investigate the DP5 statistics. The GUI visually displays the atomic probabilities, helping the chemist identify potential regions of the molecule that may be incorrect and determine possible modifications (Figure 3).

Results

A major challenge in the development of DP5 involved constructing a method to assess the efficacy of the system. As

the DP5 probability is not a physical property that can be measured, it is not straightforward to compare the DP5 probability assigned to a molecule with an experimental value. In this study two rigorous evaluation methods were devised to assess and improve the real-world effectiveness of the DP5 probability.

The database of 5140 organic molecules from NMRShiftDB was used in comprehensive leave-one-out style cross validation study summarised in Figure 4 (see SI 3.2). In this study DP5 analysis of correct and incorrect proposed candidate structures was simulated by permuting the experimental data between the structures in the dataset to form correct and incorrect pairs of structure and spectra. This analysis is particularly powerful as negative examples could be synthesised from real world data, avoiding more unreliable methods involving generating fake experimental or calculated spectra. This analysis was used to develop DP5 and was repeated for many different formulations of the DP5 probability (see supporting information section S4.1). The results for the final DP5 system can be seen in Figure 4. To ensure the computational feasibility of this analysis, only a single DFT optimised conformer was considered for each pair of structure and spectra. Full conformational analysis here would require significant additional computational resources, we assume that any subsequent decreases in accuracy in the DP5 probabilities here will affect both the correct and incorrect pairs equally, and hence will not change the final conclusions sufficiently to justify the substantial extra expense. In all other experiments structures were subject to full conformational analysis (as is standard in the final program). The final DP5 methodology was then evaluated against a series of thirteen real world structure reassignment problems from the literature, molecules S1a-S13b presented in Figure 5. The results of this study are shown in Figure 6. As the final and most subtle test, the DP5 probability was evaluated against the same dataset of 42 relative stereochemistry problems used to evaluate DP4-AI.[31] With an average of 3.49 stereocentres per molecule and a diverse range of natural-product-like carbon skeletons, this dataset provides rigorous evaluation of DP5 for many real world applications. The results of this analysis are displayed in Figure 7.

Discussion

Results from the combinatorial analysis of the 5140 molecules from NMRShiftdb are presented in Figure 4. In case 1, all incorrect pairs of structure and spectra with maximum errors <10ppm are considered equally. This represents the situation where an experienced chemist should be able to accurately and reliably predict whether a chemically reasonable proposed structure is likely to be correct or incorrect based on the DFT-NMR prediction errors alone. There is very little overlap between the DP5 probability distributions for the correct and incorrect structure proposals, the modal DP5 probability for correct structures being the maximum possible value (see supporting information section S2.8). The incorrect structures display the opposite pattern, with the modal value at close to zero. This result highlights the DP5 probability's ability to differentiate reliably between correct and incorrect structures using DFT prediction errors as well as an experienced chemist in situations where the incorrect and correct structures are both structurally dissimilar and produce very different spectra. When paired this way, the correct and incorrect pairs show different MAE distributions, with the incorrect pairs displaying a larger modal MAE and a greater variance. The DP5 probability would be even more useful if it could reliably differentiate incorrect structure proposals following the same MAE distributions as the correct structures. This is tested in Figure 4, case 2. The incorrect pairs are assigned weights to ensure that they follow the same MAE distribution as the correct pairs (see supporting information section S3.2). This gives an expectation number of about 5330 incorrect structure proposals with a mean error of <2ppm. This mean DFT error is same as that of the correct structures, and hence the correct and incorrect structures now produce spectra that not only very similar to each other but are also often within DFT error of each other. This represents the radically more challenging situation where the correct and incorrect structure proposals have significantly different structures but are practically indistinguishable by their DFT prediction errors. In this situation an expert chemist would have significant difficulty deciding whether a proposed structure is correct or incorrect, and in some cases this may be impossible without collecting additional information. The results of this study are particularly exciting, as despite this test proving to be more demanding, the DP5 probability is still able to correctly differentiate many correct and incorrect structures. This is shown by the DP5 probability frequency distributions, with the correct pairs maintaining a strong peak at the maximum possible value and the incorrect pairs having significant density towards zero. This shows that DP5 will typically assign lower probabilities to incorrect structures even when they display similar spectra to the correct structure. It is able to do this because the internal statistical model takes into account the structure of the proposal, which is not possible in traditional error analysis.

