RSC INTEREST GROUP
NMR DISCUSSION

# Structure 2026: Automated Interpretation of Spectroscopy Data

**Friday 27 March 2026 • The Discovery Centre, AstraZeneca, Cambridge**

| | |
|---|---|
| **9.30** | **Registration and Coffee** |
| **10.30** | **Welcome and Introduction** |
| **10.35** | **Prof. Jonathan Goodman , University of Cambridge**<br>Molecules with confidence: DP4 and structural assignment |
| **11.05** | **Dr Marvin Alberts, IBM Research**<br>Spectra to Molecule: Language Models for Automated Structure Elucidation |
| **11.35** | **Dr Markus Blatter, Novartis**<br>Automated Structure Verification (ASV) in NMR - A full stack implementation at Novartis |
| **12.05** | **Lunch, Posters and Vendor Exhibition** |
| **13.30** | **Dr Jonathan Martens, Radboud University**<br>Infrared ion spectroscopy for structure elucidation - IR spectra from a mass spectrometer |
| **14.00** | **Prof. Kate Kemsley, University of East Anglia**<br>Automated Structural Assignment and Verification using AI-Predicted NMR Chemical Shifts |
| **14.30** | **Tea, Posters and Vendor Exhibition** |
| **15.00** | **Dr Richard Lewis, AstraZeneca**<br>Automatically verifying compounds in Pharmaceutical Discovery: Which approach; Which data? |
| **15.30** | **Prof. Jacqui Cole, University of Cambridge**<br>AI-driven Materials Characterisation |
| **16.00** | **Poster Prize and Close** |

ACD/Labs®

BRUKER

SciY

magritek

Jason

# Structure 2026 – Speaker Abstracts

## Molecules with confidence: DP4 and structural assignment

**Prof. Jonathan Goodman, University of Cambridge**

Calculating the NMR spectrum expected for a molecule is helpful in structural determination but only a part of the process. Starting from a calculated spectrum, what structural information can be determined with high confidence? The widely used DP4 method assigns the probability of assigning an experimental spectrum to each of a list of calculated spectra. This works very well, provided that the assumption that the correct structure is on the list is true. We have now developed DP5 which does not require this assumption: given one experimental spectrum and one trial structure, what is the probability they correspond? We are now developing a step beyond this: starting from an experimental spectrum and the wrong structure, what is the right structure? These studies are now being extended to other analytical measurements including IR spectroscopy.

# Spectra to Molecule: Language Models for Automated Structure Elucidation

**Dr Marvin Alberts, IBM Research & NCCR Catalysis, Switzerland**

Structure elucidation is integral in the day-to-day operation of any organic chemistry laboratory allowing the structure and composition of unknown substances to be determined. Most commonly this is achieved via different spectroscopic techniques. First among them Nuclear Magnetic Resonance (NMR), Infrared (IR) spectroscopy and Mass Spectrometry (MS). While the acquisition of the spectra has been largely automated, the analysis of them is not straightforward making the analysis of spectra, particularly in large quantities, a time-consuming and tedious undertaking.

We have developed AI-driven approaches to tackle these challenges. We demonstrated for the first time that molecular structures can be reliably predicted from IR spectra alone, a task previously impossible for human chemists due to the limited interpretable information present in IR spectra. This approach was extended to the deconvolution of IR spectra of mixtures, where we developed a method to predict the components present in a mixture based solely on the IR spectrum and a molecular formula. However, chemists routinely have access to multiple analytical instruments. To leverage this, we developed a multimodal framework integrating NMR and IR spectra, which achieved performance matching expert human chemists in structure elucidation tasks. Together, these advances point toward a future where AI tools can aid and assist chemists in structure elucidation, freeing them to focus on interpretation and discovery rather than routine analysis.

# Automated Structure Verification (ASV) in NMR - A full stack implementation at Novartis

**Dr Markus Blatter, Novartis**

In medicinal chemistry, fast and reliable NMR data acquisition and interpretation are essential for structure verification and decision-making. Automated Structure Verification (ASV) using NMR data is thus emerging as a key asset in accelerated workflows. We present a full-stack ASV implementation that seamlessly integrates open-access NMR workflows, optimized for speed, robustness, and user acceptance. Performance is enhanced through a discriminative metric system, enabling differentiation of closely related chemical structures. The workflow begins with experiment selection and continues through acquisition and processing parameter optimization to ensure consistent data quality. A scalable data pipeline — featuring parallel export protocols and server-based ASV nodes — provides rapid access to analytical results. To support adoption among experienced users, a tailored reporting system delivers clear and accessible ASV outputs. The modular architecture enables seamless integration of future technological advances, ensuring long-term sustainability and flexibility. Overall, we demonstrate an unattended, orchestrated workflow where all 1D and 2D NMR data generated on open-access instruments are automatically processed and analyzed by ASV (ACD/Labs). This high-throughput verification of chemical structures reduces manual workload and supports standardized, high-quality reporting.

# Infrared ion spectroscopy for structure elucidation - IR spectra from a mass spectrometer

**Dr Jonathan Martens, HFML-FELIX & Radboud University**

Infrared ion spectroscopy (IRIS) is an emerging mass spectrometry-based technique that combines the sensitivity and selectivity of mass spectrometry with the structural specificity of infrared spectroscopy. In IRIS, mass-selected ions are irradiated using a tunable infrared laser while isolated inside an ion trapping mass spectrometer. The resulting wavelength-dependent photofragmentation yields an infrared spectrum that provides direct information on functional groups and molecular connectivity, making IRIS highly complementary to conventional MS and tandem MS (MS/MS).

A key advantage of IRIS is its ability to deliver structural information directly from complex samples and at low analyte concentrations, where nuclear magnetic resonance spectroscopy is often impractical. In contrast to MS/MS fragmentation data, which remains difficult to predict reliably in silico, infrared spectra of candidate molecular structures can be accurately predicted using routine computational chemistry methods. Comparing these predicted spectra with experimental IR spectra of unknowns, provides tentative structure assignments without the need for chemical standards. More recently, machine-learning models for IR spectral prediction have further accelerated interpretation and opened routes toward more automated workflows.

