ARTICLE

Towards automatically verifying chemical structures: the powerful combination of ^{1}H NMR and IR spectroscopy †

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

DOI: 10.1039/x0xx00000x

J. Benji Rowlands,[b] Lina Jonsson,[a] Jonathan M. Goodman,[b] Peter W. A. Howe,[c] Werngard Czechtizky,[a] Tomas Leek[a] and Richard J. Lewis*[a]

Human interpretation of spectroscopic data remains key to confirming newly synthesised chemical structures. Whilst there have been advances in automated spectral interpretation, the false positive and false negative rates remain too high to replace human interpretation. One approach, Automated Structure Verification (ASV), scores observed nuclear magnetic resonance (NMR) spectra against predicted NMR spectra. We describe a method to extend this approach to infrared (IR) spectra and apply it alongside proton NMR spectra to distinguish between a challenging set of 99 similar isomer pairs. Based on relative scores, we classify each as correct, incorrect or unsolved. Our results show that IR can be used as an efficient automated method to distinguish similar isomers with an accuracy close to that of proton NMR. We further introduce a method to combine NMR and IR results and show that the combination significantly outperforms either technique alone. At a true positive rate of 90%, unsolved pairs are reduced to 0-15% using NMR and IR together compared to 27-49% using individual techniques alone. At a true positive rate of 95%, they are reduced to 15-30% from 39-70%. These results are a significant step towards efficient automated structure verification based on easily measured spectroscopy data.

Introduction

Identifying and verifying molecular structures is key to organic, synthetic, and medicinal chemistry. NMR spectroscopy is by far the most widely used method for structure elucidation.[1] This is owing to the wealth of information that NMR spectra provide about a molecule and because the spectral information content follows rules that link directly to specific features of a molecule. Improved automated methods of confirming new structures are needed to match the increasing speed and throughput of organic synthesis. Automated methods for interpreting NMR spectra fall broadly into two categories: Automated Structure Verification (ASV)[2,3] and Computer-Assisted Structure Elucidation (CASE).[4,5] The former tests candidate structures against experimental data whereas the latter approach generates the structure from the analytical data alone. The ASV approach uses less data but relies more on non-analytical information (for example the list of candidate structures proposed from knowledge of the synthetic route). The two approaches can be thought of as lying on a continuum with the same aim -– to provide the user with a single structure with a high probability of being correct. In the context of synthetic chemistry, ASV can be thought of as a similar process to a chemist running a well characterized reaction and relying on a ^{1}H NMR spectrum to confirm the product.

Among the best-established methods in the ASV category are the DP4 and DP5 probabilities.[6,7] These methods involve using density functional theory (DFT) to calculate NMR chemical shifts for each molecule in a list of candidates supplied by the user. The probability of each molecule being correct is determined via an analysis of the observed differences between the experimental and calculated chemical shifts. The DP4 probability and derivatives are regularly used to assist with structure elucidation in challenging cases.[8–11]

A related area of current research is the application of machine learning to automated structure elucidation.[12–15] These methods have the potential to be much faster than methods involving DFT calculations but require training with large amounts of (often simulated) data. Whilst early results are promising, it remains to be seen what impact machine learning will have.

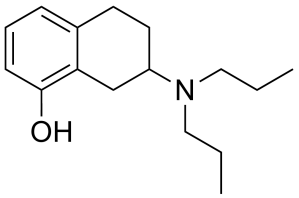
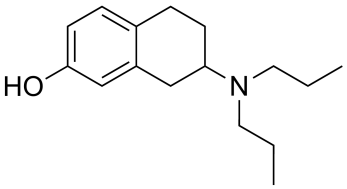
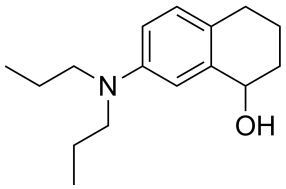
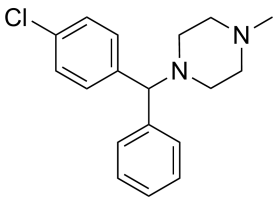
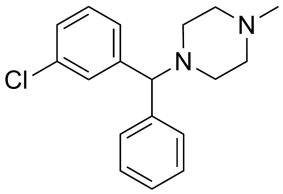
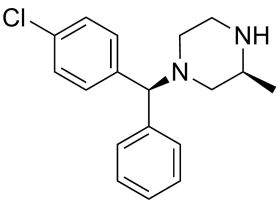
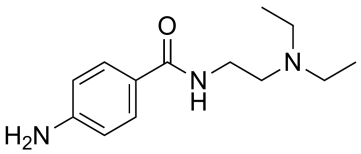
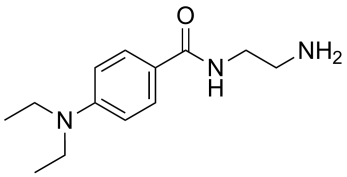
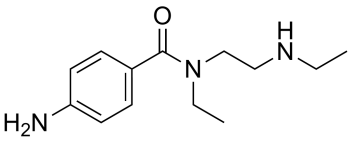
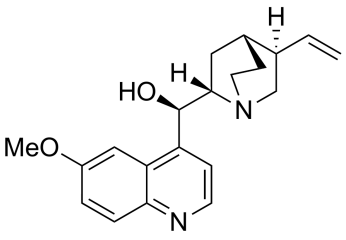
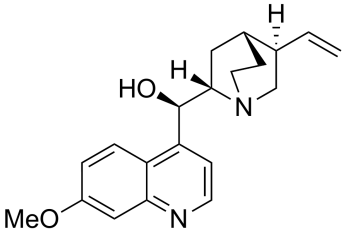
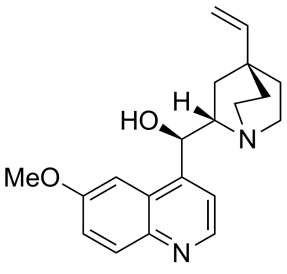
Previous work has focussed on applying automated interpretation methods to NMR data.[16–18] But these may also be applied to data which a human cannot easily interpret. IR would seem a particularly suitable technique to apply to structure determination. From a practical viewpoint, IR spectra can be collected quickly with sub-milligram amounts of material. From an information viewpoint, IR spectra originate in bond vibrations, including of bonds involving atoms not observed by NMR. Some absorptions, especially carbonyl absorptions, provide specific information about functional groups, but most of the spectrum (the fingerprint) cannot be easily related to specific functional groups. A complete structure cannot be built

[a] Department of Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (R&I), Biopharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden. Email: richard.j.lewis@astrazeneca.com
[b] Centre for Molecular Informatics, Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, CB2 1EW, U.K.
[c] Oncology Chemistry, AstraZeneca, 1 Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0AA, U.K.

