Nucleic Acids Research, 2025, **53**, D700–D708
https://doi.org/10.1093/nar/gkae1067
Advance access publication date: 22 November 2024
Database issue

The Natural Products Magnetic Resonance Database (NP-MRD) for 2025

David S. Wishart[1,2,3,4,*], Tanvir Sajed[1], Matthew Pin[5], Ella F. Poynton[5], Bharat Goel[6], Brian L. Lee[1], An Chi Guo[1], Sukanta Saha[1], Zinat Sayeeda[2], Scott Han[1], Mark Berjanskii[1], Harrison Peters[1], Eponine Oler[1], Vasuk Gautam[1], Tamara Jordan[5], Jonghyeok Kim[5], Benjamin Ledingham[5], Zachary M. Tretter[6], James T. Koller[6], Hailey A. Shreffler[7], Lillian R. Stillwell[7], Amy M. Jystad[7], Niranjan Govind[8], Jessica L. Bade[9], Lloyd W. Sumner[6], Roger G. Linington[5] and John R. Cort[7,10]

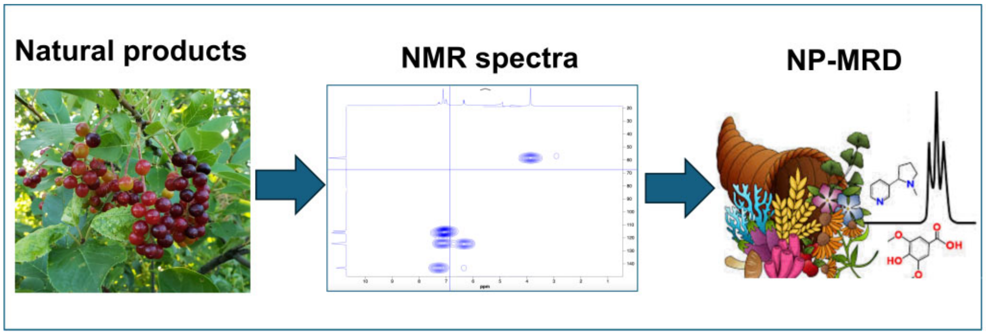
[1] Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada
[2] Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8, Canada
[3] Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, AB T6G 2B7, Canada
[4] Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, AB T6G 2H7, Canada
[5] Department of Chemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
[6] Department of Biochemistry, Bond Life Sciences Center, and Interdisciplinary Plant Group, University of Missouri, Columbia, MO 65211, USA
[7] Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA
[8] Physical Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA
[9] Chemical and Biological Signatures, National Security Directorate, Pacific Northwest National Laboratory, Richland, WA 99352, USA
[10] Institute of Biological Chemistry, Washington State University, Pullman, WA 99164, USA

[*] To whom correspondence should be addressed. Tel: +1 780 492 8574; Email: dwishart@ualberta.ca

Abstract

The Natural Products Magnetic Resonance Database (NP-MRD; https://np-mrd.org) is a comprehensive, freely accessible, web-based resource for the deposition, distribution, extraction, and retrieval of nuclear magnetic resonance (NMR) data on natural products (NPs). The NP-MRD was initially established to support compound de-replication and data dissemination for the NP community. However, that community has now grown to include many users from the metabolomics, microbiomics, foodomics, and nutrition science fields. Indeed, since its launch in 2022, the NP-MRD has expanded enormously in size, scope, and popularity. The current version of NP-MRD now contains nearly 7× more compounds (281 859 versus 40 908) and 7× more NMR spectra (5.5 million versus 817 278) than the first release. More specifically, an additional 4.6 million predicted spectra and another 11 000 spectra simulated from experimental chemical shifts were deposited into the database. Likewise, the number of NMR raw spectral data depositions has grown from 165 spectra per year to >10 000 per year. As a result of this expansion, the number of monthly webpage views has grown from 55 to 20 000 and the number of monthly visitors has increased from 7 to 2500. To address this growth and to better support the expanding needs of its diverse community of users, many additional improvements to the NP-MRD have been made. These include significant enhancements to the data submission process, notable updates to the database's spectral search utilities and useful additions to support better NMR spectral analysis/prediction. Significant efforts have also been undertaken to remediate and update many of NP-MRD's database entries. This manuscript describes these database improvements and expansion efforts, along with how they have been implemented and what future upgrades to the NP-MRD are planned.

Graphical abstract



Received: September 15, 2024. **Revised:** October 17, 2024. **Editorial Decision:** October 18, 2024. **Accepted:** October 25, 2024

Introduction

Most medicinal chemists and plant chemists define a natural product (NP) as a secondary metabolite isolated from natural sources that are produced by the pathways of secondary metabolism. However, many other fields of chemistry, biochemistry and life science have a much broader or more inclusive definition of NPs. At the Natural Products Magnetic Resonance Database (NP-MRD), we define an NP as any organic molecule (typically <5000 Da) that is fully or partially produced by living organisms. This includes any small or midsize molecule generated and/or metabolized by living organisms, from bacteria, to fungi, to plants, to invertebrates to vertebrates (including humans). Using this more inclusive definition elevates NPs to a much higher level of economic, environmental and social importance. Indeed, NPs are the source of the bulk of organic matter on earth, serving as the main ingredients to the soil we stand on, to the food we put in our mouths, and to the trees that tower above us (1). NPs are not only essential for life, they are essential to our quality of life. Indeed, NPs serve as the basis or chemical inspiration to the majority of the drugs, cosmetics, supplements, dyes, flavoring, aroma, and coloring agents used in our daily lives (2–4).

The isolation and characterization of NP structures has been of central interest to chemists and biochemists for >200 years (5,6). In addition to X-ray diffraction and microcrystal electron diffraction (7), mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy are the two techniques most commonly used to determine the structure of NPs (8,9). NMR is particularly useful as it not only allows the direct determination of the 3D structure of NPs, it can also be used to determine the absolute configurations of stereoisomers and diastereomers (10). NMR is also essential in NP dereplication (eliminating known NPs from consideration), unknown or novel compound identification and the characterization of NP extracts or mixtures (11). However, using NMR to its maximum effectiveness in NP chemistry requires large collections of high quality, well-annotated, referential NMR spectra of pure NPs along with their associated structures and chemical shift assignments (11,12). This is why the NP-MRD (https://np-mrd.org) was created.

