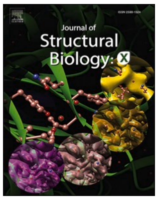
Journal of Structural Biology: X 6 (2022) 100073

Contents lists available at ScienceDirect

Journal of Structural Biology: X

journal homepage: www.sciencedirect.com/journal/journal-of-structural-biology-x



Research Article

POKY software tools encapsulating assignment strategies for solution and solid-state protein NMR data

Ira Manthey [a], Marco Tonelli [b,c], Lawrence Clos II [d], Mehdi Rahimi [e], John L. Markley [c], Woonghee Lee [e,*]

[a] Department of Chemistry, and URS Scholars Program, University of Wisconsin-Madison, Madison, WI 53706, USA
[b] National Magnetic Resonance Facility at Madison, University of Wisconsin-Madison, Madison, WI 53706, USA
[c] Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706, USA
[d] DNA Software, Plymouth, MI 48170, USA
[e] Department of Chemistry, University of Colorado Denver, Denver, CO 80204, USA

ARTICLE INFO

Keywords:
NMR Software
Assignment
Semi-Automation
NMR
POKY
Versatile Assigner

ABSTRACT

NMR spectroscopy provides structural and functional information about biomolecules and their complexes. The complexity of these systems can make the NMR data difficult to interpret, particularly for newer users of NMR technology, who may have limited understanding of the tools available and how they are used. To alleviate this problem, we have created software based on standardized workflows for both solution and solid-state NMR spectroscopy of proteins. These tools assist with manual and automated peak picking and with chemical shift assignment and validation. They provide users with an optimized path through spectral analysis that can help them perform the necessary tasks more efficiently.

1. Introduction

The Worldwide Protein Data Bank (wwPDB) and Biological Magnetic Resonance Bank (BMRB) currently contain more than 13,000 entries of NMR structures and over 15,000 entries of NMR chemical shifts from biomacromolecules respectively (Berman et al., 2007; Markley et al., 2008). Structural and functional information on proteins derived from NMR spectroscopy has furthered our understanding of normal and abnormal biological functions and supported the design of drugs and therapeutics. The complexity of NMR spectroscopy can provide a challenge to new users and veteran spectroscopists alike. The field continues to evolve with advances in instrumentation, protocols, and strategies that have the potential to reduce the workload of NMR analysis. Experienced spectroscopists generally keep up with these advances, which require adjustments in the way the data are collected and analyzed. However, less experienced researchers may need some guidance in understanding the newer optimal workflows.

To address this problem, we initially established a software platform, *Integrative NMR* (Lee et al., 2016a), whose key components, *NMRFAM-SPARKY* (Lee et al., 2015) and *PONDEROSA-C/S* (Lee et al., 2014), automate routine tasks in the workflow. We recently upgraded this software platform by incorporating modern interfaces and software programs that yield more reliable assignments. We have added routines that investigate and relate peaks in multiple spectral view windows and utilize tables of chemical shift statistics in guiding assignments. This new suite, named *POKY* (Lee et al., 2021), incorporates *I-PINE webserver* (Lee et al., 2019), *PINE-SPARKY* (Lee et al., 2009), and *PINE-SPARKY.2* (Lee et al., 2018). The *I-PINE webserver* now implements automated three-dimensional protein structure calculation routines based on backbone chemical shift-based *CS-Rosetta* (Shen et al., 2008) and NOE-based *AUDANA* (Lee et al., 2016b) methods.