A very interesting feature that these results illuminate is the value of the maximum possible DP5 probability. This value is dependent on many factors including, the dataset of atomic environments, the atomic representation chosen, and, most notably, the inherent uncertainty in the DFT NMR predictions. Using this state-of-the-art and highly-optimised set of conditions, DFT NMR predictions still have a MAE of 1.57ppm. As a result, even if a proposed structure is correct, the DP5 probability has to take into account the possible variance in NMR predictions and reflect this uncertainty. Therefore, when using this set of DFT conditions, the user can never be more than 72% confident that a structure is correct using one dimensional DFT NMR predictions alone, this important functionality has previously been missing from other structure validation systems. This value acts as a metric for assessing the accuracy of DFT NMR calculations and the reliability of the DP5 calculations. We expect that the use of even larger databases and even higher levels of theory will raise this limit. Equivalently, this can be interpreted as acknowledging that an incorrect structure could possibly produce a set of errors

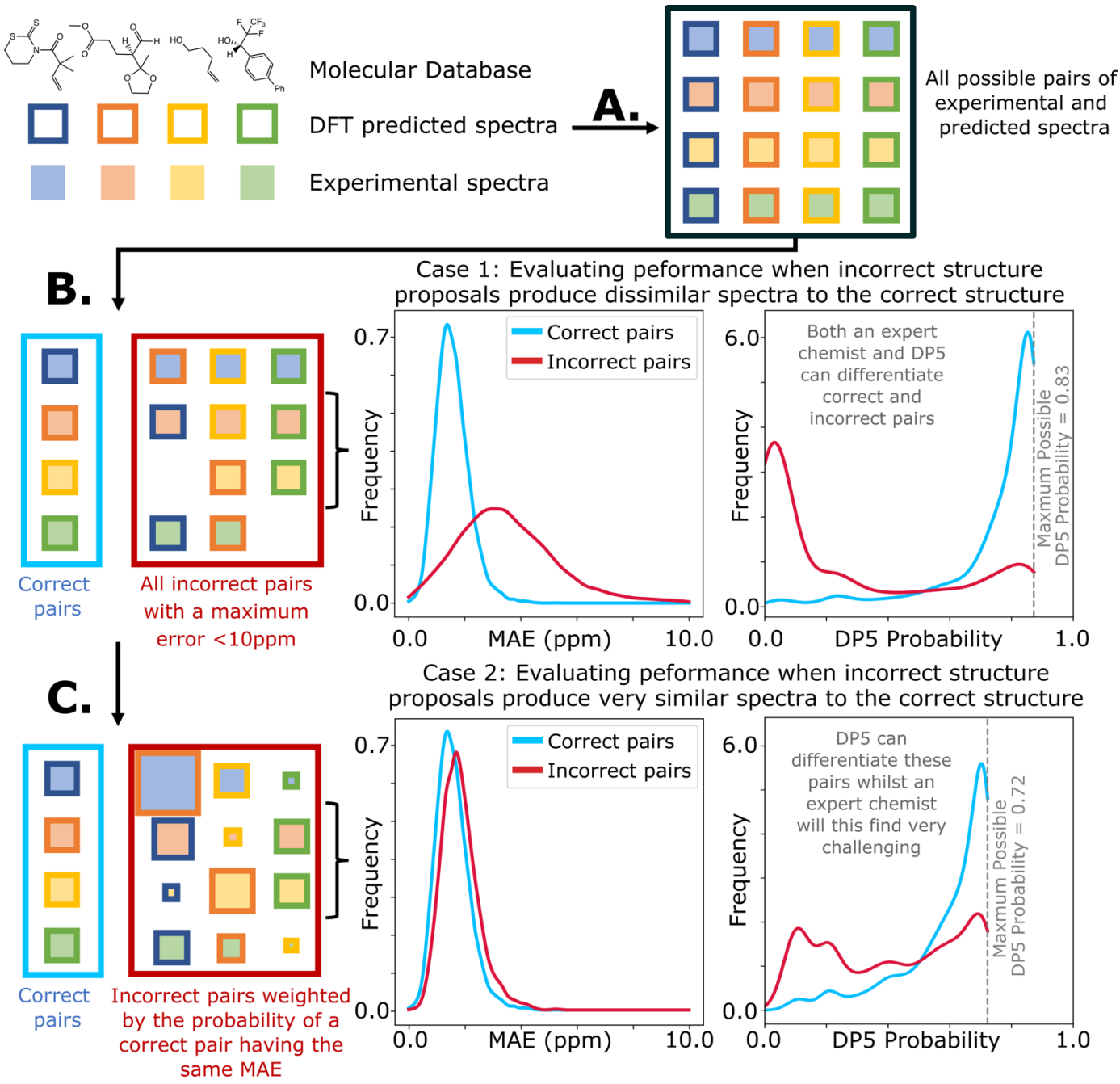

Figure 4. Schematic diagram of cross validation analysis used to evaluate the performance of DP5. A) The experimental spectra of the 5140 (n) molecules from the NMRShiftDB training set with the same number of carbon atoms are permuted to produce pairs of structures and experimental spectra. B) These pairs are separated into correct pairs, where the structure is paired to the correct spectrum and incorrect pairs where the molecule has been paired with a different spectrum. All incorrect pairs with max errors <10ppm are considered in case 1. C) In case 2 the incorrect pairs are assigned sampling weights to force the MAE distribution of the incorrect pairs to approximate that of the correct pairs, this leads to an expectation number of ~5330 incorrect combinations. All DP5 probabilities in this study are calculated using a leave-one-out scheme. (see supporting information section S3.2).

