

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

Richard Lewis

`richard.j.lewis@astrazeneca.com`

Biopharmaceuticals R&D, AstraZeneca <https://orcid.org/0000-0001-9404-8520>

Benji Rowlands

Yusuf Hamied Department of Chemistry, University of Cambridge

Lina Jonsson

Biopharmaceuticals R&D, AstraZeneca

Jonathan Goodman

University of Cambridge <https://orcid.org/0000-0002-8693-9136>

Peter Howe

Oncology Chemistry, AstraZeneca

Werngard Czechtizky

AstraZeneca

Tomas Leek

Biopharmaceuticals R&D, AstraZeneca <https://orcid.org/0000-0003-0772-4696>

Article

Keywords:

Posted Date: August 12th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-4719113/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Yes there is potential Competing 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 BR is part funded by AstraZeneca. JMG received funding from AstraZeneca to develop the IR.Cai algorithm.

Abstract

Human interpretation of spectroscopic data remains key to confirming new structures; the quest for speed and resource-efficiency suggests automating structure verification. We report that the combination of NMR and easily accessible IR data greatly improves its performance. We introduce an algorithm to quantify the similarity between experimental and calculated IR spectra and apply this to distinguish between a test set of 43 molecules and 100 similar isomeric structures. We describe a method to combine IR and ^1H NMR results measuring performance as the *structure classification characteristic area over curve* (SCC-AOC). Combination of IR and ^1H NMR significantly outperforms either technique alone (SCC-AOC 0.025 for combined data compared to IR 0.053 and ^1H NMR 0.101 and a large step towards the ideal SCC-AOC value of zero). It drives the correct classification rate of the 100 comparisons to 87% from *ca.* 80% for individual methods and brings reliable automation within grasp.

1 Introduction

Determining 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 undoubtedly 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] and Computer-Assisted Structure Elucidation (CASE) [3], [4]. 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, the ASV extreme can be thought of as similar to a chemist running a well characterized reaction and relying on a ^1H NMR spectrum to confirm the product.

Among the best-established methods in the ASV category are the DP4 and DP5 probabilities [5], [6]. These methods involve using density functional theory (DFT) to calculate NMR spectra 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 is regularly used to assist with structure elucidation in challenging cases [7], [8], [9], [10].

One area of current research interest is in the application of machine learning to automated structure elucidation [11], [12], [13]. 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 in this area.

Previous work has focussed on applying automated interpretation methods to NMR data. But these may also be applied to data which a human can't 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 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 provide atom-focussed information because the chemical shift is dominated by relatively short-range effects such as hybridization, covalent structure and 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. Indeed, the use of IR for structure determination has been reported recently for building up molecules in a fragment-based approach [15], determining regio- and stereochemistry by matching experimental and calculated IR spectra [16] and in a machine learning model of simulated data, to predict complete structures [17].

In this ASV proposal, we assume that that we know the molecular formula (for example from high resolution mass spectroscopy) and only need to distinguish between regio- and stereoisomers. We assume that we have access to a list of candidate structures which includes the correct structure. This list could be generated by the chemist or by reaction prediction software. This workflow is similar to that of a typical synthetic chemist having knowledge of the possible reaction products and testing them against the analytical data collected. An additional requirement for the method is that it should flag those cases when it cannot distinguish between candidate structures. Such *unknowns* could either be reviewed manually and/or further analysed by acquisition of other analytical data. Flagging "unknowns" is more useful for practical applications than a method which does not account for the uncertainty in the analysis.

Based on these assumptions, we propose and assess a method for ASV using a combination of ^1H NMR and IR data. We introduce an algorithm (IR.Cai) to match and score experimental and IR spectra and test this at two levels of DFT theory. 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 also analyse ^1H NMR using a commercial ASV software package (ACD/Labs). For a set of highly similar isomeric test structures, we test the hypothesis that the candidate structure that scores higher (by IR, NMR or by a combination of the scores) is the more likely to be correct, and that the larger the score difference, the greater the probability that the higher scoring

structure is correct. Any pair of candidate structures that is not differentiated by a sufficiently large score naturally falls into the “unknown” category requiring further data or interpretation.

2 Results and discussion

2.1 Verifying one of a choice of structures

Verifying a single structure without context is not an everyday task for most chemists. Classifying a single compound as “correct” or “incorrect” neglects that 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 actually do in practice in a high-throughput environment. Furthermore, by scoring and comparing similar compounds, we might expect systematic errors to cancel to a certain extent (for example, inaccurate calculation of spectroscopic data, and artefacts such as NMR peaks hidden by solvent resonances or additional peaks from residual reaction solvents).

2.2 Compounds

To evaluate the performance of different methods for structure verification, a dataset of 43 drug-like molecules was assembled. These had a molecular weight between 182 and 430 with an average of 300. For each molecule, two or three isomers were generated. 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 spectra. These were arranged into 100 pairs of the correct molecule structure and one incorrect isomeric structure.

