communications chemistry

Article

A Nature Portfolio journal

https://doi.org/10.1038/s42004-025-01812-8

NMR data processing, visualization, analysis and structure calculation with NMRFx

Ellen Koag[1], Simon G. Hulse[1], Gregory L. Helms[1], Kevin M. Call[1], Michael F. Summers[2,3], Jan Marchant[2] & Bruce A. Johnson[1]

NMR spectroscopy is applied in many scientific disciplines to derive chemical, structural, and dynamical insights into molecular systems. The utility of the technique depends on robust computational protocols for processing, visualizing, and analyzing data acquired using a wide range of experiments. Here we introduce NMRFx, a software application that integrates and augments features of our existing NMRViewJ and NMRFx Processor applications. NMRFx facilitates data processing, peak picking and assignment, chemical shift prediction, molecular structure calculation, and beyond, through a high-speed, feature-rich graphical user interface. This paper describes advances over existing software and presents a series of case studies that demonstrate its utility in diverse contexts. These case studies include the assignments of the protein ubiquitin, a 36 nucleotide RNA construct, and the natural product taccalonolide E, as well as a metabolomics study of triacylglyceride production in algal cells.

Nuclear magnetic resonance (NMR) is a powerful analytical technique with applications in various disciplines, including organic chemistry[1,2], inorganic chemistry[3], environmental chemistry[4], materials science[5], and medical imaging[6]. Since the first determination of a solution-state protein structure in 1985[7], NMR spectroscopy has been widely used in the field of structural biology[8]. The scope of NMR applications for macromolecular structure analysis continues to grow, with notable advances recently made in RNA research[9], the study of intrinsically disordered proteins[10], and the emerging field of integrated structural biology[11–13]. NMR is also used to investigate ligand binding[14], metabolomics[15,16], and molecular dynamics, for which established protocols probe motions over picosecond to multisecond time scales[17].

Central to these applications is the need for robust computational tools for data processing, visualization, and analysis. A widely used software application for the visualization and analysis of macromolecular NMR data is NMRView[18], along with its Java-based successor, NMRViewJ[19]. These highly cited programs have been instrumental in myriad investigations. A complementary program, NMRFx Processor[20], was developed more recently to exploit recent advances in computational hardware and provide a feature-rich yet accessible means of processing NMR data.

Here we introduce NMRFx, a program that merges and greatly extends the capabilities of NMRFx Processor and NMRViewJ. Additionally, NMRFx provides tools for structure calculation and allows the inclusion of plugins to enable additional functionality, with a prime example being RING NMR Dynamics[21], an application for the analysis of macromolecular dynamics by NMR. The suite of tools available in NMRFx, which facilitates a complete workflow from the free induction decay (FID) to processed and analyzed spectra, ultimately provides structural and dynamical insights. Although there are alternative programs for NMR processing[22], visualization and analysis[23–25], and structure calculation[26,27], none integrate such a wide range of features. Consequently, NMRFx is also appropriate for use in disciplines beyond structural biology. The software is designed to be user-friendly and appropriate for educational settings, allowing users to visualize and adjust various data processing routines in real time.

In this work, a summary of NMRFx's primary features is presented. Additional features for specialist use cases, such as support for ligand and pressure titrations[28] and ZZ-exchange spectroscopy[29], are described at https://nmrfx.org. The utility of the software is highlighted in four diverse case studies:

1. Assignment, secondary structure prediction, and dynamics analysis of the backbone atoms in ubiquitin.
2. Assignment of the natural product taccalonolide E.
3. Monitoring carbon fluxes in triglyceride synthesis in an alga.
4. Assignment and structure prediction of a 36-residue RNA.

[1]Structural Biology Initiative, Advanced Science Research Center at the CUNY Graduate Center, New York, NY, USA. [2]Department of Chemistry and Biochemistry, University of Maryland Baltimore County, Baltimore, MD, USA. [3]Howard Hughes Medical Institute, University of Maryland Baltimore County, Baltimore, MD, USA. e-mail: bjohnson@gc.cuny.edu

Results

Software architecture: languages and platform compatibility

NMRFx is an open-source and extensible cross-platform application. Written in the Java programming language and bundled with the Java runtime environment, NMRFx can be installed, executed, and developed on all popular operating systems. The graphical user interface (GUI) is built using the JavaFX toolkit[30]. The software is designed for interaction through the GUI for most use cases. It is also possible to issue commands from the command line, allowing for automation of repetitive tasks and delegation of computationally intensive tasks to remote resources. For example, data processing scripts can be written in the Python language, which are executed using an embedded Jython interpreter[31].

Database and file formats

A key design choice in developing NMRFx has been to prioritize support for recognized file standards over proprietary ones. This ensures seamless interaction with other programs—while NMRFx is designed as a comprehensive tool for NMR analysis, it is easy to import and export data to and from other applications, enabling specific routines to be performed outside of NMRFx. NMRFx accommodates commonly used data formats for both raw and processed NMR data, including Bruker, Varian, JEOL, and NMRPipe. Support is also provided for various data deposition and retrieval formats, such as NMR-STAR[32], NEF[33], PDB[34], and PDBx/mmCIF[35].

The NMR-STAR format is the primary method for storing NMRFx project data. Users can search for and retrieve BMRB entries within NMRFx, and NMR-STAR files can be uploaded for new depositions directly within the application. Historically, it has been common for depositions to contain the minimal required information (molecular topology and chemical shift assignments). It is hoped that this upload feature will make it less taxing for users to make more categories of data accessible, including peak lists, titration trajectories, and relaxation data. Such data is valuable for validation and for training machine learning models[36].

Numerous file formats specific to small molecules are also supported, including MOL, SDF[37], and MOL2. In addition, SMILES strings[38] can be parsed to generate small molecule structures.

Version-controlled projects

NMR projects often involve numerous intricate steps to process and analyze data, which can span months or even years. By integrating the Git version control system, NMRFx provides features to manage projects of varying complexity. A comprehensive history of all pertinent information, including processed datasets, peak lists, assignments, and window configurations, is maintained in a memory-efficient manner as a tree of snapshots called "commits". Users can navigate between commits and create multiple branches from a given commit, allowing for the exploration of alternative protocols within the same project without the risk of losing previous work. Noteworthy commits, such as the state of a project at the point of deposition or publication, can also be tagged for easy recall.

Plugins

NMRFx supports the integration of plugins—comprehensive tools written in Java that can be developed and distributed independently of the main program. A notable example of this is RING NMR Dynamics, an application designed for analyzing various types of macromolecular NMR dynamics data, including CPMG, CEST, and $R_{1\rho}$ experiments, as well as model-free analysis[21]. Although RING can function as a standalone application, its integration as a plugin within NMRFx enables communication between the two programs, creating a seamless pathway from raw NMR data to dynamic insights. Plugin users can process, peak pick, and assign spectra with the tools of NMRFx (*vide infra*), and subsequently transfer peak information directly to RING for analysis. Additionally, selecting data entries within RING instructs NMRFx to display the relevant spectral region.

Machine learning support

As with virtually all scientific disciplines, it has been demonstrated that machine learning methodologies can aid the field of NMR in numerous ways[39]. To facilitate the use of trained deep-learning models, the TensorFlow library—one of the most widely used frameworks—is embedded within NMRFx[40]. Furthermore, NMRFx also ships with the Tribuo library, which supports a wide range of classical machine learning models[41]. This enables the seamless introduction of novel models into NMRFx; some examples, including secondary structure prediction of both proteins and RNA, which already ship with the application, are described below.

NMR processing

NMRFx offers a diverse array of operations for processing NMR datasets of arbitrary dimension. Commonly employed operations such as apodization, zero-filling, Fourier transformation, and phase correction are supplemented with operations for specialized use cases, including baseline correction, peak suppression, and reconstruction of non-uniformly sampled (NUS) datasets. In most cases, an appropriate processing routine is automatically generated through analysis of the dataset's acquisition parameters, allowing the user to focus on fine-tuning the routine. Wherever possible, each operation is parallelized across multiple CPU cores to ensure computational resources are leveraged to their fullest extent. Routines can be modified within the GUI by interacting with the Processor accordion—a series of expandable tabs ordered by data dimension, which outlines the operations to be executed. Operations can be adjusted by interacting with GUI elements that modify the associated parameters. For many operations, the effects of their addition/adjustment are instantly depicted, helping users design optimal processing schemes, while also making NMRFx a valuable educational tool.

Figure 1 shows a screenshot of NMRFx after processing a NUS HNCACB dataset. The processing accordion is expanded to show the options for GRINS (GRINS is not SMILE), a NUS reconstruction algorithm developed within the Johnson group, which has similarities to SMILE reconstruction[42], hence its recursive acronym. In-house implementations of iterative soft thresholding (IST)[43] and NESTA[44] are also available within NMRFx for NUS reconstruction. Interacting with the Processor accordion updates a Python script, which is executed when the user clicks `Process`; the script used to produce the spectrum in Fig. 1 is provided in Code Listing S1 of the Supplementary Information (SI).

Visualization of spectra

Correlating data from different NMR experiments is often essential to obtain insights into the chemical system under study[45]. A notable advancement in the original NMRView software was the ability to simultaneously visualize multiple spectra and/or different arrangements of spectrum dimensions in separate windows[18]. This concept has been retained in NMRFx; a virtually unlimited number of top-level windows can be created, with each window able to support numerous spectrum charts arranged within a grid. Correlated cross-hair cursors facilitate the identification of common spectral features across charts.

This capability, while powerful, can be time-consuming to set up. To streamline the process, NMRFx now supports layout files written in YAML, which enable window configurations to be loaded automatically. Users can create custom layouts for their specific requirements. An example used for the RNA case study is shown in Code Listing S2 of the SI.

