D368–D376 *Nucleic Acids Research, 2023, Vol. 51, Database issue* *Published online 7 December 2022*
https://doi.org/10.1093/nar/gkac1050

Biological Magnetic Resonance Data Bank

Jeffrey C. Hoch[1,*], Kumaran Baskaran[1], Harrison Burr[1], John Chin[1], Hamid R. Eghbalnia[1], Toshimichi Fujiwara[2], Michael R. Gryk[1], Takeshi Iwata[2], Chojiro Kojima[2,3], Genji Kurisu[2], Dmitri Maziuk[1], Yohei Miyanoiri[2], Jonathan R. Wedell[1], Colin Wilburn[1], Hongyang Yao[1] and Masashi Yokochi[2]

[1]Department of Molecular Biology and Biophysics, UConn Health, Farmington, CT 06030-3305, USA, [2]Protein Data Bank Japan, Institute for Protein Research, Osaka University, Suita, Osaka 565-0871. Japan and [3]Graduate School of Engineering Science, Yokohama National University, Yokohama 240-8501, Japan

Received October 04, 2022; Revised October 20, 2022; Editorial Decision October 21, 2022; Accepted October 23, 2022

ABSTRACT

The Biological Magnetic Resonance Data Bank (BMRB, https://bmrb.io) is the international open data repository for biomolecular nuclear magnetic resonance (NMR) data. Comprised of both empirical and derived data, BMRB has applications in the study of biomacromolecular structure and dynamics, biomolecular interactions, drug discovery, intrinsically disordered proteins, natural products, biomarkers, and metabolomics. Advances including GHz-class NMR instruments, national and trans-national NMR cyberinfrastructure, hybrid structural biology methods and machine learning are driving increases in the amount, type, and applications of NMR data in the biosciences. BMRB is a Core Archive and member of the World-wide Protein Data Bank (wwPDB).

INTRODUCTION

NMR is one of the most versatile analytic methods for analyzing matter, using atomic nuclei as embedded reporters. Applications of NMR to biomolecular science include structure, dynamics, and interactions of biomacromolecules in the solid and liquid solution state, the identity and composition of complex mixtures of small molecule metabolites, and the structure of natural products. BMRB is the open, international online repository for all types of biomolecular NMR data.

BMRB was founded by John Markley and Eldon Ulrich at the University of Wisconsin in 1988. Initial depositions were made by curators manually entering chemical shifts and assignments extracted from publications. With the development of the Adit-NMR deposition system, the primary means of accessions became depositor-driven. In 2006, BMRB joined the World-wide Protein Data Bank (wwPDB) (1) as a member and Core Archive. BMRB facilitates accession and validation of NMR-based biomacromolecular structures deposited to PDB via the OneDep system. BMRBj operating from Osaka University processes depositions in addition to serving as a mirror site. Another mirror site operates from CERM (https://bmrb.cerm.unifi.it/), at the University of Florence, IT. In 2020, BMRB moved to UConn Health in Farmington, CT.

BMRB depositions are linked to corresponding entries in PDB, the European Nucleotide Archive (2) (EMBL-EBI), the National Center for Biotechnology Information (3) (NCBI) and UniProt (4).

In addition to collaborations with wwPDB member organizations RCSB (5), PDBe (6), PDBj, (7) and EMDB (8), BMRB has close collaborations with the National Center for Biomolecular NMR Data Processing and Analysis (9) (NMRbox.org), the Center for High-Throughput Computing and the University of Wisconsin-Madison, the Collaborative Computing Project for NMR (10) (CCPN), Rutgers University, and the University of Frankfurt.

BMRBj (7), formerly known as PDBj-BMRB at Osaka, Japan, is a satellite BMRB repository that accepts NMR experimental data via three deposition servers, (i) OneDep depositions containing NMR experimental data from Asian countries and regions, (ii) BMRBdep deposition server at Osaka and (iii) SMSDep deposition server, which accepts NMR derived structures of biologically interesting small molecules that are ineligible for OneDep deposition due to the limit on molecular size. 10% of BMRB entries have been processed at Osaka site since 2002 (7). The BMRB and BMRBj sites share annotation tools and communicate closely on issues, assuring uniformity of annotation quality using either site.

MATERIALS AND METHODS

Accessions

Data housed at the BMRB are deposited, archived, and disseminated in the NMR-STAR file format (11). NMR-STAR is a variant of the STAR file format (12–14) intro-

[*]To whom correspondence should be addressed. Tel: +1 860 6798; Email: hoch@uchc.edu

duced by Hall *et al*. in the early 1990s. STAR supports both tabular data as well as key-value pairs and is the format in use by the PDB. STAR also provides for optional grouping and stacking of similar data elements in what are referred to as save frames. NMR-STAR uses IUPAC (15) atom nomenclature. BMRB also supports the NMR Exchange Format (NEF) (16), which is also a variant of the STAR format, and conforms to definitions from the Collaborative Computing Project for NMR (CCPN) data model (10).

The BMRB maintains the NMR-STAR data dictionary, the details of which can be found on the website (https://bmrb.io/dictionary/). The current dictionary version (3.2) consists of more than 6500 defined data tags organized across >100 save frame categories corresponding to eight super-groups. These data definitions cover the various types of data archived at the BMRB: chemical shifts, relaxation, and hydrogen exchange rates (including nuclear Overhauser effects, NOE's, and scalar or dipolar coupling constants). They also cover the various types of metadata supporting the depositions: molecular assembly definitions, sample conditions and experiment setup, author information, along with cross-references to other databases of interest. The NMR-STAR dictionary is also used by external software tools which read and/or write NMR-STAR data.

The hierarchical nature of the STAR file format makes it suitable for representation as XML or JSON. A generic XML schema definition has been proposed for STAR files in general (17); the BMRB also disseminates entries in other formats conforming to a schema specific for the current version of the NMR-STAR data dictionary (18).

Whereas the content of the master BMRB archive and BMRBj are identical, BMRBj has explored alternative approaches to enhance reusability and interoperability of NMR experimental data using open standard and community-driven technologies. One example is to provide NMR experimental data in web standard formats, XML and RDF (https://www.w3.org/TR/rdf11-concepts/). This extended archive is now also available from the main BMRB web site (18). Based on the extended archive, BMRBj launched a portal site that can search BMRB and related databases simultaneously and display NMR experimental data as rich graphical content (19). BMRBj is leading the effort to remediate legacy depositions of NMR restraints in PDB, converting software-specific formats to the community-based NEF format.

