Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2026, **5**, 1023

Advances and perspectives in computer-assisted structure elucidation: a review

Dagny Aurich and Emma L. Schymanski *

Computer-Assisted Structure Elucidation (CASE) is a powerful yet underused approach in chemistry to determine molecular structures from experimental data without necessarily being restricted to the contents of chemical databases. This review provides a comprehensive overview of the current state of CASE, encompassing methodologies, computational techniques, applications, challenges, and future directions. The historical evolution of CASE tools is traced, highlighting key milestones and influential technologies. Moreover, the methodologies employed in CASE, including reduction and assembly methods, as well as hybrid approaches, are examined. Special attention is given to the integration of analytical data, such as NMR, MS, and IR, into CASE algorithms, along with computational techniques such as machine learning approaches. Through a series of case studies and real-world applications, the utility of CASE tools in drug discovery, natural products chemistry, environmental sciences, and metabolomics is illustrated. Despite advancements, challenges persist in handling complex molecular structures, improving algorithm accuracy, integrating heterogeneous data sources, benchmarking and reconciling diverse programming languages, alongside the mixture of open *vs.* closed source developments. Looking ahead, emerging trends and future directions in CASE are identified, including rapid developments with the adoption of deep learning and big data analytics. By providing insights into the current landscape of CASE, highlighting the challenges and proposing recommendations for future research, this review aims to stimulate further CASE innovation and collaboration.

Received 29th September 2025
Accepted 6th February 2026

DOI: 10.1039/d5dd00438a

rsc.li/digitaldiscovery

Introduction

Computer-Assisted Structure Elucidation (CASE) has the potential to stand at the forefront of modern chemistry,

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg. E-mail: emma.schymanski@uni.lu

Dagny Aurich

Dr. Dagny Aurich was a Doctoral and Postdoctoral Researcher in Environmental Cheminformatics, LCSB, University of Luxembourg. She completed her Historical Exposomics PhD in 2023 within the interdisciplinary Luxembourg Time Machine Project focussing on cheminformatics, high-resolution mass spectrometry, nontarget analysis, environmental history, and data visualization, expanding to CASE via MS for unknown chemicals in her postdoc. She holds a bachelor's degree in forensic science and a master's degree in analytical chemistry and quality assurance from Bonn-Rhein Sieg University of Applied Sciences, Germany, with research experience in toxicological forensics and Luxembourgish industry before and after her time at the University of Luxembourg.

Emma L. Schymanski

Professor Emma Schymanski is head of the Environmental Cheminformatics group at the Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg. She has a double degree in Chemistry/Environmental Engineering from UWA, Perth, completed her PhD (involving CASE via MS) at UFZ Leipzig and postdoc at Eawag, Switzerland. Her research combines cheminformatics and computational (high resolution) mass spectrometry approaches to elucidate unknowns in complex samples with non-target screening, and relate these to environmental causes of disease. An advocate for FAIR and open science, she is involved in several European and worldwide activities to improve workflows and the exchange of data.

representing a paradigm shift in the way molecular structures are determined and analysed. At its core, CASE uses the power of computational tools and algorithms to unravel the intricate architectures of chemical compounds from experimental data obtained through analytical techniques such as Nuclear Magnetic Resonance (NMR) or infrared (IR) spectroscopy and Mass Spectrometry (MS). This review aims to explore the significance of CASE in contemporary chemistry, emphasizing its role in automating and refining the structure determination process – a task historically reliant on manual interpretation of analytical data. By accelerating the discovery and characterization of new compounds, helping elucidate complex molecular structures, and facilitating research across diverse domains including drug discovery,[1,2] natural products chemistry,[3,4] and materials science,[5,6] CASE has the potential to make significant contributions to advance scientific knowledge and drive innovation in the field of chemistry.

Manual structure elucidation poses several challenges due to its labour-intensive and subjective nature. First, it is a time-consuming process that requires chemists to meticulously analyse experimental data, interpret spectroscopic signals, and construct plausible molecular structures. This can be particularly daunting for complex molecules or when dealing with large datasets. The number of possible structures per molecule varies depending on factors such as the presence of functional groups, stereochemistry (*e.g.*, *cis*- and *trans* isomers, chirality) or connectivity of atoms (*e.g.*, straight-chain structures like *n*-pentane and branched structures like isopentane or dimethylpropane). Additionally, interpretation of spectroscopic data and assignment of chemical shifts can be subjective, leading to potential biases and inconsistencies between different chemists or laboratories. Variability in interpretation may result in discrepancies in the proposed structures. Moreover, structure elucidation often requires specialized knowledge and expertise in spectroscopy, organic chemistry, and computational methods. Analytical techniques such as NMR and MS generate complex data containing overlapping signals, noise, and artifacts, such that deciphering these spectra and extracting meaningful information to deduce structural features is challenging, particularly for molecules with diverse functional groups or unusual bonding patterns. Ambiguity and uncertainty arise in structure elucidation where multiple structural hypotheses are consistent with the experimental data. Resolving such ambiguities requires additional experiments or computational analyses, adding complexity to the elucidation process. Moreover, human errors, such as misinterpretation of spectral peaks, misassignment of chemical shifts, or oversight of structural constraints, can occur during manual structure elucidation, leading to inaccuracies or incorrect structural assignments. Lastly, manual structure elucidation may not be scalable or suitable for high-throughput analysis, particularly in the context of large compound libraries or high-volume screening programs. Automation and computational methods offer advantages in terms of speed, throughput, and reproducibility.

Consequently, strengthening the role of computational CASE tools in automating and facilitating the structure elucidation process is essential in addressing these challenges. These tools leverage algorithms and machine learning techniques to analyse analytical data, generate structural hypotheses, and refine molecular models. CASE tools play a pivotal role in closing the loop between experimental data and structural interpretation. They can often automate tedious spectral analysis tasks like peak picking and signal assignment, reducing the time and effort needed for manual interpretation. Additionally, many CASE tools integrate diverse analytical techniques,[7,8] enabling comprehensive analysis of complex molecular structures. Through iterative refinement and validation against experimental data, these approaches enhance the accuracy and reliability of structural assignments.

One of the primary challenges faced by CASE is still underutilization, particularly in contexts where users prioritize identifying known compounds over discovering truly novel compounds. Many chemists rely on databases and existing libraries of these known compounds to match experimental data, limiting the exploration of the vast space of unknowns. Moreover, the abundance of CASE tools available, with no really established approach to "lead the way", presents a dilemma for users, as they are often overloaded with options and struggle to determine which tool best suits their needs. Compared with other computational software, CASE tools remain comparatively unknown, such that these options are rarely in the active awareness of researchers. Each tool may employ different methodologies, algorithms, and user interfaces, making it challenging to navigate the landscape of available options. Additionally, CASE tools are often underutilized because they sometimes underperform, further complicating their adoption. Another obstacle in CASE is the prevalence of proprietary software, which restricts access to source code and hinders further development and customization. Without open source tools, users cannot modify or extend the software to meet specific needs or integrate new features.

This review seeks to address this issue by providing an overview of the various tools and methodologies used in CASE, with a particular focus on the potential for CASE to support identification efforts in MS, where developments are not as mature as CASE for NMR. This review encompasses a structured exploration of the field of CASE, starting with a historical overview (chapter 2) tracing its evolution and highlighting key milestones. It then delves into the methodologies employed in CASE (chapter 3), including reduction methods, assembly methods, and hybrid approaches. Structural databases are covered in chapter 4, integration of analytical data in chapter 5, then real-world applications and case studies in chapter 6, illustrating the utility of CASE tools in various domains, from pharmaceuticals to environmental science. The potential for CASE to evolve in the coming years with the rapid evolution in ML and AI is covered in the closing chapter.

Historical overview

The evolution of CASE tools represents a journey from complex manual methods to sophisticated computational algorithms and software applications. In the early days of structure

elucidation, chemists relied heavily on manual interpretation of analytical data, painstakingly piecing together molecular structures based on observed spectral features. However, as the numbers and complexity of organic molecules under investigation increased, analytical techniques advanced and computers were developed, the limitations of manual methods became apparent. This spurred the development of the first computer programs such as those arising from the 1960s DENDRAL project (CONGEN, GENOA),[9] which aimed at automating certain aspects of structure elucidation in organic chemistry, making use of artificial intelligence (AI). DENDRAL[9,10] was one of the earliest expert systems designed to interpret mass spectra and propose molecular structures, laying the foundation for subsequent CASE tools. However, these early systems were limited by their reliance on manually encoded expert rules, restricted scalability, and a focus on relatively small and simple molecules, as well as by the limited availability and integration of experimental NMR data at the time. Many of these limitations were addressed in later approaches as computational power increased and algorithms became more sophisticated. Tools evolved to encompass a broader range of functionalities and methodologies. The integration of computational techniques with experimental data from NMR, MS and IR enabled chemists to tackle increasingly complex molecules with greater confidence and accuracy.