Peak picking

NMRFx provides the same suite of tools for locating (picking) spectral peaks as were present in NMRViewJ, with support present for 1D and multidimensional spectra, including pseudo three-dimensional (3D) spectra. It is possible to pick peaks automatically using the built-in peak picker, or interactively pick and adjust them through mouse-based control, which can be useful when peak overlap makes automated picking challenging. Peak features can be refined using non-linear regression tools in order to derive accurate quantitative information such as positions, volumes, and widths. There are recent developments in peak picking using both deep learning[46]

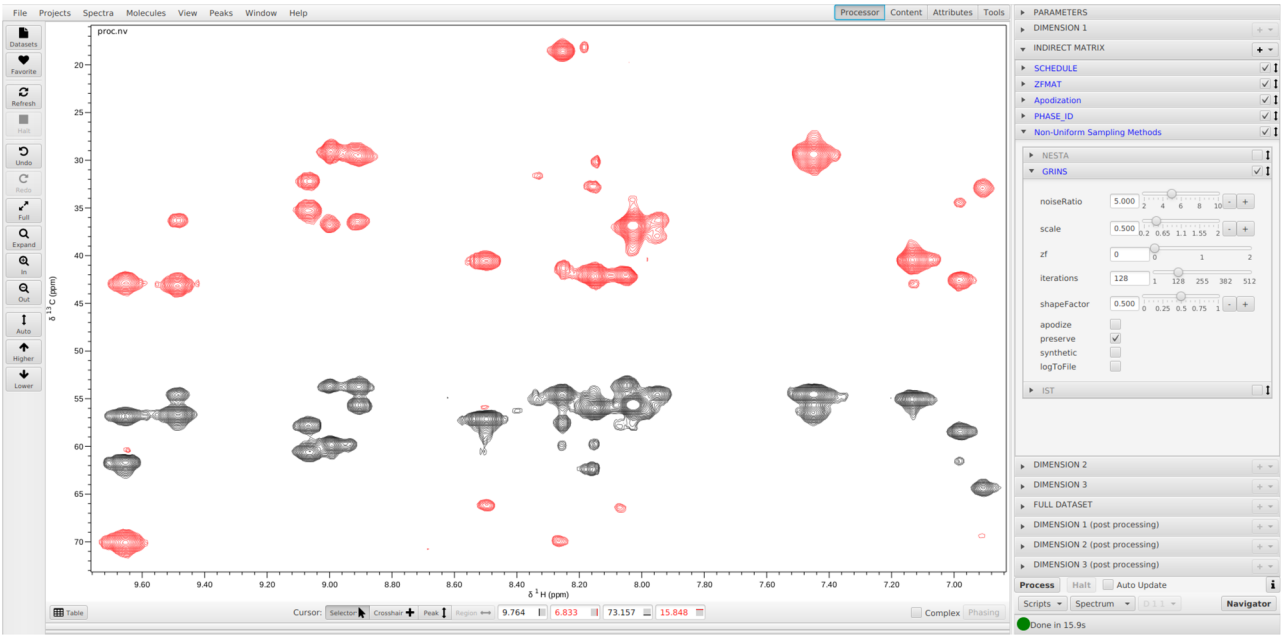

Fig. 1 | Screenshot of the NMRFx GUI. A screenshot of NMRFx after processing a NUS HNCACB dataset with the GRINS algorithm. A ^{1}H-^{13}C plane of the spectrum is displayed (δ15N = 119.33 ppm). The processing accordion, located on the right-hand side of the GUI, allows users to customize the sequence of steps involved in transforming the FID into the final spectrum. Each processing step has an accompanying drop-down menu, with that for the GRINS algorithm visible. The adjustment of processing operations updates a Python script, which can be inspected by clicking `Scripts` > `Show Script`, and executed by clicking `Process`.

and deconvolution-based approaches[47]; the development of similar capabilities is ongoing and will be available in a forthcoming release of NMRFx.

Peak assignments

NMRFx includes various tools to facilitate the process of assigning spectral peaks to atomic sites. These tools aim to automate as much of the assignment process as possible, while allowing user interaction as a means of validation, particularly for more challenging cases. Two specific tools that have found considerable use are the Peak Slider Tool[48], and RunAbout[19].

The Peak Slider Tool. The Peak Slider Tool aids in the assignment of spectra in situations where an appropriate initial guess exists as a starting point[48]. The initial guess typically comes from previous assignments of comparable structures or chemical shift predictions (*vide infra*). The use of the Peak Slider Tool involves the manual adjustment of peak-boxes to ensure agreement with the spectral peaks. When a peak-box is moved, all other peak-boxes that share an assignment, even those present in other spectra, are simultaneously updated, ensuring consistency and reducing ambiguity. Furthermore, when a peak-box is deemed to be correctly positioned, it can be frozen in place. The dimensions of other peak-boxes assigned to the same atom are also frozen, such that they may only be adjusted in the remaining free dimensions. This process is described in more detail below, in a case study related to RNA assignment.

RunAbout. The assignment of triple resonance spectra of proteins using NMRFx is facilitated by RunAbout, which was introduced as part of NMRViewJ[19]. In NMRFx, the tool has been optimized to use the new GUI toolkit and to be more flexible in terms of the experiments it supports, including those involving direct detection of ^{13}C and ^{15}N. In addition, more functions have been incorporated that automate the assignment process, minimizing the need for manual interaction. The operation of RunAbout consists of three main operations:

1. Grouping spectral peaks: peaks are grouped into distinct "spin systems"—clusters of peaks that share the same root frequencies, typically those of the backbone $^{1}H^{N}$ and ^{15}N spins.
2. Linking spin systems: adjacent spin systems are connected to form contiguous fragments through the consideration of various criteria.
3. Matching fragments: the fragments are aligned with regions of the protein's primary sequence to create the assignments.

RunAbout allows users to quickly navigate the sparsely populated 3D spectra, focusing only on areas with significant peak groupings, akin to island-hopping on a boat, which is the program's namesake.

NMRFx's support for complex spectral layouts (*vide supra*) is exploited in RunAbout to ensure all spectra are considered in a holistic manner. When analyzing peaks and grouping them into spin systems, two orthogonal views —^{1}H-^{13}C and ^{15}N-^{13}C—are considered for the intra-residue and inter-residue spectra. These views aid the inspection and fine-tuning of the peak picking. When linking spin systems, three distinct groupings of spectra are typically presented, as illustrated in Fig. 2. The two central columns display the peaks associated with the spin system currently under consideration (i), while the two leftmost and rightmost columns show the corresponding regions for spin systems which are candidates for $i-1$ and $i+1$, respectively. Spectra containing signals for different carbon types ($^{13}C'$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$) are organized into separate rows. Several metrics are provided to help users decide whether the current candidates for $i-1$ and $i+1$ are valid. These include a score quantifying the peak correlations[49], a tally of the number of peaks which are correlated, and boolean flags indicating:

- Reciprocal match: spin system $i-1$ does not form a better match with another candidate.
- Available for matching: the spin systems considered have not yet been assigned to a fragment.
- Viable fragment: the chemical shifts are consistent with at least one contiguous amino acid pair in the primary sequence.

Additionally, the program presents the chemical shifts associated with the i spin system and the $i-1$ candidate, as well as lists, ordered by likelihood, of possible amino acid types for the spin systems based on said shifts. Based on all this information, the user can manually confirm the linkages. Alternatively, an automated protocol can use the same criteria to form links across a series of spin systems. Linked spin systems, especially when they form fragments of three or more systems, can then be matched to specific residue positions.

NMRViewJ introduced an automated protocol that combines the formation of spin system links and the matching of the linked spin systems

Fig. 2 | Screenshot of the RunAbout tool. RunAbout tool after automated assignment of ubiquitin. **a** Spectral windows are arranged with the upper row displaying HNCOCACB and HNCACB spectra, and the lower row displaying HNCOCA and HNCA spectra. The central two columns display the peaks which are associated with the current spin system under consideration (11), while the two left- and rightmost columns display those of candidates for the $i-1$ and $i+1$ residues, respectively (14 and 60). **b**, **c** Information related to the candidates: a spinner to iterate through the candidates in order of likelihood; `Sys`: the identity of the spin system; `Score`: a metric which indicates the goodness of the peak correlations[49]; `N`: a tally of the number of peak correlations; and `RAV`: indicators of reciprocity, availability, and viability. **d** Middle: chemical shifts associated with residues $i-1$ (`c`, `ca`, `cb`) and i (`H`, `N`, `C`, `CA`, `CB`). Left(right): most likely amino acid types for the $i-1(i)$ residue. **e** Contiguous spin system groupings, ordered by group length. **f** Mappings of the spin system groupings onto the protein primary structure.

to sequence positions into a single automated process[50]. This process uses a bipartite matching algorithm to generate a population of possible mappings of the spin systems to primary sequence regions. These sets of matches form an initial population with higher matching scores than would be present with a simple randomly generated population. The population of matches is then optimized by a genetic algorithm. NMRFx has an updated version of this code with improvements in the protocol used to generate the population of matches and the genetic algorithm used, and the process is integrated into the GUI. Although there are many other programs available for automated assignments, including, for example, BARASA[49], FLYA[51], and MARS[52], the close integration of this automation into the NMRFx RunAbout GUI allows the user to use all RunAbout features to inspect, modify, and extend the results of the automated analysis.

Chemical shift prediction

The assignment of spectral peaks to atomic sites can be complex and time-consuming. As a result, there has been significant interest in predicting chemical shifts based on molecular structure[53–56]. These predictions can be used as a priori estimates, which can be refined through the consideration of spectral data. NMRFx provides a suite of models designed to predict chemical shifts in proteins, RNA, and small molecules. Beyond the descriptions provided here, further information about the models can be found in the Methods section, as well as the SI.

Proteins. NMRFx uses a linear regression model for protein chemical shift prediction. Similarly to other programs[53–55], the model uses numerous features derived from primary, secondary, and tertiary structural information, including dihedral angles, residue attributes, ring-current shifts, and hydrogen bonds. Individual models are trained for each pairing of amino acid and atom type ($^{13}C^{\alpha}$, $^{13}C^{\beta}$, ^{15}N, $^{1}H^{N}$, $^{1}H^{\alpha}$, etc.) to account for the varying influence of features in each case.