Deposition systems

BMRB currently operates four distinct deposition systems, each with a specific role (Figure 1). OneDep (20) is the system jointly managed with the wwPDB for accession of NMR data in support of a biomacromolecular structure, typically proteins or nucleic acids. It accepts atomic coordinates, chemical shifts, geometrical constraints, and nuclear Overhauser effect (NOESY) peak lists. BMRBdep (21) is the deposition system for NMR data not associated with a PDB structure deposition. It accepts data for systems that do not meet the size threshold for PDB, or include data that is not supported by OneDep, for example raw time-domain NMR data. SMSDep is the deposition system for biomolecule structures that do not fit the OneDep criteria. BMRB also operates a generic, uncurated deposition system that accepts arbitrary data, BMRbig (bmrbig.org).

BMRBdep

BMRBdep is the internally developed and managed system for capturing NMR data that is not associated with a PDB structure deposition. BMRBdep is a SPA (single page application) developed using the modern Angular web application framework. It supports *de novo* depositions, using an existing deposition as a starting point, or bootstrapping a deposition via the upload of an NMR-STAR file. In addition, it supports 'off-line' deposition work - allowing deposition form to be filled even if internet access is temporarily unavailable, and saving the changes when connectivity is restored. Another notable feature is the ability to 'clone' a deposition at any point during the deposition process, which is helpful when depositing multiple datasets that share some metadata.

BMRBdep was designed to allow for future enhancements such as the ability for a depositor to access a 'profile' page of all their depositions, as well as the ability to grant access to other co-depositors or grant limited read-only access to journal reviewers prior to release. After deposition, data from BMRBdep is entered into the common BMRB annotation process which is also used for OneDep and SMSDep depositions.

OneDep

The chemical shifts, restraints, and peak lists from structure determination studies are currently deposited through the OneDep (https://deposit.wwpdb.org) system. OneDep is a unified system for deposition, biocuration, and validation of experimentally determined structures of biological macromolecules to the PDB archive. It maintained as a global collaboration by the wwPDB partners (RCSB-PDB, PDBe, PDBj, BMRB and EMDB). BMRB developed the biocuration and validation protocol for the NMR data for the OneDep system.

The OneDep system currently accepts all the NMR experimental data either in NMR-STAR v3.2 or in NMR Exchange Format (NEF) v1.2. The legacy method of uploading chemical shifts in NMR-STAR format, with restraints and peak lists in any software-specific format will be phased out, tentatively by the end of 2023. The OneDep system shares the NMR data after initial bio curation and validation with BMRB for further annotation. BMRB adds additional metadata including author and experimental details, sample conditions, etc. to the deposition and prepares a stand-alone NMR-STAR file for the BMRB archive. The wwPDB releases most of the metadata along with the structure coordinate data in CIF (22) format and experimental data with minimal metadata in NMR-STAR format. The data will be released simultaneously by both the PDB and BMRB archives subsequent to a release request from the depositor.

WwPDB has a data deposition policy that specifies certain minimal sizes for biopolymers (https://www.wwpdb.org/documentation/policy). Crystal structures of peptides with fewer than 24 residues within any polymer chain