The development of software for predicting NMR spectra in the 1980s and 1990s revolutionized the field of structure elucidation. Programs like ACD/NMR Predictor[11] and ChemDraw[12] allowed chemists to simulate NMR spectra for proposed structures, aiding in structural verification and validation. Elyashberg *et al.* highlighted the "synergistic interaction between CASE, new NMR experiments, and continuously improving methods of computational chemistry".[13] Elyashberg himself contributed to this interaction through several tools, including MASS[14] (1976), X-PERT[15,16] (1997), StrucEluc[17] (1999), and Fuzzy Structure Generation[18] (2007). Other prominent researchers in the history of CASE tools include Munk, involved in ASSEMBLE[19] (1981), COCOA[20] (1988), Assemble 2.0 (ref. 21) (2000), and HOUDINI[22] (2003); Steinbeck, who contributed to LUCY[23] (1996), SENECA[24] (2001), MAYGEN[25] (2021), and SURGE[26] (2022); Kerber with the MOLGEN[27] suite (summarized in detail in 2014), and Faulon, involved in OMG,[28] MOLSIG,[29]

Table 1 Overview of major structure generation developments, with one reference per entry where available in the name section. Further detail is given in Table S1 (SI)

Year	Name	Language	Successor	Comment
1964	CONGEN (DENDRAL)[9]	LISP	CONGEN-II, GENOA	Rarely used
~1970	CHEMICS(-F)[32]	NA		Not accessible
1976	MASS[14]	FORTRAN	SMOG	Not accessible
1981	GENOA (DENDRAL)[33]	LISP		Rarely used
1981	ASSEMBLE[19]	NA	Assemble 2.0	Superseded
1985	ACCESS[34]	NA		Not accessible
1986	DARC-EPOIS[35]	NA		Not accessible
1988	COCOA[20]	Pascal, FORTRAN	GEN, HOUDINI	Not accessible
1990	AEGIS[7]	PROLOG		Not accessible
1990	MOLGEN[36]	C	MOLGEN 3.5, 4, 5	Closed source
1991	LSD[37]	PROLOG		MacOS, Win
1995	GEN[38]	Turbo Pascal	HOUDINI	Not accessible
1996	SMOG[39]	C/C++		Open
1996	LUCY[23]	NA	SENECA	Not accessible
1997	COCON[40]	NA		Online demo
1997	X-PERT[15]	NA		Not accessible
1998	MOLGEN 4.0 (ref. 41)	C	MOLGEN 5.0	Closed
1999	StrucEluc (ACD Labs)[42]	NA		Closed
2000	Assemble 2.0 (ref. 21)	NA		Win '95,97,NT
2000	ESESOC[43]	NA		Not accessible
2001	SENECA[24]	Java		Open, Unix/Win
2003	HOUDINI[22]	Pascal, FORTRAN		Not accessible
2007	Fuzzy structure generation[18]	NA		Concept
2012	OMG[28]	Java, C	PMG	Open
2013	MolSig[29]	C		Open
2013	PMG[30]	Java		Open
2014	MOLGEN 5.0 (ref. 44)	C		Online demo
2017	MassChemSite[45]	NA		Closed
2017	SMART[46]	Python/Matlab	DeepSAT	Closed
2021	MAYGEN[25]	Java	SURGE	Open
2021	Scharnica[47]	NA		Accessible
2021	MassGenie[48]	PyTorch		NA
2022	SURGE[26]	C		Accessible
2022	MSNovelist[4]	Python		Open
2023	Mass2SMILES[49]	Python		Open (preprint)
2023	DeepSAT[50]	Python		Open

and PMG[30] during 2012 and 2013. Numerous additional tools have been developed, often building on older algorithms. These tools are summarized chronologically in Table 1 and in greater detail in SI Table S1, which details their basic principles, disadvantages, programming languages, successors (if any), and references. Information was collected from the respective method papers or selected review articles (*e.g.* Yirik and Steinbeck[31]). The purpose of Table S1 is to provide a structured historical and methodological overview of CASE tools. Due to the lack of standardized benchmarks, limited availability of performance metrics, and the prevalence of closed source or commercial systems, a quantitative comparison of accuracy or performance was not feasible and is therefore not included.

Major differences have emerged in the evolution of structure elucidation tools, such as assembly *versus* reduction or hybrid approaches, whether they work from a molecular formula or use experimental data, whether they are open or closed source, and whether structures are generated with or without relying on databases.

More recently, machine-learning approaches have begun to complement traditional CASE methodologies by learning structure-spectrum relationships directly from large datasets. Algorithms trained on vast databases of chemical structures and spectral data can now predict and interpret spectra with high accuracy. Examples for NMR include SMART[46] (Small Molecule Accurate Recognition Technology), which applies deep learning to 2D NMR (HSQC) spectra for spectral embedding and dereplication (2017), and its successor DeepSAT[50] (2023), which extends this concept toward data-driven spectral annotation and scaffold recognition. For MS, examples include MassChemSite[45] (2017), using a custom database, or MassGenie[48] (2021), which leverages PubChem.[51] Other tools, such as Scharnica[47] (2021), generate possible structures independently of databases. These methodological differences are explained in the next sections.

Methodologies in CASE

Reduction methods

Reduction methods play a role in efficiently narrowing down the vast search space of possible molecular structures. These methods systematically reduce the complexity of molecular structures by focusing on key structural elements and constraints derived from experimental data. The first tool employing reduction methods was COCOA,[20] which exemplifies the principles and advantages of this approach. COCOA uses an exhaustive recursive bond-removal strategy, systematically breaking down complex molecular structures into simpler fragments. These fragments are then analysed and recombined to propose potential structures that match the given constraints and experimental data. Breaking down the molecule in a controlled and exhaustive manner ensures that all possible structural configurations are explored. This analysis helps in identifying viable candidate structures that are consistent with the input data. Unlike assembly methods (explained in the next section), which build structures step-by-step by adding atoms or fragments, reduction methods start with a complete hypergraph that represents all possible bonds between atom pairs.[31] The size of this hypergraph is then reduced by systematically checking and applying constraints based on the presence or absence of specific substructures. As bonds are deleted and the hypergraph is simplified, structures decrease in size at each step. COCOA[20] uses atom-centred fragments to optimize storage and computational efficiency.[31] Instead of storing entire molecular structures, the tool focuses on the first neighbours of each atom, which reduces the memory requirements and enhances the speed of the structure generation process. This approach is comparable to circular fingerprints and atom signatures used in cheminformatics, where the local environment of each atom is described rather than the whole molecule. COCOA's[20] reduction method is versatile in handling required substructures and potentially overlapping substructures. Required substructures are specific fragments or motifs that must be present in the final structure, often derived from experimental data such as NMR or MS spectra. These required substructures can be incorporated into the reduction process, ensuring that all generated candidates contain these critical elements. Fig. 1 shows a simplified version of the reduction approach for perfluoro-2-methoxyacetic acid (PFMOAA).

This relatively small PFAS (per- and polyfluoroalkyl substance) was chosen because the complexity of both the molecule and the reduction process increases with molecular size, making it impractical to demonstrate the method for larger molecules. Without a large number of constraints, the method generates a vast array of possible candidates. Generally, the main disadvantage of reduction methods is the massive size of the hypergraphs.[31] For molecules with unknown structures, the size of the hyper structure can become extremely large, resulting in a corresponding increase in runtime. Hybrid methods combining reduction and assembly techniques have been developed to address this (see the section describing Hybrid approaches).

Assembly methods

Assembly methods iteratively combine smaller substructures to systematically build molecular structures. These methods use mathematical and computational algorithms to explore possible structural configurations based on given constraints and spectral data. In assembly methods, the generation process begins with a set of atoms from the molecular formula or substructures derived from experimental data.[31] It systematically forms bonds between atoms. Each time a bond is created, the resulting partial structure is checked against constraints like valences, bond multiplicity, and required fragments. If any constraint is violated, the bond is removed, and a different bond is attempted. When no more bonds can be formed without violating constraints, a valid candidate structure is identified.[31] Non-overlapping substructures can be incorporated from the start, using known molecular fragments to guide the assembly.[21] This fundamental process builds complete structures by linking smaller substructural fragments together. Several tools make use of McKay's NAUTY algorithm[52] (*e.g.* SURGE[26]) to eliminate redundant structures.[26] This algorithm