2.3 Data analysis and visualization

The IR.Cai matching algorithm, DP4* and the ¹H ASV tool from Advanced Chemistry Development, Inc. (ACD/Labs) [18] and give numbers between 0–1 related to how well the experimental spectrum matches the calculated one. 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. In order 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 *not classified* (unknown) 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 number of compound pairs classified, and the number of classified pairs that are classified correctly. Thus, if we want to be correct 95% of the time, we would expect to classify fewer compound pairs than if our requirement is to correctly classify 80% of

the time. Figure 1a illustrates the process, taking as an example Compound 1 and its first incorrect isomer from the test set.

To examine this trade-off, we devised a visualization approach inspired by the “Receiver Operating Characteristic” (ROC). We call this the *Structure Classification Characteristic* (SCC). The curve is formed by plotting the two key indicators of performance (the classification rate and the fraction of classifications which are correct, each a function of the threshold) against each other. The ideal performance is a point in the top right, where all compound pairs are classified and all of these classifications are correct. The performance of a particular method can be compared to others according to how closely it reaches this point. Figure 1b shows a schematic of an SCC plot, showing how a curve which is closer to the top right-hand corner of the plot achieves better performance for structure verification. The area bounded by the curve and the top right (area over the curve, AOC) is a numerical measure of performance. The ideal SCC curve would have an AOC of 0, meaning that the correct structure would be classified as “correct” for all compound pairs tested.

2.4 Compound pairs evaluated using IR spectra

The IR.Cai algorithm (Section 4.1) compares the calculated IR spectrum for each test structure against the relevant experimental spectrum. 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. A score between 0 and 1 is produced dependent on the degree of overlap between the calculated and experimental spectra. DFT calculations were performed at lower and higher theory levels (see methods). 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 [14]. A Lorentzian line broadening with a half width at half maximum (HWHM) of 12 cm^{-1} was used [19]. Values of 8, 10 or 12 cm^{-1} were found to make little difference to the results (Supporting Information, Figure S1). The wavenumber range examined was $1250\text{--}1600\text{ cm}^{-1}$. This region contains the portion of the fingerprint information in an IR spectrum that usually contains most peaks. The strong absorbance of DMSO-d₆ at 1100 cm^{-1} precludes the use of the lower wavenumber fingerprint region. It was not found to be beneficial to extend the range to higher wavenumbers even though many of our test compounds included carbonyl groups; 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 S4 (high level) and S5 (low level).

For each of the 100 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 Fig. 2.

The SCC curves show two broadly parallel lines at low level and high levels of theory. 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 our test isomer pairs. The AOC scores of 0.064 and 0.117 for high and low level theory respectively support the idea that IR spectra contain sufficient information to be able to distinguish many similar molecules. 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* [20] 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* [21] reported the identification of several small isomeric reaction products also by matching experimental and calculated IR spectra.

2.5 Compound pairs evaluated using ^1H 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 Methods). ^1H NMR spectra were evaluated by DP4* and by ACD/Labs' ASV program. 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 (Figure S8, Supporting Information). The results using DP4* and ACD are shown on the SCC curve in Fig. 3.

The performance of DP4* and ACD is similar to that of IR.Cai with both methods achieving similar levels of accuracy at a given classification rate. The AOC values of 0.053 and 0.10 for ACD and DP4* respectively are similar to the values obtained for high and low level IR.Cai scores. This demonstrates the power of the IR data, but may also reflect the choice of test compounds as many of the incorrect isomers are expected to have rather similar NMR spectra to the correct structures. A further strength of the IR analysis is that there is no need to peak pick the spectrum. All methods of NMR analysis are sensitive to the peak-picking algorithm, whether this is automated or by hand, but this is not true for the analysis of IR spectra as the broader peaks means that an overlap integral suffices to score the match between the spectra.

2.6 Compound pairs evaluated using a combination of ^1H NMR and IR data

In order to combine IR and NMR data, we first applied the softmax scaling procedure described in Section 4.5 to the DP4* and ACD results, to scale the scores to the same mean and standard deviation as the IR.Cai scores. The results when plotted on the SCC curve are shown in Fig. 4.

Recalling that a perfect performance is a point in the top right corner (all pairs classified and all classified correctly), the combination of IR.Cai low or high level with ^1H NMR DP4* or ACD moves the performance approximately half way towards that goal as evaluated visually and by the AOC metric. For

example, IR.Cai (high) classified around 70% of compounds with 90% accuracy, achieving an AOC of 0.064. When combined with DP4* or ACD this moves to 70% classified with 98% accuracy in both cases, and the AOC improves accordingly to 0.025 (IR.Cai + DP4*) and 0.016 (IR.Cai + ACD). Combining NMR and IR data improves the structure verification performance by a factor of 2 to 3, based on the AOC scores compared to using NMR or IR data alone, giving a valuable improvement to structure verification performance.