†Electronic supplementary Information (ESI) available. See DOI: 10.1039/x0xx00000x

from an IR spectrum by following a set of simple rules, however the fingerprint can be matched against a calculated spectrum obtained from a proposed structure. In contrast, NMR spectra

Table 1: Four example molecules from the test set used to evaluate the use of IR and NMR in ASV.

Compound	Correct structure	Incorrect isomer 1	Incorrect isomer 2
1			
		Aromatic regioisomer	Unclassified regioisomer
2			
		Aromatic regioisomer	Unclassified regioisomer
3			
		Unclassified regioisomer	Unclassified regioisomer
4			
		Aromatic regioisomer	Aliphatic regioisomer

provide atom-focussed information because the chemical shift is dominated by relatively short-range effects such as hybridization, covalent structure and the electronegativity of neighbouring groups. Given the difference in the origins of the information in NMR and IR we might expect them to provide complementary information about molecular structure. ASV methods which combine such complementary information have previously been proposed for 1D ^{1}H NMR and 2D ^{1}H-^{13}C HSQC spectra.[19] Separately, the use of IR for structure determination has been reported recently for building up molecules in a

fragment-based approach,[20] determining regio- and stereochemistry by matching experimental and calculated IR spectra,[21] in machine-learning models that can generate complete structures by supplementing NMR data with functional group identification from IR[22] or from IR data alone.[23,24]

A key but challenging use for ASV is to correctly determine reaction products. For this application, we assume that the molecular weight can be determined in a straightforward manner, and the challenge is to distinguish between two or more similar regio- or stereo-isomeric products. The similarity of the potential products presents a challenge in distinguishing them by traditional scoring methods because the spectra are usually similar, leading to similar scores which cannot be distinguished by a binary correct/incorrect classification.

In this ASV proposal, we focus on testing two hypotheses. Firstly, that comparing the scores of candidate structures is a more robust approach to ASV than scoring a single compound in isolation, and secondly that IR and proton NMR chemical shifts contain complementary structural information so that ASV benefits from the combination of both. The list of candidate structures could be generated by reaction prediction software,[25] a fast developing field. The proposal overall mimics the workflow of a synthetic chemist who uses knowledge of the possible reaction products to assess against the analytical data collected.

Based on these assumptions, we propose and assess a method for ASV comparing the scores of candidate structures using a combination of ^{1}H NMR chemical shifts and IR data. We introduce an algorithm (IR.Cai) to match and score experimental and calculated IR spectra. For NMR data analysis, we modify the peak-matching algorithm of DP4 to automatically exclude outlying shifts from the analysis, circumventing the unpredictability of the chemical shifts of exchangeable protons. Such peaks are sometimes highlighted with an asterisk, so we call this modified version DP4*. We found this modification to be necessary to obtain chemically reasonable results for molecules with labile protons. We also analyse ^{1}H NMR using a commercial ASV software package (ACD/Labs).

Using a set of highly similar isomeric test structures, we test the hypothesis that the candidate structure that scores highest (by IR, NMR or by a combination of the scores) is more likely to be correct, and that the larger the score difference, the greater the probability that the highest scoring structure is correct. Any pair of candidate structures that is not differentiated by a sufficiently large score naturally is left unsolved indicating that more data or manual interpretation is required.

We also test the hypothesis that IR and NMR chemical shifts contain complementary information. As an example, consider compound 2 and the incorrect isomers 1 & 2 (Table 1). The DP4* scores are 0.53, 0.47 and 0 and the IR.Cai scores 0.74, 0.65 and 0.62 respectively. The DP4* results exclude incorrect isomer 2 by virtue of the value of zero but only show a slight preference for the correct structure over incorrect isomer 1 (0.53 vs 0.47). Conversely the IR.Cai results do not exclude incorrect isomer 2 as strongly as DP4* but show a similar small preference for the correct structure over incorrect isomer 1. Given that IR and NMR are independent methods, using both gives greater confidence that the correct structure has been identified compared to NMR or IR alone.

Results and discussion

Verifying one of a choice of structures

Verifying a single structure as correct or incorrect, without any context or additional information is not straightforward, nor representative of everyday tasks for most chemists. In most real-world structural elucidation problems, there is additional information to take advantage of. For example, knowledge of the structures of the starting materials significantly narrows the range of possible products in a predictable way. Therefore, comparing or scoring alternative products against analytical data (the ASV approach) is not only a simpler and easier task, but also closer to what a chemist would do in practice. Furthermore, by scoring and comparing similar compounds systematic errors may partially cancel (for example, inaccurate calculation of spectroscopic data).[26]

Dataset of Test Compounds

To evaluate the performance of different methods for structure verification, a dataset of 42 drug-like molecules with molecular weights between 182 and 430 (average of 300) was assembled. For each molecule, two or three isomers representing a range of reasonable transformations were constructed manually. The isomers were chosen to include a range of transformations including changes in stereochemistry (~10%) changes in aromatic (~35%) or aliphatic (~25%) regiochemistry and changes in heteroatom position (~10%). The result of these changes was to enrich the test set with highly similar isomeric structures, which would be expected to give similar NMR chemical shifts. These were arranged into 99 pairs of the correct molecule structure and one incorrect isomeric structure.

Table 1 shows a sample of molecules and their incorrect isomers from the dataset. The full list can be found in the Supporting Information (Table S1).