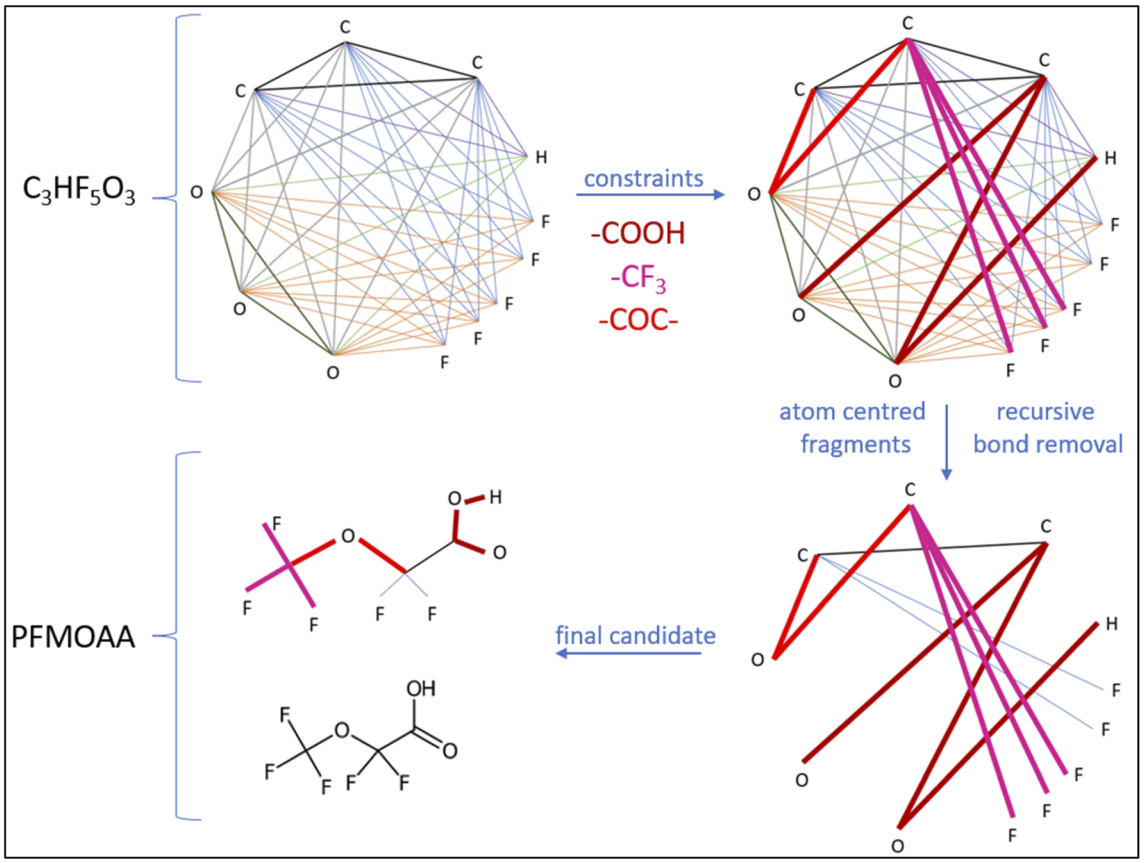

Fig. 1 Simplified reduction approach for perfluoro-2-methoxyacetic acid (PFMOAA).

calculates canonical labelling and extends structures by adding one bond at a time, ensuring each extension remains canonical. For example, during the bond formation steps in the assembly approach, NAUTY can be used to check the uniqueness of intermediate structures, avoiding the exploration of isomorphic duplicates.

Fig. 2 presents a simplified version of the assembly approach for PFMOAA. The same relatively small PFAS structure was selected (as in Fig. 1) because a greater number of substructures and consequently different assembled isomers are possible with larger examples. Several examples of tools employing this approach are listed in Table S1, highlighting their specific methodologies as well as disadvantages.

Hybrid approaches

Hybrid approaches in CASE tools integrate the strengths of both assembly and reduction methods to create more versatile and efficient structure elucidation processes. A prime example of this is GEN,[38] beginning with a hyper structure, eliminating connections that would create forbidden structures. It then assembles substructures to build new structures, filling connection matrices based on substructure information. This method efficiently handles constraints without allowing substructure overlaps, balancing the advantages of both assembly and reduction approaches.[38]

HOUDINI,[22] an improved version of GEN,[38] further refines this hybrid approach. HOUDINI relies on two main data structures: a square matrix representing all bonds in a hyper structure and a substructure representation listing atom-centred fragments. During the structure generation process, HOUDINI[22] maps these atom-centred fragments onto the hyper structure, enhancing the efficiency and accuracy of structure generation. Neither approach is available online, nor do they incorporate experimental data, instead relying solely on the molecular formula.

The MOLGEN family of structure generators are among the most time-efficient generators. MOLGEN[53] addressed several shortcomings of DENDRAL and many other tools by offering more sophisticated and time-efficient algorithms, with various versions tailored for different data inputs. MOLGEN 3.5 remains one of the fastest generators based on mathematical graph theory using just the molecular formula as an input. MOLGEN 4 (ref. 41) and the related MOLGEN-MS[54] and MOLGEN-QSPR[55] focused less on speed and more on a flexible interface with advanced restrictions (good list, bad list structures and macroatoms that could be expanded later in generation). In 2007, MOLGEN 5 (ref. 44) was released, aiming to combine the efficiency and flexibility of previous versions through a new, albeit still closed source, approach. In practice, different MOLGEN versions were better suited to different applications.[27]

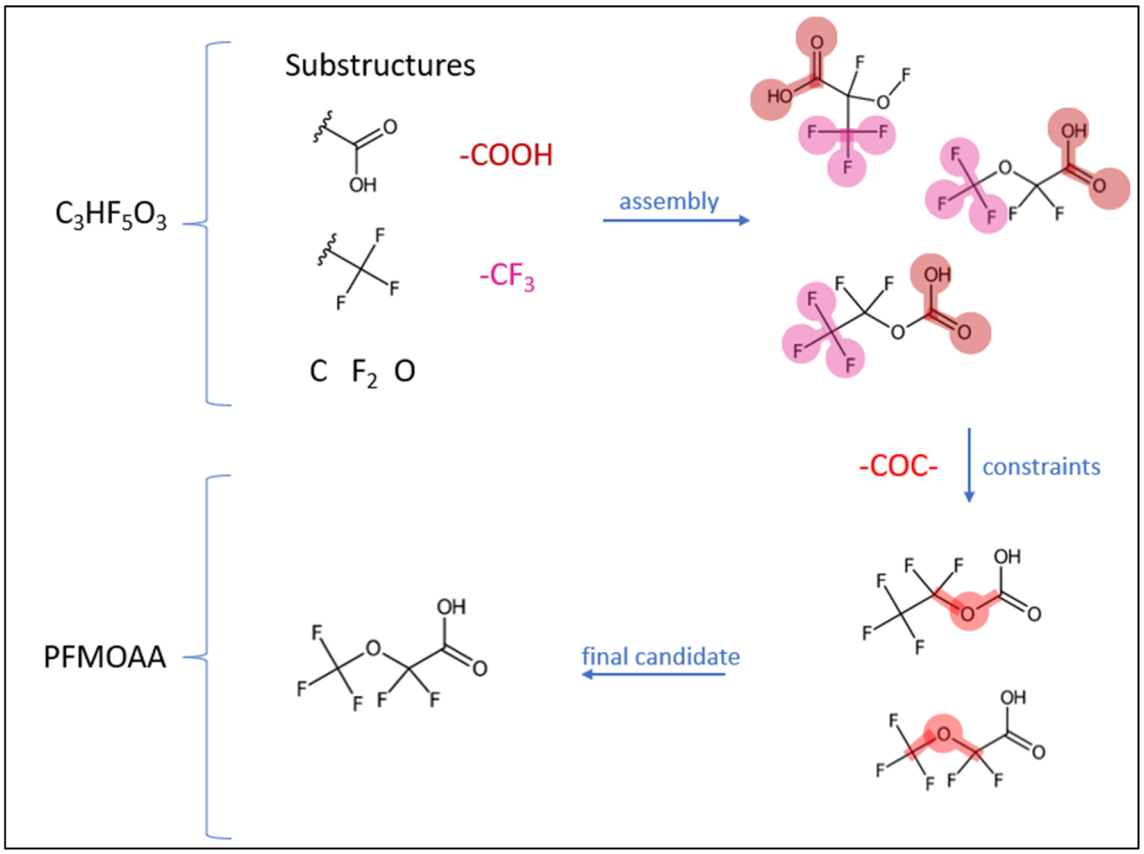

Fig. 2 Simplified assembly approach for perfluoro-2-methoxyacetic acid (PFMOAA).

In parallel, hybrid approaches have also emerged that combine experimental spectral data with data-driven, machine-learning models rather than explicit graph-based assembly. Tools such as SMART[46] and DeepSAT[50] integrate 2D NMR data directly into learned chemical representations, enabling dereplication, similarity assessment, and partial structural annotation without enumerating full molecular graphs. These approaches are typically used alongside, rather than as replacements for, traditional CASE generators, providing complementary information that can guide or constrain subsequent structure elucidation workflows.

Structural databases (chemical space)

One of the major challenges for CASE is handling the immensity of chemical space – a space that is to this date innumerable. Efforts by the Reymond group in Bern have produced large databases of all possible structures with 11,[56] 13 (ref. 57) and 17 (ref. 58) atoms, with the latter composed of C, H, N, O, S and halogens totalling 166.4 billion molecules using Nauty/GENG.[58] The Synthetically Accessible Virtual Inventory (SAVI) Database contains 1.75 billion compounds predicted to be easily synthesizable, created using a set of transforms based on CHMTRN/PATRAN using the CACTVS toolkit.[59,60] The vast majority of structures in these databases have not yet been documented or proven to exist – *i.e.*, they comply with the rules of chemical possibility (and in the case of SAVI, with a proposed synthetic route), but many have not yet been discovered or created. Other databases contain documented structures, although many of these may not have been produced in large amounts. The CAS Registry currently contains over 290 million structures,[61] increasing in the order of 10 s of millions of structures a year.[62] The largest open chemical databases, PubChem and ChemSpider, contain 122 (ref. 63) and ~128 (ref. 64) million chemicals, respectively (Sept. 2025). Smaller collections include the CompTox Chemicals Dashboard (CCD, ~1.2 million chemicals)[65] for environmental and computational toxicity efforts, or the Human Metabolome Database (HMDB, 220 945 metabolites)[66] for the human metabolome.