The high dimensionality of the supplied feature vector risks overfitting the training dataset, which would render the model ineffective at general predictions. To account for this, the model is trained using the least angle regression (LARS) algorithm[57], which acts to shrink the number of features that have a significant influence on the output. The performance of the model compared to that of the widely used SHIFTX+ program[53], on a test dataset comprising the 61 proteins, is illustrated in Fig. S1 and summarized in Table S1 of the SI. It can be seen that the overall performance of NMRFx's model is similar to SHIFTX+ across the backbone atom types. Chemical shifts of all ^{13}C, ^{15}N, and ^{1}H atoms with sufficient training data can be predicted. Scatter plots of predicted and experimental shifts and violin plots

of deviations between predicted and experimental shifts are provided in Figs. S2 and S3 of the SI.

RNA. NMRViewJ featured RNA shift predictors based on support vector regression (SVR)[58,59]. NMRFx features new models based on Factorization Machines (FM)[60]. An overview of the prediction performance of the FM compared to the SVR implementation is provided in Table S2 of the SI. Features supplied to the model are derived from the primary and secondary structure of the RNA, with considerations made for the base of interest, the two previous and successive bases in the primary structure, and any pairing partners associated with these bases. NMRFX can also predict RNA shifts using 3D RNA structures, employing models that utilize either ring-current shifts[59] or weighted distances of atoms to the target atom, similar to the method introduced by Frank et al.[61].

Small molecules. Small-molecule chemical shifts can currently be predicted by generating features akin to hierarchical ordered description of the substructure environment (HOSE) codes[62], in which each atom is represented by up to six "shells" of bonded atoms. A database of assigned shifts is then searched, and the average shift of sites with matching representations is provided as the prediction. In a future release of NMRFx, deep-learning models are also expected to be available.

Structure prediction

Structures from NOESY distance constraints. NMRFx provides tools for generating 3D molecular structures that are consistent with experimental NMR-derived distance and angle restraints. These tools have evolved from PEGASUS[63], which calculates structures in torsion-angle space—a methodology also adopted by CYANA[26] and XPLOR-NIH[27]. NMRFx's support for a wide variety of file formats affords the user flexibility in incorporating the constraints. Among these, NEF files are recommended for their superior portability with other programs. For example, initial structure predictions can be generated in NMRFx, followed by refinement using programs such as AMBER[64].

Typically, a structure calculation involves generating an ensemble of candidate structures, each optimized to adhere closely to the NMR constraints with low values of the force-field energy. Starting the structure calculations from the provided `nmrfxs` command line application allows the calculation of each structure to run in parallel in a separate operating system process (a feature not yet supported in the GUI). The various parameters for the calculation, such as the number of steps, temperatures, and forces, can be provided in the YAML format, with an example file provided in Code Listing S3 of the SI.

The structure generator was tested by calculating the 3D structures of proteins in the 100-protein NMR spectra dataset[65]. For each dataset entry, a NEF file containing NMR constraint information was provided as input, resulting in the generation of an ensemble of 200 distinct structures. A cost function was computed for each structure in the ensemble. This function quantifies the degree to which distance and angle constraints were violated, while also including repulsive terms to penalize atom overlap. The 10 structures with the lowest cost function were output to a PDBx/mmCIF file. These 10-structure ensembles were then averaged and superimposed onto the corresponding experimentally confirmed structures obtained from the PDB. The results generated by NMRFx were compared with published structures by computing the root mean squared deviations (RMSDs) between all backbone atoms in the calculated structures and the PDB structures, with the results listed Table S3 in the SI, along with the minimum and maximum distance violations of each calculation listed in Table S4. Nine examples of the results are shown in Fig. 3, demonstrating strong agreement.

2D to 3D structure mappings. NMRFx also provides support to map 2D small molecule structures to energy-optimized 3D structures using the embedded library OpenChemLib[66]. These structures aid in chemical shift prediction, facilitate the calculation of macromolecular-ligand complex structures, and assist in the assignment of NOESY cross peaks when used in conjunction with the slider tool.

Sequence display tool

NMRFx provides a sequence display tool that provides an overview of residue-specific properties in macromolecules (Fig. 4). A variety of parameters can be displayed, including: chemical shift deviations from statistical means or predicted "random coil" values[67]; CheZOD *Z*-scores, which quantify the extent of residue disorder[68]; and protein secondary structure predictions. The latter are performed with a Tensorflow deep-learning model, which predicts the relative likelihoods of four different state types (alpha-helix, H-bonded turn, bend/coil, and extended sheet).

Case Study 1: protein assignment, secondary structure analysis, and dynamics analysis

In this case study, a number of NMRFx's features are employed to perform a detailed study of the backbone atoms of ubiquitin[69].

Backbone assignment using RunAbout. Chemical shift assignments were made using RunAbout (*vide supra*). Two separate assignments were made:

1. Manual assignment: spin-system linkage and residue assignment were performed manually, making use of the associated chemical shift values and compatibility metrics described above.
2. Automated assignment: the same process was performed using the bipartite match/genetic algorithm without user intervention.

These assignments were confirmed through comparison with results published in the BMRB. A Git project was created for the manual process, with frequent commits made during the analysis; this project serves as a useful tutorial for new users of RunAbout.

HNCO, HNCACO, HNCACB, HNCOCACB, HNCA, and HNCOCA datasets were processed and peak-picked within NMRFx. The peak lists were optimized by (a) filtering out peaks that were not consistent across the spectra, (b) clustering the remaining peaks into spin systems, and (c) classifying each peak based on atom type ($^{13}C'$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$) and connectivity (intra- or inter-residue). Particular care was taken when considering the HNCO spectrum, as its generated peak list was used as a reference in order to synchronize with the other spectra.

Figure 2 depicts the result of the automated analysis. The assignments, with the exception of the final glycine residue, are fully complete and consistent with those reported previously. In general, a single complete fragment for the entire protein is not expected to be formed. Fragments (contiguous colored sequences shown in sections e and f) generally span regions bounded by a proline or a residue with missing spectral information. For more challenging assignment problems, users can make use of a combination of automated tools and manual interactions for validated results.

Predicting secondary structure. The assigned backbone chemical shifts were subjected to further analysis using the NMRFx Sequence Display Tool, as shown in Fig. 4. Shown are chemical shift deviations to random coil values, an estimate of residue-specific disorder, and a four-state prediction of secondary structure.

Dynamics analysis with RING NMR. RING NMR Dynamics, incorporated into NMRFx as a plugin, was used to derive backbone dynamic information about ubiquitin, through the calculation of amide ^{15}N R_1 and R_2 rates, along with ^{15}N-{^{1}H} NOE values. The dataset peaks were assigned to residues using the assignments computed with RunAbout. The peak-boxes were interactively aligned with the actual resonance positions in the relaxation datasets, before peak intensities were calculated in each 2D plane of the pseudo-3D datasets. At this point, RING NMR Dynamics was invoked within NMRFx via the Plugin menu, allowing for communication between the two programs. Figure 5.a shows

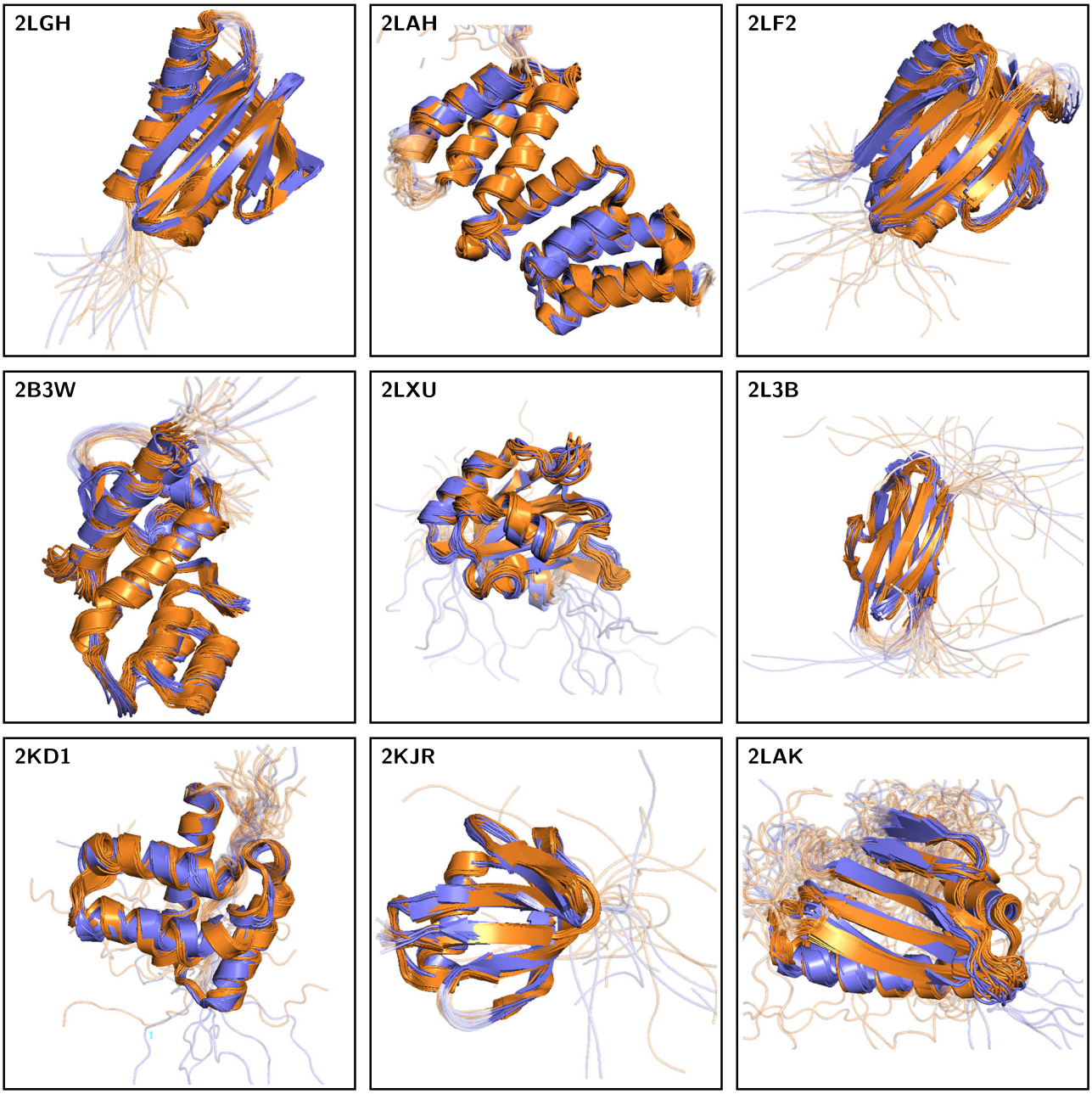

Fig. 3 | Examples of calculated protein structures. Structure recalculation results generated by NMRFx on a subset of the proteins present in the 100-protein NMR spectra dataset[65]. The 10 best structures sampled from an ensemble of 200 calculations are each shown (purple), and superimposed on the corresponding models (typically 20) from the PDB deposition (orange).

the determined R_1, R_2 and NOE values within RING. These three quantities were then used in a model-free analysis to determine order parameters (S^2) for each residue. Also visible at the top of panel a is an example fit to the peak intensities associated with valine-70 in the R_2 dataset. The spectral region of relevance to this fit is shown in NMRFx, in panel b.