Although solid-state NMR (ss NMR) spectroscopy of proteins is a rapidly developing field, the deployment of computational tools for ss NMR of proteins has lagged behind those for solution NMR. A few tools, however, are available now for the analysis of protein ss NMR data from oriented samples and from magic angle spinning (MAS). Oriented sample solid-state nuclear magnetic resonance (OS-ss NMR) has a long tradition of being used to elucidate topological restraints for membrane proteins aligned in lipid bilayers (Opella and Marassi, 2004). To assist in the analysis of OS-ss NMR data, we developed the *PISA-SPARKY* program (Weber et al., 2020), which simulates *Polar Index Slant Angle* (PISA)-wheels through exhaustive peak fitting, error analysis, and

* Corresponding author.
E-mail address: woonghee.lee@ucdenver.edu (W. Lee).

https://doi.org/10.1016/j.yjsbx.2022.100073
Received 21 March 2022; Received in revised form 4 August 2022; Accepted 23 August 2022
Available online 28 August 2022

plotting of dipolar and chemical shift waves (Marassi and Opella, 2000; Wang et al., 2000). Recent improvements in ultra-high-speed MAS technologies, which create artificial isotropic conditions, have greatly improved the signal-to-noise levels of ss NMR spectra (Polenova et al., 2015). This process is enabling ss NMR studies of large proteins, proteins that are highly dynamic, insoluble fibrils, and membrane proteins. Although automated approaches to the assignment of protein ss NMR spectra have been introduced (Hu et al., 2011; Moseley et al., 2010; Schmidt et al., 2013), they have not been widely used presumably because of the large linewidths and low peak intensities of ss NMR spectra. A flexible CLI (command-line interface) tool, *PLUQ* (PACSYlite Unified Query) and its successor, *PLUQin* (Fritzsching et al., 2016, Fritzsching et al., 2013) have proved useful in addressing these problems. *PLUQ* and *PLUQin* estimate the likelihood of amino acid types and secondary structures by querying heavy atom chemical shifts in the *PACSY DB* (Lee et al., 2012). The user still needs to determine the residue number and amino acid type from the multiple choices provided by the program.

2. Newly developed tools and methods

2.1. Overview

Our core goals in software development have been to improve communication between the user and the software, to make assignments less cumbersome while improving their accuracy, and to create a versatile analysis method that can be utilized in different contexts. Users need to know what information can be extracted form from a particular data set and how this can inform other aspects of analysis. The tools contain adjustable parameters, including the tolerance of peak positions, the sensitivity of automated peak picking, and the direction of the assignment walk, that enable their fine-tuning.

Here, we introduce new and improved tools integrated as plug-ins into the latest version of *POKY*. The tools, which are easily accessible through two-letter-code shortcuts (Table 1) or from a menu, support standardized workflows for the analysis of multidimensional solution and ss NMR spectra. The workflows, which are interactive and unrestrained, offer concrete protocols or general guidelines to users who wish to implement their own methods. We have evaluated and refined these new tools by analyzing their performance with different sets of input data.

2.2. Integrative peak picking automation by iPick and transferring tools

Because the *APES* peak picking program (Shin et al., 2008), which was integral to *Integrative NMR*, supports data from only a few solution NMR experiments, we recently developed *iPick* (Rahimi et al., 2021), a program that works with data from a wide range of 2D/3D/4D NMR experiments. We integrated the *iPick* GUI into POKY to support peak picking and peak importing (two-letter-code *iP*). Many settings are available in *iPick*, including options for changing the sensitivity of the tool, setting desired peak counts, and letting users choose peak picking software (Fig. 1). Another new feature in *iPick* is the *Reliability Score*. This is a score assigned to each peak based on the peak volume from automatic integration, the signal-to-noise ratio, and the peak linewidth. The *Reliability Score* provides a simple numeric indication of the quality of the peaks and allows peaks scoring below a user selected threshold to be removed. The *iPick* plug-in provides *Basic* and *Advanced* modes. In the *Basic* mode, one simply selects a spectrum and clicks the "Run iPick" button. The *Advanced* mode provides options for fine-tuning every aspect of the peak picking process, providing a great amount of control for the advanced user.

The *iPick* plug-in can be used in tandem with a new tool in *POKY* named *Transfer/Assign Peaks*, (two-letter-code *TP*), which looks for the presence of unpicked peaks in a given spectrum on the basis of their predicted occurrence from data from other experiments. The tool also accounts for the compression of resonance data when carrying out automated transfer between two- and three-dimensional spectra (Fig. 2).