In this presentation, I will introduce the principles of IRIS with a focus on practical implementation, applicability, and data interpretation. While the technique was originally developed using high-power, widely tunable infrared free-electron lasers available at large-scale facilities such as the FELIX Laboratory, ongoing advances in infrared laser technology are increasingly enabling IRIS on compact, table-top systems suitable for routine laboratory environments. Several case studies from biomarker discovery in inherited metabolic disorders will be presented to illustrate how IRIS resolves structural ambiguities that persist after LC-MS/MS analysis, including the differentiation of isomeric metabolites and the identification of unknown compounds.

# Automated Structural Assignment and Verification using AI-Predicted NMR Chemical Shifts

**Prof. Kate Kemsley, University of East Anglia**

In this talk, we present a fully automated framework for the structural assignment and verification of small molecules. The pipeline begins with full or partial multiplet analysis of high-resolution $^{1}H$, $^{13}C$, and HSQC experiments to extract chemical shifts, integrals, and substitution classes. Atom-level $^{1}H$ and $^{13}C$ predictions are generated using ensembles of graph convolutional neural networks of the message-passing class [1]. Trained on chemical shift annotations from hundreds of thousands of compounds, these models offer high prediction accuracy across a wide range of atomic environments. Assignment proceeds via a metaheuristic genetic algorithm, in which the optimal solution minimises the predictionobservation differences while accounting for prediction confidences, degrees of substitution, and integral information. The resultant assignment scores are expressed probabilistically using density functions that model empirical prediction-error distributions derived from a diverse reference collection of ~1500 NMR datasets. The approach is flexible and can handle spurious, poorly resolved, and missing signals in the experimental data. Further, all steps are automated and scalable to large-scale assignment and cross-verification.

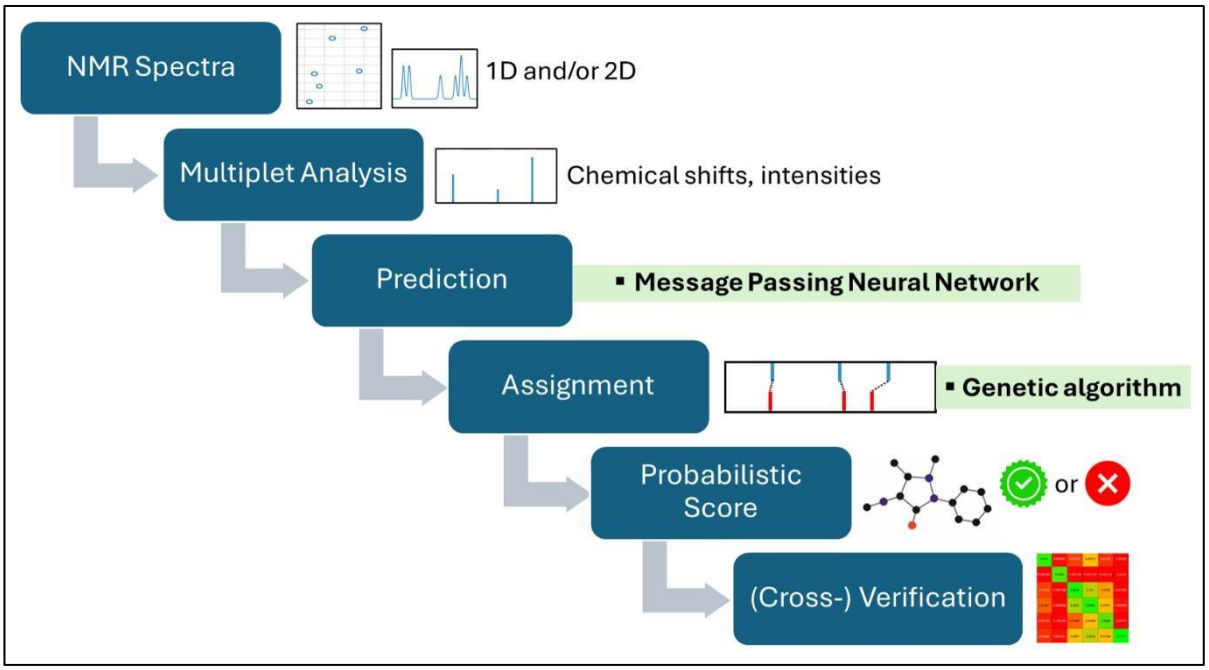



*Figure 1. Schematic of the assignment and verification pipeline*

[1] **D Williamson, S Ponte, I Iglesias, N Tonge, C Cobas, & EK Kemsley.** (2024). *Chemical shift prediction in $^{13}C$ NMR spectroscopy using ensembles of message passing neural networks (MPNNs)*, Journal of Magnetic Resonance, 368, 107795. doi: 10.1016/j.jmr.2024.107795

# Automatically verifying compounds in Pharmaceutical Discovery: Which approach; Which data?

**Dr Richard Lewis, Biopharmaceuticals R&D, AstraZeneca, Gothenburg**

Software to interpret analytical data has been in development for decades. Despite some successes they have not produced the step change in productivity needed to keep up with synthetic output that is increasingly coming from automated methods. However, current advances in AI and other methods make this field the most exciting and innovative that it has ever been.

The talk will start with an overview of methods for structure verification and determination and the use cases for each. We make the case that one of the early methods called automated structure verification (ASV) still has relevance amongst newer AI and machine-learning tools.

Previously, ASV has been exclusively applied to NMR data. Here we describe our work expanding ASV to infra-red data. Infra-red is a sensitive technique providing a structural fingerprint that requires less material than proton NMR. We describe an automated algorithm that scores experimental infrared spectra against ab initio calculated spectra. We investigate how well the infrared data can distinguish between isomeric compounds and compared this result to the performance of proton NMR. We find that infrared is as powerful as proton NMR at distinguishing the isomeric compounds and furthermore that the combined techniques outperform either proton NMR or infrared alone suggesting a complementarity in the information content of both. As part of this work, we developed a data visualization technique (the structure classification characteristic, SCC) which neatly summarizes in a visual and numeric manner how well particular analytical data can distinguish between candidate compounds. This can be applied to any analytical data that can score compounds and facilitate identification of the most valuable, time-efficient analytical data to collect.