This suggests that the information provided by the two spectroscopic methods is to some extent 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 S6 and S7). We also find that we can achieve similar results by combining the raw scores for DP4*, ACD or IR.Cai without scaling using softmax (SI, Figures S9 and S10). We believe however that the softmax scaling to put all data into the same statistical framework is in general the most robust way to combine data. Further details are given in the SI.

3 Conclusions

We present a methodology for ASV using a combination of IR and NMR data. 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 molecules that may have produced them. Rigorous testing of our methodology on a challenging dataset of 43 drug-like compounds and 100 associated isomeric decoys demonstrates that IR data can improve the performance of ASV. Tests showed that 100% of the potential comparisons (correct structure vs decoy) could be classified with an 87% correct classification rate when using a combination of NMR and IR data, achieving an AOC score of 0.025. This is an improvement on the AOC score of 0.064 using IR data alone, and represents a significant step towards the ideal case of AOC = 0. This result demonstrates that combining IR and NMR data allows more structures to be classified while maintaining the same confidence in each classification. 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 ASV software. This contrasts with ^{13}C NMR or 2D spectra, which significantly increase the time required for data acquisition and are likely to be a bottleneck in the workflow. The use of IR and ^1H NMR data together should relieve this bottleneck, and the new pinchpoint in the process may be the computational cost. DFT calculations are currently required to simulate NMR and IR spectra, but advances in machine learning methods are addressing this issue. Our process to combine low-cost inputs: 1D ^1H 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.

4 Methods

4.1 IR.Cai algorithm

The algorithm used to generate the IR scores is a modification of the previously reported Cai-factor method for automatic analysis of VCD spectra [14]. The similarity score was defined as a modification of the SimVCD integral [22], and essentially represents the normalised overlap integral between the experimental and calculated spectra. The resulting score, between 0 and 1, is a measure of the goodness of fit between the experimental and calculated spectra.

The algorithm enables a number of parameters to be set, including the wavenumber range over which the spectra are to be evaluated, and a value for line broadening applied and scaling factor. Setting the wavenumber range allows regions with strong solvent absorptions to be excluded. Calculated IR spectra give a series of sharp peaks for each conformer, so Lorentzian line broadening is applied to better match the experimental spectrum. The scaling factor allows for systematic errors to be corrected. These are due to neglect of anharmonicity in the calculations amongst other reasons [23], [24]. The algorithm can either use a fixed scaling factor, or search for a best-fit scaling factor within a narrow range. The algorithm is described in more detail in the SI.

The algorithm is available on GitHub: <https://github.com/Goodman-lab/IR.Cai>

4.2 DP4* analysis

The choice of solvent has an impact on a DP4 calculation, especially for ^1H NMR, where the shifts due to labile protons are often far away from DFT calculated values in polar solvents such as DMSO. This is important because DMSO is a very commonly used solvent in the pharmaceutical industry and academia due to its excellent solubility properties, water miscibility and low toxicity.

Initially, DP4 was used to analyse the ^1H spectra. We recognized that the statistics provided by DP4 analysis were not optimal, in particular there was a tendency to overestimate confidence in a particular result. The cause for this was traced to the tendency for labile protons attached to O or N to show experimental shifts very different from their calculated shifts. This is explained by a number of factors including pH, water content and hydrogen bonding of protic protons with the DMSO solvent, which cannot easily be accounted for by GIAO NMR shift calculation. Where there are large errors between experimental and calculated shifts, the resulting DP4 probabilities are often misleading due to the sensitivity of the DP4 probability to the assignment of experimental to calculated shifts. Therefore, we developed an alternative parameter, DP4*, which accounts for any very large errors when performing the shift assignment. This works by matching the experimental to calculated shifts to minimise the mean absolute error while maximising the number of shifts that are paired up and excluding any experimental shifts which do not have a corresponding calculated shift, within a certain threshold. Then, the DP4 probabilities are scaled to give DP4* scores according to the number of experimental shifts that were excluded when performing the assignment. Full details are given in the SI. DP4* is effective even for those molecules with several large errors in calculated shifts due to labile protons. An implementation of the DP4* algorithm is available on GitHub: <https://github.com/Goodman-lab/DP4-star>

4.3 ACD

Advanced Chemistry Development Inc. (ACD/Labs) offers an ASV tool (version 2022.2.3) which generates a “match factor”, a value between 0–1 by comparing NMR data against a structure. This has been described in the literature [25]. The match factor relies in large part on comparison of experimental chemical shifts and those calculated through a database approach. This data should act as a control for the DP4 analysis – if IR and ¹H NMR contain complementary information, then this should show both with DP4* and ACD analysis. The details of the algorithm used to calculate the match factor are not public.

4.4 Conformational search and DFT calculations

For all compounds and isomers a conformational search was performed using Macromodel (Schrodinger Inc) using the MMFF force field and a water solvent model. All conformations within an energy window of 21 kJ/mol were retained.