Table 1
Two-letter *POKY* codes that call up tools or functions described in this paper. *Command Finder* (two-letter-code *cf*) provides access to the full list of tools and features. Two-letter-codes that call up new windows can be used anywhere in *POKY*, whereas two-letter-codes that affect spectra apply only to currently selected spectra.

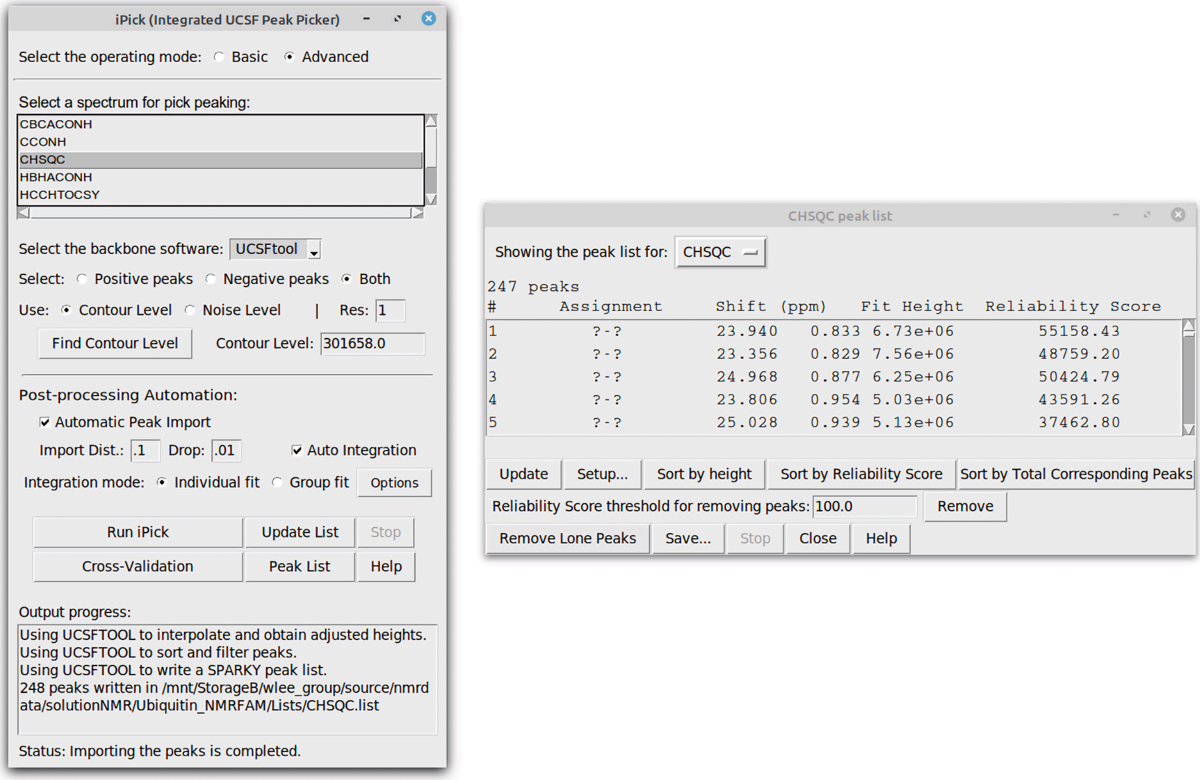
Two Letter Code	Function
cf	*Command Finder*. Opens window where users can search for tools or features and check their two-letter code for quick access.
va	*Versatile Assigner*. This tool uses standard carbon chemical shifts of amino acids to predict the residue assignments. A history of predictions also allows the tool to predict and validate sub-sequences.
pq	*PLUQin*. Opens a window for communication with the PLUQin webserver. PLUQin uses inputs of intra-residue chemical shift lists to query the PACSY database to find possible residue assignments.
TP	*Transfer/Assign Peaks*. The tool can transfer all or selected peaks between two- and three-dimensional spectra.
sp	*Strip Plot*. This tool enables views of multiple spectra or multiple views of the same spectra at once. All views are synchronized for easy exploration of multiple spectra.
lv	*LACS* (Linear Analysis of Chemical Shifts) detects offset errors in spectra and suggests chemical shift corrections.
st	*Spectrum Settings*. Allows users to change aspects of spectra, for example, chemical shift corrections found by LACS.
ir	*Reference Views*. Supports up to three separate windows with standard resonance ranges marked for amino acid atom types. The reference views are for alpha and beta carbons, all carbons, or all hydrogens. Comparison of experimental data against standard chemical shift ranges enables manual assignments to residue types.
ao	*Add Orthogonal Strip*. Adds a new strip in Strip Plot in the orthogonal plane, rotated 90° about the central vertical axis. The tool allows users to check for peaks within three-dimensional signals and check a peak's validity while using strip plot for linking or residue assignment.
dv	*Add Centered Vertical Grid Line*. Adds a vertical grid line to a spectrum in *Strip Plot*, centered in the window.
dh	*Add Centered Horizontal Grid Line*. Adds a vertical grid line to a spectrum in Strip Plot, centered in the window.
es / bs	*Generate an NMR-STAR 3.1 (es) or NMR-STAR 3.2 (bg) file.*