Receiver Operating Characteristic (ROC)

The IR.Cai matching algorithm, DP4* and the ^{1}H ASV tool from Advanced Chemistry Development, Inc. (ACD/Labs)[27] give numbers between 0-1 related to how well the experimental spectrum matches the calculated one. As our base comparison, we investigated the performance of DP4*, the ACD proton ASV tool and our IR algorithm as binary classifiers – that is to set a threshold and classify a molecule "correct" if its score is above the threshold, and "incorrect" if not. Note that the similarity of the incorrect isomers to the correct compounds makes this a particularly challenging task. Figure 1 shows the ROC curves for ACD, DP4* and IR. As anticipated due to the similarity of correct and incorrect structures it is difficult to distinguish between them using a binary classifier method. The area under the curve (AUC) is a metric for how well the method is working, with a random process scoring 0.5 and a perfect method 1.0. The AUC values at best are only halfway between a random and perfect

method with a true positive rate of, for example, 0.8 coming at the expense of a false positive rate of 0.5-0.6. This is too inaccurate for routine automatic use.

Structure Classification Characteristic (SCC)

Faced with the poor performance revealed by the ROC plots above, we focused on the difference in scores between candidate structures. For DP4* considering a list of candidate structure is already part of the process, but this is not the case for ACD or IR.Cai. For these metrics, our hypothesis was that the isomer of each test pair which scores higher is more likely to be correct, and that the score difference relates to the confidence level. To reach an acceptable level of confidence in the higher-scoring isomer, we evaluated results at different thresholds of the score difference. The pair was deemed as *unsolved* if the difference was lower than the chosen threshold. If the difference was above the threshold, then the pair was *correctly classified* (True Positive) if the correct compound scored higher, otherwise it was *incorrectly classified* (False Positive). As expected, a trade-off is seen between the true positive rate and the unsolved rate. Thus, if we wanted to be correct 95% of the time, we would choose a higher threshold which would result in a higher proportion of compound pairs unsolved. If our requirement was to correctly classify 80% of the time, we would choose a lower threshold and expect to classify more compound pairs. In a real-world scenario this choice would be dictated by the costs of being wrong about a structure and the resources required to manually evaluate unclassified pairs. Figure 2a illustrates the process, taking as an example Compound 1 and its first incorrect isomer from the test set. For this specific molecule, it is challenging to distinguish between the correct structure and its incorrect regioisomer using ^{1}H NMR shifts alone. This is therefore a good illustration of the benefits of using IR data in addition to NMR data.

To examine the trade-off, we adapted the "Receiver Operating Characteristic" (ROC) visualization to include the *unsolved* category. We call this visualization the *Structure Classification Characteristic* (SCC). The curve is formed by plotting the two key indicators of performance (the true positive rate and the

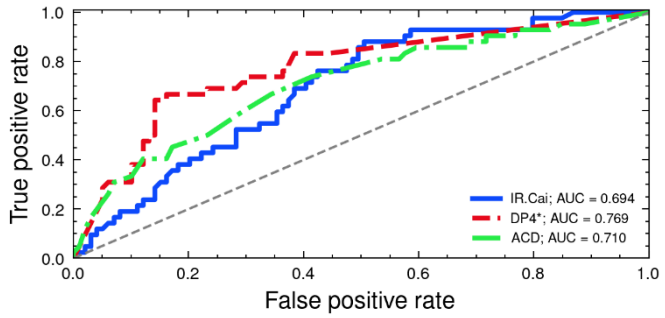
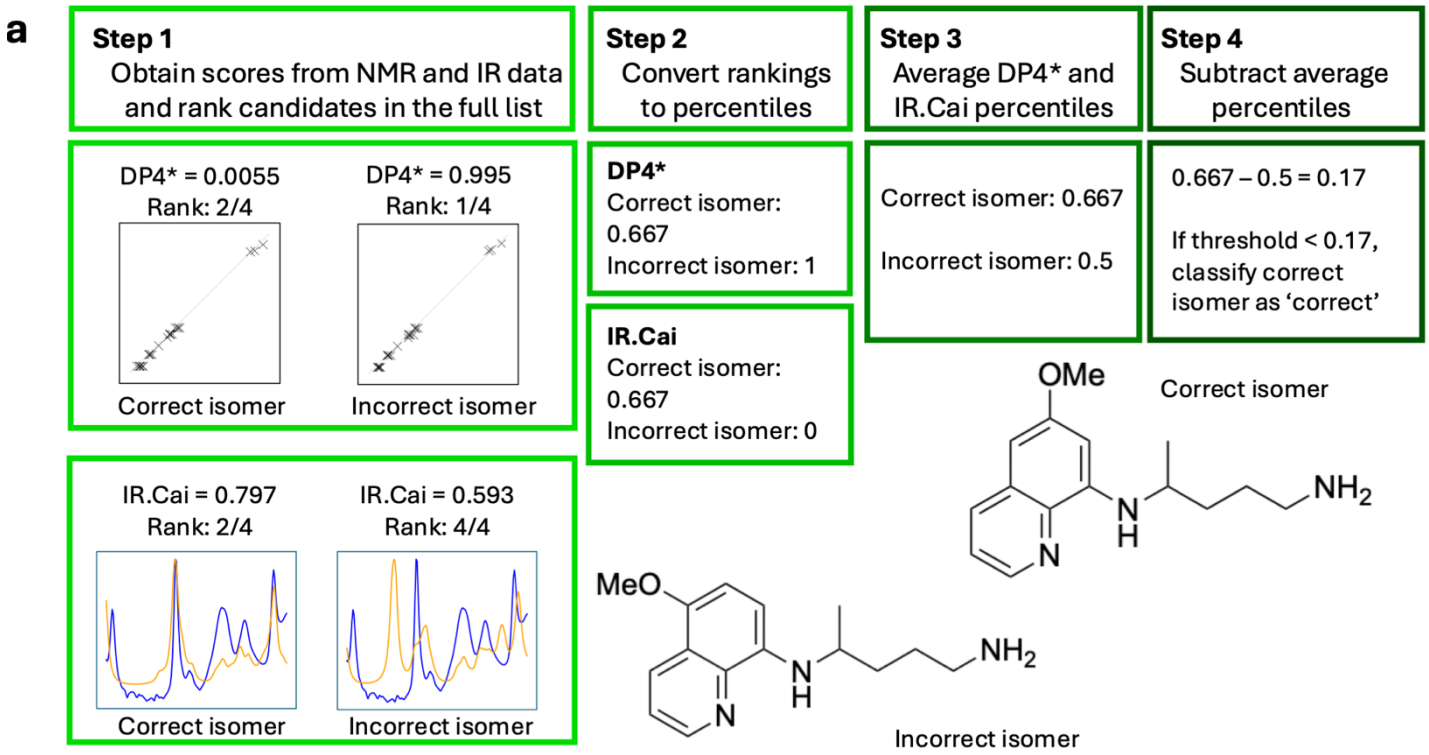


Figure 1: ROC plot based on the 42 molecules in the test set showing the performance of DP4* and IR.Cai for classifying structures as correct or incorrect. Note that the DP4* metric is comparative, i.e. it requires a list of possible candidate structures and scores them relative to one another, unlike IR.Cai and ACD which generate single molecule scores. Hence the result for DP4* is not directly comparable to the results for IR.Cai and ACD.