The NP-MRD was initially established to provide a central repository for knowledge about NPs, to aid in the archiving NMR data about NPs, to support NP dereplication, and to facilitate structure elucidation of novel NPs for the NP community (12). However, since its launch in 2022, the NP-MRD has steadily grown to become the largest open repository of NP NMR data in the world, housing the most complete collection of general information (nomenclature, structure, biology, chemistry, chemotaxonomy and NMR spectra) about NPs. Indeed, the current version of NP-MRD now contains nearly 7× more compounds (281 859 versus 40 908) and 7× more NMR spectra (5.5 million versus 817 278) than the 2022 release. This includes 19 000 sets of experimental NMR spectra (for 3700 NPs), 416 000 simulated NMR spectra (for 22 000 NPs obtained from literature-derived NMR assignments) and 5.1 million predicted NMR spectra (for 282 000 NPs). Not only has the NP-MRD's content grown, so too has its popularity and utility (Figure 1). For instance, the number of NMR spectral depositions going into the NP-MRD has ballooned by nearly 100-fold over the past 3 years. The number of monthly website visitors and webpage views has grown by >300-fold over the same period. Additionally, because the scope of the NP-MRD has been expanded to include both secondary metabolites (traditional NPs) and primary metabolites (non-traditional NPs), the community that the NP-MRD now serves has grown to include many users from the metabolomics, microbiomics, foodomics, and nutrition science fields.

To address this growth and to better support the changing needs of its diverse community of users, many important improvements to the NP-MRD have been made. These include major improvements to the NP-MRD's data submission process that now allows users to deposit NMR data in a few seconds using an intelligent drag and drop data deposition system (13). More comprehensive data format conversion (nmrML, JCAMP-DX, NMReDATA, and Bruker) and better tools for data exchange have also been added. As the database has grown and the type and number of queries have expanded, a number of significant speed-ups and improvements to the NP-MRD's spectral search utilities have also been made to address these burgeoning needs. Likewise, improvements to the accuracy of the NP-MRD's chemical shift predictions (used for generating the database's predicted NMR spectra), as well as improvements to spectral fitting and spectral simulation have been made to facilitate the NP-MRD's spectral analysis and interpretation utilities. Significant efforts have also been undertaken to remediate and update many of the NP-MRD's older database entries. More detailed descriptions regarding each of these improvements, expansions or implementations are given in the following text.

Database description/expansion

The first description of the NP-MRD, along with details regarding all the website's functions and menu options appeared in 2022 (12). The overall database design, format, navigation and layout have remained largely unchanged since that time. However, important changes have occurred to the data submission system (see next section) and modest changes have been made to the searching and utilities functions (see later sections). The most significant changes have occurred with the NP-MRD database content and size. When the first description of the NP-MRD appeared in 2022, it was remarked that the known size of the 'NP universe' was quite large, with nearly 300 000 NPs listed in various commercial repositories. However, it was also known that only a fraction of these known NPs had published or publicly accessible NMR assignments. As a result, the first release of the NP-MRD contained data for only 41 400 molecules. The majority of these compounds came from the Natural Products Atlas (NP-Atlas) (14) (20 468) and the JEOL CH-NMR-NP database (https://ch-nmr-np.jeol.co.jp/en/nmrdb/) (19 025), with smaller numbers coming from the Human Metabolome Database (HMDB) (15) (879), the Biological Magnetic Resonance Data Bank (BMRB) (16) (284) and literature backfilling efforts (185).

Interestingly, <1500 molecules (<4%) in the NP-MRD's first release had full NMR spectra with complete assignment data. Usually, these assignments were made at only one NMR spectrometer frequency. Therefore, to make the NP-MRD dataset more useful for other magnetic fields, curators took the reported chemical shift assignments for all compounds, including those from the JEOL CH-NMR-NP, HMDB and the BMRB and generated simulated NMR spectra at 10 different magnet field strengths (from 100 to 1000 MHz, in 100 MHz steps for ^{1}H data and from 25 to 250 MHz, in 25 MHz steps for ^{13}C data). This was done using the reported

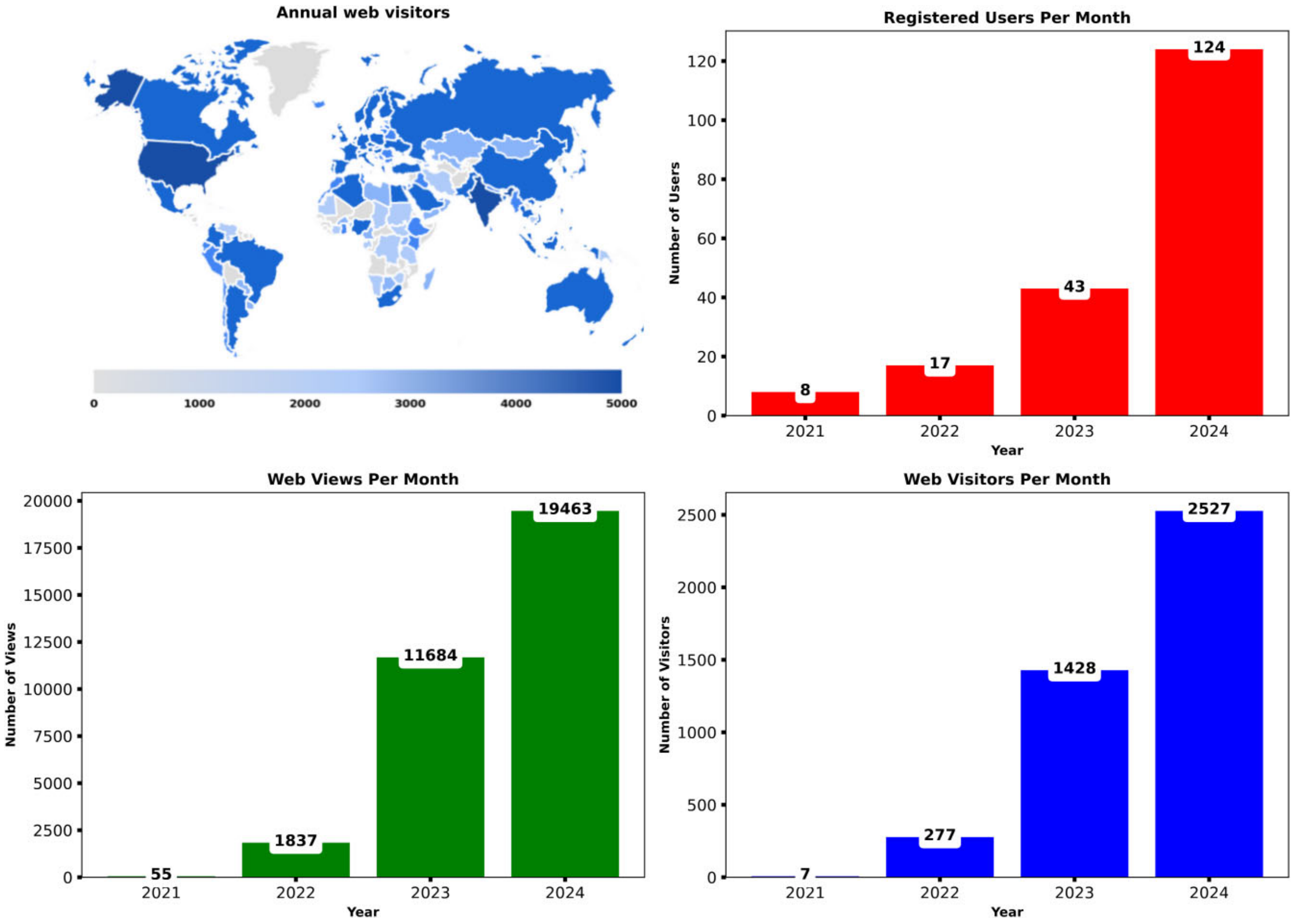


Figure 1. Growth of the NP-MRD. Panel (A) shows that >190 countries have accessed the NP-MRD. Panels (B)–(D) show the growth in registered users, web visits, and unique visitors per month.