2.3. Reference Views

Reference Views is a plugin (two-letter-code *ir*; Fig. 3) that provides synchronous assignment references for backbone carbons, aliphatic carbons and aliphatic protons by statistical analysis of chemical shifts deposited in BMRB. *Reference Views* offers up to three interactive windows in which an experimental spectral slice is compared with standard chemical shift ranges for all 20 amino acids: a *CA-CB* window (Fig. 3), an *all carbon* window, and an *all hydrogen* window. We converted the 2D probability density maps for the ^{1}H, ^{13}C, and ^{15}N chemical shifts of the 20 amino acids, previously used in the *I-PINE* algorithm, into an interactive layout that allows users in real time to cross-validate signal positions against chemical shift probability ranges. This method is far faster and more accurate than those that came before, such as having separate pages for standard carbon shifts of each amino acid or probability density maps of 20 amino acid types with large overlapping regions.

2.4. Chemical shift reference correction

LACS (Wang et al., 2005) is a tool that performs a linear analysis of chemical shifts to determine the zero offset compatible with the data; the difference between this and the standard frequency of the DSS reference, suggests an offset correction, which can be implemented through *Spectrum Settings* (two-letter-code *st*; Fig. 4B). If the spectra have been referenced to a different standard, an adjustment can be applied in



iPick (Integrated UCSF Peak Picker)

Select the operating mode: ○ Basic ● Advanced

Select a spectrum for pick peaking:

CBCACONH
CCONH
CHSQC
HBHACONH
HCCHTOCSY

Select the backbone software: UCSFtool

Select: ○ Positive peaks ○ Negative peaks ● Both

Use: ● Contour Level ○ Noise Level | Res: 1

Contour Level: 301658.0

Post-processing Automation:

☑ Automatic Peak Import

Import Dist.: .1 Drop: .01 ☑ Auto Integration

Integration mode: ● Individual fit ○ Group fit

Output progress:

```
Using UCSFTOOL to interpolate and obtain adjusted heights.
Using UCSFTOOL to sort and filter peaks.
Using UCSFTOOL to write a SPARKY peak list.
248 peaks written in /mnt/StorageB/wlee_group/source/nmrd
ata/solutionNMR/Ubiquitin_NMRFAM/Lists/CHSQC.list
```

Status: Importing the peaks is completed.