Spectral databases contain structures for which analytical data exists in sufficient amounts to be measured with the respective technique. Mass spectral libraries have grown impressively in recent years. NIST produce some of the largest mass spectral libraries, with the 2023 release of the NIST/EPA/NIH EI-MS library for electron impact MS data including 394k spectra of 347 100 compounds and the NIST Tandem Mass Spectral Library containing 2.4 million spectra of 51 501 compounds.[67] The METLIN library now contains tandem mass spectra for over 960 000 compounds,[68,69] although the compound list has not been made publicly available to assess the relevance of the compound coverage. One of the largest open mass spectral libraries is MassBank of North America (MoNA) with 2 080 139 spectra of 651 236 compounds

(including some combinatorial libraries such as LipidBlast).[70] The open NMR database NMRshiftDB/nmrshiftdb2 contains 271 816 structures, with 70 026 measured and 396 583 calculated spectra in Dec. 2025.[71] PubChem collates spectral information (or the presence of spectral information) for 1 650 108 compounds,[72] corresponding to 1 229 560 compounds with mass spectra (including the calculated LipidBlast), 659 362 with NMR spectra, 228 628 with IR spectra and 16 029 with UV spectra. Unfortunately, the majority of these entries are thumbnails or partial data (and thus unsuitable for CASE).

The challenge of matching analytical signals to a documented (or hypothetically possible) structure differs depending on the analytical technique used, with NMR generally yielding the richest source of structural information. For mass spectrometry, with generally sparser information, the number of possible structures for discrete formulae rapidly becomes unmanageable, even at relatively small masses (see Table 2). At larger masses (~400–500 Da), even the number of possible formulae, let alone the number of structures, becomes difficult to manage when including small elements without isotopic patterns such as fluorine.[73]

Analytical data integration

Most CASE tools use predominantly NMR data due to its detailed structural information, while some also incorporate MS and IR data despite the increased complexity associated with these techniques. Incorporating NMR data into CASE involves several strategies to predict and interpret spectra. This is *e.g.* shown by tools like COCON,[40] its online version WEB-COCON,[75] LSD[37] or StrucEluc.[42] Spectral prediction algorithms simulate NMR spectra based on potential molecular structures, helping to narrow down plausible candidates. Furthermore, sophisticated spectral interpretation techniques decode complex NMR patterns to extract meaningful structural information. Several comprehensive overviews on the use of CASE in NMR were published by *e.g.* Elyashberg *et al.*[76,77] and Williams *et al.*[78] The key points collected in the 2008,[76] 2011 (ref. 77) and 2016 (ref. 78) efforts were summarized in the 2021 article by Elyashberg and Argyropoulos[13] highlighting the significant evolution of CASE over 50 years, from simple prototypes to advanced tools integral to NMR spectroscopy. Initially reliant on molecular formulas and 1D/2D NMR spectra, these systems are now adaptable to various NMR experiments and other spectroscopic data, making CASE integral to NMR spectroscopy.[13] They stress that the synergy between CASE, emerging NMR techniques, and computational methods continues to enhance its capabilities, with future developments likely to include deeper integration with advanced computational tools like DFT (density functional theory) and deep learning. The inclusion of CASE in educational curricula could help prepare new generations of chemists to implement CASE approaches and help them become routine in both academic and industrial settings. Despite this readiness, the authors claim that the "golden age" of CASE is still ahead.[13]

MS and IR data integration, while less common due to the generally less detailed structural information available, follow similar principles. MS data provides molecular masses and fragmentation patterns that can confirm or refute structural hypotheses generated from NMR data. However, structure elucidation can also be performed directly from MS data. MOLGEN-MS, based on low resolution electron impact mass spectra[27,54] and MOLGEN 4.0, generated structures from a molecular formula using spectral classifiers to determine possible structural features using a "good list" and "bad list" (substructures present/absent to a given probability threshold set by the user) that was then used to constrain the generation. Coupling MOLGEN-MS with classifiers from the NIST database (which was much larger than the original training set) resulted in notable performance improvements,[79] but was only applied in a handful of cases[80] (discussed in more detail below). Since these efforts, structural elucidation with MS has developed significantly, but typically coupled to databases of structures (PubChem, ChemSpider, HMDB or others), rather than *de novo* identification based on structure generation. While manually-performed elucidation efforts generally outperformed automated methods in the early "Critical Assessment of Small Molecule Identification" (CASMI) contests (initiated in 2012),[81] computational methods improved dramatically in the years of active contest and clearly outperformed manual attempts in later years.[82] However, very few of these entries over the years used structure generation due to the poor performance relative to database lookup. While directly training better performing structure generation models using tandem mass spectrometry (MS2) spectra is likely currently still out-of-reach due to the limited availability of public training data, the development of methods leveraging latest advances in ML are underway. MSNovelist[4] leverages the success of CSI:FingerID[83] and SIRIUS,[84] which have performed well in CASMI, using compound databases by predicting a molecular fingerprint based on MS2 data.[83] MSNovelist[4] combines fingerprint prediction[83] with an encoder-decoder neural network using a Recurrent Neural Network (RNN) model with Long Short-Term Memory (LSTM) architecture to generate structures *de novo* solely from MS2 spectra. It predicted 25% of structures correctly on the first rank and retrieved 45% of structures overall in evaluations, successfully reproducing 61% of correct database annotations without having seen the structures during training.[4] This does not reach top CASMI performance level, but is closer than may have been expected. A recent effort by Brogat-Motte *et al.* also shows some potential to interpolate novel

Table 2 Number of possible structural isomers (calculated with MOLGEN5) and documented/known structural isomers in the CompTox Chemicals Dashboard (CDD) in Sept. 2017 (ref. 74)

#Carbons	#Isomers	SDF file size	#Isomers in CDD
2	9	6 kB	9
3	29	22 kB	27
4	116	108 kB	38
5	506	561 kB	35
6	2455	3176 kB	34
7	12 783	18 939 kB	40
8	>70 000	117 146 kB	[>upload limit]

structures without using predefined finite candidate set,[85] with first plausible applications likely to be transformations of existing molecules, such as the "suspect library" from GNPS (discussed further below).[86]

IR spectra offer insights into functional groups present within the molecule, aiding in the construction of accurate structural models. Tools like Scharnica,[47] AEGIS,[7] ASSEMBLE[21] or CHEMICS[8,32] make use of IR data for structure elucidation. Advancements in spectral data processing have been pivotal in enhancing the performance of CASE tools. Improved algorithms for noise reduction, peak detection, and baseline correction ensure higher quality input data for structure elucidation. Spectral validation techniques have also evolved, providing robust mechanisms to verify the consistency and accuracy of predicted structures against experimental data. Cross-validation with multiple data types (NMR, MS, IR), *e.g.* shown by Scharnica,[47] ensures that the proposed structures are not only mathematically plausible but also chemically and physically consistent.

An important consideration when performing CASE coupled with analytical data is the role of stereochemistry. As highlighted above, the number of structural isomers possible for given molecular formulae rapidly expands into unmanageable proportions; considering the number of stereoisomers possible for combinations of stereocentres in a molecule greatly expands this problem. For instance ESESOC,[43] which examines the 2D connection table to identify all stereocentres, removes all equivalent stereoisomers and then generates candidate structures, was noted to be a very time-consuming approach (Table S1). Many approaches work on such small numbers of atoms that the true combinatorial impact of stereochemistry in CASE is not yet sufficiently explored. Since MS experiments rarely yield stereochemistry information (only possible in very rare cases or with chiral chromatography), CASMI contests were often evaluated on the structural skeleton, by collapsing all candidates by the InChIKey first block (connectivity).[82] Recent MS-based CASE developments (MassGenie,[48] MASS2SMILES,[49] see Table S1) go one step further, ignoring stereochemistry altogether by using canonical (or connectivity) SMILES.

Applications and case studies

CASE tools have found diverse applications across various scientific domains, such as pharmaceuticals, metabolomics, natural products chemistry and environmental studies.

In the pharmaceutical industry, the identification and structural elucidation of small molecule impurities and degradation products is a crucial aspect enforced by regulatory agencies worldwide. Liu *et al.* provide a comprehensive review of how CASE tools, particularly MS-based techniques, are employed to address this need.[87] The review underscores the critical importance of structure elucidation in pharmaceutical development, noting that complete identification of impurities and degradation products often necessitates a combination of chromatographic, MS, and NMR techniques.[87] For *de novo* structure elucidation of compounds the authors refer to the MS2LDA tool,[88] which aids impurity identification by extracting common patterns (Mass2Motifs) from MS/MS spectra, which can indicate shared substructures between impurities and drug APIs (Active Pharmaceutical Ingredients). CASE tools that were originally developed for metabolomics and metabolite identification are now increasingly adopted in pharmaceutical settings to streamline this process.[87]

Metabolomics is a critical area where CASE tools have found application. Several studies have explored the utility of CASE tools in metabolomics, with articles by Dias *et al.*[3] or de Jonge *et al.*[89] providing comprehensive overviews of these efforts. These publications highlight the advancements and challenges in using computational tools for metabolite identification. The field has seen the development of deep learning tools such as MassGenie[48] and Mass2SMILES[49] (see Table S1), which were specifically built on metabolomics data. These tools represent significant advancements in the ability to predict molecular structures and functional groups directly from MS data, addressing one of the major bottlenecks in metabolomics: the identification of unknown metabolites. Furthermore, the open source methods OMG[28] and its parallelized version PMG[30] were designed with metabolomics applications in mind. Overall, these applications show that the synergistic interaction between human expertise and computational tools could provide a powerful approach to chemical structure elucidation.