equally or more convincing than the correct structure, just as two molecules may produce similar experimental spectra. However, this is seldom a problem in organic chemistry as in most real-world applications, there are additional constraints on the potential structures that need to be considered. However, one can sometimes be 100% confident that a structure is incorrect. For example, in robot-controlled syntheses, the particular sequence of reactions is known limiting the potential products. In most cases, a DP5 probability of 73%, combined with this additional data, will give the chemist much higher levels of certainty their structure is correct. In cases where multiple structures give high DP5 probabilities for the same spectra, this is a good indication of where DP4 can be

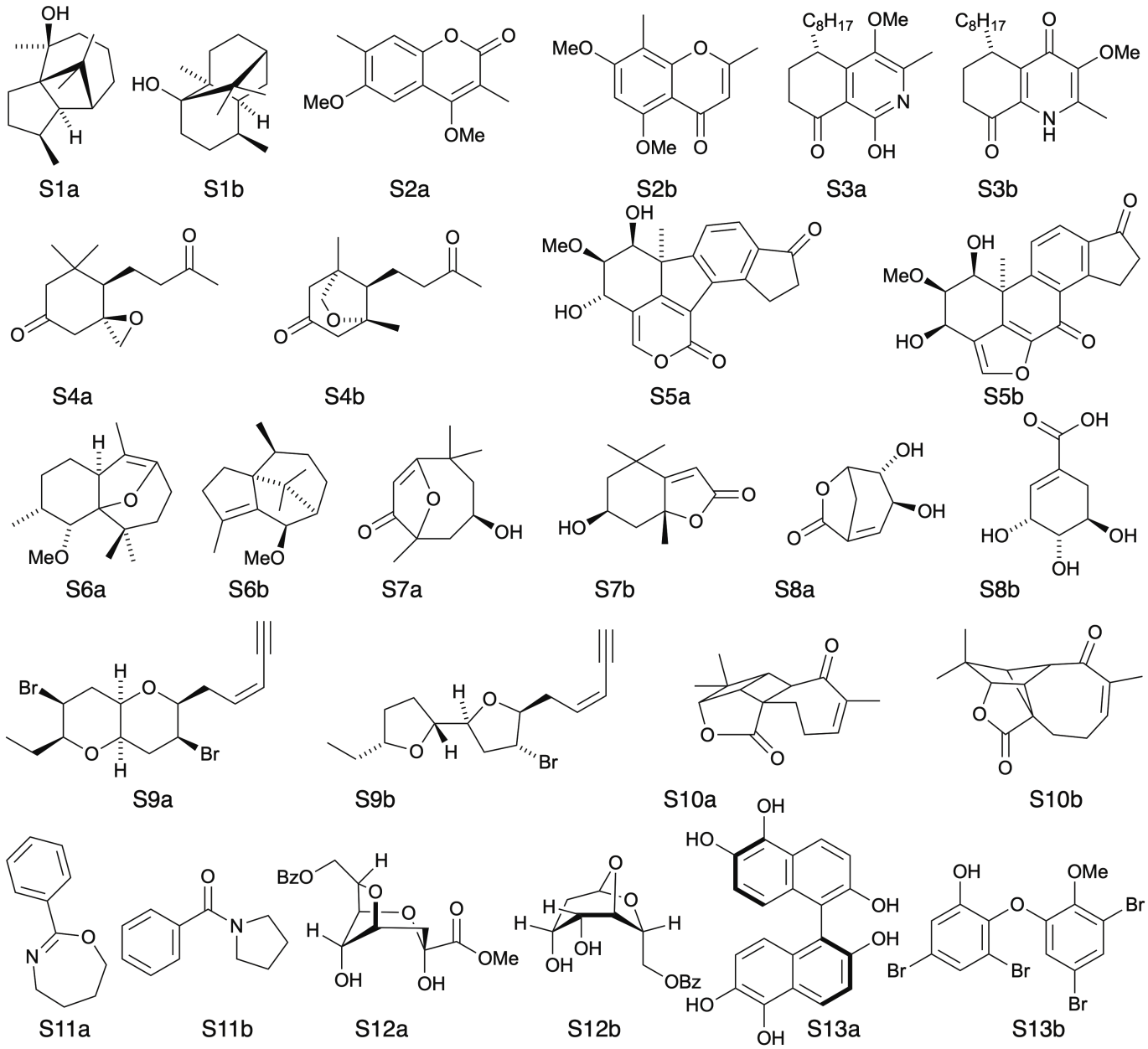

Figure 5. Test set of real-world structure reassignment problems taken from chemical literature. In each example an incorrect structure was initially published (S#a) which was later reassigned to the corresponding correct structure (S#b)

applied in conjunction with DP5 to give even more accurate relative probabilities.