IR spectrum calculations were performed using Gaussian 16 Revision B.01 [26] at two levels of theory – B3LYP/6-31G* and B3PW91/cc-pVTZ including a PCM model for DMSO. NMR calculations were performed on the minimized conformations at B3LYP/6-31G* using the mPW1PW91/6-311G(d) level of theory. Single point energies for each conformer were obtained using M062X/def2-TZVP.

All outputs of the calculations performed together with experimental proton NMR and IR spectra are available at Apollo: <<https://doi.org/10.17863/CAM.110235>>

4.5 Combining multiple methods

To combine scores from two or more methods, the scores must be statistically comparable to each other. If the score distribution of one method is significantly different than that of another method, combining the scores in a simple linear fashion may not properly show the effect of aggregating the two pieces of information. For example, if for one of the methods the scores are either 1 or 0, but for another method the scores are continuous between 0.3 and 0.4, the first of the methods may “outweigh” the information provided by the other simply due to the scores being more extreme. We therefore devised a method to ensure the scores from different methods are comparable before combining. Firstly, the softmax function [27] was applied to each molecule for one of the methods, giving a new set of scores. Then, the average standard deviation of the new scores across all molecules was calculated. Finally, softmax was applied to the scores from the method to be combined, but now the softmax temperature parameter was chosen such that the resulting scores had the same average standard deviation as the initial method. This method is similar in some respects to the previously reported temperature scaling technique for scaling the output of machine learning classifiers to give well-calibrated probabilities [28]. The main difference is that the temperature scaling technique requires a validation dataset for probability calibration, which is not available in the present study. We therefore focused on aligning the two methods (DP4*/ACD and IR.Cai) to be consistent with each other, rather than to give accurate probabilities.

As the DP4* scores were often sharply peaked in favour of one of the structures, with other scores being negligible, the DP4* scores were adjusted by taking the natural logarithm before performing this softmax procedure. This ensured that there was good differentiation between molecules which both have low, but significantly different, scores. These small scores are valuable data: a molecule with a score of 10^{-5} is vastly more likely to be correct than a molecule with a score of 10^{-10} [29]. Full details of the softmax scaling procedure are given in the SI.

Declarations

Competing Interests

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 BR is part funded by AstraZeneca. JMG received funding from AstraZeneca to develop the IR.Cai algorithm.

Author contributions

RJL devised the project with contribution from JMG, TL and WC. 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. BR devised improvements to DP4 and the softmax normalization procedure, performed calculations and data analysis. BR, RJL, PH and JMG wrote the initial draft of the manuscript. All authors contributed to the final draft.

Acknowledgements

We thank M. Priessner, J. P. Janet, A. Tomberg and G. Hulthe for their discussions and AstraZeneca for funding support for BR.

References

1. Bifulco G, Dambruoso P, Gomez-Paloma L, Riccio R (2007) Determination of Relative Configuration in Organic Compounds by NMR Spectroscopy and Computational Methods, *Chem Rev*, vol. 107, no. 9, pp. 3744–3779, Sep. 10.1021/cr030733c
2. Golotvin SS, Pol R, Sasaki RR, Nikitina A, Keyes P (Jun. 2012) Concurrent combined verification: reducing false positives in automated NMR structure verification through the evaluation of multiple challenge control structures. *Magn Reson Chem* 50(6):429–435. <https://doi.org/10.1002/mrc.3818>
3. Burns DC, Mazzola EP, Reynolds WF (2019) The role of computer-assisted structure elucidation (CASE) programs in the structure elucidation of complex natural products. *Nat Prod Rep* 36(6):919–933. 10.1039/C9NP00007K