Case Study 2: assignment of taccalonolide E

As an example of its applicability to small molecule NMR, NMRFx was employed to analyze data derived from the natural product taccalonolide E[70]. Figure 6 shows the result of automated processing and analysis of a 1D ^{1}H spectrum. The full spectrum is shown in panel a, along with a zoomed-in region in panel b, which includes annotations describing the two multiplet structures present. The multiplet analysis is facilitated by simulating the spectrum, and subsequently fine-tuning the peak features (positions, intensities, widths) using an optimization routine to ensure agreement with the experimental data. Complex and/or overlapping multiplets can be manually adjusted by the user to accurately describe their structure. Molecular structures can be viewed alongside the spectral data, with assignments made possible by selecting a multiplet and clicking on the corresponding atom (panel c). Summaries of the multiplet structures can be generated in a variety of journal-supported formats (panel d). Further to the 1D ^{1}H spectrum, 1D ^{13}C, HSQC, TOCSY, ROESY, and HMBC datasets of taccalonolide E were processed, and complete assignments—including all diastereotopic CH_2 protons—were made using the Peak Slider Tool. The initial guess of chemical shifts was generated by inputting an SDF file outlining the 3D structure of the molecule[71] into NMRFx's chemical shift prediction model.

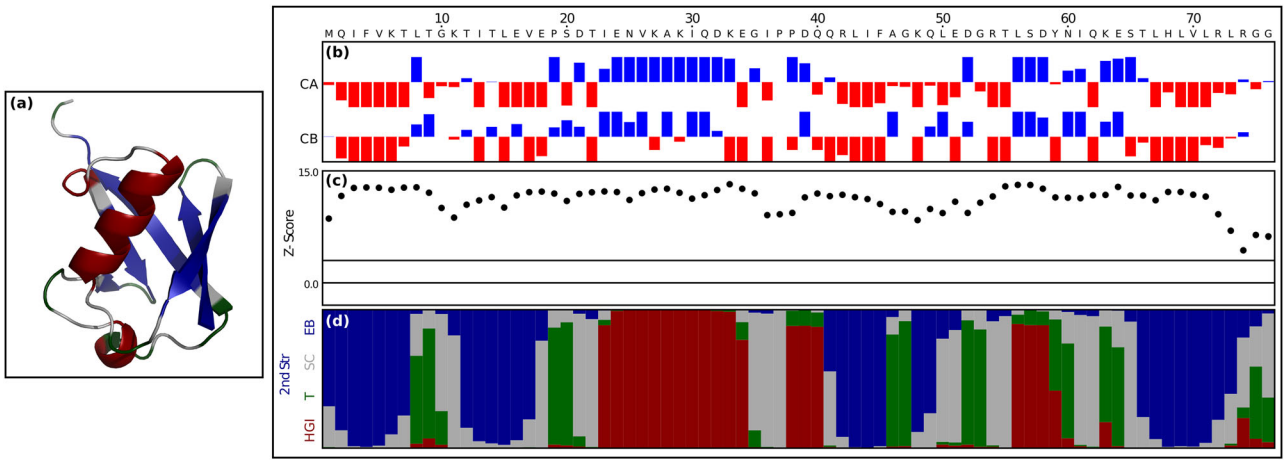

Fig. 4 | Chemical shift analysis. The sequence display tool presents a secondary structure analysis of ubiquitin. **a** Cartoon diagram of crystal structure of ubiquitin, refined at 1.8 Å resolution (PDB deposition 1UBQ)[76]. **b** Deviations of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts from estimated random coil chemical shifts[67]. **c** Plot of the CheZOD *Z*-score, an indicator of the degree of residue-specific disorder[68]. As a rule of thumb, residues for which $Z < 3$ are considered disordered ($Z = 3$ is plotted in solid black); those for which $3 < Z < 8$ are partially formed, and those for which $Z > 8$ are fully-formed. **d** Probabilities of residue classifications, using the following four-state system[77]: red: HGI (alpha-helix); green: T (hydrogen-bonded turn); gray: SC (bend/coil); blue: EB (extended sheet). The ribbon diagram of (**a**) is colored such that each residue corresponds to the classification with the highest probability.

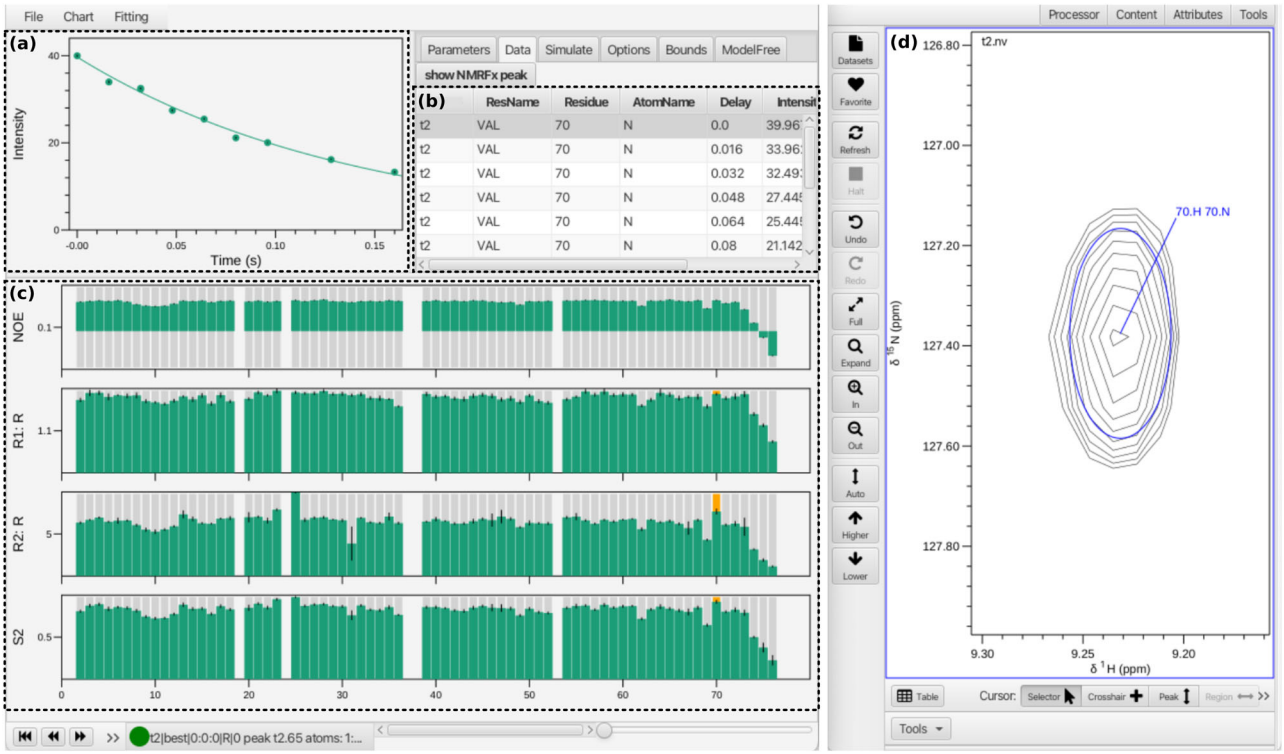

Fig. 5 | RING NMR Dynamics plugin. RING NMR Dynamics (left) being used as a plugin within NMRFx (right) to analyze ubiquitin dynamics. **a** An example exponential fit to obtain R_2 for valine-70. **b** Data table listing the experimental peak intensities seen in (**a**). **c** Estimated backbone amide ^{15}N R_1, R_2, ^{15}N-{^1H} NOE and order parameter (S^2) values. Selecting a residue entry (column highlighted in orange) shows the measured intensities in the table. **d** Selecting an entry in the table, and clicking `show NMRFx peak`, instructs NMRFx to display the relevant peak.

Case Study 3: metabolomic study of triacylglyceride production in live algal cells

NMRFx provides the Scanner Tool, a means of overlaying and analyzing multiple 1D or 2D datasets, with applications including the comparison of structural analogs, kinetics, relaxation behavior, diffusion, metabolomics, and ligand titrations. Figure 7 shows a metabolomics study of live algal cells where 1D High-Resolution Magic Angle Spinning (HRMAS) datasets were acquired by sampling, as a function of time, a culture of *Chlorella vulgaris* deprived of nitrogen and supplied with $NaH^{13}CO_3$ as the sole carbon source. This study shows the accumulation of both triacylglycerides, a model of which is provided in panel a, and sucrose.