solvent and chemical shift referencing compounds. For compounds that did not have NMR assignments (namely the NP-Atlas dataset), the NP-MRD curation staff used an early version of the PROSPRE (17) ^{1}H chemical shift predictor, a ^{13}C chemical shift predictor (CASPRE) and a draft version of NP-MRD's spectral predictor (NPSP) to predict the NMR spectra for these compounds at the 10 aforementioned magnet field strengths in water with sodium trimethylsilylpropanesulfonate (DSS) as the chemical shift referencing compound.

For the 2025 release, the NP-MRD curation team significantly expanded the number of compounds in the NP-MRD by including all non-lipid compounds in the latest version of HMDB (15), all compounds listed in the latest version of FooDB [https://www.foodb.ca/], all compounds listed in the latest version of KNApSAcK (18), all compounds listed in MiMeDB (19), all new compounds added to NP-Atlas (14), and all compounds in the latest version of the LOTUS database (20). This collection was determined to cover most of the known NPs in animals (HMDB and FooDB), microbes (NP-Atlas and MiMeDB), and plants (KNApSAcK, FooDB and LOTUS). Careful checks were made using InChI identifiers, InChI keys and nomenclature (names/synonyms) to avoid duplications, with the order of acceptance being HMDB > FooDB > MiMeDB > NP-Atlas > KNApSAcK > LOTUS. In this way, a total of 240 000 new, non-redundant molecules were added to the NP-MRD. Each new structure was annotated and classified using the NP-MRD annotation pipeline which includes passing each molecule through ChemoSummarizer (21), DataWrangler (21) and ClassyFire (22) as previously described (12).

After the compound annotations were completed, the chemical shifts for newly added NP-MRD structures were predicted using the latest version of the PROSPRE (17) ^{1}H chemical shift predictor, the latest version of our ^{13}C chemical shift predictor (CASPRE) and an updated version of NP-MRD's spectral predictor (NPSP) to predict the NMR spectra for these compounds. PROSPRE, a machine learning (ML)-based method for chemical shift prediction, has a mean absolute error (MAE) between predicted and observed ^{1}H shifts of 0.10 ppm, while CASPRE, also an ML approach, has an MAE between predicted and observed ^{13}C shifts of 1.6 ppm across multiple solvents and chemical shift referencing reagents. These spectra were predicted at the 10 designated magnet field strengths. A solvent preference predictor (trained on experimental data from the NP-MRD) was used to determine the single most likely solvent (water, methanol and chloroform) that each compound would be most compatible. This process generated a total of 4.6 million 1D NMR spectra.

In addition to these ML-predicted NMR spectra, the NP-MRD curation team has also been calculating and compiling quantum-mechanically derived chemical shifts using density functional theory (DFT) methods implemented in the ISiCLE package (23). The MAE for these DFT chemical shift calculations for ^{1}H chemical shifts is typically 0.4 ppm while the

Table 1. Comparison of the content statistics between the first release of NP-MRD and the latest release of NP-MRD

Parameter	NP-MRD 1.0	NP-MRD 2.0
Natural products	41 140	281 859
Experimental spectra	165	19 045
Simulated spectra	405 223	416 405
Predicted spectra	411 890	5 043 694
DFT-calculated chemical shifts	0	7 424
DOIs	0	3 070
Drag-and-drop submissions	0	18 683

MAE for ^{13}C chemical shift predictions is typically 4 ppm (23). A total of 7424 ^{1}H and ^{13}C chemical shifts for 3712 compounds have been generated using these DFT methods and deposited into NP-MRD. DFT calculations are very computationally expensive (often taking many hours on a high-performance computer for a single molecule), and so there are far fewer of these DFT-predicted shifts than ML-predicted shifts.

Since 2022, the NP-MRD curation team (via backfilling) has provided experimental NMR assignments for another 594 compounds with 524 sets of ^{13}C chemical shift assignments and 594 sets of ^{1}H chemical shift assignments. These deposited compounds with experimental assignments have also had their NMR spectra simulated at 10 designated magnet field strengths. This spectral simulation was done using the reported solvent and chemical shift referencing compounds. Table 1 compares the content statistics between the first release of NP-MRD and the latest release or NP-MRD.

The distinction between experimentally measured NMR spectra (usually collected in a single solvent at a single magnetic field strength), simulated spectra (which use experimental chemical shift data and experimental J-coupling data to generate NMR spectra at multiple field strengths) and predicted spectra (which use ML methods to predict chemical shifts, J-couplings and NMR spectra at multiple field strengths) is important to note. Some compounds in the NP-MRD have all three types of spectra (experimental, simulated and predicted), others only have two types of spectra (experimental and simulated), and still others only have predicted NMR spectra. The last category covers the vast majority of compounds in the NP-MRD. Within the NP-MRD, each type of spectrum is appropriately labeled and each type is selectable (individually or in bulk). The most accurate and useful spectra are obviously experimental NMR spectra. The least accurate are predicted NMR spectra. However, given the enormous improvement in NMR spectral prediction accuracy achieved over the last few years, even predicted NMR spectra are now very useful for compound identification, dereplication and characterization. Indeed, predicted chemical shifts and predicted NMR spectra are certainly far more useful than having no data at all. In some cases, they may even be more useful than published chemical shift assignments, which may contain typographic errors or misassignments (up to 5% of published assignments in our experience).

Improved data submission and entry

The NP-MRD is an archival database. This means that it is designed to accept external (and internal), user-deposited data. When the NP-MRD was first launched, two parallel paths for data submission or data deposition were undertaken: one was called literature backfilling or retrospective data entry (to be done by NP-MRD curators) and the other was called prospective data entry (to be done by NP-MRD users in the NP community). The goal of the literature backfilling process was to fill the NP-MRD with thousands of previously published NMR assignment datasets of well-known or well-studied NPs. The data backfilling process involved a manual literature review conducted by trained NP-MRD curators to identify novel NPs with appropriate NMR data and then manually enter the NMR data into the NP-MRD using a specially developed NP-MRD data deposition system. On the other hand, the goal of the prospective data entry process was to capture NMR data of newly determined or newly published NPs as they appeared in the literature. The prospective data entry was to be performed by registered NP-MRD users via the same specially developed NP-MRD data deposition interface.