4. Buevich AV, Elyashberg ME (Dec. 2016) Synergistic Combination of CASE Algorithms and DFT Chemical Shift Predictions: A Powerful Approach for Structure Elucidation, Verification, and Revision. *J Nat Prod* 79(12):3105–3116. 10.1021/acs.jnatprod.6b00799
5. Smith SG, Goodman JM (2010) Assigning Stereochemistry to Single Diastereoisomers by GIAO NMR Calculation: The DP4 Probability, *J Am Chem Soc*, vol. 132, no. 37, pp. 12946–12959, Sep. 10.1021/ja105035r
6. Howarth A, Goodman JM (2022) The DP5 probability, quantification and visualisation of structural uncertainty in single molecules. *Chem Sci* 13(12):3507–3518. 10.1039/D1SC04406K
7. Richardson J, Sharman G, Martínez-Olíd F, Cañellas S, Gomez JE (2020) Unlocking the potential of late-stage functionalisation: an accurate and fully automated method for the rapid characterisation of multiple regioisomeric products. *React Chem Eng* 5(4):779–792. 10.1039/C9RE00431A
8. Rodríguez Martín-Aragón V, Trigo Martínez M, Cuadrado C, Daranas AH, Fernández A, Medarde, Sánchez López JM (2023) OSMAC Approach and Cocultivation for the Induction of Secondary Metabolism of the Fungus *Pleotrichocladium opacum*, *ACS Omega*, vol. 8, no. 42, pp. 39873–39885, Oct. 10.1021/acsomega.3c06299
9. Zhang F-Z, Li X-M, Meng L-H, Wang B-G (2023) A new steroid with potent antimicrobial activities and two new polyketides from *Penicillium variabile* EN-394, a fungus obtained from the marine red alga *Rhodomela confervoides*. *J Antibiot (Tokyo)*. 10.1038/s41429-023-00666-3
10. Pan C et al (2023) Amoxetamide A, a new anokis inducer, produced by combined-culture of *Amycolatopsis* sp. and *Tsukamurella pulmonis*. *J Antibiot (Tokyo)*. 10.1038/s41429-023-00668-1
11. Zhang J et al (Jan. 2020) NMR-TS: de novo molecule identification from NMR spectra. *Sci Technol Adv Mater* 21(1):552–561. 10.1080/14686996.2020.1793382
12. Huang Z, Chen MS, Woroch CP, Markland TE, Kanan MW (2021) A framework for automated structure elucidation from routine NMR spectra. *Chem Sci* 12(46):15329–15338. 10.1039/D1SC04105C
13. Cortés I, Cuadrado C, Hernández A, Daranas, Sarotti AM (2023) Machine learning in computational NMR-aided structural elucidation, *Frontiers in Natural Products*, vol. 2, [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fntpr.2023.1122426>
14. Lam J, Lewis RJ, Goodman JM (2023) Interpreting vibrational circular dichroism spectra: the Cai-factor for absolute configuration with confidence. *J Cheminform* 15(1):36. 10.1186/s13321-023-00706-y
15. Pesek M, Juvan A, Jakoš J, Košmrlj J, Marolt M, Gazvoda M Database Independent Automated Structure Elucidation of Organic Molecules Based on IR, ¹H NMR, ¹³C NMR, and MS Data. *J Chem Inf Model*, 61, 3, pp.756–763, 10.1021/acs.jcim.0c01332
16. Bösel L, Dötzer R, Steiner S, Stritzinger M, Salzmann S, Riniker S (2020) Determining the Regiochemistry and Relative Stereochemistry of Small and Druglike Molecules Using an Alignment Algorithm for Infrared Spectra, *Anal Chem*, vol. 92, no. 13, pp. 9124–9131, Jul. 10.1021/acs.analchem.0c01399

17. Alberts M, Laino T, Vaucher AC (2023) Leveraging Infrared Spectroscopy for Automated Structure Elucidation. ChemRxiv
18. Spectrus Processor Suite, version 2022.2.3, Advanced Chemistry Development Inc. (ACD/Labs), Toronto, ON, Canada, www.acdlabs.com
19. Bösel L, Aerts R, Herrebout W, Riniker S (2023) Improving the IR spectra alignment algorithm with spectra deconvolution and combination with Raman or VCD spectroscopy. *Phys Chem Chem Phys* 25(3):2063–2074. 10.1039/D2CP04907D
20. Cotter E, Pultar F, Riniker S, Altmann K-H (Mar. 2024) Experimental and Theoretical Studies on the Reactions of Aliphatic Imines with Isocyanates. *Chem – Eur J* 30:e202304272. <https://doi.org/10.1002/chem.202304272>
21. Nolvachai Y et al (2021) Nov., Structure Elucidation Using Gas Chromatography – Infrared Spectroscopy/Mass Spectrometry Supported by Quantum Chemical IR Spectrum Simulations, *Anal Chem*, vol. 93, no. 46, pp. 15508–15516, 10.1021/acs.analchem.1c03662
22. Shen J, Zhu C, Reiling S, Vaz R (2010) A novel computational method for comparing vibrational circular dichroism spectra. *Spectrochim Acta Mol Biomol Spectrosc* 76(3):418–422. <https://doi.org/10.1016/j.saa.2010.04.014>
23. Merrick JP, Moran D, Radom L (2007) An Evaluation of Harmonic Vibrational Frequency Scale Factors, *J Phys Chem A*, vol. 111, no. 45, pp. 11683–11700, Nov. 10.1021/jp073974n
24. Kesharwani MK, Brauer B, Martin JML (2015) Frequency and Zero-Point Vibrational Energy Scale Factors for Double-Hybrid Density Functionals (and Other Selected Methods): Can Anharmonic Force Fields Be Avoided? *J Phys Chem A*, vol. 119, no. 9, pp. 1701–1714, Mar. 10.1021/jp508422u
25. Golotvin SS, Vodopianov E, Lefebvre BA, Williams AJ, Spitzer TD (May 2006) Automated structure verification based on ¹H NMR prediction. *Magn Reson Chem* 44(5):524–538. <https://doi.org/10.1002/mrc.1781>
26. Frisch MJ et al Gaussian 16, Revision B.01. Gaussian, Inc., Wallingford CT
27. Bridle JS (1990) Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition, in *Neurocomputing*, F. F. Soulié and J. Héroult, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 227–236
28. Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On Calibration of Modern Neural Networks, in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., in *Proceedings of Machine Learning Research*, vol. 70. PMLR, Mar. pp. 1321–1330. [Online]. Available: <https://proceedings.mlr.press/v70/guo17a.html>
29. Hinton G, Vinyals O, Dean J (2015) Distilling the Knowledge in a Neural Network

Tables

Table 1 is available in the Supplementary Files section.