Natural products have long served as a rich source of novel compounds with diverse biological activities, making them crucial targets for structure elucidation. Some CASE tools have been built and tested specifically on natural products, using primarily NMR data.[78] Notable examples include COCON[40,75] and CISOC-SES.[90] In their 2018 review, Burns *et al.* highlighted the role of CASE tools in the structure elucidation of complex natural products, addressing several critical issues.[91] Despite advancements, a concerning number of incorrect natural product structures are still reported in the literature. CASE programs can mitigate this risk by generating all possible structures consistent with the input data and ranking them by probability. These tools are effective in determining structures for complex natural products, although they may struggle with compounds containing very few protons.[91] Different CASE programs were described, emphasizing their handling of longer-range correlation peaks. These programs either provide just planar skeletal structures or use stereospecific NMR data to determine 3D structures.[91] The paper discusses additional forms of computer assistance in structure elucidation, including the growing use of theoretical DFT calculations to determine 3D structures and predict chemical shifts. Burns *et al.* concluded with suggestions for improving CASE programs and proposed a challenge match between current CASE program developers to further enhance their capabilities and accuracy.[91]

Environmental samples are generally too complex for NMR, leaving CASE *via* MS as the primary choice. To date, applications have been rather limited. This includes some approaches with low resolution GC-MS data based on MOLGEN-MS, which helped identify some unknowns in effect-directed analysis studies in Bitterfeld[80] and elucidate a toxic transformation product (TP) of diclofenac,[92] whose identity was confirmed *via* synthesis of

a reference standard. Some efforts with high resolution data include the elucidation of several benzotriazole TPs,[93] although final proof of many structure remained elusive due to lack of reference standards. The advent of large open structure databases such as PubChem and ChemSpider in the early 2000s alongside the developments of high resolution mass spectrometry (HRMS) has seen the field shift focus to documented chemicals, since this "known" chemical space is already challenging enough to master at present,[62] let alone the unknown. However, TPs, *i.e.* relatively slight modifications of documented structures, are likely the next domain within reach of CASE *via* MS and are the focus of several current developments such as BioTransformer 4.0 (ref. 94) and the Chemical Transformation Simulator.[95] Molecular networking approaches based on GNPS[96] have been used to generate so-called "suspect libraries" of spectra that are one node away from known spectral/structural associations, on the hypothesis that many of these may be from structurally-related compounds.[86] Recent developments[85] may help interpolate novel structures to enable CASE for these TPs, as mentioned above.

Challenges and future directions

The field of CASE has seen several advancements over the past decades, yet several challenges remain. Handling complex molecular structures is one of the primary challenges. Many CASE tools struggle with molecules that have intricate frameworks, multiple rings, or high symmetry, leading to ambiguities and errors in the generated structures (see Table S1). Improving the robustness and precision of algorithms to handle such cases more effectively is essential.

Despite progress, the accuracy of CASE algorithms remains a critical concern. Current algorithms can sometimes produce incorrect or incomplete structures, particularly when dealing with noisy or incomplete spectral data (restricted to all tools dealing with experimental data, *e.g.* LSD[37] or COCON[40,75]). Enhancing the reliability of these algorithms is crucial to ensure accurate structure elucidation. Another challenge is integrating data from different analytical techniques, such as NMR, MS, and IR, into a cohesive structure elucidation process. Each technique provides complementary information, but combining these data streams seamlessly remains difficult and, although certainly a worthy time investment, is rather low demand as it is quite rare that all three methods are available for the same question.

The performance of most CASE systems still limits their adoption. Overall, scientists are rarely satisfied with an honest, unbiased assessment of how many structural possibilities may be theoretically possible for a given CASE problem, with substituted long chains being particularly problematic due to the high number of branching/substitution possibilities. The "hard truth" of potential possibilities is combined with the great difficulty and expense in confirming potential "unknown unknowns" (which would involve synthesis or isolation of sufficient amounts for detailed analysis – both often very difficult in reality), such that these confirmation efforts are only performed in very rare cases, and focus is often placed on easier problems to solve. Different CASE systems apply a range of methods to rank their candidates, which makes it difficult to compare results, while benchmarking is also challenging (see below). Although various systems have evolved in the last two decades to quantify confidence of identification (*e.g.* the Metabolomics Society Initiative,[97] confidence levels for HRMS data,[98] HRMS data coupled with PFAS[99] or CCS[100]), an attempt by the Metabolomics Society to create a reporting system catering for the structural information and confidence applicable to both MS and NMR has so far failed to reach community consensus. Although Metz *et al.*[101] recently proposed a probability approach, the probabilities would be so low for any CASE problem considering all possible structures that this is not yet feasible for *de novo* identification efforts.

The use of different programming languages across various CASE tools (*e.g.*, LISP, FORTRAN, C, see Tables 1 and S1) presents a challenge in terms of interoperability, maintenance, and integration. This diversity complicates the ability to seamlessly combine or compare results from different software systems, hindering collaborative efforts and the development of standardized workflows. Additionally, many CASE tools face computational bottlenecks, such as long computing times and difficulties in handling overlapping substructures or duplicate fragments, let alone stereoisomers. These issues can slow down the elucidation process and reduce the efficiency of the tools, highlighting the need for more optimized and scalable algorithms.

CASE is hampered in many ways by the mix of open *versus* closed/commercial approaches and data. Closed source tools further exacerbate many challenges faced by CASE by limiting transparency and hindering collaborative development. The lack of access to the underlying algorithms and data processing methods in these tools prevents the wider scientific community from verifying results, contributing to improvements, or integrating these tools into broader workflows. For instance, the commercial license on the MOLGEN suite – one of the most efficient structure generators developed – prevented further developments following the retirement of Prof. Kerber and the distribution of the know-how away from the license holder (University of Bayreuth). While the push for open code of recent years has clearly impacted the CASE field, with many new developments now open (see Tables 1 and S1), the availability of sufficient open data to train and benchmark CASE methods is also becoming a bottleneck, with few scientists incentivised to measure, let alone contribute the resulting data to open resources, while many of the largest collections remaining licensed or closed (*e.g.*, the NIST and METLIN libraries for MS). Since the recent trend in open source code availability has come strongly from funders and institutions, it is likely that a similarly coordinated approach to incentivise large, multinational collections of measured data would be needed to contribute sufficient amounts of new data to openly available collections. Rigorous benchmarking exercises are also now generally beyond the reach of individual research groups, such that communities, networks or societies may need to consider how such efforts could be stimulated and supported.

Despite the reflections above, it is uncertain whether substantial improvements to CASE tools would lead to their widespread adoption across various scientific fields. Routine laboratories still face significant challenges in fully identifying compounds or structures documented in databases (*e.g.*, even non-target analysis on "known unknowns" is not standardized yet, making CASE tools to identify true "unknown unknowns") more of a future prospect that will remain largely underutilized for the near future.

While gathering the literature and information for this review and Table S1, several documentation and interoperability issues with many of the CASE tools became evident, particularly with older ones. Locating the original references and the exact year of publication was challenging, especially when they were published in different languages (*e.g.*, CHEMICS in Japanese) or only available in print with no online access. Additionally, comparing the computational demands of these tools was difficult due to differences in the hardware and software capabilities of the machines used at the time. Since many of these tools only run on outdated operating systems, direct comparisons between old and modern tools are complicated. Thus, while CASE has been around many years and is one of the reasons that cheminformatics as a discipline exists, it is in desperate need of modernization.

Emerging trends in CASE include the adoption of deep learning and big data analytics. Deep learning and other machine learning techniques are increasingly being integrated into CASE tools (as shown above). For instance, python tools like MassGenie[48] and MASS2SMILES[49] leverage deep neural networks to predict molecular structures and substructures from spectral data, demonstrating the potential power of these technologies in CASE. In parallel, several recent studies have applied deep learning directly to NMR spectra: transformer- and neural network-based models such as NMRMind,[102] SMART,[46] DeepSAT[50] and other preprinted architectures[103,104] map experimental spectra to molecular structures or structural features, complementing existing MS-based approaches. Both MS and NMR-based tools, however, would become more reliable for specific structural challenges and for larger, more complex molecules with additional training data. The integration of larger and more comprehensive spectral databases can allow for better matching and validation of experimental spectra against known compounds. As these spectral databases evolve, the challenge of curation and maintenance becomes critical. Most publicly available MS spectral databases are still too small to effectively train models. Additionally, a significant amount of closed-source research remains, particularly in the industry, which hinders the growth of these resources *via* collaborative efforts. As discussed above, concerted community efforts will be needed to address the lack of data issue.

Future advancements in CASE are likely to come from interdisciplinary approaches that combine insights from chemistry, computer science, and bioinformatics. Collaborations across these fields could lead to the development of more sophisticated algorithms and software capable of addressing the current limitations and pushing the boundaries of what is possible in structure elucidation. Making CASE tools more user-friendly and accessible should be an ongoing goal, but requires incentives for all sides. Simplifying the interfaces and workflows of these tools can help non-experts use them effectively, broadening their adoption and impact. Additionally, open-source initiatives and collaborative platforms could facilitate wider access and community-driven improvements.