Unfortunately, several problems became apparent when the two systems were launched in 2022. Email-based efforts to identify or encourage members of the NP community to submit their newly published data to the NP-MRD received low response rates. Likewise, a manual review of past NP literature to perform back-filling tasks proved to be much slower and much more difficult than expected. Additionally, user feedback regarding the specially designed NP-MRD deposition system indicated that the system was too slow (taking >30 min for a typical deposition), required considerable manual input (especially with regard to chemical shift assignment entry) and was prone to frequent user errors. That deposition system was also limited in its ability to allow users to embargo release dates or support other common deposition/release requests. These problems led to a significant redesign of both the retrospective backfilling and prospective data entry process.

This redesign led to, first, an improved semi-automated literature tracking system and user reminder system to facilitate backfilling. Second, a faster, easier, ‘smarter’ drag-and-drop deposition system was developed to make data deposition painless, easy and fast (the so-called ‘carrot’ approach in the carrot-and-stick motivation theory). Third, arrangements with several journals (including the *Journal of Natural Products*) were made to ‘strongly’ encourage or require data deposition to the NP-MRD prior to publication (the so-called ‘stick’ approach).

More specifically, the revamped NP-MRD deposition system now includes: (i) automated literature tracking with natural language processing (NLP) tools (based on ML) to identify new papers describing new NPs; (ii) NLP to extract key data (compound names, source organisms, etc.) from article titles and abstracts; (iii) an automated email system that sends personally curated data requests and deposition links for all new NP articles; (iv) a flexible data deposition framework that accepts data from published articles, pre-submissions and private repositories; (v) extensive quality control and standardization tools to ensure that deposited data is correct, complete, and well standardized for uploading to the NP-MRD; (vi) data conversion tools for Bruker 1D NMR data and open data formats (nmrML, NMReData and JCAMP-DX); (vii) security features to protect against the distribution of malware; (viii) embargo management to allow depositors to control the release date for deposited data, and provide private links that can be shared with reviewers; (ix) a unified single exchange format (nmrML) for all data types (raw data, assignment

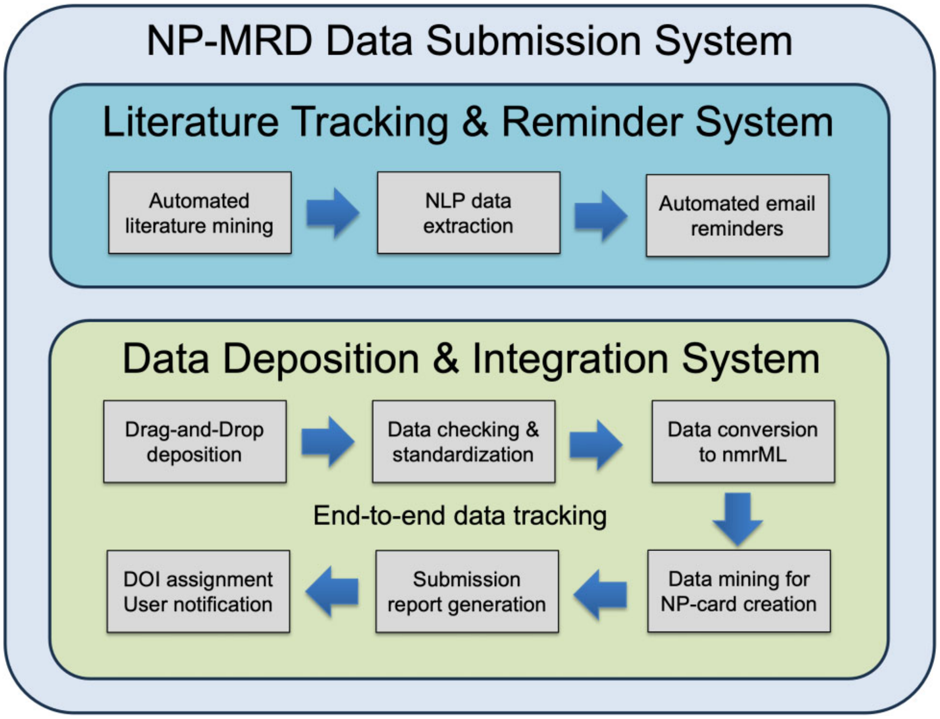

NP-MRD Data Submission System
Literature Tracking & Reminder System
Automated literature mining
NLP data extraction
Automated email reminders
Data Deposition & Integration System
Drag-and-Drop deposition
Data checking & standardization
Data conversion to nmrML
End-to-end data tracking
DOI assignment User notification
Submission report generation
Data mining for NP-card creation

Figure 2. A flow diagram depicting the two components of the NP-MRD's new data submission system, including the literature tracking and reminder system (top panel) and the data deposition and integration system (bottom panel).

data, peak lists, calculated chemical shift data, etc.) with standardization and validation tools to simplify data ingestion to NP-MRD; (x) end-to-end tracking of all deposited data, and detailed submission reports to administrators and users to provide NP-MRD accession numbers and highlight any issues with data ingestion; (xi) extensive tools for metadata extraction and validation from raw data, reducing data entry requirements for end users and improving data accuracy; (xii) linked secure account management between the deposition system and the NP-MRD database; (xiii) interface harmonization with the NP-MRD website; and (xiv) automated assignment of digital object identifiers (DOIs) for compounds with user-submitted NMR data. A flow diagram depicting the main components of the new data deposition/submission system is shown in Figure 2.