To further test the efficacy of DP5 analysis, the system was evaluated against thirteen real-world examples of structures originally incorrectly published in the literature and later reassigned, these structures are presented in Figure 5.[1,58,61–65] The results of this study are striking and are presented in Figure 6. In all cases the DP5 probabilities of the incorrect structures are equal to or close to zero. This illustrates that DP5 can reliably pick out structures that are likely to be incorrect, and will suggest to the user in these cases that the proposed structures may require revision. On the other hand, DP5 typically assigns probabilities close to the maximum to the correct structures. This will inform the user these structures are likely to be correct. There are three examples where the DP5 probabilities of the correct and incorrect structure are both equal to zero (S12, S2, S13). This result is not surprising and does not show a weakness, but rather a distinct advantage of the DP5 probability over DP4. Being a single structure probability, DP5 is questioning the initial structure and the revision independently, if both structures are improbable DP5 can assign low probabilities to both. In these situations, when all the candidate structures are unlikely, DP4 probabilities must still sum to one and DP4 will typically randomly show overconfidence in one of the structures. This behaviour was clearly displayed when the analysis was repeated using DP4. Only in these three cases did DP4 assign any confidence to the incorrect structures and in two cases (S12 and S2) DP4 assigned the most confidence to the incorrect structures. The low DP5 probabilities for S8, S9 and S13, suggest DP4 may have also assigned these structures correctly by

chance. These results highlight the consequences of the underlying assumptions of the DP4 methodology. For DP4 probabilities to be reliable, the correct structure must be present in the list of candidates. When this is true, DP4 is often more accurate than DP5 as it is more sensitive to slight differences in NMR spectra and as more information is available within the calculation. This makes DP4 the perfect system when the correct structure is guaranteed to be in the list of proposals. However, in cases where none of the candidate structures may be correct, only the DP5 probability can reflect this and can be calculated to assess the reliability of the DP4 calculation.

By analysing the DFT-prediction errors observed for these examples (see supporting information section 7), we can gain insight into how the DP5 probabilities have been assigned. In examples such as S3 and S4, the correct structures show significantly smaller errors than the incorrect structures. In these situations a trained chemist would be able to distinguish these structures using the DFT-NMR prediction errors. These results show DP5 also reliably assigns probabilities to these structures confirming the findings of the first analysis of the molecules from NMRShiftdb. In examples such as S5 and S8, both the correct and incorrect structures display a number of atoms with large errors. In these cases, analysis of the DFT errors may allow a chemist to qualitatively decide which structure they think is correct. However, previous work in this field has shown that large errors occur frequently in correct structures, and as a result statistical distributions with wide tails must be used to describe DFT-NMR prediction errors accurately.[12,55] In addition, it is also very well known that the shapes of the DFT-NMR prediction error distributions vary in complex ways with the structure of the atomic environment.[55,66] As a result, it is often difficult to draw accurate and reliable conclusions about structure proposals based on the sizes of the prediction errors alone. In contrast, DP5 quantitatively analyses the DFT-NMR prediction errors whilst taking into account how the underlying distribution will be affected by the structure of the atomic environment in question. This behaviour is clearly displayed in the atomic probabilities calculated for these examples, atoms with larger errors sometimes display lower probabilities than those with smaller errors, and vice versa (see supporting information section 7). Finally, by utilising equations 3 and 4 to combine the atomic probabilities, the DP5 Probability reliably reflects the overall probability for the whole molecule in a way a simple error analysis cannot.

These results demonstrate some key behaviours of the DP5 Probability and in addition confirm with real world examples the conclusions made during the analysis of the NMRShiftdb molecules. These results also suggest utilising DP5 analysis may have prevented these incorrectly assigned structures from being published.

To confirm the finding in the second stage of the analysis of the NMRShiftdb molecules that DP5 can reliably differentiate correct and incorrect structure proposals indistinguishable by their errors, DP5 was evaluated against a set of 42 real world relative stereochemistry elucidation problems originally used to test DP4-AI. (See supporting information section 5.1).[31]