Figures

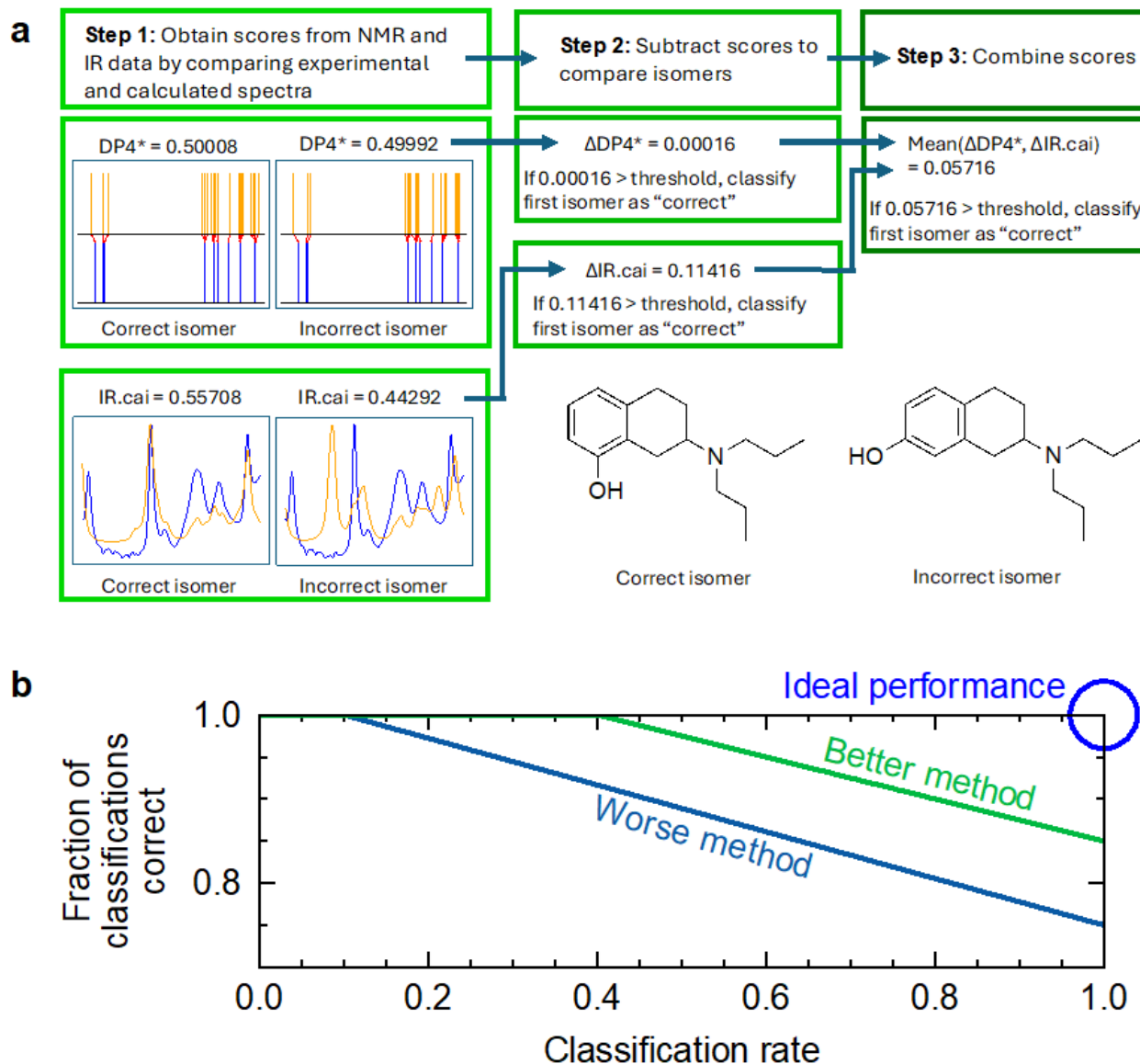


Figure 1

a Scheme showing how the DP4* and IR.Cai scores are used to classify a pair of isomers, where one of them is known to be correct. The values displayed in the figure are the softmax-scaled values obtained using the procedure described in Section 4.5. **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 right corner of the plot, denoted by the blue circle. This ideal SCC curve would have an AOC (area over curve) of 0.