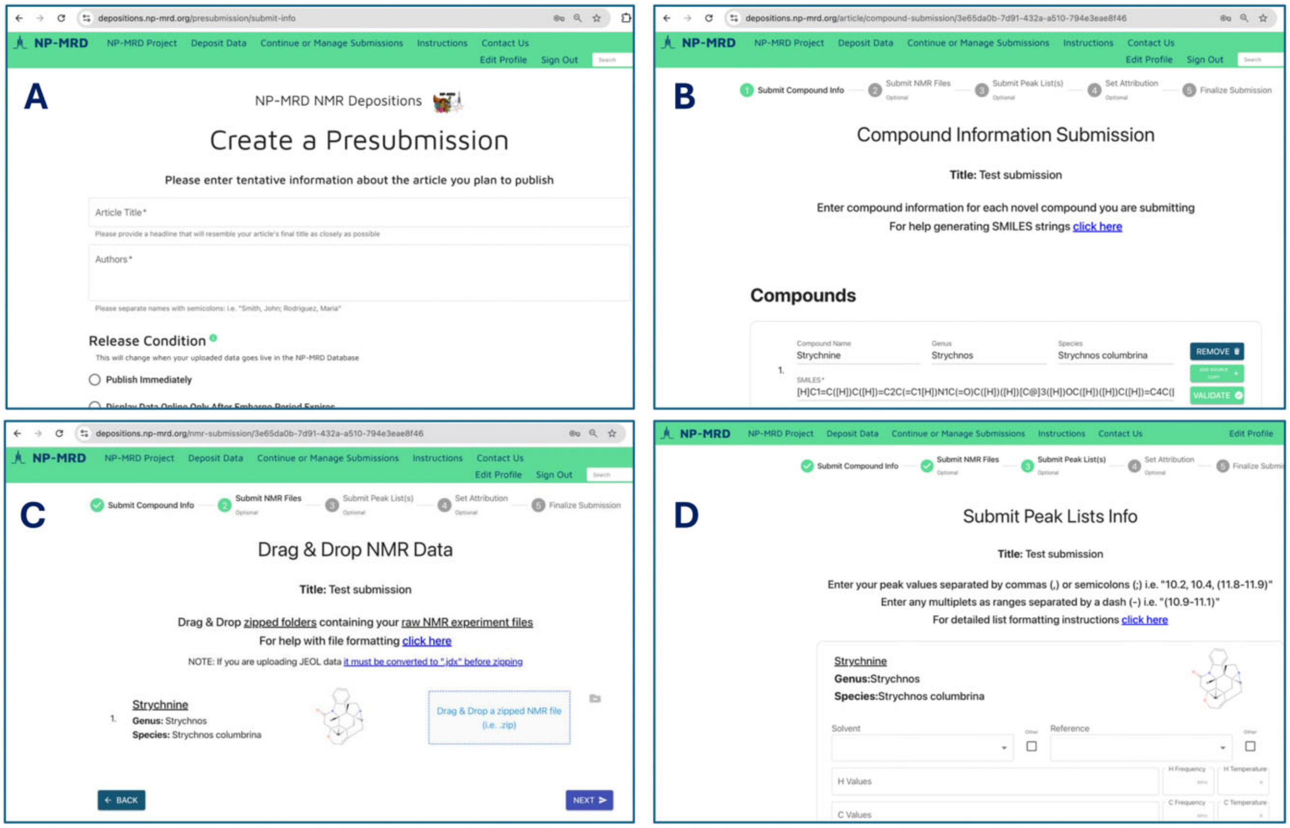
The new deposition platform not only supports all of these tasks and workflows, it has also dramatically shortened the deposition times from >30 min per compound to <5 min per compound. The consequences of these changes to the deposition system have been quite dramatic. Indeed, the number of NMR spectral depositions has grown from a 160 spectra per year in 2022 to >10 000 per year for 2024 (so far).

The new NP-MRD data deposition system supports the deposition of 1D (^{1}H and/or ^{13}C NMR data) as well as 2D (hetero or homonuclear data) from both pure NPs as well as NP mixtures and extracts. To date, the 'pure compound' deposition tool has been used to deposit data on 3413 compounds and 18 683 NMR spectra. The mixture deposition tool has been used to deposit 15 mixtures and 15 NMR mixture spectra. The user interface for pure compound deposition obviously differs from the deposition interface for the mixtures. Screenshots showing the NP-MRD data deposition interface and a portion of the step-wise process for pure compound submission are shown in Figure 3.

Better database searching

The nearly 10-fold growth in the size of the NP-MRD has obviously led to much more useful NP data being available, browsable or searchable within the NP-MRD website. However, it has also led to significant challenges regarding database searching and querying—especially with regard to performance. Not only has the number of structures grown substantially (by hundreds of thousands), the number of chemical shifts in the NP-MRD has grown even more (millions). The total number of searchable entities in the NP-MRD now numbers in the tens of millions and given that many searches are often performed over a range of values or categories (chemical shifts, masses, formulas and biological origins), the number of combined searches or search combinations quickly becomes overwhelming and, therefore, incredibly slow.

To address these issues, several database redesign efforts were undertaken. First, the database was reorganized to support rapid look-ups through the creation of composite indices that indexed on frequently searched columns, such as spectral type, NMR-detected nucleus and spectrometer frequency. Second, each chemical shift in the database was converted into integer bins that were closest to the nearest 0.01 ppm for ^{1}H chemical shifts and 0.1 ppm for ^{13}C chemical shifts. For example, a ^{1}H chemical shift of 1.27 ppm is in the bin '127'. These allowed the database to use integers for searching chemical shifts, rather than floating point values, speeding up the search significantly. Third, the chemical shift tables are cached and loaded onto solid state drives (or RAM) to accelerate the query process. Fourth, several layers of data filtering were added to each of the search functions. The intent of these filtering functions is to limit the scope or size of searches, so that spectral searches could be more targeted and, therefore, faster. For instance, users can now select or filter NP-MRD compounds based on their biological origins (plant,



depositions.np-mrd.org/presubmission/submit-info

NP-MRD NMR Depositions

Create a Presubmission

Please enter tentative information about the article you plan to publish

Article Title*

Please provide a headline that will resemble your article's final title as closely as possible

Authors*

Please separate names with semicolons: i.e. "Smith, John; Rodriguez, Maria"

Release Condition

This will change when your uploaded data goes live in the NP-MRD Database

○ Publish Immediately

○ Display Data Online Only After Embargo Period Expires

depositions.np-mrd.org/article/compound-submission/3e65da0b-7d91-432a-a510-794e3eae8f46

Compound Information Submission

Title: Test submission

Enter compound information for each novel compound you are submitting

For help generating SMILES strings click here

Compounds

1. Compound Name: Strychnine; Genus: Strychnos; Species: Strychnos columbrina

SMILES*: [H]C1=C([H])C([H])=C2C(=C1[H])N1C(=O)C([H])([H])[C@]3([H])OC([H])([H])C([H])=C4C([

depositions.np-mrd.org/nmr-submission/3e65da0b-7d91-432a-a510-794e3eae8f46

Drag & Drop NMR Data

Title: Test submission

Drag & Drop zipped folders containing your raw NMR experiment files

For help with file formatting click here

NOTE: If you are uploading JEOL data it must be converted to ".jdx" before zipping

1. Strychnine
Genus: Strychnos
Species: Strychnos columbrina

Drag & Drop a zipped NMR file (i.e. .zip)

Submit Peak Lists Info

Title: Test submission

Enter your peak values separated by commas (,) or semicolons (;) i.e. "10.2, 10.4, (11.8-11.9)"

Enter any multiplets as ranges separated by a dash (-) i.e. "(10.9-11.1)"

For detailed list formatting instructions click here

Strychnine
Genus: Strychnos
Species: Strychnos columbrina

Solvent □ Other
Reference □ Other
H Values | H Frequency MHz | H Temperature K
C Values | C Frequency MHz | C Temperature K

Figure 3. Screenshots of the new NP-MRD data deposition interface with panels highlighting some of the key submission tools, including the (**A**) pre-submission page, (**B**) compound information submission page, (**C**) drag and drop interface for NMR data submission and (**D**) peak submission interface.

animal, fungal, microbial – or combinations of these), magnetic field strength, assignment type (experimental, simulated, predicted – or combinations of these), nucleus of interest (^{1}H or ^{13}C), solvent type, molecular weight or molecular weight range and chemical formula. Similarly, the use of more efficient 'joins' in the querying process and other minor architectural optimizations were found to greatly improve the search efficiency.