The talk concludes by showing how machine learning models give valuable input to the ASV process by suggesting reaction products or suggesting alternative products consistent with the analytical data. We describe the currently implementation at AstraZeneca and how we are expanding this to IR data.

# AI-driven Materials Characterisation

**Prof. Jacqueline Cole, University of Cambridge**

This talk showcases the development and application of machine-learning models that automate the (a) spectral classification of raw data from molecular spectroscopy; (b) determination of size and shape of nanostructures from small-angle scattering data.

The former demonstrates the power and efficiency of convolutional neural networks (CNN) to automate the spectral classification of raw data acquired from Fourier transform infra-red (FTIR), 1H and 13C nuclear magnetic resonance (NMR) spectroscopy. The raw spectrum of an unknown organic chemical is the sole input to the CNN model.

The latter presents the development of a new transformer model, SASformer, that can classify the size and shape of a nanostructure using a raw small-angle scattering (SAS) profile as the sole input. This overcomes a key scientific bottleneck in the materials characterisation of nanostructures because conventional SAS data analysis requires the human to present guesses of the shape and size at the input stage to data analysis. The SASformer can automatically use the raw SAS profile of an unknown nanostructure to inform the human of the likely shape and size prior to conventional SAS data analysis. This data-driven scattering-model classification and parameter regression enables new capabilities in structure determination of nanomaterials as well as enhancing efficiency in SAS data analysis.

Overall, this portfolio of AI-driven classification of spectroscopy and scattering data offers several contributions towards the full automation of materials characterisation.

[1] G. Jung, S. G. Jung, J. M. Cole, “Automatic materials characterization from infrared spectra using convolutional neural networks”, Chem. Sci., 2023, 14, 3600-3609.

[2] S. Liu, J. M. Cole, “Automated Determination of the Molecular Substructure from Nuclear Magnetic Resonance Spectra Using Neural Networks”, J. Chem. Inf. Model. 2025, 65, 16, 8435-8447

[3] B. Yildirim, J. Doutch, J. M. Cole, “Multi-task scattering-model classification and parameter regression of nanostructures from small-angle scattering data”, Digital Discovery, 2024, 3, 694-704

# Structure 2026 – Poster Abstracts

## Approaches to automate NMR ID reference spectra generation and assay by NMR

**Steve Coombes[1], Matt Goodwin[1] & William Bourke[1]**

1. Chemical Development, Pharmaceutical Technology & Development, Operations, AstraZeneca, Macclesfield, SK10 2NA United Kingdom

Whilst data acquisition for NMR assay and characterisation experiments are fully automated within our labs, the data analysis process is still fairly manual. For NMR assay, it takes time to accurately integrate individual signals, collate MW and standard purity information and transcribed the values into the calculation to generate an analyte assay value. For NMR ID characterisation (i.e., generating reference $^{1}H$ and $^{13}C$ NMR spectra), the time taken to fully assigning a complete 2D NMR dataset depends on the skill of the user and the (increasingly) complex nature of the data. If we can speed up these data analysis processes, it gives us the opportunity to greatly reduce the amount of time it takes to both complete, and check these datasets, therefore, we have investigated approaches to automate and improve the quality of data processing in both areas.

We will share details about how we have automated data transfer and assay calculation in our corporate ELN and our efforts (successful and otherwise!) to automate peak picking and spectral assignment of full 2D datasets.

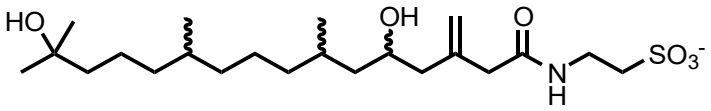
# Machine Learning and Data Science for the Assignment of Relative Stereochemistry of Flexible Natural Products

**Richard Cox[1], Max Deering[1], Malcolm George[1], Ben Morgan[1], Calvin Yiu[1], Andrew McCluskey[1], Varinder Aggarwal[1], Craig Butts[1]**

1. University of Bristol, School of Chemistry, Cantocks Close, Bristol, BS8 1TS

Determining the stereochemistry of natural products remains a significant challenge in structural elucidation. Conformationally flexible systems are particularly challenging and often require incredibly time-consuming methods such as X-ray crystallography or total synthesis. While J-based configurational analysis offers a solution, it requires the meticulous measurement of NMR parameters which is often labour-intensive and their interpretation error-prone.[1,2]

To simplify stereochemical analysis and overcome these limitations, we have developed an approach that integrates data science and machine learning to analyse high-dimensional relationships within experimental NMR data. This method enables accurate relative stereochemical assignment for flexible moieties prevalent in polyketide natural products, including 1,3-hydroxymethyl, dimethyl, and diol systems, with proof-of-concept extensions to 1,2- and 1,4-systems. Our approach has been validated on experimental data of compounds with known stereochemistry. We apply this approach to the Copepodamide family of marine natural products.[3]




1. N. Matsumori, D. Kaneno, M. Murata, H. Nakamura and K. Tachibana, J. Org. Chem., 1999, **64**(**3**), 866
2. T. Suyama, W. Gerwick, K. McPhail, Bioorg. Med. Chem., 2011, **19**(**22**), 6675
3. E. Selander, J. Kubanek, M. Hamberg, H. Pavia, Proc. Nat. Acad. Sci., 2015, **112**(**20**), 6395

# Identifying intermediates using ab initio and inline spectroscopic data

**Eugénie Fournier and Maxime Tarrago**
Pharmaceutical Technology & Development, AstraZeneca, Macclesfield, United Kingdom.

Controlling reactive intermediates is critical in route design and process development, yet many reagents are unstable and transient, limiting characterization of processes. The use of mid-infrared (IR)/Raman spectra predicted from first principles is established for spectral analysis and assignment; with the emergence of complimentary machine-learning methods (Alcolea Palafox, 2018; Neese et al., 2020; Sorrentino et al., 2025). We aimed to (i) develop a method combining IR and Raman spectroscopy with Quantum Mechanics (QM) to enable confident, in-situ identification of reactive and non-isolable reagents in complex mixtures, (ii) validate the method on real projects by demonstrating correct peak-reagents assignment and clarified kinetics, and (iii) demonstrate process impact.