CASE tools have found applications in fields like pharmaceuticals, natural products chemistry, metabolomics and environmental science, though their use remains limited. Despite notable progress and promising trends such as deep learning, big data analytics, and improved user accessibility, significant challenges persist. Many applications are still only successful for small or carefully-selected cases, and are not broadly applicable. Expanding the current offerings through advanced computational techniques, better data integration, and interdisciplinary collaboration could be key to broader adoption of CASE tools across various scientific and industrial domains. However, it is uncertain whether there is sufficient demand within the scientific community to drive these advancements. ML-based developments are improving rapidly, but are dependent on large amounts of novel data, while experimentalists have relatively few incentives to contribute valuable measurements to open resources to support ML developments. Benchmarking efforts suffer from the same lack of data. At this stage, CASE *via* NMR seems to be enjoying rapid developments, and although several key new breakthroughs are now available for CASE *via* MS, their performance is not yet sufficient for routine use. As long as challenges in identifying known structures persist – which is still the case for MS experiments – fully automated new structure generation with CASE tools remains a future prospect.

Author contributions

Dagny Aurich: writing – original draft (lead), review and editing. Emma Schymanski: funding acquisition, writing – original draft, review and editing (lead).

Conflicts of interest

The authors have no conflict of interest to declare.

Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review. The collated information supporting this article have been included as part of the supplementary information (Table S1). Supplementary information is available. See DOI: https://doi.org/10.1039/d5dd00438a.

Acknowledgements

We thank the Environmental Cheminformatics group members and other colleagues and collaborators for all discussions related to this work. DA and ELS acknowledge funding support from the Luxembourg National Research Fund (FNR) for project

A18/BM/12341006 and the University of Luxembourg Institute for Advanced Studies (IAS) for the Audacity project "LuxTIME".

References

1 M. Ahlqvist, C. Leandersson, M. A. Hayes, I. Zamora and R. A. Thompson, Software-aided structural elucidation in drug discovery, *Rapid Commun. Mass Spectrom.*, 2015, **29**(21), 2083–2089, DOI: 10.1002/rcm.7364.

2 T. Kind and O. Fiehn, Advances in structure elucidation of small molecules using mass spectrometry, *Bioanal Rev*, 2010, **2**(1), 23–60, DOI: 10.1007/s12566-010-0015-9.

3 D. A. Dias, O. A. H. Jones, D. J. Beale, *et al.*, Current and Future Perspectives on the Structural Identification of Small Molecules in Biological Systems, *Metabolites*, 2016, **6**(4), 46, DOI: 10.3390/metabo6040046.

4 M. A. Stravs, K. Dührkop, S. Böcker and N. Zamboni, MSNovelist: *de novo* structure generation from mass spectra, *Nat. Methods*, 2022, **19**(7), 865–870, DOI: 10.1038/s41592-022-01486-3.

5 M. E. Elyashberg and A. J. Williams. *Computer–Based Structure Elucidation from Spectral Data: The Art of Solving Problems*. Vol 89. Springer, Berlin Heidelberg. 2015. DOI: 10.1007/978-3-662-46402-1.

6 J. L. Faulon. Stochastic Generator of Chemical Structure. 1. *Application to the Structure Elucidation of Large Molecules*. ACS Publications. DOI: 10.1021/ci00021a031.

7 H. J. Luinge and J. H. Van Der Maas, AEGIS, an algorithm for the exhaustive generation of irredundant structures, *Chemom. Intell. Lab. Syst.*, 1990, **8**(2), 157–165, DOI: 10.1016/0169-7439(90)80131-O.

8 S. ichi Sasaki, H. Abe, Y. Hirota, *et al.*, CHEMICS-F: A Computer Program System for Structure Elucidation of Organic Compounds, *J. Chem. Inf. Comput. Sci.*, 1978, **18**(4), 211–222, DOI: 10.1021/ci60016a007.

9 R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum and J. Lederberg, DENDRAL: A case study of the first expert system for scientific hypothesis formation, *Artif Intell*, 1993, **61**(2), 209–261, DOI: 10.1016/0004-3702(93)90068-M.

10 R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum and J. Lederberg. *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*. McGraw-Hill Book Co; 1980.

11 *NMR Prediction | 1H,C 13C, 15N, 19F, 31P NMR Predictor*. ACD/Labs. 2024, accessed May 3, 2024. https://www.acdlabs.com/products/spectrus-platform/nmr-predictors/.

12 D. A. Evans, History of the Harvard ChemDraw Project, *Angew Chem Int Ed*, 2014, **53**(42), 11140–11145, DOI: 10.1002/anie.201405820.

13 M. Elyashberg and D. Argyropoulos, Computer Assisted Structure Elucidation (CASE): Current and future perspectives, *Magn. Reson. Chem.*, 2020, **59**(7), 669–690, DOI: 10.1002/mrc.5115.

14 V. V. Serov, M. E. Elyashberg and L. A. Gribov, Mathematical synthesis and analysis of molecular structures, *J. Mol. Struct.*, 1976, **31**(2), 381–397, DOI: 10.1016/0022-2860(76)80018-X.

15 M. E. Elyashberg, E. R. Martirosian, YuZ. Karasev, H. Thiele and H. Somberg, X-PERT: a user-friendly expert system for molecular structure elucidation by spectral methods, *Anal. Chim. Acta*, 1997, **337**(3), 265–286, DOI: 10.1016/S0003-2670(96)00391-1.

16 M. E. Elyashberg, YuZ. Karasev, E. R. Martirosian, H. Thiele and H. Somberg, Expert systems as a tool for the molecular structure elucidation by spectral methods. Strategies of solution to the problems, *Anal. Chim. Acta*, 1997, **348**(1–3), 443–463, DOI: 10.1016/S0003-2670(97)00229-8.

17 ACD/Labs. CASE NMR Software | Structure Elucidator Suite™. Structure Elucidator Suite™. 2025, accessed September 4, 2025. https://www.acdlabs.com/products/spectrus-platform/structure-elucidator-suite/.

18 M. E. Elyashberg, K. A. Blinov, S. G. Molodtsov, A. J. Williams and G. E. Martin, Fuzzy Structure Generation: A New Efficient Tool for Computer-Aided Structure Elucidation (CASE), *J. Chem. Inf. Model.*, 2007, **47**(3), 1053–1066, DOI: 10.1021/ci600528g.

19 C. A. Shelley and M. E. Munk, Case, a computer model of the structure elucidation process, *Anal. Chim. Acta*, 1981, **133**(4), 507–516, DOI: 10.1016/S0003-2670(01)95416-9.

20 B. D. Christie and M. E. Munk, Structure generation by reduction: a new strategy for computer-assisted structure elucidation, *J. Chem. Inf. Comput. Sci.*, 1988, **28**(2), 87–93, DOI: 10.1021/ci00058a009.

21 M. Badertscher, A. Korytko, K. P. Schulz, *et al.*, Assemble 2.0: a structure generator, *Chemom. Intell. Lab. Syst.*, 2000, **51**(1), 73–79, DOI: 10.1016/S0169-7439(00)00056-3.

22 A. Korytko, K. P. Schulz, M. S. Madison and M. E. Munk, HOUDINI: A New Approach to Computer-Based Structure Generation, *J. Chem. Inf. Comput. Sci.*, 2003, **43**(5), 1434–1446, DOI: 10.1021/ci034057r.

23 C. Steinbeck, LUCY—A Program for Structure Elucidation from NMR Correlation Experiments, *Angew Chem. Int. Ed. Engl.*, 1996, **35**(17), 1984–1986, DOI: 10.1002/anie.199619841.

24 C. Steinbeck, SENECA: A Platform-Independent, Distributed, and Parallel System for Computer-Assisted Structure Elucidation in Organic Chemistry, *J. Chem. Inf. Comput. Sci.*, 2001, **41**(6), 1500–1507, DOI: 10.1021/ci000407n.

25 M. A. Yirik, M. Sorokina and C. Steinbeck, MAYGEN: an open-source chemical structure generator for constitutional isomers based on the orderly generation principle, *J. Cheminf.*, 2021, **13**(1), 48, DOI: 10.1186/s13321-021-00529-9.

26 B. D. McKay, M. A. Yirik and C. Steinbeck, Surge: a fast open-source chemical graph generator, *J. Cheminf.*, 2022, **14**(1), 24, DOI: 10.1186/s13321-022-00604-9.

27 A. Kerber, R. Laue, M. Meringer, C. Rücker and E. Schymanski. *Mathematical Chemistry and Chemoinformatics: Structure Generation, Elucidation, and Quantitative Structure - Property Relationships. ISBN: 978-3-11-030007-9*. De Gruyter; 2014.