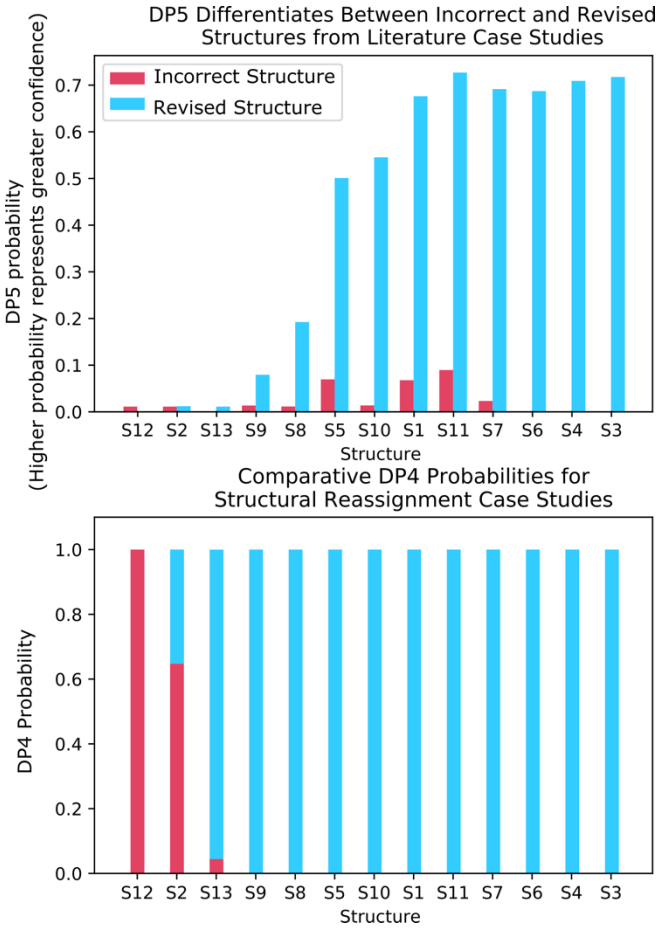

Figure 6. (top) DP5 probabilities calculated for the thirteen incorrectly published structures and corresponding revised structures. DP5 assigns much greater confidence to the revised structures and also displays the three cases where both the initially proposed and revised structures are equally improbable. (bottom) DP4 probabilities calculated for the same thirteen examples. These results show how the DP5 probability can be used to test the reliability of a DP4 calculation, as only DP5 can discern if any of the structure proposals are likely to be correct.

Relative stereochemistry elucidation is a particular challenge as the average difference between spectra of diastereomers (~1-2ppm) is similar to the average error in state-of-the-art DFT calculations. As in the second stage of the analysis of the molecules from NMRShiftdb, this results in diastereomers being extremely challenging to distinguish even by an expert chemist using traditional analysis with the DFT errors alone. This test is significantly different to those already presented. In the previous examples, a single structure is proposed for a single spectra. In contrast, when resolving relative stereochemistry, multiple candidate structures (diastereomers) are proposed for the same spectra and the correct structure is assumed to be one of the candidates. This is the situation for which DP4 was designed; DP5 does not make this assumption. Despite this, Figure 7 shows the performance of DP5 and DP4 is similar in this