Successful implementation of this technique was demonstrated on two projects, with computed-experimental agreement leading to identification of intermediate signals. The first project focused on obtaining the intermediate structure by comparing experimental IR ν(C=O) near 1780 cm-1 across several candidates. Using energy barrier analysis, a structure was selected and the kinetics obtained. The objective of the second project was to obtain kinetics for the degradation of intermediates into carbon monoxide to de-risk the process at large scale. Raman spectroscopy captured rapid formation of intermediates followed by slower decay. The computed peak intensity separation between the two intermediates closely matched the experimental spacing, enabling confident peak-species assignment despite spectral overlap. The QM-IR/Raman method enables confident, in-situ identification of non-isolable intermediates helping to clarify mechanistic pathways. This approach has substantial potential to reduce reliance on reagents isolation, cutting timelines from weeks to days.

**References**

Alcolea Palafox, M. (2018) DFT computations on vibrational spectra: Scaling procedures to improve the wavenumbers. Physical Sciences Reviews, 3(6). https://doi.org/10.1515/psr2017-0184.

Neese, F., Wennmohs, F., Becker, U. and Riplinger, C. (2020). The ORCA quantum chemistry program package. The Journal of Chemical Physics, 152(22), p.224108. https://doi.org/10.1063/5.0004608

Sorrentino, S., Gussoni, A., Calcagno, F., Pasotti, G., Avagliano, D., Rivalta, I., Garavelli, M. & Polli, D. (2025) Mol2Raman: a graph neural network model for predicting Raman spectra from SMILES representations. Digital Discovery, Advance Article. https://doi.org/10.1039/D5DD00210A.

# A protocol for validating small molecule structure assignment using calculated 13C NMR chemical shifts with quantum mechanics and MOE

**Andrew Henry, Joe Leonard, Alain Ajamian and Paul Labute**
Chemical Computing Group

Structural assignment of newly synthesized compounds or validation of newly assigned natural products with close isomeric relationships can be quite challenging, especially when the variations in the carbon framework configuration or stereochemistry is epimeric. Here we present a streamlined protocol for calculating and analyzing $^{13}C$ chemical shifts of close structurally related compounds. The calculated chemical shifts are then compared with experimental $^{13}C$ NMR values to determine and validate the correct structural assignment. The steps in the protocol are as follows: 1) conduct conformational search using LowModeMD, 2) refine conformations using a QM method (e.g. Gaussian), 3) calculate shieldings for each conformation with Gaussian and convert to chemical shifts, 4) determine the weighted Boltzmann distribution for $^{13}C$ chemical shifts, 5) compare the calculated $^{13}C$ NMR chemical shifts of multiple compound candidates with experimentally derived $^{13}C$ values to identify the best match using the NMR Spectral Analysis application in MOE.

# Assessing the reliability of automated structure verification by NMR based on information content and structural possibilities

**Peter W.A. Howe**
Early Oncology Chemistry & DMPK, AstraZeneca, 1 Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0AA

In automated structure verification by NMR, a sample is usually classified by matching predicted data to observed data. If the match is close, then the sample is classified as correct while if the match is poor then it is classified as incorrect. However, it is now clear that this approach is insufficient, because there is considerable overlap in match values between structures which are correct (true positives) and structures which are incorrect (false positives). A more robust approach is to assess several structures at once against the data - these structures can be generated manually, automatically, or by reaction prediction software. However, it is impossible to cover the expanse of possible structures without significant prior knowledge such as the synthetic route. This poster proposes an alternative method of assessing ASV, which is to consider the information content of the NMR data used in the ASV match calculation. The first measure is the Crews' Rule ratio, a commonly used metric in natural products discovery, which states that molecules with a H/C ratio of less than 1 are challenging to identify. A second useful metric is the proportion of atoms in the molecule that are observable by NMR scaled by the number of atoms. This measures the information available (chemical shifts) in relation to the number of potential structures. These two metrics are easily calculated from the molecular formula. Example molecules will be used to demonstrate the application of these metrics in assessing the potential use of ASV by NMR for confirming their structures.

# Vibrational Circular Dichroism (VCD) Spectroscopy for the Rapid Detection of Synthetic Adulteration.

**A. M. Jackson*[1], J. R. Bows[2], G. R. Nash[1]**

*[1]Natural Sciences, Faculty of Environment, Science and Economy, University of Exeter, EX4 4QF, United Kingdom*
*[2]PepsiCo R&D, Leicester, LE4 1ET, United Kingdom*
**Corresponding Author: aj675@exeter.ac.uk*

Monoterpenes like limonene are among the most abundant volatile constituents in the human diet and are the major components of citrus and conifer essential oils (EOs). With an estimated 80% of EOs adulterated globally, robust quality control processes are becoming increasingly important. Enantiomeric composition serves as a valuable authenticity marker because synthetic material cannot replicate the natural stereochemical signatures found in plant material. Conventional methods of enantiomeric quantification, including chromatography based techniques, are accurate but expensive and time consuming. Vibrational Circular Dichroism (VCD) spectroscopy has emerged as a promising alternative for EO authentication.[1] Although limonene's VCD profile is well documented qualitatively,[2] quantitative applications are limited. Prior reports of VCD based enantiomeric excess (ee) determination for other chiral monoterpenes have achieved ≤1% ee precision,[3] yet noise levels remain a limiting factor for VCD analysis, which typically rely on long acquisition times to achieve adequate signal to noise ratios (SNR). This constraint has limited the adoption of VCD to in-situ analysis of authenticity.

Using limonene as a model system, we demonstrate that combining automated SNR-based VCD spectral region selection with integrated multivariate linear modelling enhances quantitative performance compared to traditional univariate peak-based approaches. Predictive models were built using both 30-minute and 5-minute acquisition data. The 30-minute data yielded excellent model performance ($R^2 > 0.999$, RMSE = 1.58% ee), and notably, the 5-minute acquisition retained 99.7% of this performance while delivering a 6-fold increase in throughput. These results support the feasibility of near real-time VCD analysis. Moreover, because the multivariate framework integrates multiple independent spectral features, this approach is readily extensible to systems where several enantiomeric markers contribute to product authenticity. Overall, this work points to a broader opportunity for multi-marker profiling in EOs and represents a promising path towards practical, scalable chiral analysis across the food, flavour, and fragrance sectors.