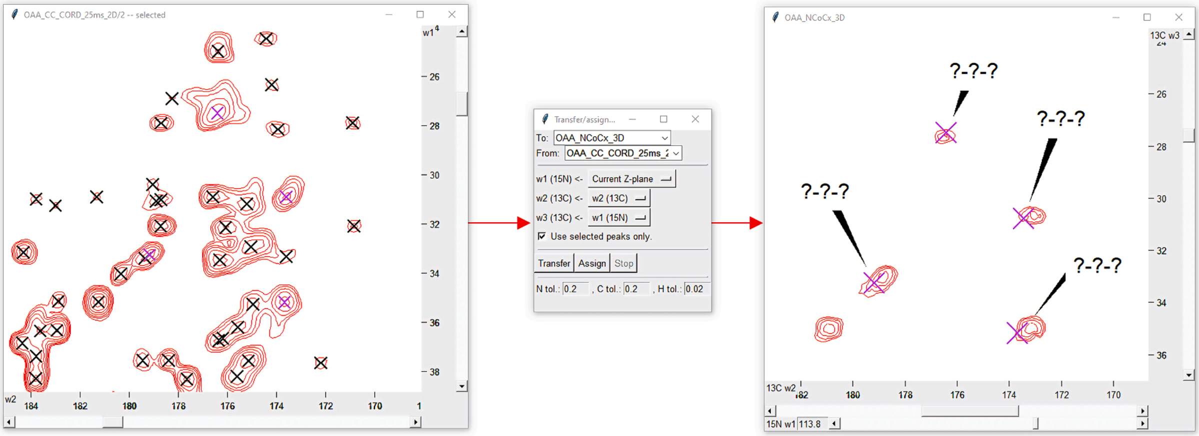
CHSQC peak list

Showing the peak list for: CHSQC

```
247 peaks
#        Assignment         Shift (ppm)    Fit Height  Reliability Score
1           ?-?          23.940   0.833 6.73e+06         55158.43
2           ?-?          23.356   0.829 7.56e+06         48759.20
3           ?-?          24.968   0.877 6.25e+06         50424.79
4           ?-?          23.806   0.954 5.03e+06         43591.26
5           ?-?          25.028   0.939 5.13e+06         37462.80
```

Reliability Score threshold for removing peaks: 100.0

Fig. 1. (*Left*) Screen shot showing the *Advanced* mode of the *iPick plug-in*, which provides various options for fine-tuning the process. *Auto Integration*, which is checked in this picture, leads to the calculation of a *Reliability Score* value for each peak. (*Right*) Screen shot showing a portion of the final peak list along with the calculated *Reliability Score* for each peak. This window opens automatically when *iPick* has completed peak picking. For an efficient workflow, first click on the "*Sort by Reliability Score*" button and then navigate down the list to check low values by double-clicking on the peak number; this takes the user to the actual peak. These results can help the user set the reliability score threshold level for actual peaks. Then clicking the "*Remove*" button removes all picked peaks lower than that threshold.



Transfer/assign...

To: OAA_NCoCx_3D

From: OAA_CC_CORD_25ms_2

w1 (15N) <- Current Z-plane

w2 (13C) <- w2 (13C)

w3 (13C) <- w1 (15N)

☑ Use selected peaks only.

N tol.: 0.2 , C tol.: 0.2 , H tol.: 0.02

Fig. 2. *Transfer/Assign Peaks* (two-letter-code *TP*) is a tool for transferring peaks from one spectrum to another, including between 2D and 3D spectra. Once the spectra to be matched have been selected, the user can choose which dimensions should be matched. Transferred peaks are automatically given the label "?-?-?" in order to make it easy for users to see which peaks have been added. The tool allows the user to add peaks to the plane of a dimension currently being viewed in *POKY*, so that transfers can be verified visually. When using this feature with a 3D spectrum, the *Transfer* window (two-letter-code *TP*) tool ensures that the dimensionality of the window opened matches that of the target spectrum.

Spectral Properties window (two-letter-code *st*): for example, −2 ppm and −40.48 ppm offsets for TMS and adamantane, respectively. As of June 9, 2022, of the ss NMR entries deposited in BMRB, 312 have used DSS, 19 TMS, and 144 adamantane as the reference. *POKY* contains a plugin (two-letter-code *lv*) that implements *pyLACS*, our new PYTHON implementation of LACS, to seamlessly analyze a chemical shift table for possible offset correction. Unlike with NMRFAM-SPARKY, the operation is carried out without the need for an Internet connection.