The data were collected with and without broadband ^{13}C decoupling. The Scanner Tool is able to apply the same processing operation to all imported raw FIDs, and the resulting spectra can then be grouped for visualization as desired. Here, all decoupled spectra (panel b, red) are plotted separately to the coupled spectra (panel c, black). Within these

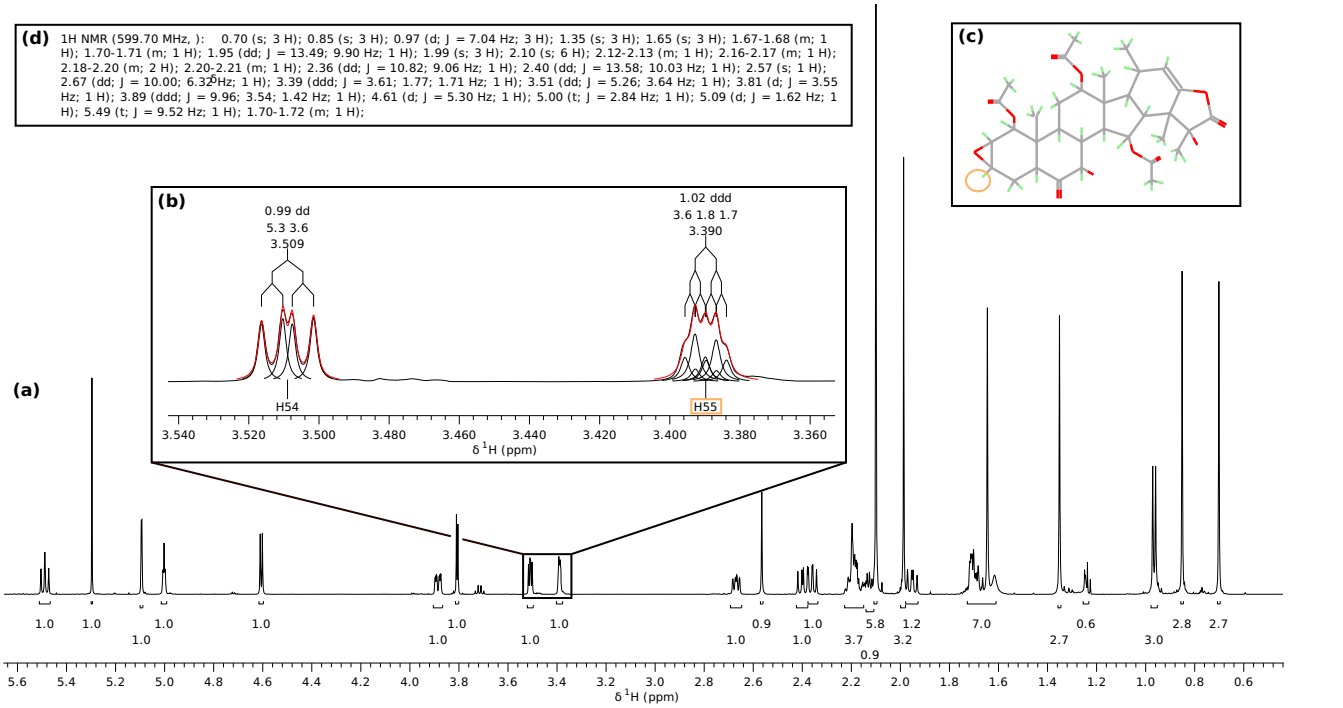

Fig. 6 | Natural product assignment. Analysis of a 1D ^{1}H spectrum of the natural product taccalonolide E using NMRFx. **a** A wide view (5.8–0.5 ppm) of the spectrum, with multiplet integrals denoted to 1 decimal place. **b** A zoomed-in view (3.54–3.36 ppm) of the same spectrum, showing two multiplet structures. The experimental spectrum is plotted (thin black line), along with the result of fitting the multiplets, with individual peak components (thick black lines) and their sum (red lines) shown. Relative integrals (2 decimal places), J-couplings (Hz), the central frequency (ppm), and a tree diagram outlining the couplings are also shown. **c** Display of the molecular structure of taccalonolide E. It is possible to visualize links between multiplet structures and atomic sites by interacting with the spectrum, with one such link highlighted in orange. **d** A generated summary of the spectrum, in J. Org. Chem. format.

two groups, the spectra are ordered by both sample time and replicate number.

Integrals can be computed across the spectra by defining regions of interest. For each region, a new column is formed in the Scanner Tool's data table (panel d). As an example, see the rightmost visible column, which lists integrals of the fatty acid chain methyl groups, lying in the range 0.95 ppm to 0.85 ppm. Integrals can be visualized and fit using built-in tools, or exported for external analysis. Example plots are presented for saturated CH_2 (panel e), omega-3 CH_3 (panel f), and the CH_2 corresponding to sucrose C6 (panel g).

For the saturated CH_2, the increase in integral upon decoupling indicates that most of the carbon incorporated into the saturated fatty acids is derived from $^{13}CO_2$ fixed during photosynthesis. Conversely, the minimal increase in integral of the omega-3 CH_3 signal upon decoupling indicates that the majority of the carbon comprising the polyunsaturated fatty acids is derived from cellular carbon pools established before addition of the labeled bicarbonate (carbon recycling). The plot of the sucrose integrals shows rapid and exponential incorporation of labeled carbon, which reaches a steady state after about 16 h.

In addition to generic scatter and box plotting tools, the Scanner Tool provides specific plotting functionality for diffusion data as well as for tracing the trajectory of resonances in 2D ligand titration studies.

Case Study 4: assignment and structure calculation of a 36 nt RNA construct

NMRFx has been used by the Summers laboratory to assign the aromatic and $^{1}H1'$ chemical shifts of more than a dozen RNA constructs with sizes of up to 60 nt. This discussion focuses on a 36 nt RNA construct derived from stem loop C of the MMLV 5′-Leader (SLC^{A}), which consists of two helices separated by a non-canonical k-turn[72,73]. A suite of ^{1}H-^{1}H NOESY, ^{1}H-^{1}H TOCSY, and ^{1}H-^{13}C HMQC datasets were processed within NMRFx, prior to chemical shift assignment. The chemical shift assignment strategy employed here used several NMRFx features, including RNA secondary structure prediction, chemical shift prediction, the Peak Slider Tool, and the spectral layout tool.

Chemical shift predictions of the construct were generated using NMRFx's Factorization Machine model, which requires the secondary structure of the RNA as input (*vide infra*). Since no known secondary structure was available, it was predicted from the primary sequence using a deep-learning model bundled with NMRFx; the result is presented in Fig. 8.

Simulated peak lists were then generated based on the predicted correlations in each of the experiments. Several peak-box generation routines are available in NMRFx for various types of experiments. As an example, for the ^{1}H-^{1}H NOESY, a model is available that predicts atoms pairing likely to give rise to NOE cross peaks based on secondary structure. This model was trained on a database that was assembled by scanning and processing a library of 3D RNA structures accessed from the PDB.

The ^{1}H-^{1}H NOESY, ^{1}H-^{1}H TOCSY, and ^{1}H-^{13}C HMQC spectra each contain several regions of interest. NMRFx was configured to display the aromatic-$^{1}H1'$ and $^{1}H1'$-$^{1}H1'$ regions of the NOESY and TOCSY spectra, along with the aromatic and $^{1}H1'$-$^{13}C1'$ regions of the HMQC spectrum in a compact fashion, as shown in Fig. 9. The Peak Slider Tool was then activated to perform the assignments. Despite differences between the predicted chemical shifts and the spectrum, as seen in Fig. 10, the predictions were robust enough to facilitate several initial assignments, particularly for outliers such as 33.H2, 11.H2, 7.H8, and 20.H8. These initial assignments served as a foundation, enabling the assignment of most aromatic and $^{1}H1'$ chemical shifts in under a day with high confidence. Figure 9 illustrates the project's status post-analysis, with red peak-boxes indicating frozen assignments. Further experiments would be necessary to resolve uncertainties in the remaining ribose chemical shifts. It is straightforward to retrospectively add additional datasets to the project and expand the peak network to facilitate further assignment.

Fig. 7 | Scanner Tool with metabolism data. Using the Scanner Tool to analyze a series of 1D ^{1}H HRMAS spectra of live algal cells sampled over 38 h, after being supplied $NaH^{13}CO_3$ as the sole carbon source. **a** Structure of the model triacylglyceride with saturated C16:0, monounsaturated C18:1, and polyunsaturated C18:3 fatty acid chains. **b** Broadband ^{13}C decoupled ^{1}H spectra with fatty acid resonance positions labeled. **c** Non-decoupled ^{1}H spectra acquired at the same points over the 38-h period. **d** Scanner Tool control panel. N.B. The `C13` column has been selected (blue box), which instructs the Scanner Tool to separate datasets with different values (i.e., those that were acquired with/without decoupling). Plots of spectrum integral as a function of time for the two groups of spectra, for the following spectral regions: **e** saturated CH_2, **f** omega-3 CH_3, **g** sucrose C6 CH_2. The sucrose data in (**g**) is presented alongside a fit of an inverse exponential of the form $I = A + B(1-e^{-Ct})$, with $A = 0.0045$, $B = 0.1077$, $C = 0.149\ s^{-1}$.

After assigning the spectra, a 3D structure calculation was performed based on distance restraints, dihedral angle restraints, and residual dipolar couplings (RDCs) derived from the assignments. The CMA-ES algorithm[74], a derivative-free optimizer, was used to minimize the deviation between the predicted RDCs—derived from a fit based on Singular Value Decomposition—and the experimental values. An ensemble of 50 structures was calculated; Fig. S4 of the SI shows the structure with the closest agreement between predicted and experimental RDC values. Information on violations associated with the ten best structures is provided in Table S5 of the SI.

Conclusions

NMRFx is an integrated application that provides access to a wide range of features, making it a valuable means of analyzing NMR data in structural biology and beyond, as highlighted by its application in four diverse case studies. The Java implementation allows cross-platform compatibility and enables efficient use of multi-core hardware. Support for Python and Java extensions (including plugins) allows users to extend this open-source program with new capabilities, and its support for standardized file formats allows it to be used in conjunction with other applications. NMRFx is under active development, and we expect that its existing and forthcoming tools, especially those supported by its integrated deep-learning capabilities, will facilitate protocols using both standard and novel analysis methods.