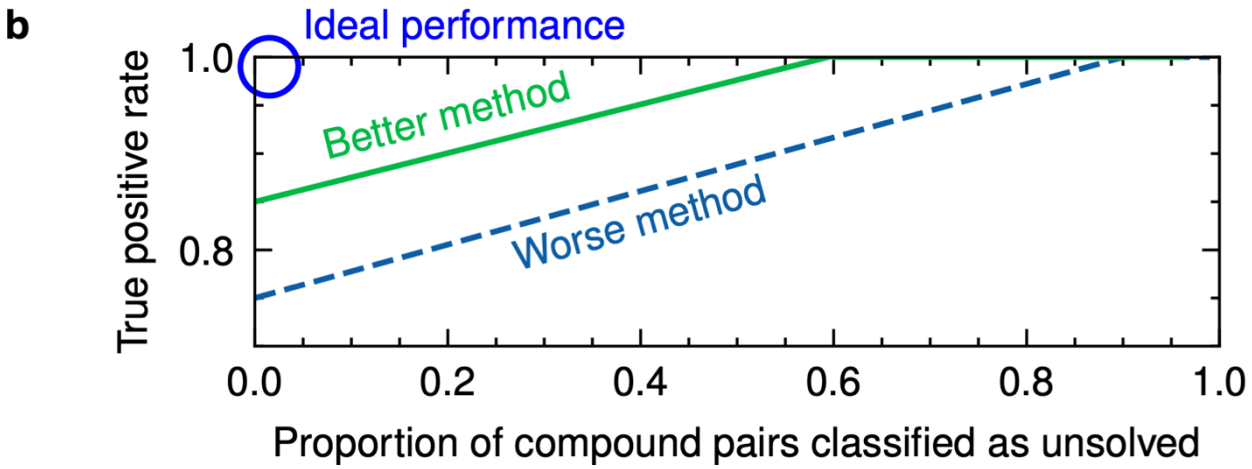


Figure 2: **a** Scheme showing how DP4* and IR.Cai scores are used to classify a pair of isomers, where one of them is known to be correct. Information from the DP4* and IR.Cai scores is aggregated using the average percentile-rank method described in the text. **b** Illustration of the structure comparison characteristic (SCC) plot. The green line is more useful for structure classification than the blue line, as the green method can correctly classify a higher proportion of molecules. The ideal result would be a point in the top left corner of the plot, denoted by the blue circle. This ideal SCC curve would have a CA (classification area) of 1.

proportion of compound pairs unsolved, both a function of the threshold) against each other. The ideal performance is a point in the top left, where no compound pairs are unsolved, and the correct structure always scores highest (unsolved proportion is 0 and true positive rate is 1). The performance of a particular method can be compared to others according to how closely it reaches this point. Figure 2b shows a schematic of an SCC plot, showing that a curve which is closer to the top left-hand corner of the plot achieves better performance for structure verification. The area under the curve (classification area, CA) is a numerical measure of performance. The ideal SCC curve would have a CA of 1, meaning that the correct structure would be classified as "correct" for all compound pairs tested. As analogous to the AUC for a ROC plot, a random process would have a CA of 0.5.

Combining multiple data types

When using a single data type (NMR or IR) the data can be used directly to generate the SCC plot. Combining data types, however, is non-trivial if the match scores have different distributions and scales. In such cases, there is precedent for using a ranking-based method instead of the raw scores to combine multiple different metrics. So-called 'consensus' scoring of various different scoring functions has been reviewed in relation to virtual screening of ligands.[28]

To combine IR and NMR data, we used a percentile-rank combination procedure similar to that described by Hsu and Taksa.[29] These authors noted that combination using ranks performs better than combination using raw scores if the raw scores have different distributions. The procedure first ranks each candidate structure using the relevant IR and NMR match scores from highest (top scoring structure) to lowest (bottom scoring structure). These ranks are then converted to percentile-ranks using the formula:

$$p_i = \frac{N - r_i}{N - 1},$$

where N is the length of the candidate list and p_i and r_i are the percentile and absolute ranks respectively of candidate i. This has the effect of converting the absolute ranks (which are not comparable between lists of different length) to a score on a 0—1 scale. For each candidate structure, the mean of the percentile-ranks of both spectroscopic modalities is calculated to give a combined percentile-rank in which both data types are weighted equally. If a structure is correct, we would expect it to achieve a relatively high NMR match score as well as a relatively high IR match score, and it should therefore rank highly in *both* lists. The average percentile-rank method should be effective at highlighting those structures which rank well with both data types, as well as giving lower scores to structures which score highly with one data type but poorly with the other.

A limitation applying to small list sizes is that several candidates may have the same average percentile-rank. To distinguish these candidates, we add a small additional term to the average percentile-ranks, given by:

$$\Delta_i = \epsilon \cdot \bar{z}_i,$$

where $\epsilon = 10^{-8}$ is a small parameter and $\bar{z}_i$ is the *mean z-score* for candidate i. The z-score is given by:

$$z_i = \frac{x_i - \mu}{\sigma},$$

where x_i is the raw score, μ is the mean score and σ is the standard deviation of the scores. The effect of the additional Δ term is simply to break ties using raw score information; the parameter ϵ is chosen so that the term is too small to affect the rank order of untied candidates. The specific value of ϵ does not alter rank order, provided that it is small relative to the percentile-rank scores (i.e. 10^{-3} or lower).