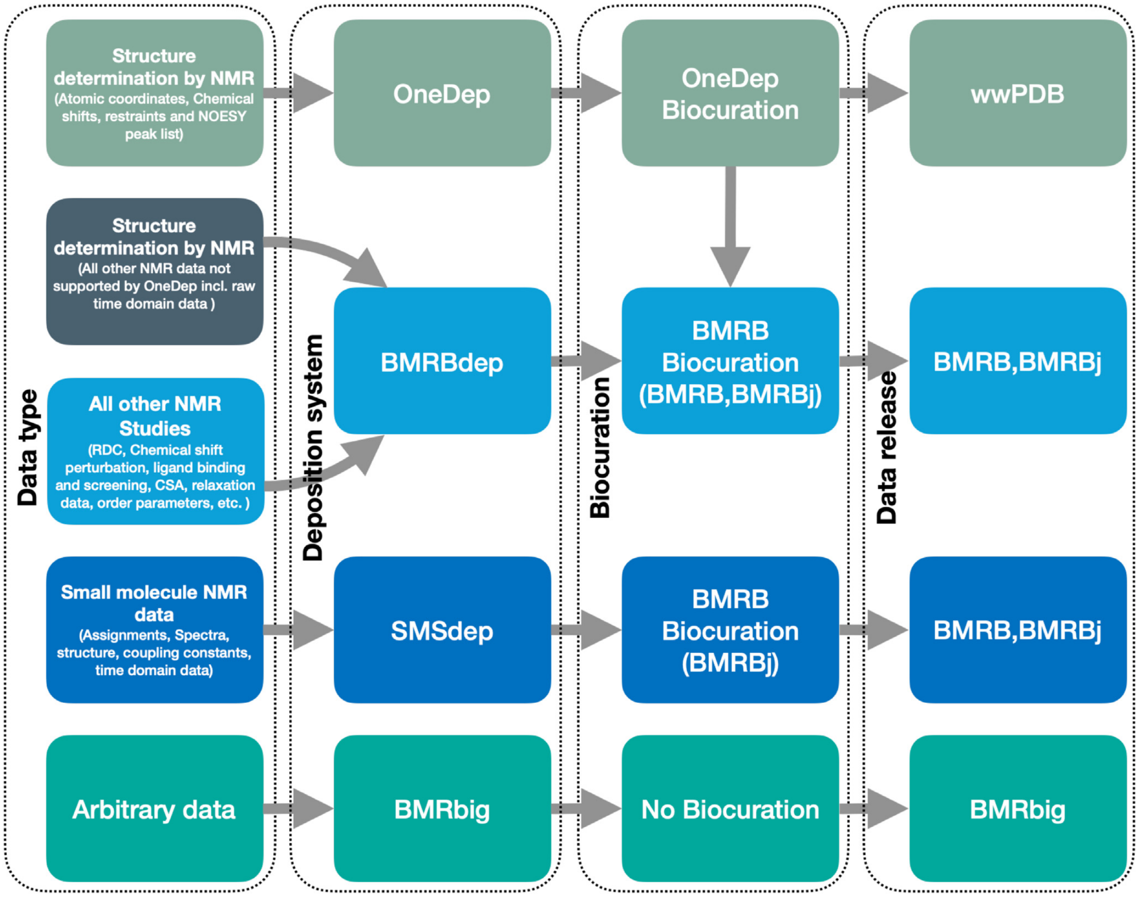

Data type
Deposition system
Biocuration
Data release
Structure determination by NMR
(Atomic coordinates, Chemical shifts, restraints and NOESY peak list)
OneDep
OneDep Biocuration
wwPDB
Structure determination by NMR
(All other NMR data not supported by OneDep incl. raw time domain data)
BMRBdep
BMRB Biocuration (BMRB,BMRBj)
BMRB,BMRBj
All other NMR Studies
(RDC, Chemical shift perturbation, ligand binding and screening, CSA, relaxation data, order parameters, etc.)
Small molecule NMR data
(Assignments, Spectra, structure, coupling constants, time domain data)
SMSdep
BMRB Biocuration (BMRBj)
BMRB,BMRBj
Arbitrary data
BMRbig
No Biocuration
BMRbig

Figure 1. BMRB supports deposition through OneDep, BMRBdep, SMSdep and BMRBbig.

that do not meet the criteria wwPDB criteria should be deposited at the Cambridge Crystallographic Data Centre (23) (CCDC, http://www.ccdc.cam.ac.uk/products/csd/deposit/). NMR structures of such molecules should be submitted to BMRB through SMSDep (http://smsdep.protein.osaka-u.ac.jp/bmrb-adit/).

SMSDep

SMSDep was developed by BMRB for biomolecular structures that cannot be deposited via OneDep, in particular: peptides with fewer than 24 residues in any polymer chain. SMSDep system is operated by BMRB-Japan (BMRBj), with structure validation carried out by PDBj and NMR data curation by BMRBj. SMSDep is the only option available for depositing experimentally determined structures of small peptides from NMR studies.

Content

The BMRB archive is dominated by assigned chemical shifts (frequencies of nuclear resonances) for ^{1}H, ^{13}C and ^{15}N nuclei in proteins and nucleic acids. There are currently >10 868 000 assigned chemical shifts corresponding to 15451 studies in the archive. Additional data includes nuclear magnetic relaxation rates, hydrogen exchange rates, and scalar internuclear coupling constants. There are experimental time-domain data sets and experimental reference spectra for more than 1200 common small-molecule metabolites. The BMRB archive includes a richly annotated collection of small molecule entries that are uniquely identified by ALATIS-derived atom nomenclature IDs. Additionally, more than 650 standard small-molecule spectra are complemented by GISSMO-derived spin system matrices, thus enabling 1D NMR spectral simulation at arbitrary spectrometer field strengths.

Shift statistics

BMRB hosts more than 11 million experimentally measured chemical shifts from biologically important molecules including proteins, nucleic acids and metabolites. Figure 2 depicts the chemical shifts content of BMRB by molecule and atom type. Up-to-date and detailed statistics on content of BMRB in tablular form is available from the 'Query grid' link on the search page at https://bmrb.io/search. Other precomputed query data is also available.

Services

In addition to data accessions via the OneDep (for biomacromolecular structures determined by NMR), BMRBdep (for biomolecule NMR data not associated with a macromolecular structure), and SMSDep (for small

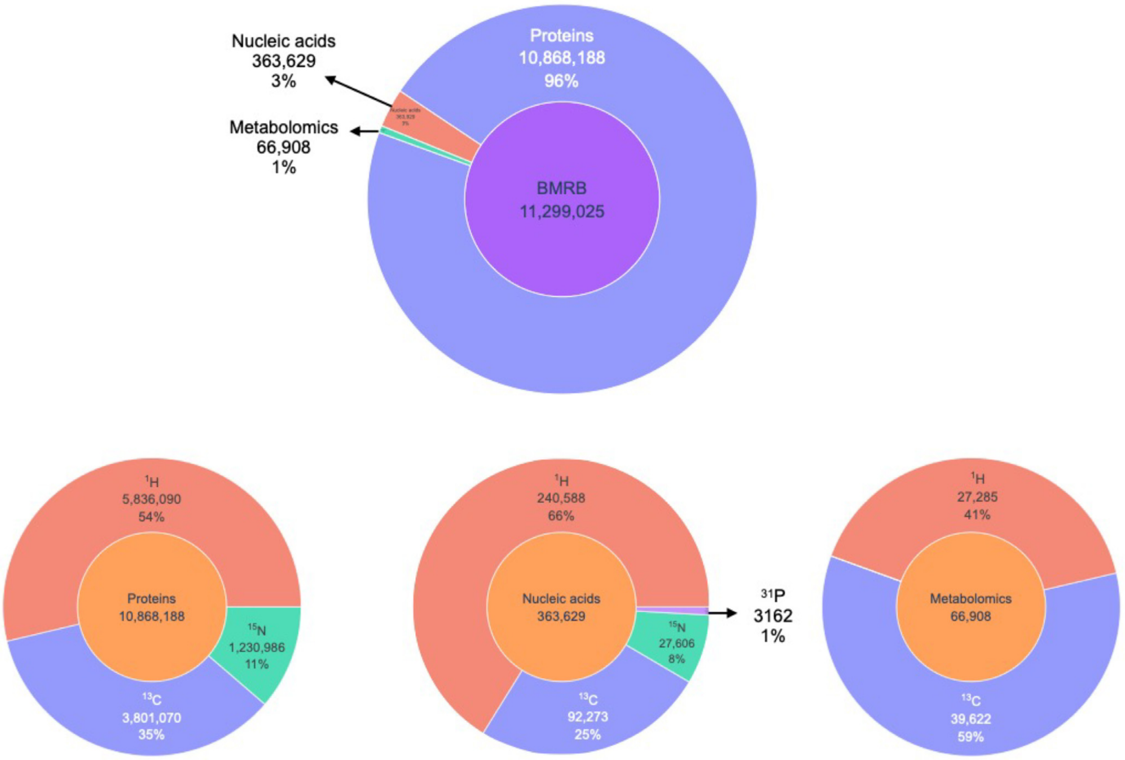


Figure 2. Chemical shift statistics for BMRB as of 14 September 2022. The chemical shifts for nuclei other than ^{1}H, ^{13}C, ^{15}N and ^{31}P are included in the totals, but not shown in the chart because of their small contribution.