Methods

NUS processing example

The spectrum presented in Fig. 1 was generated by processing one of the datasets used in the NUScon competition[75]. The NUScon datasets were constructed by injecting synthetic peaks into experimentally acquired, uniformly sampled NMR datasets, and subsequently applying exponentially-biased sampling schedules to form NUS datasets. Tables 2 and 4 of the NUScon publication[75] provide an outline of all the NUS datasets. The dataset presented in Fig. 1 is the HNCACB dataset of Protein A, with synthetic peak set 1, and 8.8% sampling. The data were provided as a series of NMRPipe time-domain files along with a schedule file; NMRFx provides support to directly import such files. The full Python script used for the processing is given in the SI.

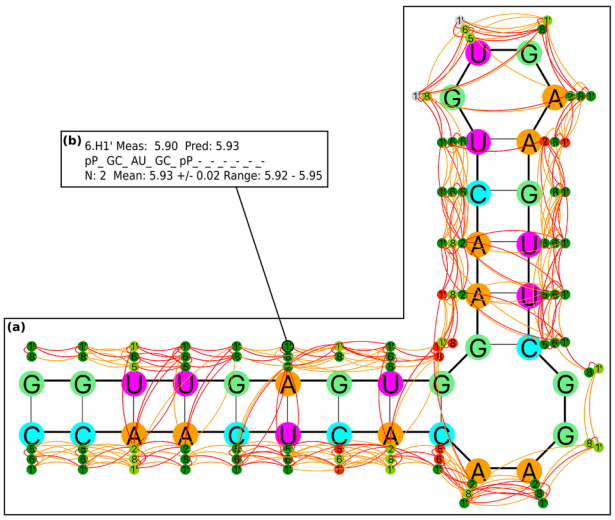

Fig. 8 | RNA Secondary Structure Tool. The secondary structure of the 36 nt RNA construct considered in Case Study 4. **a** The predicted secondary structure, generated using NMRFx's deep-learning model. Users can customize the types of atom displayed (circles with atom numbers). Curved lines connect atoms with predicted or measured NOE cross peaks; in this instance, the connections have been validated through use of the Slider Tool. The atoms are colored based on their assignment status: gray circles indicate unassigned atoms, while other colors (dark-green, light-green, orange, red) reflect the degree of agreement between the assigned and predicted chemical shift values. Orange/red colorings do not necessarily indicate incorrect assignments, but they may warrant increased scrutiny. **b** Clicking on an atom opens a pop-up window displaying the chemical shift prediction ($6 \cdot H1'$ in this example). Top: measured versus predicted chemical shift value. Middle: attributes input into the prediction model, which considers information about the base of interest, and the two closest bases on either side, along with H-bonded partners. `A` adenine, `C` cytosine, `G` guanine, `U` uracil, `P` purine base, `p` pyrimidine base. Bottom: Statistics on examples present in the training dataset with an identical set of attributes to the atom of interest.

Chemical shift prediction

To train the protein shift prediction model, a dataset of 1480 non-paramagnetic proteins comprising standard residues with matching BMRB and PDB entries was used. For testing, a dataset of 61 proteins previously used for ShiftX2 was used[53]. The RMSD and MAE values of the predicted shift values versus the experimentally assigned values for the 61 protein test dataset are listed in Table S1 of the SI. Furthermore, Table S6 of the SI lists the values of the attributes used for prediction of chemical shifts for different atom classes in glutamine residues as an example. During training, similar attribute sets with varying values are derived for each atom-amino acid pairing.

The RNA chemical shift predictor was trained with data assembled as previously described[59]. The current dataset consists of 371 RNA molecules that have assignments in the BMRB and associated PDB structures. Attributes comprising secondary structure motifs (loops, bulges, commonly occurring tetraloops) and nucleotide base pairing information were extracted and organized into training files. FM models were trained with stochastic gradient descent as implemented in the Tribuo library[41]. A table in the SI provides an evaluation of the model by cross-validated RMSD for each atom type (^{1}H, ^{13}C, ^{15}N), along with a comparison of the previously presented SVR model[59]. The updated FM model robustly handles a significant increase in training data points while maintaining high performance.

Structure calculation of the 100-protein dataset

For each protein in the 100-protein structure dataset[65], a NEF file containing NMR constraints was processed by NMRFx using the following command, issued from a UNIX terminal:

```
nmrfxs batch -n 200 -k 10 -a <nef-file>
```

where `<nef-file>` is a placeholder for the NEF file. The command generates an ensemble of 200 structures (`-n 200`), retains the ten with the lowest target function value (`-k 10`), and aligns the structures (`-a`). The output is a single PDBx/mmCIF file containing the structures of the refined

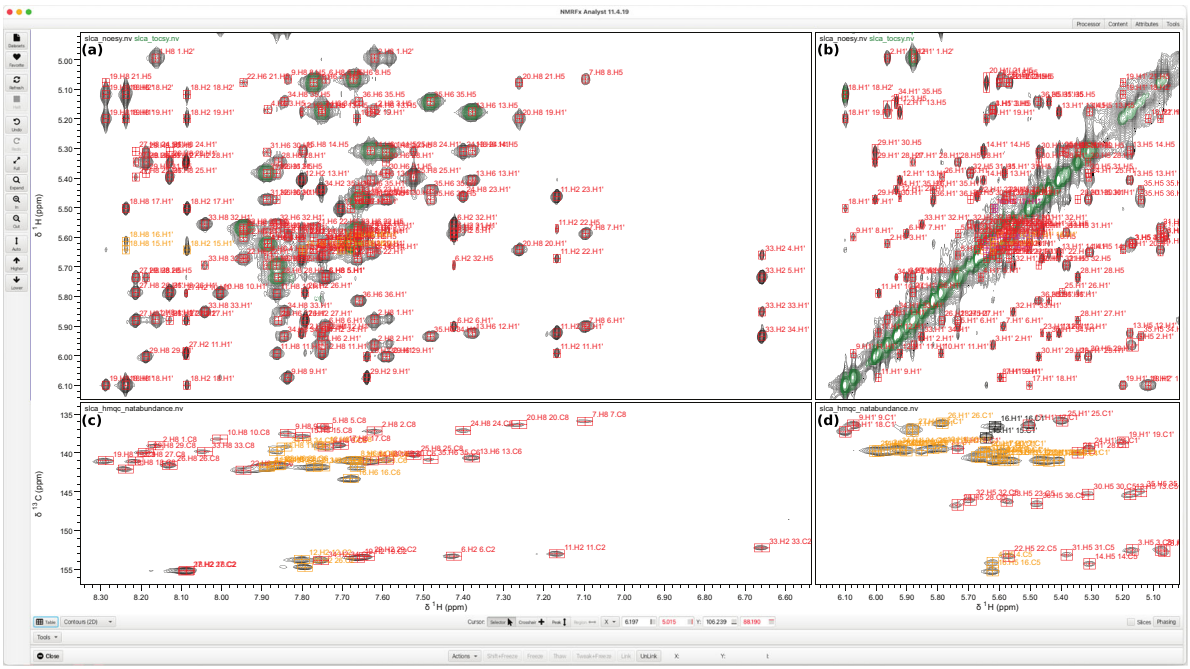

Fig. 9 | Spectral layout. The use of NMRFx's layout capability to visualize spectra related to the 36 nt RNA construct. Displays of both the ^{1}H-^{1}H NOESY (black) and ^{1}H-^{1}H TOCSY (green) spectra, with the visible region configured to show peaks associated with **a** aromatic-$^{1}H1'$ and **b** $^{1}H1'$-$^{1}H1'$ correlations. The ^{1}H-^{13}C HMQC spectrum, with the visible region configured to display peaks corresponding to **c** aromatic and **d** $^{1}H1'$-$^{13}C1'$ correlations. The charts are synchronized, such that when the user navigates one spectrum, the linked axes in the other spectra are automatically updated. The compact display of multiple regions of interest can be taken advantage of when the peak Slider Tool is employed (see also Fig. 10). The figure shows an advanced stage in the assignment process using the Slider Tool, with a majority of the assignments completed (red boxes).

