

Cite this: *Chem. Sci.*, 2025, 16, 21590 All publication charges for this article have been paid for by the Royal Society of ChemistryReceived 5th September 2025  
Accepted 8th October 2025

DOI: 10.1039/d5sc06866e

rsc.li/chemical-science

# Towards automatically verifying chemical structures: the powerful combination of $^1\text{H}$ NMR and IR spectroscopy

J. Benji Rowlands,<sup>a</sup> Lina Jonsson,<sup>a</sup> Jonathan M. Goodman,<sup>b</sup> Peter W. A. Howe,<sup>c</sup> Werngard Czechtizky,<sup>a</sup> Tomas Leek<sup>a</sup> and Richard J. Lewis<sup>a\*</sup>

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.<sup>1</sup> 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)<sup>2,3</sup> and Computer-Assisted Structure Elucidation (CASE).<sup>4,5</sup> 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  $^1\text{H}$  NMR spectrum to confirm the product.

Among the best-established methods in the ASV category are the DP4 and DP5 probabilities.<sup>6,7</sup> 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.<sup>8–11</sup>

A related area of current research is the application of machine learning to automated structure elucidation.<sup>12–15</sup> 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.<sup>16–18</sup> But these may also be applied to data which a human cannot easily interpret. IR would

<sup>a</sup>Department of Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (R&I), Biopharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden. E-mail: richard.j.lewis@astrazeneca.com

<sup>b</sup>Centre for Molecular Informatics, Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, CB2 1EW, UK

<sup>c</sup>Oncology Chemistry, AstraZeneca, 1 Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0AA, UK



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 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  $^1\text{H}$  NMR and 2D  $^1\text{H}$ - $^{13}\text{C}$  HSQC spectra.<sup>19</sup> Separately, the use of IR for structure determination has been reported recently for building up molecules in a fragment-based approach,<sup>20</sup> determining regio- and stereochemistry by matching experimental and calculated IR spectra,<sup>21</sup> in machine-learning models that can generate complete structures by supplementing NMR data with functional group identification from IR<sup>22</sup> or from IR data alone.<sup>23,24</sup>

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,<sup>25</sup> 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  $^1\text{H}$  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  $^1\text{H}$  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).<sup>26</sup>

### 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 Four example molecules from the test set used to evaluate the use of IR and NMR in ASV

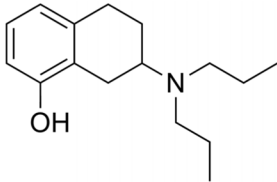
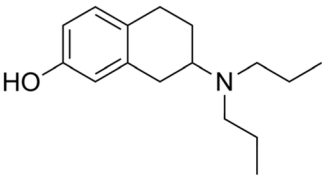
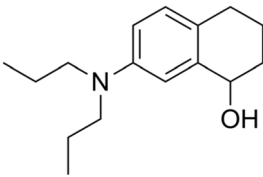
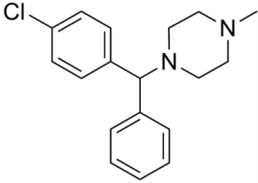
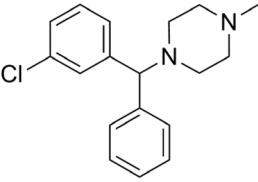
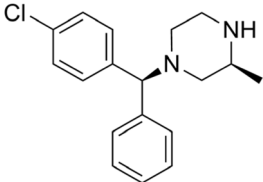
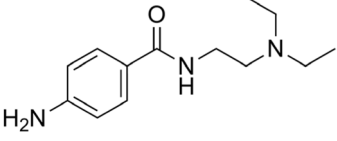
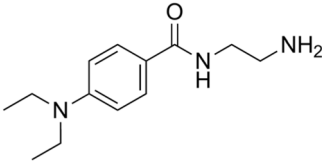
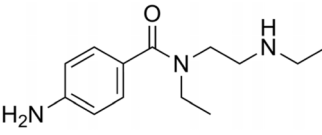
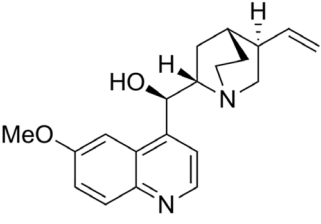
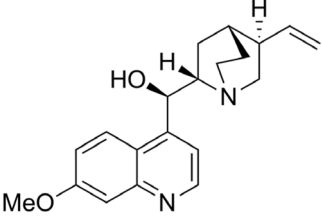
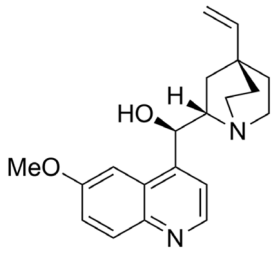
Compound	Correct structure	Incorrect isomer 1	Incorrect isomer 2
1			
		Aromatic regioisomer	Unclassified regioisomer
2			
		Aromatic regioisomer	Unclassified regioisomer
3			
		Unclassified regioisomer	Unclassified regioisomer
4			
		Aromatic regioisomer	Aliphatic regioisomer