1. Said, M. E. A. *et al.* Isolation of the major chiral compounds from Bubonium graveolens essential oil by HPLC and absolute configuration determination by VCD. *Chirality* **29**, 70–79 (2017).
2. Guo, C. *et al.* Fourier transform vibrational circular dichroism from 800 to 10,000 cm−1: Near-IR-VCD spectral standards for terpenes and related molecules. *Vib. Spectrosc.* **42**, 254–272 (2006).
3. Urbanová, M., Setnička, V. & Volka, K. Measurements of concentration dependence and enantiomeric purity of terpene solutions as a test of a new commercial VCD spectrometer. *Chirality* **12**, 161–294 (2000).

**John Bows is an employee of PepsiCo, Inc. The authors thank PepsiCo, Inc. for part funding this research. The views expressed in this abstract are those of the authors and do not necessarily reflect the position or policy of PepsiCo, Inc.**

# Generative Machine Learning for Automating Structure Elucidation in Synthesis

**Zheqi Jin[1], Mohammad Golbabaee[2], Craig Butts*[1]**

[1] School of Chemistry, University of Bristol, Cantock's Cl, Bristol, BS8 1TS
[2] School of Engineering Mathematics and Technology, University of Bristol, Ada Lovelace Building, Tankard's Cl, Bristol, BS8 1TW

**Keywords:** Machine Learning, Nuclear Magnetic Resonance, Structure Elucidation, Graph Neural Networks

Accurate elucidation of molecular structures from nuclear magnetic resonance (NMR) spectroscopy is of fundamental importance in chemical synthesis and drug discovery. Current methods, such as Computer-Aided Structure Elucidation (CASE)[1], typically rely on rule-based generation of large ensembles of candidate structures, which are often difficult to distinguish using ranking metrics. Machine learning methods offer an alternative approach by enabling models to learn the relationship between NMR spectroscopic data and molecular structures, thereby facilitating automated structure elucidation.

In this work, we present the inverse-IMPRESSION platform, which comprises four interconnected IMPRESSION-G2 Graph Transformer Network models[2]. Each model employs distinct node and edge features and is responsible for a specific stage of the structure elucidation workflow (Figure 1). Together, these components operate as a unified system to generate 2D molecular structures with assigned bond orders. The platform is trained using simulated NMR data for molecules containing up to 30 heavy atoms (H, C, N, O, and F), incorporating both chemical shifts and scalar coupling correlations. A noise-augmentation strategy is employed to generate multiple candidate structures for each prediction event, which are subsequently ranked according to their frequency of occurrence to identify the best-fit structure. Using this approach, a Top-1 accuracy of 77.6% is achieved on simulated test molecules. When applied directly to experimental NMR spectra, the platform attains a reconstruction accuracy of 47.4%.

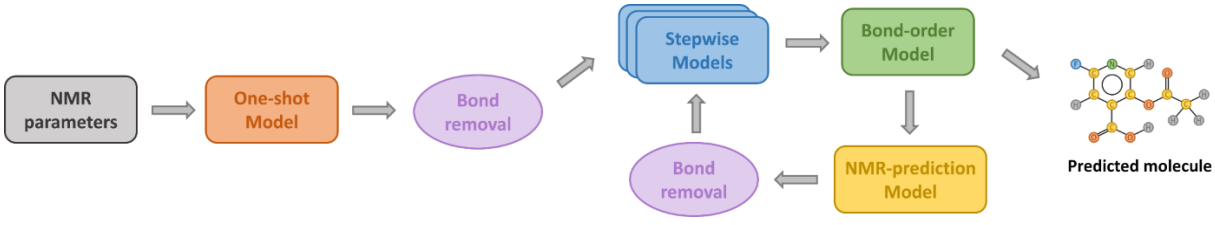



Fig 1. The workflow and architecture of the inverse-IMPRESSION platform

1. Elyashberg M, Argyropoulos D. Computer assisted structure elucidation (CASE): current and future perspectives. Magnetic Resonance in Chemistry. 2021;59(7):669-90.
2. Gerrard W, Bratholm LA, Packer MJ, Mulholland AJ, Glowacki DR, Butts CP. IMPRESSION–prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. Chemical Science. 2020;11(2):508-15.

# Small molecule automated structure verification using solely HSQC and HMBC datasets.

**Eric Hughes[1] and Alan Kenwright[1,2]**

We demonstrate the use of the simpleNMR[3] suite of software tools to facilitate the automated, verified assignment of 2D NMR spectra for small organic molecules. We show, in certain circumstances, the minimum dataset required is an HSQC and an HMBC. Using only HSQC data, the accuracy of the assignment relies on the quality of the carbon shift prediction. Introduction of the HMBC data allows quaternary carbons to be located for a near complete assignment but, more importantly, by taking advantage of the HMBC correlations across the molecule, the accuracy of the assignment can be clearly judged and, using an automated simulated annealing refinement step, an improved solution can be found if necessary.

This approach not only gives a much higher level of confidence in structure verification; it also ensures that correct assignments are provided for reporting in the literature and brings the realisation of a fully automated structure verification workflow a step closer.

1. Department of Chemistry, Durham University, Stockton Road, Durham DH1 3LE.
2. Retired.
3. E. Hughes, A. M. Kenwright, *Magn Reson Chem* 2024, 62(7), 556.
https://doi.org/10.1002/mrc.5441

simplenmr.pythonanywhere.com

# *Peering Inside the Gut*: Real-Time Spatially Resolved Analysis of Pharmaceutical Degradation in the Human GIT

**Trey T. Koev[1], Joseph Mauracheea[1], Fred J. Warren[2] and Matthew Wallace[1]**

1. School of Chemistry, Pharmacy and Pharmacology, University of East Anglia, Norwich Research Park, NR4 7TJ
2. Food, Microbiome and Health, Quadram Institute Bioscience, Norwich Research Park, NR4 7UQ

Understanding how pharmaceutical materials degrade within the human gastrointestinal (GI) tract is critical for optimising drug delivery, bioavailability, and therapeutic outcomes. Conventional *in vitro* or bulk sampling methods often fail to capture the complex, dynamic environment of the GI system, leaving a critical gap between formulation design and *in vivo* performance. We present a novel platform enabling real-time, spatially resolved monitoring of pharmaceutical material degradation within a simulated human GI tract. By integrating advanced chemical shift imaging NMR with *in situ* data acquisition, our approach provides unprecedented insight into where, when, and how degradation occurs along the digestive pathway. This capability not only reveals site-specific variations in excipient stability and dissolution but also informs predictive models of parenteral drug delivery. The methodology opens new avenues for personalised medicine, improved formulation strategies, and regulatory science, bridging a longstanding disconnect between laboratory testing and real-world physiological conditions.