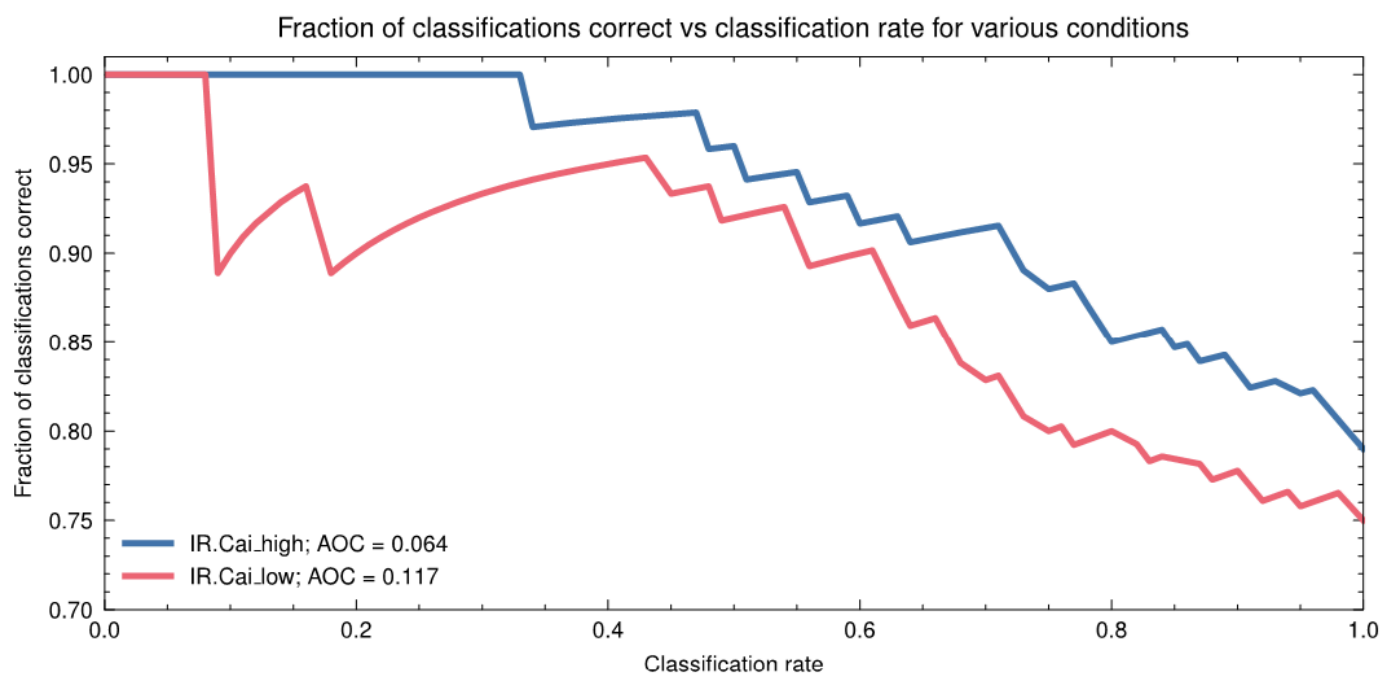


Figure 2

Structure classification characteristic (SCC) curve 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 lower AOC indicates better performance for the higher theory level.