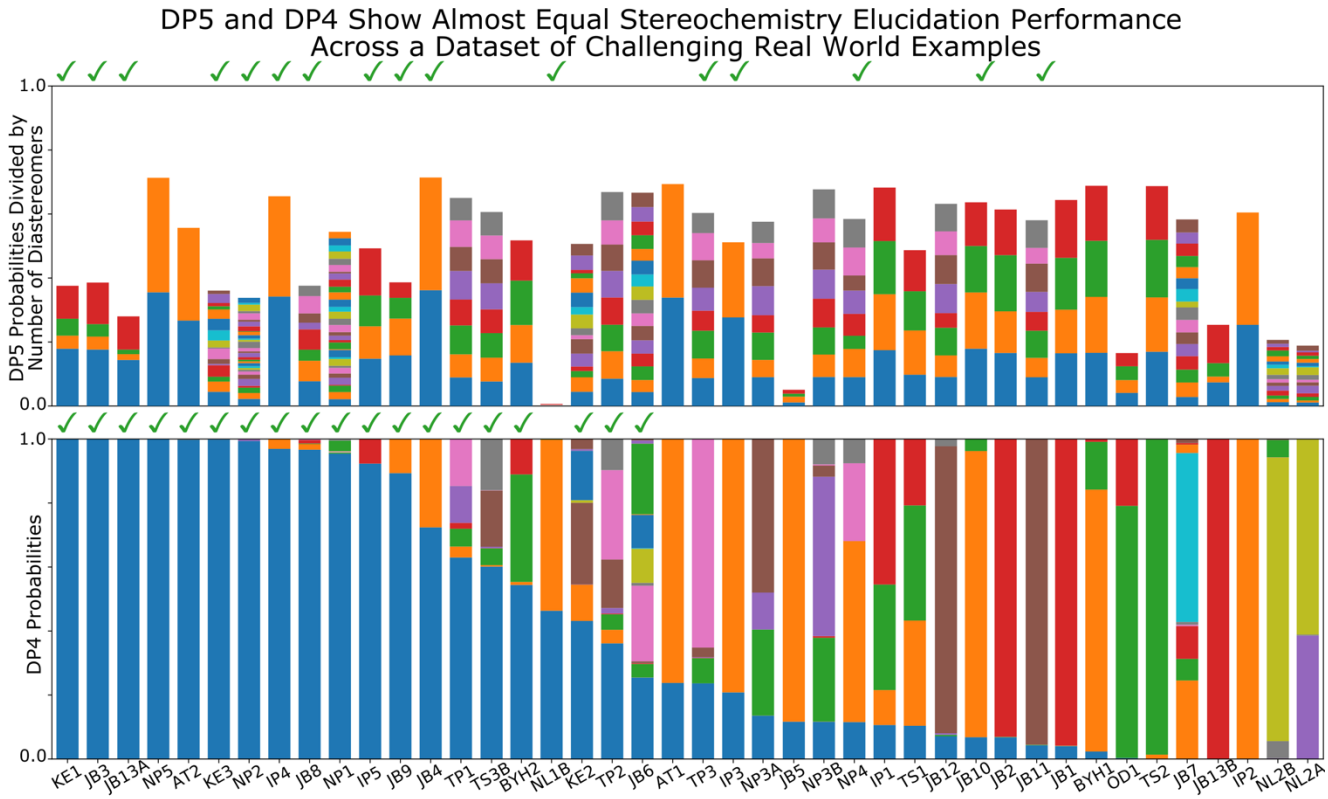

Figure 7. Results of DP5 (top) and DP4 (bottom) calculations on a dataset of 42 challenging real world stereochemistry elucidation examples (see supporting information section 5.1 for structures). In both plots probabilities calculated for each diastereomer are stacked in the same order with matching colours, the correct diastereomer is always represented by the blue bar at the bottom of the stack. The checkmarks above each plot indicate molecules correctly assigned by each program. The DP5 probabilities have been divided by the number of diastereomers for each molecule, this ensures the total sum of these probabilities is within the 0-1 range (see supporting information section 5.1 for unnormalized results). These results show the two systems display similar stereochemistry elucidation performance, DP4 assigns 19 molecules correctly, whilst DP5 assigns 16 correctly, the probabilities of assigning as many

regime: DP4 assigns 19 molecules correctly whilst DP5 assigns 16 correctly, the probabilities of assigning as many molecules in this dataset correctly by chance are ~0.0001 and ~0.01 respectively, showing the significance of these results. Interestingly, whilst the number of molecules correctly assigned by the two systems is very similar, the behaviours of the two systems are distinct. DP5 typically assigns similar probabilities to all of the diastereomers, whilst DP4 often shows greater confidence in a smaller number. These behaviours clearly reflect the questions DP5 and DP4 have been designed to answer. DP5 is individually comparing each diastereomer to chemical space. DP5 assigns similar probabilities to diastereomers as they are typically more similar to each other than they are to other molecules across chemical space. This can be seen when comparing the MAEs between spectra of random molecules, ~50-100ppm with those seen for diastereomers, typically ~1-2ppm. In contrast, DP4 is directly comparing the diastereomers against each other, leading to a greater variation in the probabilities. This is often beneficial for DP4 as it always assumes the correct structure is in the list of proposals, whilst DP5 does not. These results again highlight the DP5 gives reliable overall probabilities without this assumption, while DP4 shows additional sensitivity based on this assumption. In the second stage of the analysis of the molecules from NMRShiftdb, DP5 was shown to typically assign much lower probabilities to incorrect structure proposals than to correct structures even when these displayed similar MAEs (~2ppm). In contrast, in the stereochemical examples, whilst DP5 again assigns a significant number of the molecules correctly, the calculated probabilities are more similar. The key observation here is that whilst differences between the spectra of the correct and incorrect structures are similarly small in these two situations, in the case of diastereomers, the structures of the proposals are also more similar. Therefore DP5 should be expected to assign similar probabilities to diastereomers. This illustrates a powerful behaviour of DP5, as the proposals become more similar to each other in structure, the probabilities DP5 assigns also become more similar. This behaviour is more useful than the alternative of assigning a probability of 1.0 to correct structure proposals and zero to everything else, as this allows the user to assess if their proposals are more or less similar to the true structure. This also clearly demonstrates that DP5 Probability can be interpreted as a probability. If two structures are very similar and yield similar spectra, their probabilities of being correct when considering DFT-prediction errors only should necessarily be similar. In

certain situations the reliability of the DP4 Probability can be further increased, for example when the user has developed a custom statistical model for working with a particular class of molecules or when multiple sets of spectral data are available.
These results also demonstrate how DP4 and DP5 can and should be used together in cases where the user knows their proposal has multiple stereocentres. In examples: JB5, OD1, NL1B, NL2A and NL2B, where the DP5 probabilities for all the diastereomers are low (the total height of each bar in DP5 results in Figure 7 can be interpreted as the average confidence the diastereomers) this suggests that the DP4 probabilities are likely to be less reliable. There are a number of reasons why the DP5 probability may be low in this way, for example, the correct structure may be missing from the list, there may be a problem with the NMR assignment or there may be an impurity in the spectrum. The results of this test are very exciting and the application of the DP5 probability in situations with multiple structure proposals is now being more thoroughly explored.
These examples show how DP5 can serve as a valuable tool whenever a new molecule is made, increasing confidence when proposed structures are correct, highlighting cases where they are not and also playing a stern jury when an improbable but correct structure has been proposed.