1.

# Automated Absolute Configuration Determination of Solifenacin by Raman Optical Activity

**Dr. Monika Dengler, Melanie Kueffer, Petra Hoffman,**
**Dr. Alessia Portieri, Dr. Carin R. Lightner**

Enantios

The determination of absolute configuration (AC) is a critical requirement in pharmaceutical development, particularly for chiral active pharmaceutical ingredients where stereochemistry directly impacts efficacy and safety. We present an automated workflow for absolute configuration determination of solifenacin using Enantios' Raman Optical Activity (ROA) technology. Solifenacin, a chiral muscarinic receptor antagonist used in the treatment of overactive bladder, serves as a representative complex drug molecule with multiple conformations and flexible side chains.

The method integrates high-sensitivity ROA measurements with automated spectral processing and computational comparison to density functional theory (DFT) simulations. A dedicated analysis pipeline enables spectral matching, and objective configuration assignment with minimal user intervention. The workflow reduces manual data handling and interpretation, delivering reliable results in a streamlined "one-click" format.

The automated approach demonstrates robust discrimination between enantiomeric forms of solifenacin and highlights the capability of ROA as a powerful tool for absolute configuration determination in pharmaceutical research and quality control. This work illustrates the potential of automated ROA platforms to close a critical measurement gap in chiral drug development and manufacturing.

# Large-Scale Neural Chemical Shift Prediction from 1.3 Million Experimental NMR Spectra

**J. Benji Rowlands[1], Richard J. Lewis[2], Peter Howe[3], Jonathan M. Goodman[1,*]**

[1]Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, UK
[2]Medicinal Chemistry, R&D, AstraZeneca, Gothenburg, Sweden
[3]Oncology Chemistry, AstraZeneca, Cambridge, UK
*Corresponding author: jmg11@cam.ac.uk

Accurate nuclear magnetic resonance (NMR) chemical shift prediction remains limited by the small size and narrow chemical diversity of atom-assigned training datasets. Here, we address that bottleneck by leveraging the new NMRexp resource[1] with weak supervision, training CASCADE-style graph neural networks[2] with an optimal-assignment loss on approximately 1.2 million unassigned $^{13}C$ spectra and 1.4 million unassigned $^{1}H$ spectra alongside conventional assigned data. This joint training strategy substantially improves performance over assigned-only baselines: on held-out NMRexp data, $^{13}C$ mean absolute error (MAE) falls from 1.2375 to 0.7039 ppm and $^{1}H$ MAE from 0.2835 to 0.0933 ppm, while $^{13}C$ performance on the assigned Exp22K benchmark[2] improves from 0.8150 to 0.5938 ppm. These gains also translate into better structure verification performance, with hard-decoy top-1 retrieval improving from 0.8882 to 0.9430 for $^{13}C$ and from 0.7069 to 0.9025 for $^{1}H$. Taken together, these results show that large-scale unassigned literature spectra can provide an effective training signal for more accurate and practically useful NMR prediction models.

## References

[1] Jun-Jie Wang, Yongqi Jin, Chen-Yu Zhi, Yu-Jie Liu, Xu-Hao Huang, Fanjie Xu, Xiaohong Ji, Xi Fang, Haoyi Tao, Weinan E, Linfeng Zhang, Guolin Ke, and Rong Zhu. NMRexp: A database of 3.3 million experimental NMR spectra. *Sci. Data*, 12(1):1954, December 2025.

[2] Abhijeet Bhadauria, Zhitao Feng, Mihai Popescu, and Robert Paton. CASCADE-2.0: Real Time Prediction of 13C-NMR Shifts with sub-ppm Accuracy. *ChemRxiv*, 2025(0728).

# MolDeTr: A Chemistry-Informed Deep Learning Model for Next-Generation Automated Analysis of $^{1}$H NMR Spectra

**Nicolas Schmid[1,2,3], Marc Wanner[4], Giulia Fischetti[1,5], Andreas Henrici[1], Mohsen Meshkian[1], Simon Bruderer[6], Rudolf M Füchslin[1,7], Bjoern Heitmann[6], Jan Dirk Wegner[2], Roland K O Sigel[3], Dirk Wilhelm[1]**

1. Institute of Applied Mathematics and Physics (IAMP) Zurich University of Applied Sciences (ZHAW)
2. Department of Mathematical Modeling and Machine Learning (DM3L) University of Zurich (UZH)
3. Department of Chemistry University of Zurich (UZH)
4. ETH Zurich
5. Ca' Foscari University of Venice
6. Bruker Switzerland AG
7. European Centre for Living Technology (ECLT)