molecules) services, BMRB provides a web-based service for querying the BMRB archive, application programming interfaces (APIs) for querying the archive and manipulating NMR data in NMR-STAR format, and endpoints for batch downloads of the archive. Snapshots of the archive are maintained on the NMRbox (nmrbox.org) platform. BMRB also provides a CS-Rosetta service for determining protein structure from NMR data via a dedicated high-performance computing cluster. The BMRB web site also provides rich statistics on the archive content, documentation of NMR data standards, and reference and educational resources related to biomolecular NMR. The BMRbig (bmrbig.org) platform provides a persistent archive for arbitrary data on biomolecules, and serves as a transitional repository enroute to full BMRB deposition. BMRB also hosts the ALATIS server (https://alatis.bmrb.io) for generating unique INCHI IDs. BMRB provides software resources through the GitHub (https://github.com/bmrb-io) repository for developers. The source code for the NMR-STAR parser (PyNMRSTAR : https://pypi.org/project/pynmrstar/) and data visualization tools in Python(PyBMRB: https://pypi.org/project/pybmrb/)and R(RBMRB https://CRAN.R-project.org/package=RBMRB) are available through BMRB GitHub.

Curated data types

The BMRB archive accepts arbitrary data types, however most major types of biomolecular NMR data are validated as part of the curation process. The most important are assigned chemical shifts, i.e. the nuclear resonance frequencies of specific atoms in the biomolecule. Other types include distance restraints between hydrogen atoms, derived from nuclear Overhauser effects (NOE); nuclear relaxation rates including spin-lattice (R_1), spin-spin (R_2) and heteronuclear NOE; and unassigned chemical shifts (as peak tables in NMR-STAR or text format).

Accessing content

BMRB services are primarily accessed through the BMRB web site (https://bmrb.io—see next section) but BMRB data is made available in a variety of ways to facilitate community access. These support both interactive use as well as programmatic access. For programmatic access, BMRB provides data via an API, an rsync endpoint, a Globus endpoint, and via ReBoxitory. The BMRB API supports programmatic queries against the BMRB database. Querying for entries based on tag values (searches) is supported, as is looking up corresponding entries from related databases (PDB entries, UniProt entries), searching through the chemical shifts, searching for structures based on a set of chemical shifts, and many other queries. (https://github.com/bmrb-io/BMRB-API)

Additional bulk access to data (both NMR-STAR data, as well as primary data such as time domain data) is supported via rsync and Globus endpoints. Both rsync and Globus are software tools that enable a user to select one or more folders of data to synchronize between a local computer and the BMRB archive. rsync requires a manual step to update the local copy of data, whereas Globus can keep data synchronized continuously. rsync is a standard Unix tool and is included by default on nearly all Linux and Mac operating systems (though also available on Windows). Globus requires downloading the Globus utilities and creating an account before accessing BMRB data, but it is commonly used in academic settings for accessing and transferring data, and as such many researchers already have the Globus software installed and configured. Within the Globus software, searching for 'BMRB' will return the available BMRB data sets within the Globus network. More details on using rsync or Globus to access BMRB data are available on the BMRB web site (https://bmrb.io/data_library/rsync.shtml, https://bmrb.io/data_library/globus.shtml)

Another means of data access is via ReBoxitory - a 'data lake' provided automatically to all users of the NMRbox Platform as a Service (PaaS). ReBoxitory provides versioned copies of multiple public domain databases (BMRB, PDB, AlphaFold, etc.) via the file system within NMRbox. NMRbox users can simply browse to */reboxitory/$year/$month/$database* to access the data. For BMRB, the data is available in the 'BMRB' folder. $year and $month are variables which should be substituted with a 4-digit year and 2-digit month to access a given snapshot of the data. Available data includes the entire macromolecule and small molecule archives in NMR-STAR format, any corresponding raw data, as well as dumps of the relational database tables in use by BMRB for these databases.

Web site

The BMRB web site underwent a major modernization project in 2020, the results of which are now public. The redesigned site includes a significantly improved navigation panel which takes up less space on the screen and also reorganizes content for easier navigation according to intuitive categories, has a mobile version of the navigation that automatically displays on mobile devices, loads faster, and has a new home page designed to help users discover available BMRB services. As part of this modernization, the BMRB has also moved from its traditional bmrb.wisc.edu domain to the new bmrb.io domain and has rebranded with a much more modern and heuristic logo that matches the redesign of the BMRB web site.

Primary changes that users will notice are the new design, the new home page which highlights primary BMRB resources in a tiled grid, and the improved navigation menu. The primary sections of the new BMRB menu help different types of users (from students to established domain experts) to quickly locate the resources they are looking for. The new primary menu categories are About, Deposit, Search, Visualize, Analyze, Data and Learn. When these menu categories are highlighted (or selected, in the case of a mobile device) they expand to show the various resources available within that category. Furthermore, resources can be categorized under additional headers within these categories, such as the various links in the 'Data preparation' section of the Deposit category.

Despite the major improvements to the navigation and appearance of the site, no legacy URLs were broken in the process of upgrading the site. This means that bookmarks and published links to BMRB were operative during the update process.