Compound Pairs Evaluated Using IR Spectra

The IR.Cai algorithm compares the calculated IR spectrum for each test structure against the relevant experimental spectrum. This is achieved by calculating the IR spectrum for a given test structure using DFT, applying Lorentzian line broadening with full-width at half-maximum (FWHM) of 12 cm^{-1} to simulate the broad peaks found in experimental spectra.[30] The match score between the experimental and calculated spectra is then found by using the formula:

$$\mathrm{IR.Cai} = \frac{\sum_i c_i e_i}{\sqrt{\sum_i c_i^2 \sum_i e_i^2}},$$

which gives a convenient match score in the range 0—1. A score closer to 1 indicates a better match between the calculated and experimental spectra, and therefore a higher probability that the suggested structure is correct. Whilst manual analysis of IR spectra usually focuses on key peaks above 1500 cm^{-1}, this algorithm can look at the whole spectrum, including the details of the fingerprint region. DFT calculations were performed at two levels of theory (B3LYP/6-31G* and B3PW91/cc-pVTZ including a PCM model for DMSO; hereinafter referred to as lower and higher theory levels respectively). Details of the approximate runtime of the DFT calculations may be found in the Section 1.5 of the SI. A fixed scaling factor of 0.97 (low level theory) and 0.98 (high level theory) was used as these were found optimal in our earlier work.[31] A Lorentzian line broadening with a full width at half maximum (FWHM) of 12 cm^{-1} was used.[30] Values of 8, 10 or 12 cm^{-1} were found to make little difference to the results (Supporting Information, Figure S1). Whilst the IR.Cai algorithm can examine the whole wavenumber range, in this work the region examined was 1250-1600 cm^{-1}. This region contains the portion of the fingerprint information in an IR spectrum that usually contains most peaks. The strong absorbance of DMSO-d_6 at 1100 cm^{-1} precludes the use of the lower wavenumber fingerprint region. No improvement was found in extending the range to higher wavenumbers even though many of our test compounds included carbonyl groups; we believe this may be because H-bonding and other interactions make it difficult to accurately calculate the stretching frequencies. The scores from the IR.Cai algorithm for all compounds and isomers are shown in the supporting information, Tables S6 (high level) and S7 (low level).

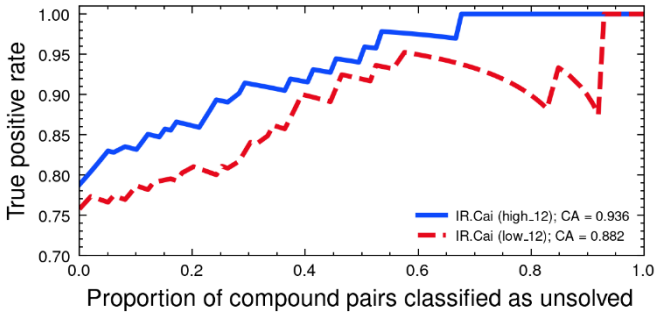


Figure 3: Structure classification characteristic (SCC) curve based on the 42 molecules in the test set using IR.Cai scores measuring the degree of overlap between calculated and experimental spectra. IR.Cai_high and IR.Cai_low here refer to IR.Cai scores calculated with IR spectra computed at the B3PW91/cc-pVTZ (high level) and B3LYP/6-31G* (low level) levels of theory respectively. The position of the SCC curve and higher CA indicates better performance for the higher theory level. We therefore use the higher level of theory for the results in the rest of this work.

For each of the 99 test pairs formed by comparing a correct structure with each of its incorrect isomers individually, the score difference was calculated. Analysis using the methods described above results in the SCC curves shown in Figure 3.

The SCC curves show two broadly parallel lines at low level and high levels of theory with the higher level of theory performing

better as defined by the CA. This is likely a reflection of better IR prediction at the higher theory level. Indeed the average IR.Cai values for the correct isomers was 0.8098 (high) vs 0.7844 (low). This algorithmic method of comparing isomers based simply on the comparison of experimental and calculated IR spectra is effective for distinguishing isomers. For example, applying the criterion of being correct 90% of the time, the method can distinguish 55-70% of the isomer pairs depending on theory level. If we wish to be correct 95% of the time, the method can still distinguish 40-50% of the isomer pairs. Noting that the CA for the SCC curves is a measure of performance as AUC is for the ROC curves, the CA scores of 0.936 and 0.882 (for high- and low-level theory respectively) are a significant improvement on value of 0.694 obtained for the absolute ASV method shown in the ROC curves in Figure 1. This supports the hypotheses both that IR spectra contain sufficient information to be able to distinguish many similar molecules and that comparing compounds is a more effective method for ASV. A strength of the IR analysis is that there is no need to peak pick the spectrum. NMR analysis is sensitive to the peak-picking algorithm, but for the analysis of IR spectra the broader peaks mean that an overlap integral suffices to score the match between the spectra.

To those unfamiliar with IR, it may come as a surprise how well this simple and sensitive technique is able to distinguish between structural isomers. This is however in keeping with our own observations and those reported by others. For example, Cotter *et al*[32] reported the successful identification of reaction products of amines and isocyanates using experimental and calculated IR spectra. It was not possible to distinguish the products using regular NMR methods. Nolvachai *et al*[33] reported the identification of several small isomeric reaction products also by matching experimental and calculated IR spectra.

Compound Pairs Evaluated Using ^{1}H NMR Data

We adapted DP4 to allow for labile protons (which are challenging for DFT methods to predict) to give a metric we name DP4* (see SI Section 3). ^{1}H NMR spectra were evaluated by DP4* and by ACD/Labs' ASV program.[27] As peak picking was not part of our evaluation, we peak picked the spectrum manually ignoring minor impurities and solvent and picking peaks close to the residual DMSO and water resonances. In a few cases, peaks were completely hidden by solvent and no allowance was made for this – *i.e.* they are missing from our peak picked spectrum. The peak listing after manual peak picking is given in Section 10 of the SI. We note that using the automated peak picking routine available in the ACD/Labs software resulted in only a moderate degradation in performance. This suggests that the approach of using just the chemical shifts to perform the DP4* analysis is amenable to automation in combination with a suitable automatic peak-picking routine. The scores from DP4* and ACD (automatic and manual peak picking) for all compounds and isomers are shown in the SI, Tables S5, S8 and S9.