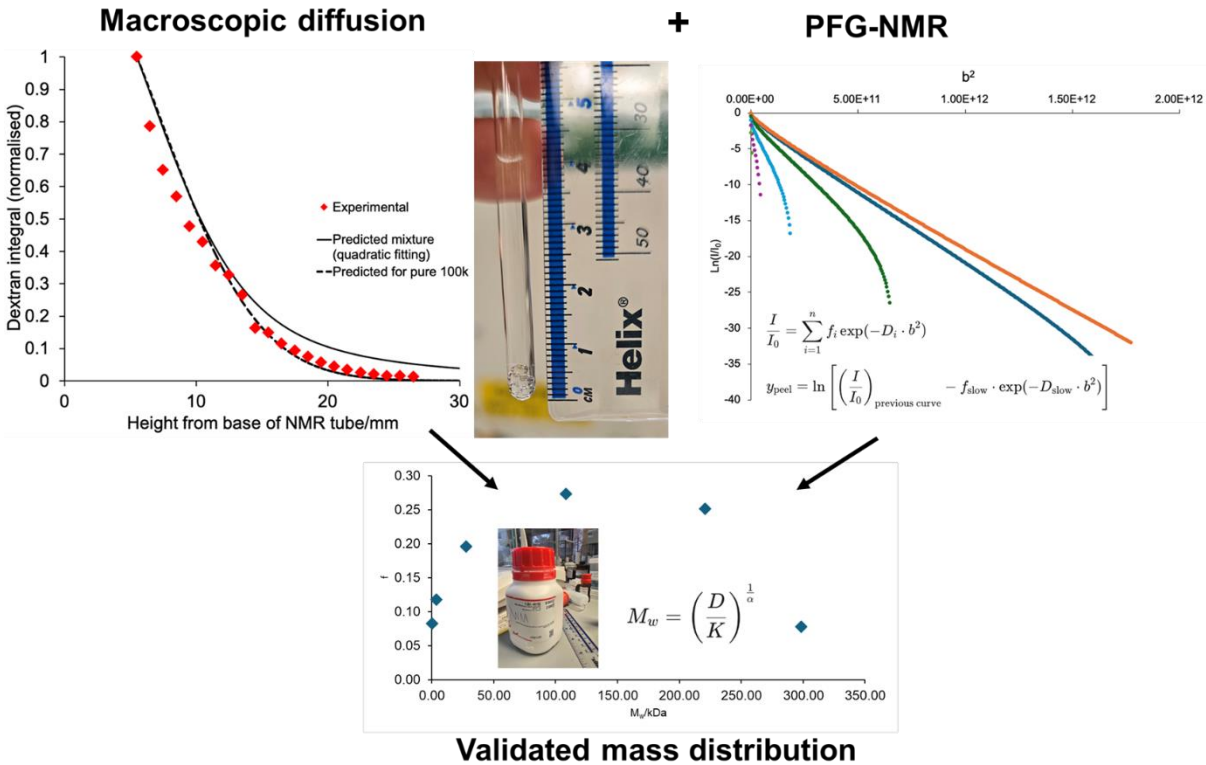
Accurate interpretation of one-dimensional proton nuclear magnetic resonance ($^{1}$H NMR) spectra remains a rate-limiting step in molecular structure elucidation, particularly when severe signal overlap, strong spin coupling, and instrumental distortions mask key features. Existing automated approaches depend on computationally intensive and sensitive iterative quantum-mechanical fitting and still require expert oversight. Here we introduce MolDeTr, a chemistry-informed deep-learning framework derived from the detection-transformer architecture that unifies peak picking, multiplet identification, and extraction of chemical shifts, scalar coupling constants, relaxation-dependent decay times, and proton counts in a single-network pass. The method targets prototypical spin systems of single-component small molecules in 1D $^{1}$H NMR, with up to ten distinct groups of chemically equivalent spins (multiplets). MolDeTr is trained exclusively on synthetic spectra generated by spin-dynamics simulations and augmented with realistic experimental artefacts, enabling it to generalize to unseen compounds—including experimental spectra with crowded and strongly coupled multiplets—without reference standards or prior spin-system knowledge. Unlike structure-conditioned shift-prediction or calculation models , e.g., density functional theory (DFT), that assume the molecular structure is known, MolDeTr addresses the spectrum-conditioned inverse problem and extracts spin-system parameters directly from measured 1D $^{1}$H NMR spectra, thereby substantially improving chemical-shift prediction precision by one to two orders of magnitude compared to existing structure-conditioned approaches. Benchmarking against challenging experimental $^{1}$H NMR spectra of small molecules shows median absolute errors of 0.89 Hz for chemical shifts and 0.20 Hz for coupling constants, while absolute proton counts are predicted with 93.5 % accuracy, outperforming state-of-the-art spectrum analysis software and experienced spectroscopists. Owing to its physics-guided, modality-agnostic design, MolDeTr can be extended to other spectroscopic and diffraction techniques. By eliminating iterative fitting and expert intervention, MolDeTr offers a scalable route to fully automated spectral analysis, accelerating molecular discovery across the chemical sciences.

# The polymer onion: a new view on identification of polymer sizes using PFG-NMR

**Vangelis Theodoratos and Matthew Wallace**
School of Chemistry, Pharmacy and Pharmacology, University of East Anglia, Norwich, UK, NR4 7TJ

Identifying a polymer's molecular weight *via* techniques such as size exclusion chromatography is a time-consuming, resource-intensive task. By utilizing a "peeling off" technique for the analysis of PFG-NMR datasets, we can calculate the molecular weight distribution of a polymer with simple quadratic fitting *via* an Excel spreadsheet. This self-consistent method streamlines the characterisation process and provides accuracy within one order of magnitude. Furthermore, we can verify any estimation of weight or diffusion coefficients from PFG-NMR and test them by studying the macroscopic diffusion of the sample up an NMR tube.

This simple, cheap and fast approach yields approximate distributions of diffusion coefficients without complicated iterations. We have fitted a semilogarithmic plot of signal intensity (I) vs gradient parameters (b) to a quadratic function using the LINEST function of Excel, where the x term accounts for the polydispersity of the sample. We have used the quadratic function to generate a plot of log(I) vs $b^2$. By extending this plot to very high values of $b^2$, the plot becomes almost equal to the slope and intercept of the slowest component.[1] This component is subtracted from the quadratic function and the process is repeated until a full set of diffusion coefficients and populations is obtained. The results can then be validated *via* macroscopic diffusion by chemical shift imaging (CSI).[2]




1. S. D. Foss, *Biometrics*, 1969, **25**, 580-584.
2. S. Monaco, J. A. Angulo and M. Wallace, *J. Am. Chem. Soc.*, 2023, **145**, 16391.

# Reference-free structure elucidation of Fludipink: A challenging Industrial Impurity using gas-phase IR spectra and NMR shifts combined with DFT predictions.