Furthermore, because users may wish to perform different types of chemical shift or spectral searches, different search functions were also made available to reduce both the complexity and scope of each search. Some individuals may have only partial chemical shift sets, a few or even a single chemical shift and just want a list of potential matching/similar compounds. Other individuals may have a complete list of chemical shifts – for a single nucleus type (say ^{1}H) and want a list of specific top scoring chemical hits. Still others may have a mass (or mass range) and a set of partial shifts and want a list of top scoring chemical hits. And yet others may have heteronuclear single quantum coherence or heteronuclear multiple quantum coherence data and want compound matches with paired ^{1}H and ^{13}C shifts. In each case, users may want to have different shift tolerances based on the precision of their measurements or the perceived accuracy of the selected database set. To accommodate these different queries, several new (named) query types were created, including a 'Partial' search option, which is used for partial spectral matching, 'Pure Compound' which is used for exact spectral matching and 'Mixture' which is used for searching compound mixtures.

The implementation of these database, architecture, hardware and programmatic changes allowed search speeds to drop from 2–3 min per search to under 10 s per search against experimental NMR chemical shifts. These changes have also made the database more robust to continued expansion and to the expected growth over the coming 5 years.

Improved database utilities

Since the first release of the NP-MRD, a number of new and improved spectral utilities have been added to database. These utilities are intended to provide functions for NP-MRD users to facilitate spectral assignment, structure determination, compound dereplication and spectral generation or modeling. These include a dedicated ^{1}H chemical shift predictor (PROSPRE), a dedicated ^{13}C chemical shift predictor (CASPRE), a format conversion tool (NMRdex or NMR data exchange) and a spectral viewer (JSpectraViewer or JSV) application programming interface (API). PROSPRE (17), an ML-based method for chemical shift prediction, has an MAE between predicted and observed ^{1}H shifts of 0.10 ppm. CASPRE, an ML-based ^{13}C chemical shift predictor, has an MAE between predicted and observed ^{13}C shifts of 1.6 ppm across multiple solvents and chemical shift referencing reagents. NMRdex is a format conversion tool to facilitate NMR data exchange. It converts 1D Bruker NMR data to 1D nmrML. It also converts nmrML data to JCAMP-DX (24), as well as NMReDATA (25) to nmrML and nmrML to NMReDATA. The JSV JavaScript library has been made available for embedded usage by users. The source code along with its static assets can be downloaded as a 'minified' file for easy and quick integration into existing NMR projects. Both the 'SpectraViewer1D' and 'SpectraViewer2D' APIs may be utilized for visualizing spectral data, interactive atom assignment tables as well as a molecular structure viewer powered by JSMol (26) from nmrML files. The package is compatible with most modern JavaScript engines, such as V8 (used by Google Chrome, Microsoft Edge and Opera), SpiderMonkey (used by Mozilla Firefox) and JavaScriptCore (used by Safari).

Database implementation, curation and FAIRness

NP-MRD was developed using Ruby 2.5.1 and the Ruby on Rails web framework (http://rubyonrails.org, version 4.2.11). It employs a MariaDB relational database (https://www.mariadb.com, version 10.4.6) to manage various aspects of the 'back-end' data. This includes entity relationships, external references, chemical descriptions, chemical structures, spectra and chemical shift values. NP-MRD's design follows the model-view-controller architecture, which is optimal for separating internal data logic from user input and data presentation (15). This means the entire codebase for NP-MRD is highly modular and easily organized.

The core information stored within NP-MRD is converted into user-visible web pages through NP-MRD's HTML interface responder. The dedicated server hosting NP-MRD website runs the Ubuntu 20.04.6 LTS OS on 8-CPUs with maximal speed of 3.3 GHz, 64 GB of RAM and 400 GB of disk space, ensuring good scalability and rapid searching. The NP-MRD website is connected to the NP-MRD database server that stores all the information and has similar specifications with the exception that it has a 5 TB hard-drive. Regular backups (once a week) and stringent protocols safeguard data integrity on both the website and database servers.

All data internally uploaded to the NP-MRD has been vetted and validated by multiple curators. Likewise, all members of the NP-MRD curation team were required to have at least an undergraduate degree in chemistry, biochemistry or NP chemistry. To monitor the (internal) data entry process, all newly added data are entered into a centralized, password-controlled database, allowing all changes and edits to the database to be monitored, time-stamped and automatically transferred. All curation team members were also given extensive training by the lead curator(s) in NP-MRD annotation via hands-on mentoring, text instructions, peer support and tutorials. All data externally uploaded (by users) are vetted through a series of automated data checking routines. Users are notified of potential errors or missing information by email and/or real time, online warnings. Additional annotations and chemical shifts assignments are checked manually. This process is aided by an internally developed 'Curator App' which helps to accelerate or semi-automate the process.

The NP-MRD has embraced DOIs as part of its data submission ecosystem. A DOI is a unique alphanumeric string assigned to digital objects, such as research papers, datasets or NP-MRD data submissions. DOIs provide a persistent link to their location on the internet. Unlike URLs, which can change, DOIs offer permanent and reliable access, ensuring consistent retrieval of content. Some of the other advantages of DOIs lie in the fact that they simplify citation and linking in scientific writing by providing a standardized reference format and are associated with rich metadata (e.g. author, title and publication date), enhancing discoverability in databases and search engines. Additionally, DOIs enable tracking of citations, downloads and other usage metrics. DOIs are now being assigned to NP-cards with all user-submitted experimental data. Because of their widespread use in academic publishing, NP-MRD users can use these DOIs to support their manuscript submissions. In total, 3070 DOIs have been issued at the time of this writing.

The NP-MRD is FAIR compliant (27) and details regarding its 'FAIRness' are provided under the 'About NP-MRD' menu tab. With regard to findability, all entries in the NP-MRD have a unique and permanent 7-digit NP-MRD identifier. With regard to accessibility, the NP-MRD website is open and free, and all of its operations (including downloads) are compatible with and have been tested on most modern web browsers (Google Chrome, Microsoft Edge, Safari, Mozilla Firefox and Opera). To ensure interoperability, the NP-MRD's spectral data are available in a data exchange format (nmrML) that is universally readable, all textual data and metadata in the NP-MRD are written in English, all images are stored in standard PNG format and all nomenclature for compounds and spectral data follows standard ontologies or vocabularies used to describe these entities. An extensive and well-annotated data download section is also provided with files available in standard CSV, JSON and XML formats. To ensure re-usability, all of the data in the NP-MRD are well documented, carefully curated and extensively sourced with clear information on provenance and provide rich information about the data creation context.