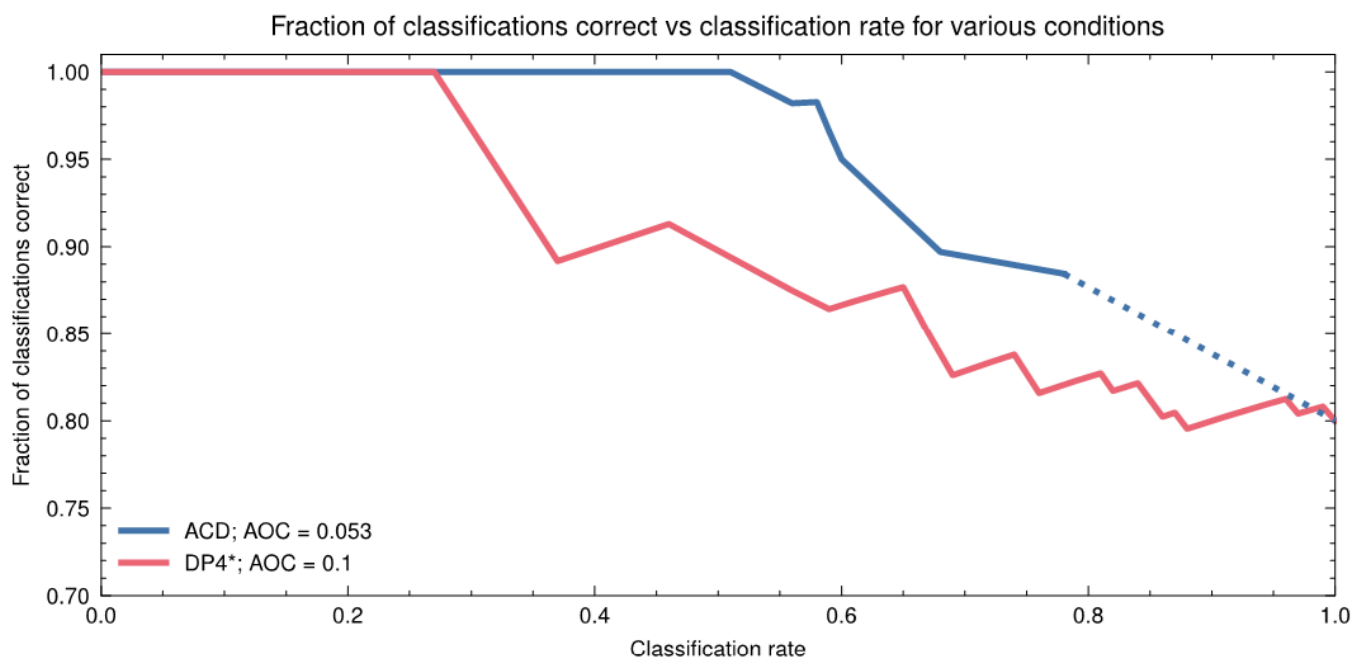


Figure 3

Structure classification characteristic (SCC) curve 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 of the comparisons. 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.

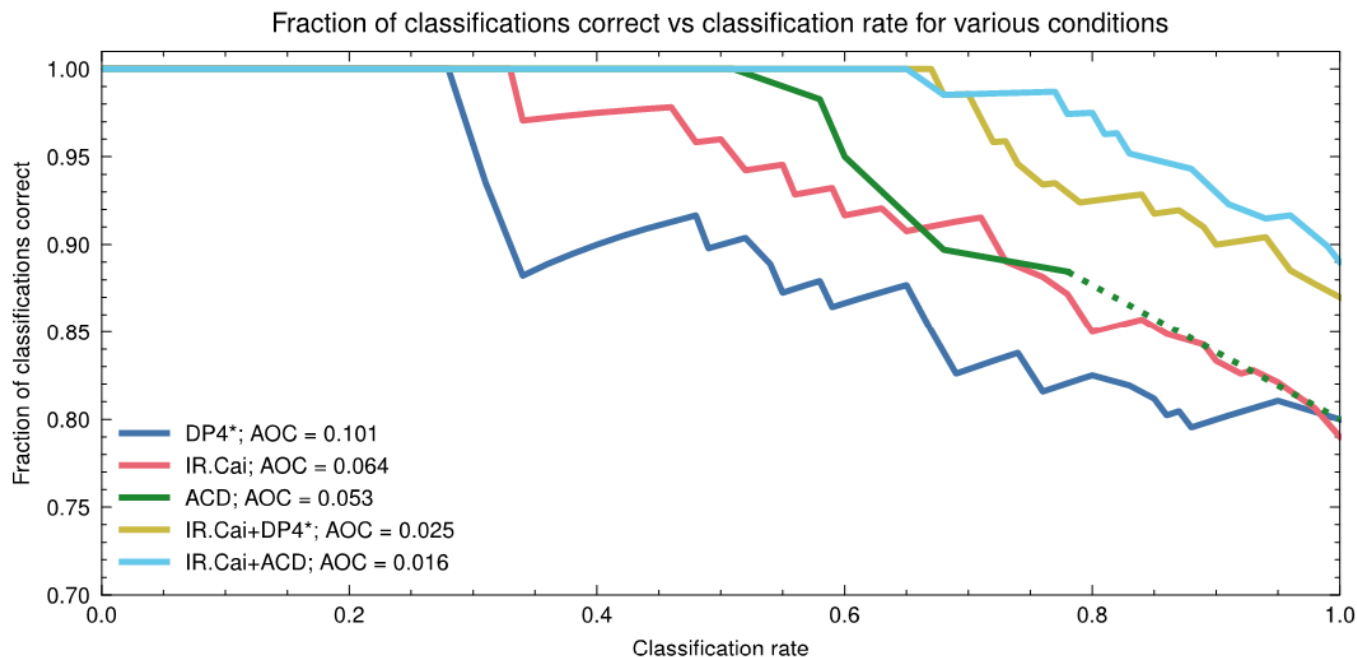


Figure 4

SCC curve for high-level IR, DP4*, ACD and the combinations of DP4* and ACD with IR. The DP4* and ACD scores are scaled to the IR scores following the scaling procedure described in Section 4.5. Corresponding SCC curves for combination with low-level IR are shown in the supporting information (Figure S5).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Table1.docx](#)
- [SupportingInformationFinal.docx](#)