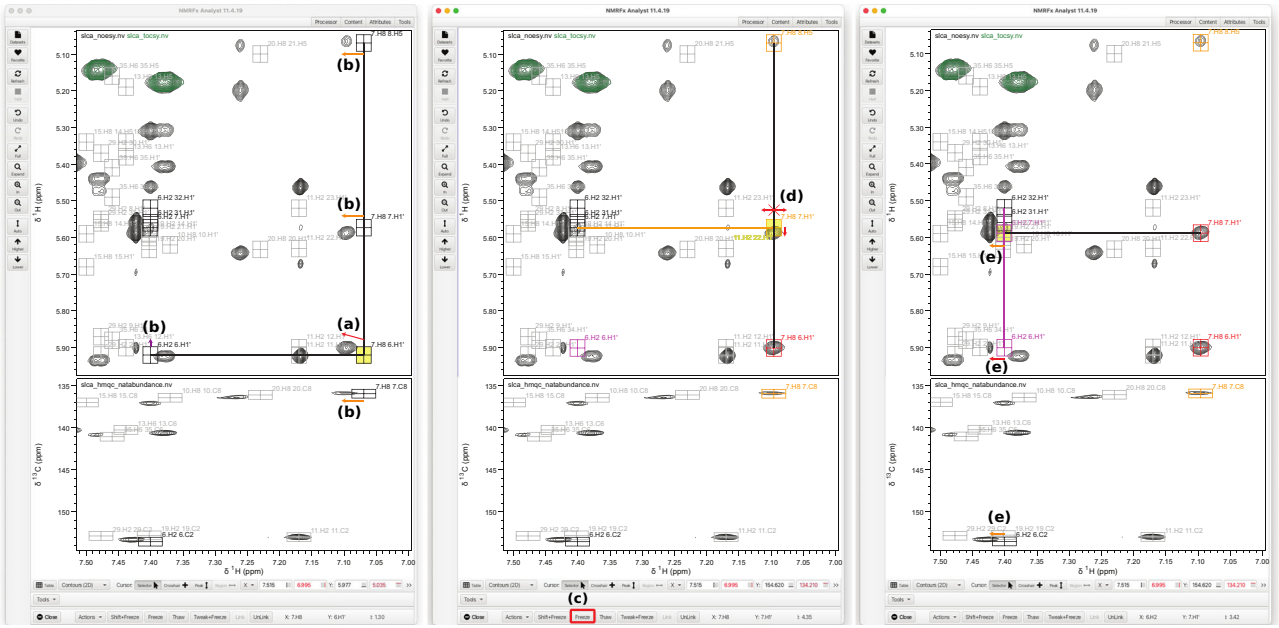

Fig. 10 | Peak Slider Tool. Assignment of peaks associated with the 36 nt RNA construct using the Peak Slider Tool. In the top row of charts, both the ^{1}H-^{1}H NOESY (black) and ^{1}H-^{1}H TOCSY (green) spectra are displayed, while the bottom row of charts shows the ^{1}H-^{13}C HMQC spectrum. **a** The peak-box corresponding to the predicted NOE between atoms 7.H8 and 6 · *H1′* is interactively adjusted based on close agreement with a peak in the observed NOESY spectrum. **b** Peak-boxes sharing an assignment are simultaneously updated. **c** Good agreement across the network of linked peaks suggests correct placement of this peak-box, which is therefore frozen in place by clicking the `Freeze` button. **d** Linked peak-boxes can no longer be adjusted in the frozen dimensions, but still can be in other dimensions, which is indicated by purple and orange colorings, indicating whether the peak can be moved in the *x*- or *y*-axis, respectively. **e** Assignment of atom 6.H2 is facilitated by two partially frozen peak-boxes, which can only be adjusted in the *x*-axis.

10-structure ensemble. No alterations to the method were made to optimize the calculation for any individual protein.

Algae metabolomics data processing

The green alga *Chlorella vulgaris* UTEX 395 was used in Case Study 3. Details of the culturing conditions and the ^{13}C labeling strategy, as well as the data acquisition, can be found in the Supplementary Information. The resulting FIDs were imported into NMRFx's Scanner Tool using a TSV file comprising the following columns: the path to the raw data; the time of data collection since introducing $NaH^{13}CO_3$; the replicate number; and whether ^{13}C decoupling was applied. The last decoupled dataset, which featured the most intense signals, was apodized with 0.3 Hz of line broadening, zero-filled, Fourier transformed, and phased. The baseline was corrected using manually selected regions and a third-order polynomial. All of the other FIDs were then processed using the same routine, and the spectra were displayed in stacked mode. The spectra were aligned using the TMSP peak at 0 ppm as the reference. Normalization was not performed because the cell pellets were of uniform cell count, and the triacylglyceride accumulation mirrored quantification obtained by gas chromatography fatty acid methyl ester (GC-FAME) analysis.

Data availability

The following Zenodo repositories contain data and tutorials of relevance to this work: 1. https://doi.org/10.5281/zenodo.17468978—NMRFx project using the Scanner Tool to investigate metabolism in wheat. 2. https://doi.org/10.5281/zenodo.17469095—NMRFx project for the assignment of taccalonolide E. 3. https://doi.org/10.5281/zenodo.17468856—NMRFx project for the 36 nt RNA case study 4. https://doi.org/10.5281/zenodo.17468622—NMRFx tutorial project for ubiquitin assignment using RunAbout.

Code availability

The NMRFx source code is available under the GNU General Public License, v3.0 at https://github.com/nanalysis/nmrfx. Executable versions of the software are available at http://nmrfx.org. Extensive documentation is provided at https://docs.nmrfx.org/.

Received: 22 August 2025; Accepted: 17 November 2025;
Published online: 05 December 2025

References

1. Claridge, T. D. W. *High-Resolution NMR Techniques in Organic Chemistry* 3rd edn (Elsevier Science, 2016).
2. Jacobsen, N. E. *NMR Data Interpretation Explained: Understanding 1D and 2D NMR Spectra of Organic Compounds and Natural Products* (Wiley, 2017).
3. Iggo, J. A. & Luzyanin, K. *NMR Spectroscopy in Inorganic Chemistry* 2nd edn (Oxford University Press, 2020).
4. Simpson, M. J. & Simpson, A. J. *NMR Spectroscopy: A Versatile Tool for Environmental Research* eMagRes Books. (John Wiley & Sons, 2014).
5. Bakhmutov, V. I. *Solid-State NMR in Materials Science: Principles and Applications* (CRC Press, 2011).
6. Vlaardingerbroek, M. T. & Boer, J. A. *Magnetic Resonance Imaging: Theory and Practice* 3rd edn (Springer Berlin, 2003).
7. Williamson, M. P., Havel, T. F. & Wüthrich, K. Solution conformation of proteinase inhibitor IIA from bull seminal plasma by ^{1}H nuclear magnetic resonance and distance geometry. *J. Mol. Biol.* **182**, 295–315 (1985).
8. Cavanagh, J., Skelton, N.J., Fairbrother, W.J., Rance, M. & Palmer III, A.G. *Protein NMR Spectroscopy: Principles and Practice* (Academic Press, 2010).
9. MaruŠiČ, M., Toplishek, M. & Plavec, J. NMR of RNA—structure and interactions. *Curr. Opin. Struct. Biol.* **79**, 102532 (2023).
10. Abyzov, A., Mandelkow, E., Zweckstetter, M. & Rezaei-Ghaleh, N. Fast motions dominate dynamics of intrinsically disordered tau protein at high temperatures. *Chemistry* **29**, e202203493 (2023).

11. Gronenborn, A. M. Integrated multidisciplinarity in the natural sciences. *J. Biol. Chem.* **294**, 18162–18167 (2019).
12. Hoff, S. E., Zinke, M., Izadi-Pruneyre, N. & Bonomi, M. Bonds and bytes: the odyssey of structural biology. *Curr. Opin. Struct. Biol.* **84**, 102746 (2024).
13. Schwalbe, H. et al. The future of integrated structural biology. *Structure* **32**, 1563–1580 (2024).
14. Nitsche, C. & Otting, G. NMR studies of ligand binding. *Curr. Opin. Struct. Biol.* **48**, 16–22 (2018).
15. Markley, J. L. et al. The future of NMR-based metabolomics. *Curr. Opin. Biotechnol.* **43**, 34–40 (2017).
16. Gowda, G.A.N. & Raftery, D. (eds). *NMR-Based Metabolomics: Methods and Protocols* Vol. 2037 of *Methods in Molecular Biology* (Springer, 2019).
17. Palmer III, A. G. NMR characterization of the dynamics of biomacromolecules. *Chem. Rev.* **104**, 3623–3640 (2004).
18. Johnson, B.A. Using NMRView to visualize and analyze the NMR spectra of macromolecules. in *Protein NMR Techniques* (ed. Downing, A. K.) 313–352 (Humana Press, 2004).
19. Johnson, B.A. From raw data to protein backbone chemical shifts using NMRFx processing and NMRViewJ analysis. in *Protein NMR: Methods and Protocols* (ed. Ghose, R.) 257–310 (Springer New York, 2018).
20. Norris, M., Fetler, B., Marchant, J. & Johnson, B. A. NMRFx processor: a cross-platform NMR data processing program. *J. Biomol. NMR* **65**, 205–216 (2016).
21. Beckwith, M. A., Erazo-Colon, T. & Johnson, B. A. RING NMR dynamics: software for analysis of multiple NMR relaxation experiments. *J. Biomol. NMR* **75**, 9–23 (2021).
22. Delaglio, F. et al. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
23. Skinner, S. P. et al. CcpNmr AnalysisAssign: a flexible platform for integrated NMR analysis. *J. Biomol. NMR* **66**, 111–124 (2016).
24. Lee, W., Tonelli, M. & Markley, J. L. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **31**, 1325–1327 (2015).
25. Lee, W., Rahimi, M., Lee, Y. & Chiu, A. POKY: a software suite for multidimensional NMR and 3D structure calculation of biomolecules. *Bioinformatics* **37**, 3041–3042 (2021).
26. Güntert, P. Automated NMR structure calculation with CYANA. in *Protein NMR Techniques* (ed. Kristina Downing, A.) 353–378 (Humana Press, 2004).
27. Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, M. G. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160**, 65–73 (2003).
28. Xu, X., Gagné, D., Aramini, J. M. & Gardner, K. H. Volume and compressibility differences between protein conformations revealed by high-pressure NMR. *Biophys. J.* **120**, 924–935 (2021).
29. Silvestrini, M. L. et al. Gating residues govern ligand unbinding kinetics from the buried cavity in HIF-2α PAS-B. *Protein Sci.* **33**, e5198 (2024).
30. Chin, S., Vos, J. & Weaver, J. *The Definitive Guide to Modern Java Clients with JavaFX: Cross-Platform Mobile and Cloud Development Updated for JavaFX 21 and 23* (Apress, 2024).
31. Juneau, J., Baker, J., Ng, V. Soto, L. & Wierzbicki, F. *The Definitive Guide to Jython* (Apress, 2010).
32. Ulrich, E. L. et al. NMR-STAR: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments. *J. Biomol. NMR* **73**, 5–9 (2019).
33. Gutmanas, A. et al. NMR exchange format: a unified and open standard for representation of NMR restraint data. *Nat. Struct. Mol. Biol.* **22**, 433–434 (2015).
34. Bernstein, F. C. et al. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542 (1977).
35. Westbrook, J. D. et al. PDBx/mmCIF ecosystem: foundational semantic tools for structural biology. *J. Mol. Biol.* **434**, 167599 (2022).
36. Higman, V. A., Płoskoń, E., Thompson, G. S. & Vuister, G. W. Perspective: On the importance of extensive, high-quality and reliable deposition of biomolecular NMR data in the age of artificial intelligence. *J. Biomol. NMR* **78**, 193–197 (2024).
37. Dalby, A. et al. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **32**, 244–255 (1992).
38. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
39. Chen, D., Wang, Z., Guo, D., Orekhov, V. & Qu, X. Review and prospect: Deep learning in nuclear magnetic resonance spectroscopy. *Chem. - A Eur. J.* **26**, 10391–10401 (2020).
40. Martín, A. et al. TensorFlow: large-scale machine learning on heterogeneous systems, Software available from tensorflow.org. (2015).
41. Oracle. Tribuo: machine learning library in Java. https://tribuo.org/. Accessed 31 July 2025.
42. Ying, J., Delaglio, F., Torchia, D. A. & Bax, A. Sparse multidimensional iterative lineshape-enhanced (SMILE) reconstruction of both non-uniformly sampled and conventional NMR data. *J. Biomol. NMR* **68**, 101–118 (2017).
43. Hyberts, S. G., Milbradt, A. G., Wagner, A. B., Arthanari, H. & Wagner, G. Application of iterative soft thresholding for fast reconstruction of NMR data non-uniformly sampled with multidimensional Poisson gap scheduling. *J. Biomol. NMR* **52**, 315–327 (2012).
44. Sun, S., Gill, M., Li, Y., Huang, M. & Byrd, A. R. Efficient and generalized processing of multidimensional NUS NMR data: the NESTA algorithm and comparison of regularization terms. *J. Biomol. NMR* **62**, 105–117 (2015).
45. Berger, S. & Braun, S. *200 and More NMR Experiments: A Practical Course* (John Wiley & Sons, 2004).
46. Li, D.-W., Hansen, A. L., Yuan, C., Bruschweiler-Li, L. & Brüschweiler, R. DEEP picker is a deep neural network for accurate deconvolution of complex two-dimensional NMR spectra. *Nat. Commun.* **12**, 5229–5229 (2021).
47. Buchanan, C. et al. UnidecNMR: automatic peak detection for NMR spectra in 1-4 dimensions. *Nat. Commun.* **16**, 449 (2025).
48. Marchant, J., Summers, M. F. & Johnson, B. A. Assigning NMR spectra of RNA, peptides and small organic molecules using molecular network visualization software. *J. Biomol. NMR* **73**, 525–529 (2019).
49. Bishop, A. C., Torres-Montalvo, G., Kotaru, S., Mimun, K. & Wand, A. J. Robust automated backbone triple resonance NMR assignments of proteins using Bayesian-based simulated annealing. *Nat. Commun.* **14**, 1556 (2023).
50. Johnson, B.A. From raw data to protein backbone chemical shifts using NMRFx processing and NMRViewJ analysis. in *Protein NMR: Methods and Protocols, Methods in Molecular Biology* (ed. Ghose, R.) 257–310 (Springer New York, 2018).
51. Schmidt, E. & Guntert, P. A new algorithm for reliable and general NMR resonance assignment. *J. Am. Chem. Soc.* **134**, 12817–29 (2012).
52. Jung, Y.-S. & Zweckstetter, M. Mars - robust automatic backbone assignment of proteins. *J. Biomol. NMR* **30**, 11–23 (2004).
53. Han, B., Liu, Y., Ginzinger, S. W. & Wishart, D. S. SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR* **50**, 43–57 (2011).