Conclusion and future directions

As highlighted here, the NP-MRD has undergone significant growth (by nearly a factor of 10) in the past 3 years. Even more dramatic growth has been seen in the number of user-generated data depositions, with a nearly 100-fold increase. This has even been exceeded by the number of page views, which has seen a nearly a 300-fold increase. This growth has led to a number of unexpected challenges. However, as outlined in this manuscript, they have been successfully addressed leading to the creation of a much more user-friendly, faster, more accurate, and more resilient NP data resource. The decision to include both primary and secondary metabolites within the NP-MRD has made the NP-MRD more broadly appealing. Likewise, the decisions to draw from many well-known and well-regarded databases to obtain more diverse chemical data has led to a much richer, more complete data resource. The decision to address data deposition bottlenecks with a more intelligent, more intuitive system has led to a number of improvements. Data deposition has now been greatly simplified, which has played a large role in the significant increase in external user depositions. Data backfilling has been greatly improved and accelerated, which has increased the volume of experimental assignment data available in the NP-MRD. Data search speeds and search capabilities have been significantly enhanced, making that database far more useful for structure discovery and matching. A number of key NMR utilities and spectral predictors have also been enhanced, all of which are improving the look, content, user-friendliness and utility of the database. However, much still remains to be done.

In addition to these database coordination activities, some of the other planned improvements to the NP-MRD include developing much more sophisticated automation for database operation and data deposition. This automation is needed for the NP-MRD to efficiently scale up its operation as rates of deposition and retrieval increase. This will include the development of automated processes and APIs for improved querying and data quality review to allow team members to easily update and correct database entries. New support will also be added for handling and depositing large NP datasets. Additionally, we will implement automated peak assignment for deposited NMR data using modifications to the newly developed chemical shift and J-coupling predictors. Greater use of ML, chemical ontologies [ChemFOnt (28) and large language models] will also be used to help improve the harvesting of archival and newly published NP descriptive data from the scientific literature. To allow better identification or more facile discovery of novel NPs, the NP-MRD will be expanded to include generative chemical language models (CLMs) of predicted NPs (29). It is expected that these new CLMs will generate nearly 70 million biofeasible NP structures (30). Improved spectral prediction tools (NMR, retention time and mass spectral predictors) and the corresponding predictions will be added to support *de novo* or computer-aided structure elucidation using the predicted CLM NPs.

Another key development will be a focus on adding interoperability to the NP-MRD with other spectral and NP databases. Indeed, the NP-MRD is not the only archival NMR database available. Other NMR data deposition and analysis platforms, with somewhat different mandates and goals, are available to the NMR community, such as NMRShiftDB (31) and NMRXiv (32). Over the coming year, the NP-MRD team will be working with these databases to increase data sharing and to help develop more uniform and consistent data submission processes. Additionally, given the role that MS now plays in NP structure elucidation and characterization, the NP-MRD will start accepting and including reference MS data of NP compounds in the coming year. So rather than being known as the Natural Products Magnetic Resonance Database, the NP-MRD may 'essentially' become the Natural Product Material Reference Database. The inclusion of MS data will begin by federating the NP-MRD's NMR data with MS data from other NP databases, including GNPS (33), MoNA (https://mona.fiehnlab.ucdavis.edu/), MassBank (34), FooDB and HMDB. This will involve creating automated tools to query these databases for spectral matches as well as developing protocols for distributing and receiving cross-links between resources. It will also involve harvesting open MS data from other resources and predicting MS data using a variety of tools developed by our team (35). NP-MRD will make all its data, including MS data, freely available to both its partner databases and the NP user community.

Data availability

All data from NP-MRD are freely accessible through the NP-MRD website at https://np-mrd.org and the dedicated NP-MRD download page. The platform is open and free to use, with data downloads compatible with all modern web browsers. Download files are available in XML, JSON, nmrML, TXT, CSV, SDF and JCampDX formats. Additionally, NP data can be accessed in multiple formats, including SMILES, SDF, MOL, PDB, InChI and InChIKey. NMR spectra are available in JCampDX, nmrML and vendor formats. Images are available in PNG format, and all textual data and metadata are written in English. The data in NP-MRD are released under a Creative Commons Attribution BY and NC license.

Acknowledgements

We wish to thank Dr Marcia LeVatte for her assistance in preparing the manuscript.

Funding

Natural Sciences and Engineering Research Council of Canada; Canada Foundation for Innovation; National Center for Complementary and Integrative Health (NCCIH); Office of Dietary Supplements (ODS) [U24AT010811]. Funding for open access charge: National Institutes of Health.

Conflict of interest statement

None declared.

References

1. Bar-On,Y.M., Phillips,R. and Milo,R. (2018) The biomass distribution on earth. *Proc. Natl Acad. Sci. U.S.A.*, **115**, 6506–6511.
2. Atanasov,A.G., Zotchev,S.B., Dirsch,V.M., International Natural Product Sciences Taskforce and Supuran,C.T. (2021) Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.*, **20**, 200–216.
3. González-Manzano,S. and Dueñas,M. (2021) Applications of natural products in food. *Foods*, **10**, 300.
4. Liu,J.-K. (2022) Natural products in cosmetics. *Nat. Prod. Bioprospect.*, **12**, 40.
5. Duranton,F., Jankowski,J., Więcek,A. and Argilés,À. (2016) On the discovery of UREA. Identification, synthesis and observations that let to establishing the first uraemic retention solute. *G. Ital. Nefrol.*, **33**, 33.S66.16.
6. Katz,L. and Baltz,R.H. (2016) Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biotechnol.*, **43**, 155–176.
7. Danelius,E., Halaby,S., van der Donk,W.A. and Gonen,T. (2021) MicroED in natural product and small molecule research. *Nat. Prod. Rep.*, **38**, 423–431.
8. Breton,R.C. and Reynolds,W.F. (2013) Using NMR to identify and characterize natural products. *Nat. Prod. Rep.*, **30**, 501–524.
9. Bouslimani,A., Sanchez,L.M., Garg,N. and Dorrestein,P.C. (2014) Mass spectrometry of natural products: current, emerging and future technologies. *Nat. Prod. Rep.*, **31**, 718–729.
10. Kong,L.-Y. and Wang,P. (2013) Determination of the absolute configuration of natural products. *Chin. J. Nat. Med.*, **11**, 193–198.
11. Halabalaki,M., Vougogiannopoulou,K., Mikros,E. and Skaltsounis,A.L. (2014) Recent advances and new strategies in the NMR-based identification of natural products. *Curr. Opin. Biotechnol.*, **25**, 1–7.
12. Wishart,D.S., Sayeeda,Z., Budinski,Z., Guo,A.C., Lee,B.L., Berjanskii,M., Rout,M., Peters,H., Dizon,R., Mah,R., *et al.* (2022) NP-MRD: the Natural Products Magnetic Resonance Database. *Nucleic Acids Res.*, **50**, D665–D677.
13. Pin,M., Poynton,E.F., Jordan,T., Kim,J., Ledingham,B., van Santen,J.A., Yang,V., Maras,A., Tavangar,P., Gautam,V., *et al.* (2023) A data deposition platform for sharing nuclear magnetic resonance data. *J. Nat. Prod.*, **86**, 2554–2561.
14. van Santen,J.A., Poynton,E.F., Iskakova,D., McMann,E., Alsup,T.A., Clark,T.N., Fergusson,C.H., Fewer,D.P., Hughes,A.H., McCadden,C.A., *et al.* (2022) The Natural Products Atlas 2.0: a

database of microbially-derived natural products. *Nucleic Acids Res.*, **50**, D1317–D1323.