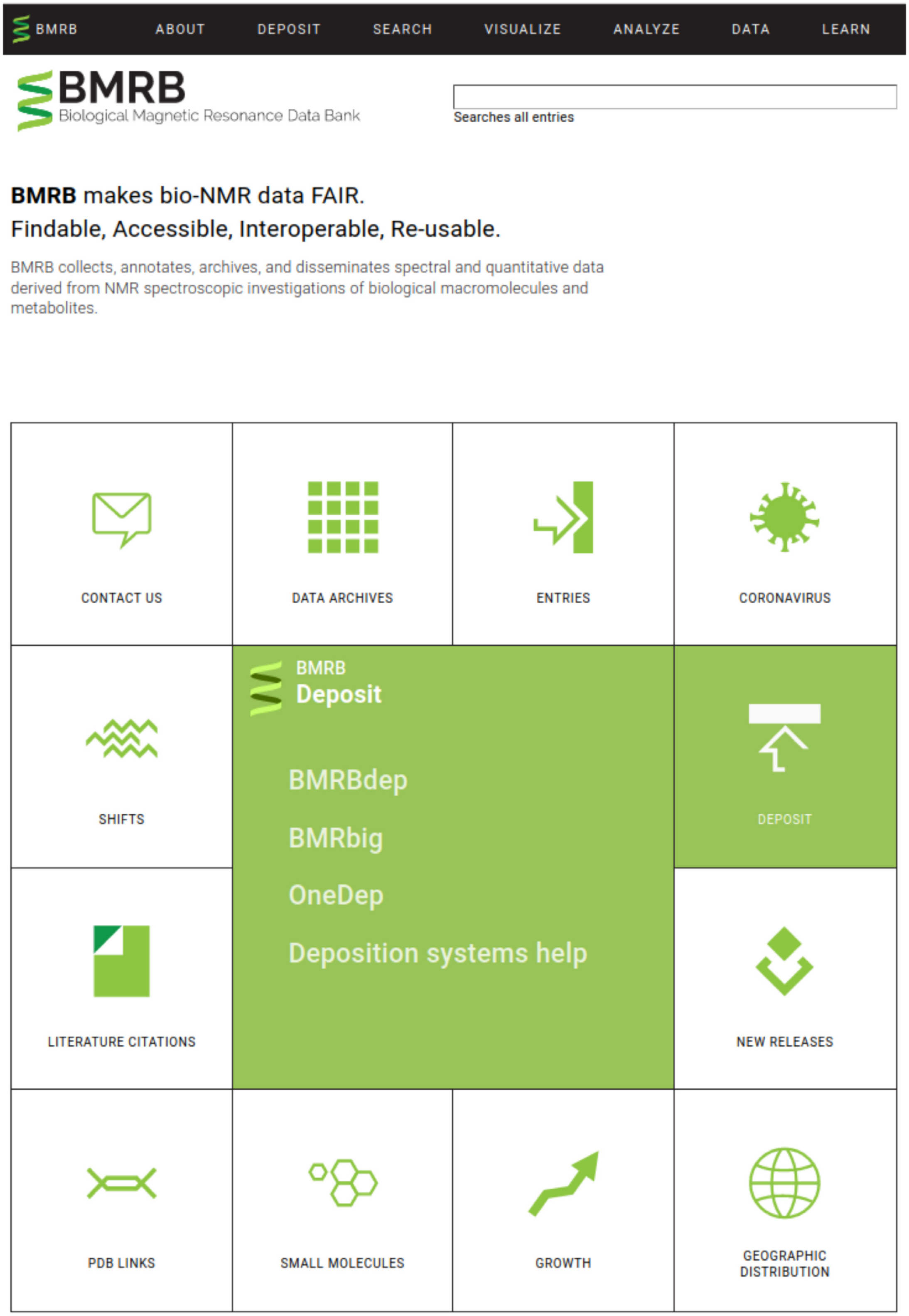
In addition to the site redesign, BMRB is continually working on new features, such as an improved deposition statistics page (https://bmrb.io/bmrb/deposition_stats.shtml) and an interactive map showing the geographic distribution of BMRB entries (https://bmrb.io/bmrb/deposition-map.shtml). User suggestions are welcomed at any time, and can be made by clicking the 'Support' button on the lower right corner of the page and submitting a request to BMRB staff.

Figure 3 depicts the new home page of the BMRB, showing the revised navigation menu as well as the new home page grid. Here the 'Deposit' section of the grid has been clicked, which has updated the center tile to show information on available BMRB deposition options.

Small molecules

BMRB has been the repository of record for small molecule structure data that do not satisfy the PDB structure deposition criteria - for example, oligopeptides with less than 24 residues. In recent years, BMRB has become a pioneering repository for spectra of small molecule by incorporating several innovations. The current archive of standards contains 1343 small molecules comprising of 656 metabolites and 665 fragments from the Maybridge fragment library. More than 1200 compounds, includes many key mammalian metabolites and drug-like molecular fragments used in ligand screening are represented. Small molecule data includes theoretically derived chemical shifts for 219 compounds. Additionally, BMRB contains between one to three sets of chemical shifts, corresponding to different conditions, for 487 lignins. BMRB is undertaking additional innovations to improve the quality and interoperability of open data related to small molecules, such as metabolites, drugs, natural products, food additives, and environmental contaminants. Small molecules are provided a unique identifier based on the implementation of an extended version of the IUPAC International Chemical Identifier (InChI) system called ALATIS (24,25) (https://alatis.bmrb.io/). The implementation utilizes the three-dimensional structure of a compound to generate reproducible compound identifiers (standard InChI strings) and universally reproducible designators for all constituent atoms of each compound. Another factor that has contributed to improved quality is the GISSMO (26,27) mechanism for the specification and exchange of small molecule NMR data in a robust way by parameterizing the description of the experimental data.

ALATIS addresses the critical problem of deriving a complete, consistent, and unique representation for describing a compound based on its 3D structure in a format that can be sorted and compared. During this process the compound itself and all its constituent atoms will receive unique and reproducible names – including the protons. ALATIS uses an automated algorithm for the bijective mapping between the descriptor and 3D structure to avoid human error, and it also maintain a low time complexity order in order to



BMRB makes bio-NMR data FAIR.
Findable, Accessible, Interoperable, Re-usable.

BMRB collects, annotates, archives, and disseminates spectral and quantitative data derived from NMR spectroscopic investigations of biological macromolecules and metabolites.

Figure 3. Landing/home page for BMRB.

mitigate very long computation times for molecules with a large number of atoms. The algorithm generates a standard InChI string as the unique identifier and assigns unique atom identifiers (numbers) to each atom in the molecule. Because of this unique mapping, the software can also take this enhanced InChI string and generate a 3D model for the compound with all atoms reproducibly identified.