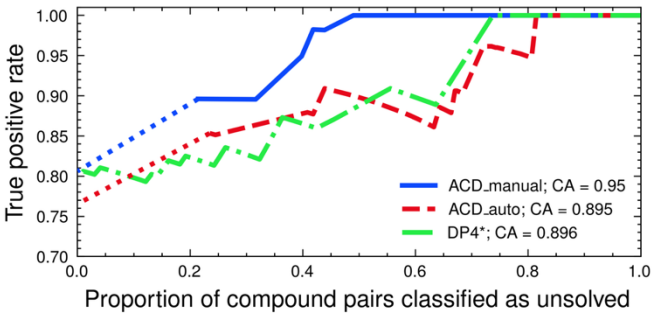


Figure 4: Structure classification characteristic (SCC) curve based on the 42 molecules in the test set using raw DP4* and ACD scores. Some of the ACD scores were identical for the correct structure and an incorrect isomer, so it was not possible to classify all pairs. The dotted section of the line for ACD therefore represents the expectation of random guessing to choose the correct isomer for molecules which had the same score. Results using ACD's automatic peak-picking procedure are shown (ACD_auto) as well as results using manually peak-picked spectra (ACD_manual). Note that using ACD's automatic peak picking results in only a mild degradation in performance.

The results using DP4* and ACD are shown on the SCC curve in Figure 4. The performance of DP4* and ACD is similar to that of IR.Cai with both methods achieving similar levels of accuracy at a given unsolved proportion. The CA values of 0.949 and 0.896 for ACD and DP4* respectively are similar to the values obtained for high and low level IR.Cai scores.

Only ^{1}H chemical shifts and relative integrals were used for structure verification. Proton NMR spectra also contain rich and possibly complementary information from proton-proton J-couplings, but these are currently much harder to interpret automatically. First, second-order effects (so-called strong coupling) make it challenging to measure couplings between protons with similar chemical shifts. Second, J-couplings can only be calculated from accurate 3D structures using developments of the Karplus equation[34] or DFT calculations,[35] far more resource-demanding than shielding constant calculations.[36,37] Recent progress with machine learning methods suggests they can deliver DFT accuracy for $^3J_{HH}$[38] and, assuming this can be extended to $^2J_{HH}$ and $^4J_{HH}$, it may be possible to incorporate this into ASV methods by combining them with spectral fitting and density-matrix calculations to resolve second-order effects.

Compound Pairs Evaluated Using a Combination of ^{1}H NMR and IR Data

Data combination using the previously described percentile-rank procedure produces the SCC curves shown in Figure 5.

Recalling that a perfect performance is a point in the top left corner (all pairs classified correctly and no pairs unsolved), the combination of IR.Cai with ^{1}H NMR DP4* or ACD moves the performance approximately half way towards that goal as evaluated visually and by the CA metric. For example, IR.Cai (high) solved around 73% of isomer pairs at a 90% True Positive rate, achieving a CA of 0.936. When combined with DP4* or ACD this improves to 85% (DP4*) to 100% (ACD) solved at the same True Positive rate. The CA improves accordingly to 0.966 (IR.Cai + DP4*) and 0.979 (IR.Cai + ACD). Similarly, at a True Positive rate of 95%, IR.Cai solved 50% of isomer pairs, rising to 70-85%

when combined with DP4* or ACD. Combining NMR and IR data therefore improves the structure verification performance by a factor of 2 to 3, based on the difference to the ideal CA score of 1, compared to using NMR or IR data alone.

This valuable improvement to the structure verification procedure suggests that the information provided by the two spectroscopic methods is complementary. If the methods were providing similar information, we might expect the SCC curve of the combination to lie between the performance of the two methods individually. As a control and as expected we find that neither IR (low) and IR (high) nor ACD and DP4* can be combined to show an improvement (SI, Figures S5 and S6). We also find that we can achieve similar results by combining the raw scores for DP4*, ACD or IR.Cai without using the percentile-rank method (SI, Figures S7 and S8). We believe, however, that the percentile-rank combination procedure is the most appropriate and least arbitrary way of combining data. Further details are given in the SI.

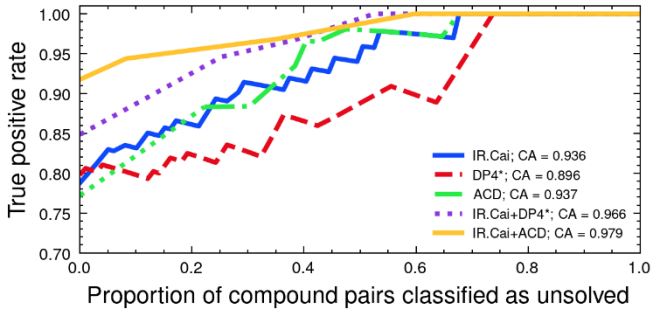


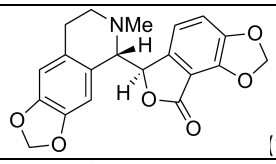
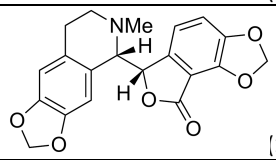
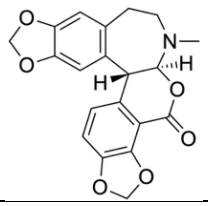
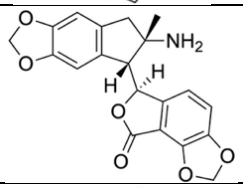
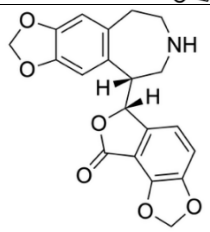
Figure 5: SCC curve based on the 42 molecules in the test set for high-level IR, DP4*, ACD and the combinations of DP4* and ACD with IR. The combination lines are obtained using the percentile-rank procedure described in the text. Corresponding SCC curves for combination with low-level IR are shown in the supporting information (Figure S5)