15. Wishart,D.S., Guo,A., Oler,E., Wang,F., Anjum,A., Peters,H., Dizon,R., Sayeeda,Z., Tian,S., Lee,B.L., *et al.* (2022) HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res.*, **50**, D622–D631.
16. Hoch,J.C., Baskaran,K., Burr,H., Chin,J., Eghbalnia,H.R., Fujiwara,T., Gryk,M.R., Iwata,T., Kojima,C., Kurisu,G., *et al.* (2023) Biological Magnetic Resonance Data Bank. *Nucleic Acids Res.*, **51**, D368–D376.
17. Sajed,T., Sayeeda,Z., Lee,B.L., Berjanskii,M., Wang,F., Gautam,V. and Wishart,D.S. (2024) Accurate prediction of ^{1}H NMR chemical shifts of small molecules using machine learning. *Metabolites*, **290**, 14.
18. Afendi,F.M., Okada,T., Yamazaki,M., Hirai-Morita,A., Nakamura,Y., Nakamura,K., Ikeda,S., Takahashi,H., Altaf-Ul-Amin,M., Darusman,L.K., *et al.* (2012) KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol.*, **53**, e1.
19. Wishart,D.S., Oler,E., Peters,H., Guo,A., Girod,S., Han,S., Saha,S., Lui,V., LeVatte,M., Gautam,V., *et al.* (2023) MiMeDB: the Human Microbial Metabolome Database. *Nucleic Acids Res.*, **51**, D611–D620.
20. Rutz,A., Sorokina,M., Galgonek,J., Mietchen,D., Willighagen,E., Gaudry,A., Graham,J.G., Stephan,R., Page,R., Vondrášek,J., *et al.* (2022) The LOTUS initiative for open knowledge management in natural products research. *Elife*, **11**, e70780.
21. Wishart,D.S., Feunang,Y.D., Marcu,A., Guo,A.C., Liang,K., Vázquez-Fresno,R., Sajed,T., Johnson,D., Li,C., Karu,N., *et al.* (2018) HMDB 4.0: the Human Metabolome Database for 2018. *Nucleic Acids Res.*, **46**, D608–D617.
22. Djoumbou Feunang,Y., Eisner,R., Knox,C., Chepelev,L., Hastings,J., Owen,G., Fahy,E., Steinbeck,C., Subramanian,S., Bolton,E., *et al.* (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.*, **8**, 61.
23. Yesiltepe,Y., Nuñez,J.R., Colby,S.M., Thomas,D.G., Borkum,M.I., Reardon,P.N., Washton,N.M., Metz,T.O., Teeguarden,J.G., Govind,N., *et al.* (2018) An automated framework for NMR chemical shift calculations of small organic molecules. *J. Cheminform.*, **10**, 52.
24. Lancashire,R.J. (2007) The JSpecView project: an open source Java viewer and converter for JCAMP-DX, and XML spectral data files. *Chem. Cent. J.*, **1**, 31.
25. Pupier,M., Nuzillard,J.-M., Wist,J., Schlörer,N.E., Kuhn,S., Erdelyi,M., Steinbeck,C., Williams,A.J., Butts,C., Claridge,T.D.W., *et al.* (2018) NMReDATA, a standard to report the NMR assignment and parameters of organic compounds. *Magn. Reson. Chem.*, **56**, 703–715.
26. Hanson,R.M., Prilusky,J., Renjian,Z., Nakane,T. and Sussman,J.L. (2013) JSmol and the next-generation web-based representation of 3D molecular structure as applied to proteopedia. *Isr. J. Chem.*, **53**, 207–216.
27. Wilkinson,M.D., Dumontier,M., Aalbersberg,I..J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.W., da Silva Santos,L.B., Bourne,P.E., *et al.* (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
28. Wishart,D.S., Girod,S., Peters,H., Oler,E., Jovel,J., Budinski,Z., Milford,R., Lui,V.W., Sayeeda,Z., Mah,R., *et al.* (2022) ChemFOnt: the chemical functional ontology resource. *Nucleic Acids Res.*, **51**, D1220–D1229.
29. Skinnider,M.A., Stacey,R.G., Wishart,D.S. and Foster,L.J. (2021) Chemical language models enable navigation in sparsely populated chemical space. *Nat. Mach. Intell.*, **3**, 759–770.
30. Tay,D.W.P., Yeo,N.Z.X., Adaikkappan,K., Lim,Y.H. and Ang,S.J. (2023) 67 million natural product-like compound database generated via molecular language processing. *Sci. Data*, **10**, 296.
31. Steinbeck,C., Krause,S. and Kuhn,S. (2003) NMRShiftDB—constructing a free chemical information system with open-source components. *J. Chem. Inf. Comput. Sci.*, **43**, 1733–1739.
32. Sharma,N., Bach,F., Herres-Pawlis,S., Jung,N., Liermann,J., Nainala,V.C., Razum,M., Koepler,O., Kuhn,S. and Steinbeck,C. (2022) nmrXiv—Open, FAIR and consensus-driven NMR spectroscopy data repository and analysis platform. NFDI4Chem.
33. Wang,M., Carver,J.J., Phelan,V.V., Sanchez,L.M., Garg,N., Peng,Y., Nguyen,D.D., Watrous,J., Kapono,C.A., Luzzatto-Knaan,T., *et al.* (2016) Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.*, **34**, 828–837.
34. Horai,H., Arita,M., Kanaya,S., Nihei,Y., Ikeda,T., Suwa,K., Ojima,Y., Tanaka,K., Tanaka,S., Aoshima,K., *et al.* (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.*, **45**, 703–714.
35. Wang,F., Pasin,D., Skinnider,M.A., Liigand,J., Kleis,J.-N., Brown,D., Oler,E., Sajed,T., Gautam,V., Harrison,S., *et al.* (2023) Deep learning-enabled MS/MS spectrum prediction facilitates automated identification of novel psychoactive substances. *Anal. Chem.*, **95**, 18326–18334.

Received: September 15, 2024. **Revised:** October 17, 2024. **Editorial Decision:** October 18, 2024. **Accepted:** October 25, 2024