GISSMO (https://gissmo.bmrb.io) provides a mechanism for the specification and exchange of NMR data in a robust way by parameterizing the description of the experimental data. Parameterization of experimental data is specialized work that entails expert modelling and detailed specification, but it is another key necessary step for advancing data quality and usability. Associated NMR-STAR metadata for each entry serves to define the calculated parameters and associates them with computer-readable information describing data collection, such as sample properties (concentration, pH, temperature, buffer), NMR instrument manufacturer, model, field-strength, type of probe, etc. The quality of the derived spin system model is evaluated by calculating the RMSD between the experimental spectrum and the spectrum calculated from the spin system at the corresponding magnetic field.

The website for GISSMO maintains tools for simulating 1H NMR spectra of mixtures of small molecules and for automated peak analysis of 1H NMR spectra of biological fluids or cell extracts to identify compounds present. Once the spin system matrix has been accurately modelled to represent an experimental NMR spectrum collected at a given magnet field strength, the matrix can be used to model spectra at any desired magnetic field strength. The ^{1}H NMR spectrum of each compound is simulated at a variety of magnetic fields (40, 60, 80, 90, 100, 200, 300, 400, 500, 600, 700, 750, 800, 900, 950, 1000, 1100 and 1300 MHz). Every entry modelled by the GISSMO modelling process can be downloaded in NMR-STAR and NMReDATA data format.

The search functions of BMRB, available under the ‘Search’ menu item in the top banner of the website, includes a peak query function under the heading ‘chemical shift search’. Peak query is a recent service provided by the BMRB that has garnered significant usage by the community. The web form allows the user to enter a list of ^{13}C, ^{15}N or ^{1}H chemical shift values. The values are used to search entries, with the highest priority given to the match with the largest number of peaks, and the second highest priority to the closeness of the matches in ppm. The service is available for both metabolomics and macromolecular entries.

Educational/reference

The BMRB web site hosts an extensive collection of documents defining standards for nomenclature, chemical shift referencing, and data formats, as well as background tutorials on biomolecular NMR. It also provides many links to external web services pertinent to biomolecular NMR.

CS-Rosetta server

BMRB developed and released a CS-Rosetta (28) service in 2010 and has continued to maintain and improve it. Until 2022, the service used the computational resources of the CHTC (Center for High Throughput Computing) and OSG (Open Science Grid), when it was migrated to a dedicated cluster at UConn. The service accepts chemical shifts in either NMR-STAR or TALOS format and permits selection of the number of structures to generate. The server then uses the HTCondor (https://htcondor.org) high throughput computing software to distribute the workload across a large number of machines and computing cores. This massive parallelism has allowed BMRB to provide results in <24 h, even for large numbers of structures. With recent improvements and a dedicated computational cluster, results are now provided in as little as 4 hours.

While the standard submission form supports options such as whether or not to remove flexible tails, whether any disulfide bonds are present, and whether or not to exclude homologous structures, the BMRB CS-Rosetta server also provides an advanced submission mode that allows submission of constraints (NOE or otherwise derived) in the Rosetta constraints format, and residual dipolar couplings (RDCs).

BMRbig

BMRbig is a BMRB project designed to accommodate the acquisition of diverse data (not just NMR data) beyond the types currently curated and annotated by BMRB. Unlike BMRBdep, it allows for the deposition of arbitrary data and collects the appropriate metadata provided by the depositor. To facilitate timely release of data into the public domain, BMRbig provides the ability for depositors to release data nearly immediately without undergoing a lengthy annotation process. It also supports updating deposited data, and appending additional data to existing depositions. A versioned history is maintained of all changes, and future improvements will allow reviewing any particular release of a BMRbig entry.

RESULTS

BMRB has enabled the development of dozens of software packages that are used by thousands of investigators every day. These include TALOS and PECAN, for determining secondary structure of proteins; Sparta (29), ShiftX2 (30), CAMSHIFT, UCBShift (29) and 4DSPOT (30) for prediction of chemical shifts from protein structure; Chemical Shift Index, for identifying disordered regions of proteins; CS-Rosetta (28,30) and CHESHIRE (31) for determining protein structures from chemical shifts; CCPN, for the analysis of protein NMR spectra; FLYA (32) and PINE (33), for assignment of protein NMR spectra; CYANA (34) and XPLOR-NIH (35). for computing macromolecular structures from NMR data; LACS and PANAV for referencing protein chemical shifts against standards. Collectively, the publications describing these software packages have been cited >16 000 times (Google Scholar).

As BMRB grows, the utility of the archive extends beyond discerning general trends to identifying unusual structural or dynamical features of biomacromolecules. BMRB has crossed the threshold where extreme values no longer are merely statistical outliers, but now are abundant and capable of revealing hidden properties. A recent example (36) is an investigation of shift outliers as sentinels of hydrogen bonds between amide N–H groups (H-donor) and aromatic rings (H-acceptor). First postulated in proteins by Perutz and Levitt (37), federation of BMRB chemical shift data with PDB structures revealed evidence of hundreds of amide-aromatic H-bonds.

Although BMRB content is dominated by protein data, applications of NMR to RNA are increasingly important. Examples include applications to hybrid/integrative structural biology (38) and investigating RNA as potential drug targets (39).

DISCUSSION

Emerging and future applications of BMRB will enable deeper insights into biomacromolecular dynamics and disorder, including LLPS (liquid-liquid phase separation) propensity, dynamics from relaxation rates and chemical shifts, conformational propensities of disordered proteins and nucleic acids, the detection and characterization of binding interactions (37,39,40), personalized medicine (41) (drug metabolism and pharmacokinetics, diagnostics), and microbial metabolomics (42–45). As a rich source of curated data, BMRB will foster the application of advances in machine learning (ML) to these challenge areas. BMRB was an early driver of applications of neural networks to biomolecular NMR (46–48). Curated biomolecular NMR data is essential for realizing the potential of deep learning for unlocking latent knowledge present in biomolecular NMR data.

DATA AVAILABILITY

The URL for BMRB is https://bmrb.io. BMRB maintains an open-source GitHub repository (https://github.com/bmrb-io).

ACKNOWLEDGEMENTS

We are grateful for Professor John Markley and Dr Eldon Ulrich for establishing BMRB and guiding it for nearly four decades.