Conclusions

In conclusion, we have developed a new measure to quantify molecular structural uncertainty, the DP5 probability. This work represents a leap forward in quantification of structural uncertainty as instead of a comparative dimensionless parameter, DP5 quantifies the probability of a structure being correct. This system was rigorously evaluated by a cross validation study and it was found that DP5 could perform as well as a human in classifying correct and incorrect structure proposals and in some cases could classify structures indistinguishable to a chemist. DP5 was evaluated against thirteen real-world examples of structures that were incorrectly published and subsequently revised in the literature. In all these challenging cases, DP5 expressed the maximum concern for the incorrect structures and was on average 41% more confident in the revised structures. DP5 was finally evaluated against 42 real world stereochemistry elucidation examples, displaying almost equal performance to DP4. The DP5 probability can be calculated fully automatically and so should find wide applications in uses cases such as high throughput reaction screening, automated chemical synthesis and drug discovery. In addition, DP5 may be run on a single molecule basis and the results explored utilizing the GUI, helping to guide the development of complex syntheses. This work also suggests how DP5 may be developed to help further to accelerate chemical discovery. The DP5 probability has been evaluated here with ^{13}C NMR data, DP5 is currently being extend to utilise different types of spectral data, and a DFT free version of DP5 is also being explored. In addition, utilizing DP5 alongside generative models and other machine learning methods to automatically guide structure determination is an attractive possibility. The DP5 system is available as open-source software at https://github.com/Goodman-lab/DP5

Acknowledgements

We thank EPSRC (A.H.) for financial support. We thank Dr Kristaps Ermanis for his advice and guidance on this project in addition to his continued work maintaining the DP4-AI repository.

Conflicts of interest

There are no conflicts to declare.

Notes and references

1 K. C. Nicolaou and S. A. Snyder, *Angew. Chemie - Int. Ed.*, 2005, **44**, 1012–1044.

2 K. C. Nicolaou, H. Zhang and A. Ortiz, *Angew. Chemie Int. Ed.*, 2009, **48**, 5642–5647.

3 K. C. Nicolaou, A. Ortiz and H. Zhang, *Angew. Chem. Int. Ed. Engl.*, 2009, **48**, 5648–52.

4 A. G. Kutateladze and T. Holt, *J. Org. Chem*, 2019, **84**, 51.

5 A. V. Buevich and M. E. Elyashberg, *J. Nat. Prod.*, 2016, **79**, 3105–3116.

6 P. Kessler and M. Godejohann, *Magn. Reson. Chem.*, 2018, **56**, 480–492.

7 J.-M. Nuzillard and B. Plainchont, *Magn. Reson. Chem.*, 2018, **56**, 458–468.

8 G. Barone, D. Duca, A. Silvestri, L. Gomez-Paloma, R. Riccio and G. Bifulco, *Chem. - A Eur. J.*, 2002, **8**, 3240.

9 G. Barone, L. Gomez-Paloma, D. Duca, A. Silvestri, R. Riccio and G. Bifulco, *Chem. - A Eur. J.*, 2002, **8**, 3233.

10 G. Lauro and G. Bifulco, *European J. Org. Chem.*, 2020, **2020**, 3929–3941.

11 C. S. Kim, J. Oh and T. H. Lee, *Arch. Pharm. Res.*, 2020, 43, 1114–1127.

12 S. G. Smith and J. M. Goodman, *J. Am. Chem. Soc.*, 2010, **132**, 12946–12959.

13 N. Grimblat, M. M. Zanardi and A. M. Sarotti, *J. Org. Chem.*, 2015, **80**, 12526–12534.

14 N. Grimblat, J. A. Gavín, A. Hernández Daranas and A. M. Sarotti, *Org. Lett.*, 2019, **21**, 4003–4007.

15 L. A. Maslovskaya, A. I. Savchenko, E. H. Krenske,

S. Chow, T. Holt, V. A. Gordon, P. W. Reddell, C. J. Pierce, P. G. Parsons, G. M. Boyle, A. G. Kutateladze and C. M. Williams, *Chem. – A Eur. J.*, 2020, **26**, 11862–11867.

16 I. E. Ndukwe, X. Wang, N. Y. S. Lam, K. Ermanis, K. L. Alexander, M. J. Bertin, G. E. Martin, G. Muir, I. Paterson, R. Britton, J. M. Goodman, E. J. N. Helfrich, J. Piel, W. H. Gerwick and R. T. Williamson, *Chem. Commun.*, 2020, **56**, 7565–7568.

17 S. R. Lee, D. Lee, M. Park, J. C. Lee, H. J. Park, K. S. Kang, C. E. Kim, C. Beemelmanns and K. H. Kim, *J. Nat. Prod.*, 2020, **83**, 354–361.

18 J. S. An, J. Y. Lee, E. Kim, H. Ahn, Y. J. Jang, B. Shin, S. Hwang, J. Shin, Y. J. Yoon, S. K. Lee and D. C. Oh, *J. Nat. Prod.*, 2020, **83**, 2776–2784.

19 P. Yan, G. Li, C. Wang, J. Wu, Z. Sun, G. E. Martin, X. Wang, M. Reibarkh, J. Saurí and K. R. Gustafson, *Org. Lett.*, 2019, **21**, 7577–7581.