28 J. E. Peironcely, M. Rojas-Chertó, D. Fichera, *et al.*, OMG: Open Molecule Generator, *J. Cheminf.*, 2012, **4**(1), 21, DOI: 10.1186/1758-2946-4-21.
29 P. Carbonell, L. Carlsson and J. L. Faulon, Stereo Signature Molecular Descriptor, *J. Chem. Inf. Model.*, 2013, **53**(4), 887–897, DOI: 10.1021/ci300584r.
30 M. M. Jaghoori, S. S. T. Q. Jongmans, F. De Boer, *et al.*, PMG: Multi-core Metabolite Identification, *Electron Notes Theor Comput Sci.*, 2013, **299**, 53–60, DOI: 10.1016/j.entcs.2013.11.005.
31 M. A. Yirik and C. Steinbeck, Chemical graph generators, *PLoS Comput. Biol.*, 2021, **17**(1), e1008504, DOI: 10.1371/journal.pcbi.1008504.
32 K. Funatsu and S. S. ichi, Recent Advances in the Automated Structure Elucidation System, CHEMICS. Utilization of Two-Dimensional NMR Spectral Information and Development of Peripheral Functions for Examination of Candidates, *J. Chem. Inf. Comput. Sci.*, 1996, **36**(2), 190–204, DOI: 10.1021/ci950152r.
33 R. E. Carhart, D. H. Smith, N. A. B. Gray, J. G. Nourse and C. Djerassi, Applications of artificial intelligence for chemical inference. 37. GENOA: a computer program for structure elucidation utilizing overlapping and alternative substructures, *J. Org. Chem.*, 1981, **46**(8), 1708–1718, DOI: 10.1021/jo00321a037.
34 W. Bremser and W. Fachinger, Multidimensional spectroscopy, *Magn. Reson. Chem.*, 1985, **23**(12), 1056–1071, DOI: 10.1002/mrc.1260231208.
35 J. E. Dubois, A. Panaye and R. Attias, DARC system: notions of defined and generic substructures. Filiation and coding of FREL substructure (SS) classes, *J. Chem. Inf. Comput. Sci.*, 1987, **27**(2), 74–82, DOI: 10.1021/ci00054a007.
36 A. Kerber, R. Laue and D. Moser, Ein strukturgenerator für molekulare graphen, *Anal. Chim. Acta*, 1990, **235**, 221–228, DOI: 10.1016/S0003-2670(00)82078-4.
37 J. M. Nuzillard and M. Georges, Logic for structure determination, *Tetrahedron*, 1991, **47**(22), 3655–3664, DOI: 10.1016/S0040-4020(01)80878-4.
38 S. Bohanec, Structure Generation by the Combination of Structure Reduction and Structure Assembly, *J. Chem. Inf. Comput. Sci.*, 1995, **35**(3), 494–503, DOI: 10.1021/ci00025a017.
39 M. S. Molchanova, V. V. Shcherbukhin and N. S. Zefirov, Computer Generation of Molecular Structures by the SMOG Program, *J. Chem. Inf. Comput. Sci.*, 1996, **36**(4), 888–899, DOI: 10.1021/ci950393z.
40 T. Lindel, J. Junker and M. Köck, Cocon: From NMR Correlation Data to Molecular Constitutions, *J Mol Model*, 1997, **3**(8), 364–368, DOI: 10.1007/s008940050052.
41 A. Kerber, R. Laue, T. Grüner and M. Meringer, MOLGEN 4.0. MATCH, *Commun Math Comput Chem*, 1998, **37**, 205–208.
42 K. A. Blinov, D. Carlson, M. E. Elyashberg, *et al.*, Computer-assisted structure elucidation of natural products with limited 2D NMR data: application of the StrucEluc system, *Magn. Reson. Chem.*, 2003, **41**(5), 359–372, DOI: 10.1002/mrc.1187.
43 J. Hao, L. Xu and C. Hu, Expert system for elucidation of structures of organic compounds (ESESOC): —Algorithm on stereoisomer generation, *Sci China Ser B Chem*, 2000, **43**(5), 503–515, DOI: 10.1007/BF02969496.
44 R. Gugisch, A. Kerber and A. Kohnert, *et al.*, MOLGEN 5.0, A Molecular Structure Generator. in *Advances in Mathematical Chemistry and Applications*. Vol 1. Bentham Science Publishers, 2015, pp. 113–138. https://www.benthamdirect.com/content/books/9781681081977.chapter-6.
45 H. Yao, Y. Liu, S. Tyagarajan, *et al.*, Enabling Efficient Late-Stage Functionalization of Drug-Like Molecules with LC-MS and Reaction-Driven Data Processing, *Eur. J. Org Chem.*, 2017, **2017**(47), 7122–7126, DOI: 10.1002/ejoc.201701573.
46 C. Zhang, Y. Idelbayev, N. Roberts, *et al.*, Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research, *Sci. Rep.*, 2017, **7**(1), 14243, DOI: 10.1038/s41598-017-13923-x.
47 M. Pesek, A. Juvan, J. Jakoš, J. Košmrlj, M. Marolt and M. Gazvoda, Database Independent Automated Structure Elucidation of Organic Molecules Based on IR, 1H NMR, 13C NMR, and MS Data, *J. Chem. Inf. Model.*, 2021, **61**(2), 756–763, DOI: 10.1021/acs.jcim.0c01332.
48 A. D. Shrivastava, N. Swainston, S. Samanta, I. Roberts, M. Wright Muelas and D. B. Kell, MassGenie: A Transformer-Based Deep Learning Method for Identifying Small Molecules from Their Mass Spectra, *Biomolecules*, 2021, **11**(12), 1793, DOI: 10.3390/biom11121793.
49 D. Elser, F. Huber and E. Gaquerel, Mass2SMILES: deep learning based fast prediction of structures and functional groups directly from high-resolution MS/MS spectra, *BioRxiv*, 2023, preprint, DOI: 10.1101/2023.07.06.547963.
50 H. W. Kim, C. Zhang, R. Reher, *et al.*, DeepSAT: Learning Molecular Structures from Nuclear Magnetic Resonance Data, *J. Cheminf.*, 2023, **15**(1), 71, DOI: 10.1186/s13321-023-00738-4.
51 S. Kim, J. Chen, T. Cheng, *et al.*, PubChem 2023 update, *Nucleic Acids Res.*, 2023, **51**(D1), D1373–D1380, DOI: 10.1093/nar/gkac956.
52 B. D. McKay and A. Piperno, Practical graph isomorphism, II, *J Symb Comput*, 2014, **60**, 94–112, DOI: 10.1016/j.jsc.2013.09.003.
53 C. Benecke, R. Grund, A. Kerber, R. Laue and T. Wieland, Chemical Education *via* MOLGEN, *J. Chem. Educ.*, 1995, **72**(5), 403, DOI: 10.1021/ed072p403.
54 A. Kerber, R. Laue, M. Meringer and K. Varmuza, MOLGEN-MS: Evaluation of low resolution electron impact mass spectra with MS classification and exhaustive structure generation, *Adv Mass Spectrom*, 2001, **15**(1), 939–940.
55 C. Rücker, M. Meringer and A. Kerber, QSPR Using MOLGEN-QSPR: The Challenge of Fluoroalkane Boiling Points, *J. Chem. Inf. Model.*, 2005, **45**(1), 74–80, DOI: 10.1021/ci0497298.
56 T. Fink and J. L. Reymond, Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers)

and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery, *J. Chem. Inf. Model.*, 2007, **47**(2), 342–353, DOI: 10.1021/ci600423u.

57 L. C. Blum and J. L. Reymond, 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13, *J. Am. Chem. Soc.*, 2009, **131**(25), 8732–8733, DOI: 10.1021/ja902302h.

58 L. Ruddigkeit, R. Van Deursen, L. C. Blum and J. L. Reymond, Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17, *J. Chem. Inf. Model.*, 2012, **52**(11), 2864–2875, DOI: 10.1021/ci300415d.

59 H. Patel, W. D. Ihlenfeldt, P. N. Judson, *et al.*, SAVI, *in silico* generation of billions of easily synthesizable compounds through expert-system type rules, *Sci Data*, 2020, **7**(1), 384, DOI: 10.1038/s41597-020-00727-4.

60 H. Patel, W. D. Ihlenfeldt and P. N. Judson, *et al.*, Synthetically Accessible Virtual Inventory (SAVI) Database Download Page. Published online 2020. DOI: 10.35115/37N9-5738.

61 American Chemical Society. CAS REGISTRY®. 2025, accessed September 4, 2025. https://www.cas.org/cas-data/cas-registry.

62 H. P. H. Arp, D. Aurich, E. L. Schymanski, K. Sims and S. E. Hale, Avoiding the Next Silent Spring: Our Chemical Past, Present, and Future, *Environ. Sci. Technol.*, 2023, **57**(16), 6355–6359, DOI: 10.1021/acs.est.3c01735.

63 NCBI/NLM/NIH. PubChem Website. 2025, accessed September 4, 2025. https://pubchem.ncbi.nlm.nih.gov/.

64 Royal Society of Chemistry. ChemSpider Blog. 2025, accessed September 4, 2025. https://blogs.rsc.org/chemspider/.

65 US EPA. CompTox Chemicals Dashboard. 2025, accessed September 4, 2025. https://comptox.epa.gov/dashboard/.

66 D. S. Wishart, A. Guo, E. Oler, *et al.*, HMDB 5.0: the Human Metabolome Database for 2022, *Nucleic Acids Res.*, 2022, **50**(D1), D622–D631, DOI: 10.1093/nar/gkab1062.

67 National Institute of Standards and Technology. NIST Mass Spectrometry Data Center. https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:start, accessed 4 Sept. 2025. https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:start.

68 Scripps. METLIN Classic. https://metlin.scripps.edu, accessed 4 Sept. 2025. https://metlin.scripps.edu/auth-login.html.

69 J. Xue, C. Guijas, H. P. Benton, B. Warth and G. Siuzdak, METLIN MS2 molecular standards database: a broad chemical and biological resource, *Nat. Methods*, 2020, **17**(10), 953–954, DOI: 10.1038/s41592-020-0942-5.

70 FiehnLab. MassBank of North America, accessed September 4, 2025. https://mona.fiehnlab.ucdavis.edu/.

71 S. Kuhn, H. Kolshorn, C. Steinbeck and N. Schlörer, Twenty years of nmrshiftdb2: A case study of an open database for analytical chemistry, *Magn. Reson. Chem.*, 2024, **62**(2), 74–83, DOI: 10.1002/mrc.5418.