FUNDING

B.M.R.B. is supported by the US National Institutes of Health [R01GM109046]; BMRBj was supported by grants from the Database Integration Coordination Program from the department of NBDC program, Japan Science and Technology Agency [JPMJND2205]; Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED [22ama121001, in part] Funding for open access charge: Hoch discretionary funds.

Conflict of interest statement. None declared.

REFERENCES

1. wwPDB consortium (2019) Protein data bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.
2. Harrison,P.W., Ahamed,A., Aslam,R., Alako,B.T., Burgin,J., Buso,N., Courtot,M., Fan,J., Gupta,D. and Haseeb,M. (2021) The european nucleotide archive in 2020. *Nucleic Acids Res.*, **49**, D82–D85.
3. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
4. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
5. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chen,L., Crichlow,G.V., Duarte,J.M., Dutta,S., Fayazi,M. and Feng,Z. (2022) RCSB protein data bank: celebrating 50 years of the PDB with new tools for understanding and visualizing biological macromolecules in 3D. *Protein Sci.*, **31**, 187–208.
6. Varadi,M., Anyango,S., Appasamy,S.D., Armstrong,D., Bage,M., Berrisford,J., Choudhary,P., Bertoni,D., Deshpande,M. and Leines,G.D. (2022) PDBe and PDBe-KB: providing high-quality, up-to-date and integrated resources of macromolecular structures to support basic and applied research and education. *Protein Sci.*, **31**, e4439.
7. Bekker,G.J., Yokochi,M., Suzuki,H., Ikegawa,Y., Iwata,T., Kudou,T., Yura,K., Fujiwara,T., Kawabata,T. and Kurisu,G. (2022) Protein data bank japan: celebrating our 20th anniversary during a global pandemic as the asian hub of three dimensional macromolecular structural data. *Protein Sci.*, **31**, 173–186.
8. Lawson,C.L., Patwardhan,A., Baker,M.L., Hryc,C., Garcia,E.S., Hudson,B.P., Lagerstedt,I., Ludtke,S.J., Pintilie,G. and Sala,R. (2016) EMDataBank unified data resource for 3DEM. *Nucleic Acids Res.*, **44**, D396–D403.
9. Maciejewski,M.W., Schuyler,A.D., Gryk,M.R., Moraru,I.I., Romero,P.R., Ulrich,E.L., Eghbalnia,H.R., Livny,M., Delaglio,F. and Hoch,J.C. (2017) NMRbox: a resource for biomolecular NMR computation. *Biophys. J.*, **112**, 1529–1534.
10. Skinner,S.P., Fogh,R.H., Boucher,W., Ragan,T.J., Mureddu,L.G. and Vuister,G.W. (2016) CcpNmr analysisassign: a flexible platform for integrated NMR analysis. *J. Biomol. NMR*, **66**, 111–124.
11. Ulrich,E.L., Baskaran,K., Dashti,H., Ioannidis,Y.E., Livny,M., Romero,P.R., Maziuk,D., Wedell,J.R., Yao,H., Eghbalnia,H.R. *et al.* (2019) NMR-STAR: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments. *J. Biomol. NMR*, **73**, 5–9.
12. Hall,S.R. and Cook,A.P.F. (1995) STAR dictionary definition language: initial specification. *J. Chem. Inf. Comput. Sci.*, **35**, 819–825.
13. Hall,S.R. and Spadaccini,N. (1994) The STAR file: detailed specifications. *J. Chem. Inf. Comput. Sci.*, **34**, 505–508.
14. Hall,S.R. (1991) The STAR file: a new format for electronic data transfer and archiving. *J. Chem. Inf. Comput.*, **31**, 326–333.
15. Markley,J.L., Bax,A., Arata,Y., Hilbers,C.W., Kaptein,R., Sykes,B.D., Wright,P.E. and Wüthrich,K. (1998) Recommendations for the presentation of NMR structures of proteins and nucleic acids–IUPAC-IUBMB-IUPAB inter-union task group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. *Eur. J. Biochem.*, **256**, 1–15.
16. Gutmanas,A., Adams,P.D., Bardiaux,B., Berman,H.M., Case,D.A., Fogh,R.H., Guntert,P., Hendrickx,P.M., Herrmann,T., Kleywegt,G.J. *et al.* (2015) NMR exchange format: a unified and open standard for representation of NMR restraint data. *Nat. Struct. Mol. Biol.*, **22**, 433–434.
17. Gryk,M.R. (2021) Deconstructing the STAR file format. *Balisage Ser. Markup Technol.*, **26**, https://doi.org/10.4242/balisagevol26.gryk01.
18. Yokochi,M., Kobayashi,N., Ulrich,E.L., Kinjo,A.R., Iwata,T., Ioannidis,Y.E., Livny,M., Markley,J.L., Nakamura,H., Kojima,C. *et al.* (2016) Publication of nuclear magnetic resonance experimental data with semantic web technology and the application thereof to biomedical research of proteins. *J. Biomed. Semantics*, **7**, 16–11.
19. Kinjo,A.R., Bekker,G.J., Wako,H., Endo,S., Tsuchiya,Y., Sato,H., Nishi,H., Kinoshita,K., Suzuki,H., Kawabata,T. *et al.* (2018) New tools and functions in data-out activities at protein data bank japan (PDBj). *Protein Sci.*, **27**, 95–102.
20. Young,J.Y., Westbrook,J.D., Feng,Z., Sala,R., Peisach,E., Oldfield,T.J., Sen,S., Gutmanas,A., Armstrong,D.R., Berrisford,J.M. *et al.* (2017) OneDep: unified wwPDB system for deposition,

biocuration, and validation of macromolecular structures in the PDB archive. *Structure*, **25**, 536–545.