Case study: Choosing from a large set of incorrect isomers

To investigate the generality of the proposed system, we applied it to a larger set of incorrect isomers. A pipeline was designed using REINVENT4[39] to obtain structures with high Tanimoto similarity (calculated using Morgan fingerprints of length 2048 and radius 2) to an input molecule. Full details and all structures are given in the SI. In our case, we performed this procedure using one of the larger molecules from our test set (molecule 43, MW367) to augment the number of incorrect isomers. The 18 additional incorrect isomers which had the highest Tanimoto similarity to the correct structure were selected. Calculations were performed using the same settings as described earlier in the text to obtain DP4*, IR.Cai and ACD scores for all of the new structures. The structures of the top 3 candidates ranked by DP4* and IR.Cai are shown in Table 2; the full results for all the isomers are shown in the SI (Table S4). The percentile-rank combination procedure was used to combine the information from the DP4* and IR.Cai scores. This procedure identified the correct structure as the most likely one, whereas it ranks 3rd for both DP4* and IR.Cai when used alone. The average combined percentile-rank is highest for the correct structure (0.909 compared to 0.864 for the next best candidate) as the correct structure scores relatively highly for both DP4* *and* IR.Cai. Although incorrect structures may score higher than the correct structure for one of the methods, for orthogonal data it is unlikely that an incorrect structure would score higher for both methods. The percentile-rank method identifies the structure that scores well for both types of spectra as the most likely candidate.

Whilst it is difficult from the structures shown in Table 2 to identify structural features that influence the performance of NMR or IR identification, it is interesting to note that both techniques can distinguish between diastereoisomers. Note that the ACD software does not score these compounds well likely due to the unusual geminal J-coupling of the methylenedioxy groups (peak listing in Section 10 and ACD scores in Table S10 of the SI).

Table 2: The top 3 candidates as ranked by DP4* and IR.Cai for the expanded compound 43 test set, sorted by descending average percentile-rank. The average percentile-rank procedure ranks the correct structure (candidate 0) the highest, despite its diastereoisomer (candidate 3) being ranked highest by DP4*.

Candidate (average percentile rank)	Structure (Tanimoto similarity to correct structure, Morgan fingerprints)	DP4* rank / 23	IR.Cai rank / 23
0 (**correct**) (0.909)	 (1.0)	3	3
3 (0.864)	 (1.0)	1	7
5 (0.864)	 (0.72)	7	1
22 (0.818)	 (0.59)	18	2
12 (0.750)	 (0.60)	2	21

ARTICLE

Conclusions

We present a method for ASV using a combination of IR and NMR data to score compounds and focusing on differences between the scores. The similarity metric between calculated and experimental IR spectra, IR.Cai, is combined with the DP4* score for NMR assignment to create a new measure of the correspondence between spectra and candidate molecules that may have produced them. Testing the method on a challenging dataset of 42 drug-like compounds and 99 closely related incorrect isomers demonstrates that IR data can improve the performance of ASV. 100% of the potential comparisons (correct structure vs incorrect isomer) could be solved with an 85% true positive rate when using a combination of NMR and IR data, with a CA score of 0.966. This is an improvement on the CA score of 0.936 using IR data alone and represents a significant step towards the ideal case of CA = 1. This result demonstrates that combining IR and NMR data allows a higher proportion of similar isomers to be distinguished while maintaining the same confidence in each classification. We validate these results on a larger set of automatically-generated candidate molecules and show that the consensus choice of molecule of DP4* and IR.Cai picks the correct isomer when either technique alone fails. While the relative contributions of NMR and IR to distinguishing isomeric structures will naturally be influenced by the test set chosen, these results clearly demonstrate the complementarity of the two forms of spectroscopy. The ease with which IR spectra can be collected from small amounts of material makes it an attractive proposition for use in the ASV approach. This contrasts with ^{13}C NMR spectra, which increase the time required for data acquisition and are likely to be a bottleneck in the workflow. The use of IR and ^{1}H NMR data together should avoid this potential bottleneck. Although DFT calculations are currently required to simulate NMR and IR spectra, advances in machine learning methods are beginning to address this.[38,40–44] Similarly advances in synthesis prediction tools means that the automatic generation of a realistic set of possible reaction products is becoming a reality. Our process to combine low-cost inputs: 1D ^{1}H NMR and IR spectra, provides a foundation for even faster structure verification by extracting useful information from multiple inexpensive techniques, and will enable the acceleration of chemical discovery.

Author contributions

RJL devised the project with contribution from JMG, TL and WC. WC and TL obtained funding. JMG wrote the IR.Cai algorithm. LJ collected experimental data and performed calculations together with RJL. LJ and RJL performed initial data analysis. PH suggested improved analysis methods. JBR devised DP4* and investigated the optimal combination procedure, performed calculations and data analysis. JBR, RJL, PH and JMG wrote the initial draft of the manuscript. All authors contributed to the final draft.

Conflicts of interest

RJL, LJ, PH, WC and TL are employed by AstraZeneca and RJL, PH, WC and TL own shares in the company. The PhD studentship of JBR is part funded by AstraZeneca. JMG received funding from AstraZeneca to develop the IR.Cai algorithm.

Data availability

Data for this article, including recorded IR and NMR spectra for all compounds as well as DFT calculation files are available at Apollo: https://doi.org/10.17863/CAM.110235

Acknowledgements

We thank M. Priessner and G. Hulthe for helpful discussions and AstraZeneca for funding support for JBR.
JBR is grateful for funding from an EPSRC CASE award (ref: EP/W524633/1)

This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk).

Notes and references

1 G. Bifulco, P. Dambruoso, L. Gomez-Paloma and R. Riccio, *Chem. Rev.*, 2007, **107**, 3744–3779.
2 P. Keyes, G. Hernandez, G. Cianchetta, J. Robinson and B. Lefebvre, *Magn. Reson. Chem.*, 2009, **47**, 38–52.
3 S. S. Golotvin, R. Pol, R. R. Sasaki, A. Nikitina and P. Keyes, *Magn. Reson. Chem.*, 2012, **50**, 429–435.