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

### Receiver operating characteristic (ROC)

The IR.Cai matching algorithm, DP4\* and the <sup>1</sup>H ASV tool from Advanced Chemistry Development, Inc. (ACD/Labs)<sup>27</sup> 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. Fig. 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



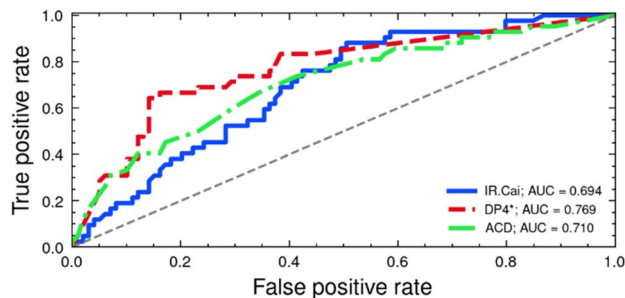


Fig. 1 ROC plot based on the 42 molecules in the test set showing the performance of DP4\* and IR.Cai for classifying structures as correct or incorrect. Note that the DP4\* metric is comparative, *i.e.* it requires a list of possible candidate structures and scores them relative to one another, unlike 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.

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. Fig. 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  $^1\text{H}$  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 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. Fig. 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.<sup>28</sup>

To combine IR and NMR data, we used a percentile-rank combination procedure similar to that described by Hsu and Taksa.<sup>29</sup> 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:

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

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

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





Fig. 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.

where  $x_i$  is the raw score,  $\mu$  is the mean score and  $\sigma$  is the standard deviation of the scores. The effect of the additional  $\Delta$  term is simply to break ties using raw score information; the parameter  $\varepsilon$  is chosen so that the term is too small to affect the rank order of untied candidates. The specific value of  $\varepsilon$  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  $12\text{ cm}^{-1}$  to simulate the broad peaks found in experimental spectra.<sup>30</sup> The match score between the experimental and calculated spectra is then found by using the formula:

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

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  $1500\text{ cm}^{-1}$ , this







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

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).

## 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





- 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, <https://www.acdlabs.com>.
- 28 M. Feher, *Drug Discovery Today*, 2006, **11**, 421–428.
- 29 D. Frank Hsu and I. Taksa, *Inf. Retr.*, 2005, **8**, 449–480.
- 30 L. Bösel, 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. Cheminf.*, 2023, **15**, 36.
- 32 E. Cotter, F. Pultar, S. Riniker and K.-H. Altmann, *Chem.–Eur. J.*, 2024, **30**, 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. Cheminf.*, 2024, **16**, 20.
- 40 I. Cortés, C. Cuadrado, A. Hernández Daranas and A. M. Sarotti, *Frontiers in Natural Products*, 2023, **2**, 1122426.
- 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.*, 2024, **64**(12), 4613–4629.