21. Markley,J.L., Ulrich,E.L., Berman,H.M., Henrick,K., Nakamura,H. and Akutsu,H. (2008) BioMagResBank (BMRB) as a partner in the worldwide protein data bank (wwPDB): new policies affecting biomolecular NMR depositions. *J. Biomol. NMR*, **40**, 153–155.
22. Westbrook,J.D., Young,J.Y., Shao,C., Feng,Z., Guranovic,V., Lawson,C.L., Vallat,B., Adams,P.D., Berrisford,J.M., Bricogne,G. *et al.* (2022) PDBx/mmCIF ecosystem: foundational semantic tools for structural biology. *J. Mol. Biol.*, **434**, 167599.
23. Groom,C.R., Bruno,I.J., Lightfoot,M.P. and Ward,S.C. (2016) The cambridge structural database. *Acta Crystallogr. B*, **72**, 171–179.
24. Dashti,H., Wedell,J.R., Westler,W.M., Markley,J.L. and Eghbalnia,H.R. (2019) Automated evaluation of consistency within the pubchem compound database. *Sci. Data*, **6**, 190023.
25. Dashti,H., Westler,W.M., Markley,J.L. and Eghbalnia,H.R. (2017) Unique identifiers for small molecules enable rigorous labeling of their atoms. *Sci. Data*, **4**, 170073.
26. Dashti,H., Wedell,J.R., Westler,W.M., Tonelli,M., Aceti,D., Amarasinghe,G.K., Markley,J.L. and Eghbalnia,H.R. (2018) Applications of parametrized NMR spin systems of small molecules. *Anal. Chem.*, **90**, 10646–10649.
27. Dashti,H., Westler,W.M., Tonelli,M., Wedell,J.R., Markley,J.L. and Eghbalnia,H.R. (2017) Spin system modeling of nuclear magnetic resonance spectra for applications in metabolomics and small molecule screening. *Anal. Chem.*, **89**, 12201–12208.
28. Shen,Y., Vernon,R., Baker,D. and Bax,A. (2009) De novo protein structure generation from incomplete chemical shift assignments. *J. Biomol. NMR*, **43**, 63–78.
29. Li,J., Bennett,K.C., Liu,Y., Martin,M.V. and Head-Gordon,T. (2020) Accurate prediction of chemical shifts for aqueous protein structure on “real world” data. *Chem. Sci.*, **11**, 3180–3191.
30. Han,B., Liu,Y., Ginzinger,S.W. and Wishart,D.S. (2011) SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR*, **50**, 43–57.
31. Cavalli,A., Salvatella,X., Dobson,C.M. and Vendruscolo,M. (2007) Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 9615–9620.
32. Lopez-Mendez,B. and Guntert,P. (2006) Automated protein structure determination from NMR spectra. *J. Am. Chem. Soc.*, **128**, 13112–13122.
33. Bahrami,A., Assadi,A.H., Markley,J.L. and Eghbalnia,H.R. (2009) Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. *PLoS Comput. Biol.*, **5**, e1000307.
34. Wurz,J.M., Kazemi,S., Schmidt,E., Bagaria,A. and Guntert,P. (2017) NMR-based automated protein structure determination. *Arch. Biochem. Biophys.*, **628**, 24–32.
35. Schwieters,C.D., Bermejo,G.A. and Clore,G.M. (2018) Xplor-NIH for molecular structure determination from NMR and other data sources. *Protein Sci.*, **27**, 26–40.
36. Baskaran,K., Wilburn.,C.W., Wedell,J.R., Koharudin,L.M.I., Ulrich,E.L., Schuyler,A.D., Eghbalnia,H.E., Gronenborn,A.M. and Hoch,J.C. (2021) Anomalous amide proton chemical shifts as signatures of hydrogen bonding to aromatic sidechains. *Magn. Resonance*, **2**, 765–775.
37. Levitt,M. and Perutz,M.F. (1988) Aromatic rings act as hydrogen bond acceptors. *J. Mol. Biol.*, **201**, 751–754.
38. Wang,Y.X., Zuo,X., Wang,J., Yu,P. and Butcher,S.E. (2010) Rapid global structure determination of large RNA and RNA complexes using NMR and small-angle X-ray scattering. *Methods*, **52**, 180–191.
39. Sreeramulu,S., Richter,C., Berg,H., Wirtz Martin,M.A., Ceylan,B., Matzel,T., Adam,J., Altincekic,N., Azzaoui,K., Bains,J.K. *et al.* (2021) Exploring the druggability of conserved RNA regulatory elements in the SARS-CoV-2 genome. *Angew. Chem. Int. Ed Engl.*, **60**, 19191–19200.
40. Berg,H., Wirtz Martin,M.A., Niesteruk,A., Richter,C., Sreeramulu,S. and Schwalbe,H. (2021) NMR-based fragment screening in a minimum sample but maximum automation mode. *J. Vis. Exp.*, **172**, e62262.
41. Dong,C., Honrao,C., Rodrigues,L.O., Wolf,J., Sheehan,K.B., Surface,M., Alcalay,R.N. and O’Day,E.M. (2022) Plasma metabolite signature classifies male LRRK2 parkinson’s disease patients. *Metabolites*, **12**, 149.
42. Hertel,J., Fassler,D., Heinken,A., Weiss,F.U., Ruhlemann,M., Bang,C., Franke,A., Budde,K., Henning,A.K., Petersmann,A. *et al.* (2022) NMR metabolomics reveal urine markers of microbiome diversity and identify benzoate metabolism as a mediator between high microbial alpha diversity and metabolic health. *Metabolites*, **12**, 308.
43. Tong,L., Feng,Q., Lu,Q., Zhang,J. and Xiong,Z. (2022) Combined (1)H NMR fecal metabolomics and 16S rRNA gene sequencing to reveal the protective effects of gushudan on kidney-yang-deficiency-syndrome rats via gut-kidney axis. *J. Pharm. Biomed. Anal.*, **217**, 114843.
44. Mengucci,C., Nissen,L., Picone,G., Malpuech-Brugere,C., Orfila,C., Ricciardiello,L., Bordoni,A., Capozzi,F. and Gianotti,A. (2022) K-clique multiomics framework: a novel protocol to decipher the role of gut microbiota communities in nutritional intervention trials. *Metabolites*, **12**, 736.
45. Somerville,G.A., Parrett,A.A., Reed,J.M., Gardner,S.G., Morton,M. and Powers,R. (2022) Human serum alters the metabolism and antibiotic susceptibility of staphylococcus aureus. *J. Proteome Res.*, **21**, 1467–1474.
46. Shen,Y. and Bax,A. (2015) Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. *Methods Mol. Biol.*, **1260**, 17–32.
47. Shen,Y. and Bax,A. (2010) SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR*, **48**, 13–22.
48. Hoch,J.C. (2019) If machines can learn, who needs scientists? *J. Magn. Reson.*, **306**, 162–166.