54. Shen, Y. & Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* **48**, 13–22 (2010).
55. Li, D.-W. & Brüschweiler, R. PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. *J. Biomol. NMR* **54**, 257–265 (2012).
56. L. Ptaszek, A., Li, J., Konrat, R., Platzer, G. & Head-Gordon, T. UCBShift 2.0: Bridging the gap from backbone to side chain protein chemical shift prediction for protein structures. *J. Am. Chem. Soc.* **146**, 31733–31745 (2024).
57. Hastie, T., Tibshirani, R. & Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd edn. Springer series in statistics. (Springer, 2009).
58. Barton, S., Heng, X., Johnson, B. A. & Summers, M. F. Database proton NMR chemical shifts for RNA signal assignment and validation. *J. Biomol. NMR* **55**, 33–46 (2013).
59. Brown, J. D., Summers, M. F. & Johnson, B. A. Prediction of hydrogen and carbon chemical shifts from RNA using database mining and support vector regression. *J. Biomol. NMR* **63**, 39–52 (2015).
60. Rendle, S. Factorization machines. In *Proc. 2010 IEEE International Conference on Data Mining* 995–1000 (IEEE, 2010).
61. Frank, A. T., Law, S. M. & Brooks, C. L. A simple and fast approach for predicting ^{1}H and ^{13}C chemical shifts: toward chemical shift-guided simulations of RNA. *J. Phys. Chem. B* **118**, 12168–12175 (2014).
62. Bremser, W. Hose - a novel substructure code. *Anal. Chim. Acta* **103**, 355–365 (1978).
63. Johnson, B. A. & Sugg, E. E. Determination of the three-dimensional structure of iberiotoxin in solution by proton nuclear magnetic resonance spectroscopy. *Biochemistry* **31**, 8151–8159 (1992).
64. Case, D.A. et al. Recent developments in Amber biomolecular simulations. *J. Chem. Inform. Model.* **65**, 7835–7843 (2025).
65. Klukowski, P. et al. The 100-protein NMR spectra dataset: a resource for biomolecular NMR data analysis. *Sci. Data* **11**, 30–30 (2024).
66. OpenChemLib, Open source Java-based chemistry library, https://github.com/Actelion/openchemlib.
67. Nielsen, J. T. & Mulder, F. A. A. POTENCI: prediction of temperature, neighbor and pH-corrected chemical shifts for intrinsically disordered proteins. *J. Biomol. NMR* **70**, 141–165 (2018).
68. Nielsen, J.T. & Mulder, F.A.A. There is diversity in disorder—in all chaos there is a cosmos, in all disorder a secret order. *Front. Mol. Biosci.* **3**, 4 (2016).
69. Driscoll, P., Thompson, G. & Harris, R. The ubiquitin NMR resource—Archive 1 [backbone experiments]. (2025), https://doi.org/10.5281/zenodo.14791182.
70. Mühlbauer, A., Seip, S., Nowak, A. & Tran, V. Five novel taccalonolides from the roots of the Vietnamese Plant Tacca paxiana. *Helv. Chim. Acta* **86**, 2065–2072 (2003).
71. PubChem compound CID 56672430, Taccalonolide E, https://pubchem.ncbi.nlm.nih.gov/compound/56672430.
72. D'Souza, V., Dey, A., Habib, D. & Summers, M. F. NMR structure of the 101-nucleotide core encapsidation signal of the Moloney murine leukemia virus. *J. Mol. Biol.* **337**, 427–442 (2004).
73. Marchant, J., Bax, A. & Summers, M. F. Accurate measurement of residual dipolar couplings in large RNAs by variable flip angle NMR. *J. Am. Chem. Soc.* **140**, 6978–6983 (2018).
74. Auger, A. & Hansen, N. A restart CMA evolution strategy with increasing population size. In *Proc. 2005 IEEE Congress on Evolutionary Computation* Vol. 2, 1769–1776 (IEEE, 2005).
75. Pustovalova, Y. et al. NUScon: a community-driven platform for quantitative evaluation of nonuniform sampling in NMR. *Magn. Reson.* **2**, 843–861 (2021).
76. Vijay-Kumar, S., Bugg, C. E. & Cook, W. J. Structure of ubiquitin refined at 1.8Å resolution. *J. Mol. Biol.* **194**, 531–544 (1987).
77. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

Acknowledgements

B.A.J. would like to thank the many research groups who have provided data, made suggestions, and asked questions that helped us improve NMRFx. And special thanks are due to Jannalie Taylor and Vincent Fiack for help with software engineering and to Shibani Bhattacharya for collecting the ubiquitin relaxation data. S.G.H. would like to thank Prof. Adam Schuyler, UConn Health, for useful discussions regarding the NUScon datasets. G.L.H. would like to thank Susan Mooberry and April Risinger, University of Texas Health San Antonio, for providing the sample of taccalonolide E, and Rob Gardner (deceased), Bill Hiscox, Washington State University, Todd Pedersen, Brent Peyton, and Robin Gerlach, Montana State University, for assistance with growing the algae cultures and acquiring HRMAS data. G.L.H. would also like to thank the Washington State University Center for NMR Spectroscopy for providing instrument time to acquire data for both the taccalonolide E and algal metabolomics projects. This work was supported in part by grants from the National Institute of General Medical Sciences of the National Institutes of Health, R01 GM123012 and RM1 GM145397 to B.A.J., and the National Institute of Allergy and Infectious Diseases of the National Institutes of Health, U54 AI170660 to B.A.J., J.M., and M.F.S., and R01 AI150498 to M.F.S., and from the Howard Hughes Medical Institute to M.F.S. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

Conceptualization: B.A.J. Methodology: B.A.J. Software: E.K., S.G.H., K.M.C., B.A.J. Validation: E.K., S.G.H., G.L.H., M.F.S., J.M., B.A.J. Formal analysis: E.K., S.G.H., B.A.J., J.M. Investigation: E.K., S.G.H., G.L.H., J.M., B.A.J. Data curation: E.K., S.G.H., G.L.H., J.M., B.A.J. Resources: G.L.H., M.F.S., J.M., B.A.J. Writing—original draft: G.L.H.,E.K., S.G.H., J.M., B.A.J. Writing—review and editing: G.L.H., E.K., S.G.H., J.M., M.F.S., B.A.J. Visualization: E.K., S.G.H., B.A.J., J.M. Supervision: B.A.J. Project administration: B.A.J. Funding acquisition: M.F.S., J.M., B.A.J.

Competing interests

B.A.J. has worked as a consultant to, and owns stock in, Nanalysis Corp. While the software described in this manuscript is open source, Nanalysis owns the commercial rights to the software.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42004-025-01812-8.

Correspondence and requests for materials should be addressed to Bruce A. Johnson.

Peer review information *Communications Chemistry* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.