**Teun van Wieringen,[1] Agisilaos Chantzis,[2] Louise Bacon,[2] Mansoor Saeed,[2] Jos Oomens,[1] Giel Berden,[1] Simon Perry,[2] Jonathan Martens[1]**

1. HFML-FELIX, Nijmegen, The Netherlands; Radboud University, Nijmegen, The Netherlands
2. Syngenta, Bracknell, United Kingdom

Quantum chemistry offers us powerful tools for predicting spectral properties that can give additional confidence in the identification of unknown molecular structures when compared to experimental datasets.[1-3] This reference-free approach is crucial when traditional spectral databases lack entries or established standards for novel or complex compounds are unavailable. In this poster, we describe the characterisation of a coloured synthetic impurity, referred to as fludipink (570 amu) isolated from an industrial synthetic batch sample. Due to limited sample quantity and purity, direct $^{13}$C-NMR assignments were unfeasible. Instead, 2D-NMR (HSQC and HMBC) was used to map out the carbon backbone. Additionally, mass spectrometry and Infrared Ion Spectroscopy (IRIS) were used to further characterise the impurity. IRIS is a gas-phase spectroscopy technique, that measures the IR spectrum of mass-selected ions in an ion trapping mass spectrometer.

Based on the analytic data available, a set of potential candidate structures was proposed. These candidates were evaluated by comparing the experimental IRIS and NMR data with density functional theory (DFT) predicted spectra, enabling the selection of a final candidate. This workflow demonstrates the strength of combining multiple spectroscopic methods with computational chemistry to solve complex molecular characterisation challenges.

References:

(1) van Outersterp, R. E.; Houthuijs, K. J.; Berden, G.; Engelke, U. F.; Kluijtmans, L. A. J.; Wevers, R. A.; Coene, K. L. M.; Oomens, J.; Martens, J. Reference-standard free metabolite identification using infrared ion spectroscopy. *Int. J. Mass spectrom.* **2019**, *443*, 77-85. DOI: 10.1016/j.ijms.2019.05.015.
(2) Houthuijs, K. J.; Berden, G.; Engelke, U. F. H.; Gautam, V.; Wishart, D. S.; Wevers, R. A.; Martens, J.; Oomens, J. An *In Silico* Infrared Spectral Library of Molecular Ions for Metabolite Identification. *Anal. Chem.* **2023**, *95* (23), 8998-9005. DOI: 10.1021/acs.analchem.3c01078.
(3) Xin, D.; Sader, C. A.; Chaudhary, O.; Jones, P.-J.; Wagner, K.; Tautermann, C. S.; Yang, Z.; Busacca, C. A.; Saraceno, R. A.; Fandrick, K. R.; et al. Development of a 13C NMR Chemical Shift Prediction Procedure Using B3LYP/cc-pVDZ and Empirically Derived Systematic Error Correction Terms: A Computational Small Molecule Structure Elucidation Method. *J. Org. Chem.* **2017**, *82* (10), 5135-5145. DOI: 10.1021/acs.joc.7b00321.

# Latent Diffusion-Based 3D Molecular Recovery from Vibrational Spectra

**Wenjin Wu,[1] Aleš Leonardis,[1] Linjiang Chen,[1,2] Jianbo Jiao[1]**

[1]University of Birmingham and [2]University of Science and Technology of China

Infrared (IR) spectroscopy, a type of vibrational spectroscopy, is widely used for molecular structure determination and provides critical structural information for chemists. However, existing approaches for recovering molecular structures from IR spectra typically rely on one-dimensional SMILES strings or two-dimensional molecular graphs, which fail to capture the intricate relationship between spectral features and three-dimensional molecular geometry. Recent advances in diffusion models have greatly enhanced the ability to generate molecular structures in 3D space. Yet, no existing model has explored the distribution of 3D molecular geometries corresponding to a single IR spectrum. In this work, we introduce **IR-GeoDiff**, a latent diffusion model that recovers 3D molecular geometries from IR spectra by integrating spectral information into both node and edge representations of molecular structures. We evaluate IR-GeoDiff from both spectral and structural perspectives, demonstrating its ability to recover the molecular distribution corresponding to a given IR spectrum. Furthermore, an attention-based analysis reveals that the model is able to focus on characteristic functional group regions in IR spectra, qualitatively consistent with common chemical interpretation practices.

Link to project page: https://wenjin886.github.io/IR-GeoDiff/

# HQSpectrum: Automated Simulation and Analysis of NMR spectra

**A. Zech, S. Álvarez-Barcia, A. Vazigin, N. Enenkel, H. Rittich, F. Covito, J. B. Kleine Büning, M. Hodecker, P. Pinski, P. Schmitteckert, I. Schwenk, S. Zanker**

HQS Quantum Simulations GmbH, Rintheimer Str. 23, 76131 Karlsruhe, Germany

Accurate simulation of NMR spectra is vital for the rigorous evaluation of experimental findings. This is especially true for low-field benchtop NMR spectrometers, but it also applies to high-field spectra. We present HQSpectrum,[i] our cloud-based software platform[ii] for NMR spectrum simulation and analysis. Given a set of chemical shifts and J-coupling constants, the spectrum simulation engine efficiently calculates NMR spectra. By utilizing symmetry considerations and clustering approximations that minimize truncation errors, spectral calculations are possible even for larger or complex molecules featuring tens of spins.

Chemical shifts and J-coupling constants available on the platform were calculated employing an in-house workflow based on density functional theory, which automatically explores the conformational flexibility of the compounds of interest. Additionally, we have developed a modern, fragment-based approach for fast prediction of chemical shifts and J-couplings alike, inspired by the seminal HOSE code method. Trained on NMR parameters from thousands of molecules calculated with the aforementioned approach, this model predicts chemical shifts and J-coupling constants in a matter of seconds and is likewise accessible on HQSpectrum.

Our solution resolves individual nuclear spin contributions, effectively disentangling overlapping signals and thus enabling accurate interpretation of spectra. Peak positions can be automatically adjusted to better align with experimental data, and linewidths can be tuned accordingly. A quantitative similarity score between the simulated and experimental spectrum can be computed, serving as an objective measure of fit quality. These features are especially valuable for spectra acquired on benchtop NMR devices, where broad multiplets and peak overlap are common challenges. Thus, HQSpectrum can assist classical problems such as structure elucidation as well as more complex tasks like quality control and mixture analysis.[iii]

---

[i] https://quantumsimulations.de/hqspectrum
[ii] https://cloud.quantumsimulations.de/hqspectrum
[iii] https://arxiv.org/pdf/2506.15426