4 D. C. Burns, E. P. Mazzola and W. F. Reynolds, *Nat. Prod. Rep.*, 2019, **36**, 919–933.
5 A. V. Buevich and M. E. Elyashberg, *J. Nat. Prod.*, 2016, **79**, 3105–3116.
6 S. G. Smith and J. M. Goodman, *J. Am. Chem. Soc.*, 2010, **132**, 12946–12959.
7 A. Howarth and J. M. Goodman, *Chem. Sci.*, 2022, **13**, 3507–3518.
8 J. Richardson, G. Sharman, F. Martínez-Olid, S. Cañellas and J. E. Gomez, *React. Chem. Eng.*, 2020, **5**, 779–792.
9 V. Rodríguez Martín-Aragón, M. Trigal Martínez, C. Cuadrado, A. H. Daranas, A. Fernández Medarde and J. M. Sánchez López, *ACS Omega*, 2023, **8**, 39873–39885.
10 F.-Z. Zhang, X.-M. Li, L.-H. Meng and B.-G. Wang, *J. Antibiot. (Tokyo)*, DOI:10.1038/s41429-023-00666-3.
11 C. Pan, H. Ikeda, M. Minote, T. Tokuda, T. Kuranaga, T. Taniguchi, N. Shinzato, H. Onaka and H. Kakeya, *J. Antibiot. (Tokyo)*, DOI:10.1038/s41429-023-00668-1.
12 J. Zhang, K. Terayama, M. Sumita, K. Yoshizoe, K. Ito, J. Kikuchi and K. Tsuda, *Sci. Technol. Adv. Mater.*, 2020, **21**, 552–561.
13 Z. Huang, M. S. Chen, C. P. Woroch, T. E. Markland and M. W. Kanan, *Chem. Sci.*, 2021, **12**, 15329–15338.
14 F. Hu, M. S. Chen, G. M. Rotskoff, M. W. Kanan and T. E. Markland, *ACS Cent. Sci.*, 2024, **10**, 2162–2170.
15 L. Yao, M. Yang, J. Song, Z. Yang, H. Sun, H. Shi, X. Liu, X. Ji, Y. Deng and X. Wang, *Anal. Chem.*, 2023, **95**, 5393–5401.
16 J. A. Lumley, G. Sharman, T. Wilkin, M. Hirst, C. Cobas and M. Goebel, *SLAS Discov. Adv. Sci. Drug Discov.*, 2020, **25**, 950–956.
17 S. S. Golotvin, E. Vodopianov, B. A. Lefebvre, A. J. Williams and T. D. Spitzer, *Magn. Reson. Chem.*, 2006, **44**, 524–538.
18 M. E. Elyashberg, A. J. Williams and G. E. Martin, *Prog. Nucl. Magn. Reson. Spectrosc.*, 2008, **53**, 1–104.
19 S. S. Golotvin, E. Vodopianov, R. Pol, B. A. Lefebvre, A. J. Williams, R. D. Rutkowske and T. D. Spitzer, *Magn. Reson. Chem.*, 2007, **45**, 803–813.
20 M. Pesek, A. Juvan, J. Jakoš, J. Košmrlj, M. Marolt and M. Gazvoda, *J. Chem. Inf. Model.*, 2021, **61**, 756–763.
21 L. Böselt, R. Dötzer, S. Steiner, M. Stritzinger, S. Salzmann and S. Riniker, *Anal. Chem.*, 2020, **92**, 9124–9131.
22 E. Chacko, R. Sondhi, A. Praveen, K. L. Luska and R. A. V. Hernandez, *ChemRxiv*, 2024, preprint, DOI: 10.26434/chemrxiv-2024-37v2j.
23 M. Alberts, F. Zipoli and T. Laino, *Digit. Discov.*, 2025, **4**, 1936–1943.
24 W. Wu, A. Leonardis, J. Jiao, J. Jiang and L. Chen, *J. Phys. Chem. A*, 2025, **129**, 2077–2085.
25 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.
26 S. G. Smith and J. M. Goodman, *J. Org. Chem.*, 2009, **74**, 4597–4607.
27 NMR Workbook Suite, version 2022.2.3, Advanced Chemistry Development Inc. (ACD/Labs), Toronto, ON, Canada, www.acdlabs.com, .
28 M. Feher, *Drug Discov. Today*, 2006, **11**, 421–428.
29 D. Frank Hsu and I. Taksa, *Inf. Retr.*, 2005, **8**, 449–480.
30 L. Böselt, R. Aerts, W. Herrebout and S. Riniker, *Phys. Chem. Chem. Phys.*, 2023, **25**, 2063–2074.
31 J. Lam, R. J. Lewis and J. M. Goodman, *J. Cheminformatics*, 2023, **15**, 36–36.
32 E. Cotter, F. Pultar, S. Riniker and K.-H. Altmann, *Chem. – Eur. J.*, 2024, **30**, e202304272–e202304272.
33 Y. Nolvachai, S. Salzmann, J. S. Zavahir, R. Doetzer, S. Steiner, C. Kulsing and P. J. Marriott, *Anal. Chem.*, 2021, **93**, 15508–15516.
34 C. A. G. Haasnoot, F. A. A. M. de Leeuw and C. Altona, *Tetrahedron*, 1980, **36**, 2783–2792.
35 T. Bally and P. R. Rablen, *J. Org. Chem.*, 2011, **76**, 4818–4830.
36 T. Helgaker, M. Watson and N. C. Handy, *J. Chem. Phys.*, 2000, **113**, 9402–9409.
37 I. Alkorta and J. Elguero, *Int. J. Mol. Sci.*, 2003, **4**, 64–92.
38 C. Yiu, B. Honoré, W. Gerrard, J. Napolitano-Farina, D. Russell, I. M. L. Trist, R. Dooley and C. P. Butts, *Chem. Sci.*, 2025, **16**, 8377–8382.
39 H. H. Loeffler, J. He, A. Tibo, J. P. Janet, A. Voronov, L. H. Mervin and O. Engkvist, *J. Cheminformatics*, 2024, **16**, 20.
40 I. Cortés, C. Cuadrado, A. Hernández Daranas and A. M. Sarotti, *Front. Nat. Prod.*
41 Y. Guan, S. V. S. Sowndarya, L. C. Gallegos, P. C. S. John and R. S. Paton, *Chem. Sci.*, 2021, **12**, 12012–12026.
42 C. McGill, M. Forsuelo, Y. Guan and W. H. Green, *J. Chem. Inf. Model.*, 2021, **61**, 2594–2609.
43 C. Cobas, *Magn. Reson. Chem.*, 2020, **58**, 512–519.
44 C. M. K. Stienstra, L. Hebert, P. Thomas, A. Haack, J. Guo and W. S. Hopkins, *J. Chem. Inf. Model.*, DOI:10.1021/acs.jcim.4c00378.