20 A. M. Sarotti, *Org. Biomol. Chem.*, 2013, **11**, 4847–4859.

21 S. Kuhn and N. E. Schlörer, *Magn. Reson. Chem.*, 2015, **53**, 582–589.

22 Y. Guan, S. V. S. Sowndarya, L. C. Gallegos, P. C. St. John and R. S. Paton, *Chem. Sci.*, , DOI:10.1039/D1SC03343C.

23 O. A. von Lilienfeld, *Int. J. Quantum Chem.*, 2013, **113**, 1676–1689.

24 B. Huang and O. A. Von Lilienfeld, *J. Chem. Phys.*, 2016, **145**, 161102.

25 F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke and K. R. Müller, *Nat. Commun.*, 2017, **8**, 1–10.

26 F. A. Faber, A. S. Christensen, B. Huang and O. A. Von Lilienfeld, *J. Chem. Phys.*, 2018, **148**, 241717.

27 A. S. Christensen, L. A. Bratholm, F. A. Faber and O. Anatole Von Lilienfeld, *J. Chem. Phys.*, 2020, **152**, 044107.

28 J. Townsend, C. P. Micucci, J. H. Hymel, V. Maroulas and K. D. Vogiatzis, *Nat. Commun.*, 2020, **11**, 1–9.

29 K. Ermanis, K. E. B. Parkes, T. Agback and J. M. Goodman, *Org. Biomol. Chem.*, 2016, **14**, 3943–3949.

30 K. Ermanis, K. E. B. Parkes, T. Agback and J. M. Goodman, *Org. Biomol. Chem.*, 2017, **15**, 8998–9007.

31 A. Howarth, K. Ermanis and J. M. Goodman, *Chem. Sci.*, 2020, **11**, 4351–4359.

32 2009 Schrodinger, LLC, New York, NY, .

33 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 490–519.

34 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 520–552.

35 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 553–586.

36 T. A. Halgren and R. B. Nachbar, *J. Comput. Chem.*, 1996, **17**, 587–615.

37 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 616–641.

38 T. A. Halgren, *J. Comput. Chem.*, 1999, **20**, 720–729.

39 I. Kolossváry and W. C. Guida, *J. Comput. Chem.*, 1999, **20**, 1671–1684.

40 I. Kolossváry and W. C. Guida, *J. Am. Chem. Soc.*, 1996, **118**, 5011–5019.

41 Gaussian 09, Revision A.02, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.

42 F. London, *J. Phys. le Radium*, 1937, **8**, 397–409.

43 K. Wolinski, J. F. Hinton and P. Pulay, *J. Am. Chem. Soc.*, 1990, **112**, 8251–8260.

44 R. Ditchfield, *J. Chem. Phys.*, 1972, **56**, 5688–5691.

45 C. Adamo and V. Barone, *J. Chem. Phys.*, 1998, **108**, 664–675.

46 W. J. Hehre, K. Ditchfield and J. A. Pople, *J. Chem. Phys.*, 1972, **56**, 2257–2261.

47 P. C. Hariharan and J. A. Pople, *Theor. Chim. Acta*, 1973, **28**, 213–222.

48 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297.

49 F. Weigend, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1057.

50 E. Cancès, B. Mennucci and J. Tomasi, *J. Chem. Phys.*, 1997, **107**, 3032–3041.

51 A. D. Becke, *Phys. Rev. A*, 1988, **38**, 3098–3100.

52 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785–789.

53 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2008, **120**, 215–241.

54 K. R. M. O. A. von L. A.S. Christensen, F. A. Faber, B. Huang, L.A. Bratholm, A. Tkatchenko, *GitHub Repos.*

55 K. Ermanis, K. E. B. Parkes, T. Agback and J. M. Goodman, *Org. Biomol. Chem.*, 2019, **17**, 5886–5890.

56 J. Li, J. K. Liu and W. X. Wang, *J. Org. Chem.*, 2020, **85**, 11350–11358.

57 M. Rupp, R. Ramakrishnan and O. A. Von Lilienfeld, *J. Phys. Chem. Lett.*, 2015, **6**, 3309–3313.

58 K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2015, **6**, 2326–2331.

59 B. Huang and O. A. von Lilienfeld, *Nat. Chem.*, 2017, **12**, 945–951.

60 W. Gerrard, L. A. Bratholm, M. J. Packer, A. J. Mulholland, D. R. Glowacki and C. P. Butts, *Chem. Sci.*, 2020, **11**, 508–515.

61 A. G. Kutateladze, E. H. Krenske and C. M. Williams, *Angew. Chemie*, 2019, **131**, 7181–7186.

62 B. S. Dyson, J. W. Burton, T. I. Sohn, B. Kim, H. Bae and D. Kim, *J. Am. Chem. Soc.*, 2012, **134**, 11781–11790.

63 E. E. Podlesny and M. C. Kozlowski, *J. Nat. Prod.*, 2012, **75**, 1125–1129.

64 B. Verbraeken, J. Hullaert, J. van Guyse, K. Van Hecke, J. Winne and R. Hoogenboom, *J. Am. Chem. Soc.*, 2018, **140**, 17404–17408.

65 M. W. Lodewyk, C. Soldi, P. B. Jones, M. M. Olmstead, J. Rita, J. T. Shaw and D. J. Tantillo, *J. Am. Chem. Soc.*, 2012, **134**, 18550–18553.

66 K. G. Andrews and A. C. Spivey, *J. Org. Chem.*, 2013, **78**, 22, 11302–11317