72 NCBI/NLM/NIH. PubChem Table of Contents Classification Browser. 2025, accessed September 4, 2025. https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72.

73 T. Kind and O. Fiehn, Metabolomic database annotations *via* query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm, *BMC Bioinf.*, 2006, **7**(1), 234, DOI: 10.1186/1471-2105-7-234.

74 E. Schymanski, A. Williams and J. Hollender. [SETAC FTM] Identifying Complex Mixtures in the Environment with Cheminformatics and Non-targeted High Resolution Mass Spectrometry. Zenodo. 2017. DOI: 10.5281/ZENODO.17055527.

75 M. Köck, T. Lindel and J. Junker, Incorporation of 4J-HMBC and NOE Data into Computer-Assisted Structure Elucidation with WebCocon, *Molecules*, 2021, **26**(16), 4846, DOI: 10.3390/molecules26164846.

76 M. E. Elyashberg, A. J. Williams and G. E. Martin, Computer-assisted structure verification and elucidation tools in NMR-based structure elucidation, *Prog. Nucl. Magn. Reson. Spectrosc.*, 2008, **53**(1–2), 1–104, DOI: 10.1016/j.pnmrs.2007.04.003.

77 M. E. Elyashberg, A. Williams and K. Blinov. *Contemporary Computer-Assisted Approaches to Molecular Structure Elucidation*. RSC; 2011. https://books.google.lu/books?id=unMoDwAAQBAJ.

78 A. Williams, G. Martin and D. Rovnyak. *Modern NMR Approaches to the Structure Elucidation of Natural Products: Volume 2: Data Acquisition and Applications to Compound Classes*. Royal Society of Chemistry; 2016. https://books.google.lu/books?id=1L-2DQAAQBAJ.

79 E. L. Schymanski, C. Meinert, M. Meringer and W. Brack, The use of MS classifiers and structure generation to assist in the identification of unknowns in effect-directed analysis, *Anal. Chim. Acta*, 2008, **615**(2), 136–147, DOI: 10.1016/j.aca.2008.03.060.

80 C. Meinert, E. Schymanski, E. Küster, R. Kühne, G. Schüürmann and W. Brack, Application of preparative capillary gas chromatography (pcGC), automated structure generation and mutagenicity prediction to improve effect-directed analysis of genotoxicants in a contaminated groundwater, *Environ. Sci. Pollut. Res.*, 2010, **17**(4), 885–897, DOI: 10.1007/s11356-009-0286-2.

81 E. Schymanski and S. Neumann, CASMI: And the Winner is?, *Metabolites*, 2013, **3**(2), 412–439, DOI: 10.3390/metabo3020412.

82 E. L. Schymanski, C. Ruttkies, M. Krauss, *et al.*, Critical Assessment of Small Molecule Identification 2016: automated methods, *J. Cheminf.*, 2017, **9**(1), 22, DOI: 10.1186/s13321-017-0207-1.

83 K. Dührkop, H. Shen, M. Meusel, J. Rousu and S. Böcker, Searching molecular structure databases with tandem mass spectra using CSI:FingerID, *Proc Natl Acad Sci*, 2015, **112**(41), 12580–12585, DOI: 10.1073/pnas.1509788112.

84 K. Dührkop, M. Fleischauer, M. Ludwig, *et al.*, SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information, *Nat. Methods*, 2019, **16**(4), 299–302, DOI: 10.1038/s41592-019-0344-8.

85 L. Brogat-Motte, R. Flamary, C. Brouard, J. Rousu and F. D'Alché-Buc. in Learning to Predict Graphs with Fused Gromov-Wasserstein Barycenters, ed. Chaudhuri K., Jegelka S., Song L., Szepesvari C., Niu G. and Sabato S., *Proceedings of the 39th International Conference on Machine Learning. Vol 162. Proceedings of Machine Learning Research*. PMLR, 2022, pp. 2321–2335. https://proceedings.mlr.press/v162/brogat-motte22a.html.

86 W. Bittremieux, N. E. Avalon, S. P. Thomas, *et al.*, Open access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics, *Nat. Commun.*, 2023, **14**(1), 8488, DOI: 10.1038/s41467-023-44035-y.

87 Y. Liu, E. P. Romijn, G. Verniest, K. Laukens and T. De Vijlder, Mass spectrometry-based structure elucidation of small molecule impurities and degradation products in pharmaceutical development, *TrAC Trends Anal Chem*, 2019, **121**, 115686, DOI: 10.1016/j.trac.2019.115686.

88 J. J. J. van der Hooft, J. Wandy, M. P. Barrett, K. E. V. Burgess and S. Rogers, Topic modeling for untargeted substructure exploration in metabolomics, *Proc Natl Acad Sci*, 2016, **113**(48), 13738–13743, DOI: 10.1073/pnas.1608041113.

89 N. F. De Jonge, K. Mildau, D. Meijer, *et al.*, Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools, *Metabolomics*, 2022, **18**(12), 103, DOI: 10.1007/s11306-022-01963-y.

90 C. Peng, S. Yuan, C. Zheng, *et al.*, Application of Expert System CISOC-SES to the Structure Elucidation of Complex Natural Products, *J. Chem. Inf. Comput. Sci.*, 1994, **34**(4), 814–819, DOI: 10.1021/ci00020a014.

91 D. C. Burns, E. P. Mazzola and W. F. Reynolds, The role of computer-assisted structure elucidation (CASE) programs in the structure elucidation of complex natural products, *Nat. Prod. Rep.*, 2019, **36**(6), 919–933, DOI: 10.1039/C9NP00007K.

92 T. Schulze, S. Weiss, E. Schymanski, *et al.*, Identification of a phytotoxic photo-transformation product of diclofenac using effect-directed analysis, *Environ. Pollut.*, 2010, **158**(5), 1461–1466, DOI: 10.1016/j.envpol.2009.12.032.

93 S. Huntscha, T. B. Hofstetter, E. L. Schymanski, S. Spahr and J. Hollender, Biotransformation of Benzotriazoles: Insights from Transformation Product Identification and Compound-Specific Isotope Analysis, *Environ. Sci. Technol.*, 2014, **48**(8), 4435–4443, DOI: 10.1021/es405694z.

94 S. Tian, Y. D. Feunang and E. Oler, *et al.*, BioTransformer 4.0 a comprehensive computational tool for small molecule metabolism prediction, *BioRxiv*, 2025, preprint, DOI: 10.1101/2025.07.28.667289.

95 C. Yuan, C. Tebes-Stevens and E. J. Weber, Prioritizing Direct Photolysis Products Predicted by the Chemical Transformation Simulator: Relative Reasoning and Absolute Ranking, *Environ. Sci. Technol.*, 2021, **55**(9), 5950–5958, DOI: 10.1021/acs.est.0c08745.

96 M. Wang, J. J. Carver, V. V. Phelan, *et al.*, Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking, *Nat. Biotechnol.*, 2016, **34**(8), 828–837, DOI: 10.1038/nbt.3597.

97 L. W. Sumner, A. Amberg, D. Barrett, *et al.*, Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI), *Metabolomics Off J Metabolomic Soc.*, 2007, **3**(3), 211–221, DOI: 10.1007/s11306-007-0082-2.

98 E. L. Schymanski, J. Jeon, R. Gulde, *et al.*, Identifying Small Molecules *via* High Resolution Mass Spectrometry: Communicating Confidence, *Environ. Sci. Technol.*, 2014, **48**(4), 2097–2098, DOI: 10.1021/es5002105.

99 J. A. Charbonnet, C. A. McDonough, F. Xiao, *et al.*, Communicating confidence of per- and polyfluoroalkyl substance identification *via* high-resolution mass spectrometry, *Environ. Sci. Technol. Lett.*, 2022, **9**(6), 473–481, DOI: 10.1021/acs.estlett.2c00206.

100 A. Celma, J. V. Sancho, E. L. Schymanski, *et al.*, Improving Target and Suspect Screening High-Resolution Mass Spectrometry Workflows in Environmental Analysis by Ion Mobility Separation, *Environ. Sci. Technol.*, 2020, **54**(23), 15120–15131, DOI: 10.1021/acs.est.0c05713.

101 T. O. Metz, C. H. Chang, V. Gautam, *et al.*, Introducing "Identification Probability" for Automated and Transferable Assessment of Metabolite Identification Confidence in Metabolomics and Related Studies, *Anal. Chem.*, 2025, **97**(1), 1–11, DOI: 10.1021/acs.analchem.4c04060.

102 X. Xue, H. Sun, J. Sun, *et al.*, NMRMind: A Transformer-Based Model Enabling the Elucidation from Multidimensional NMR to Structures, *Anal. Chem.*, 2025, **97**(41), 22603–22614, DOI: 10.1021/acs.analchem.5c03783.

103 M. Alberts, F. Zipoli and A. C. Vaucher. Learning the Language of NMR: Structure Elucidation from NMR spectra using Transformer Models. *Chemistry*. Preprint posted online August 14, 2023. DOI: 10.26434/chemrxiv-2023-8wxcz.

104 Q. Yang, B. Wu and X. Liu, *et al.*, DiffNMR: Diffusion Models for Nuclear Magnetic Resonance Spectra Elucidation. *arXiv. Preprint posted online* 2025. DOI: 10.48550/ARXIV.2507.08854.