2.5. *Versatile Assigner*

Versatile Assigner (Fig. 5) uses sub-sequence information to validate

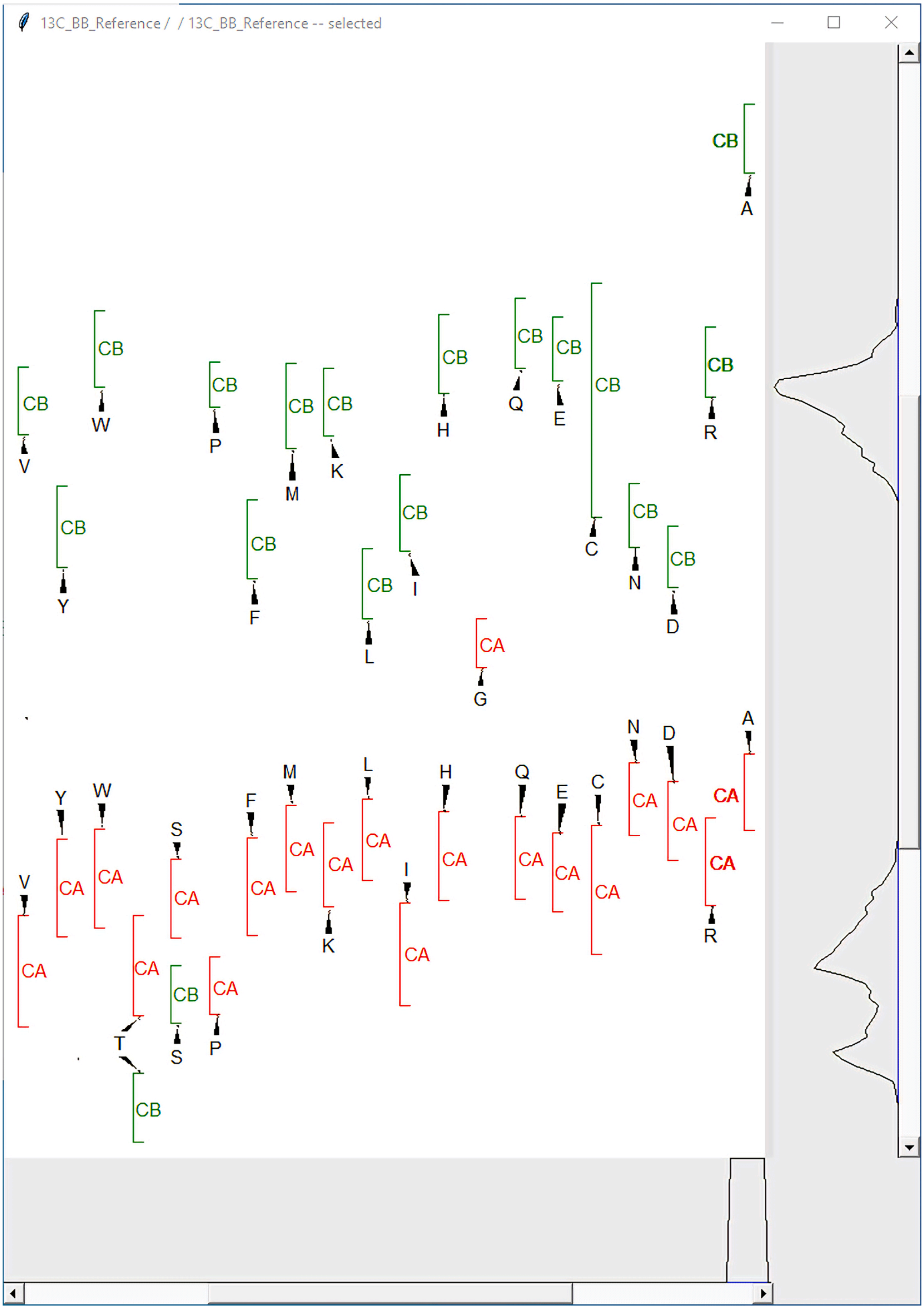
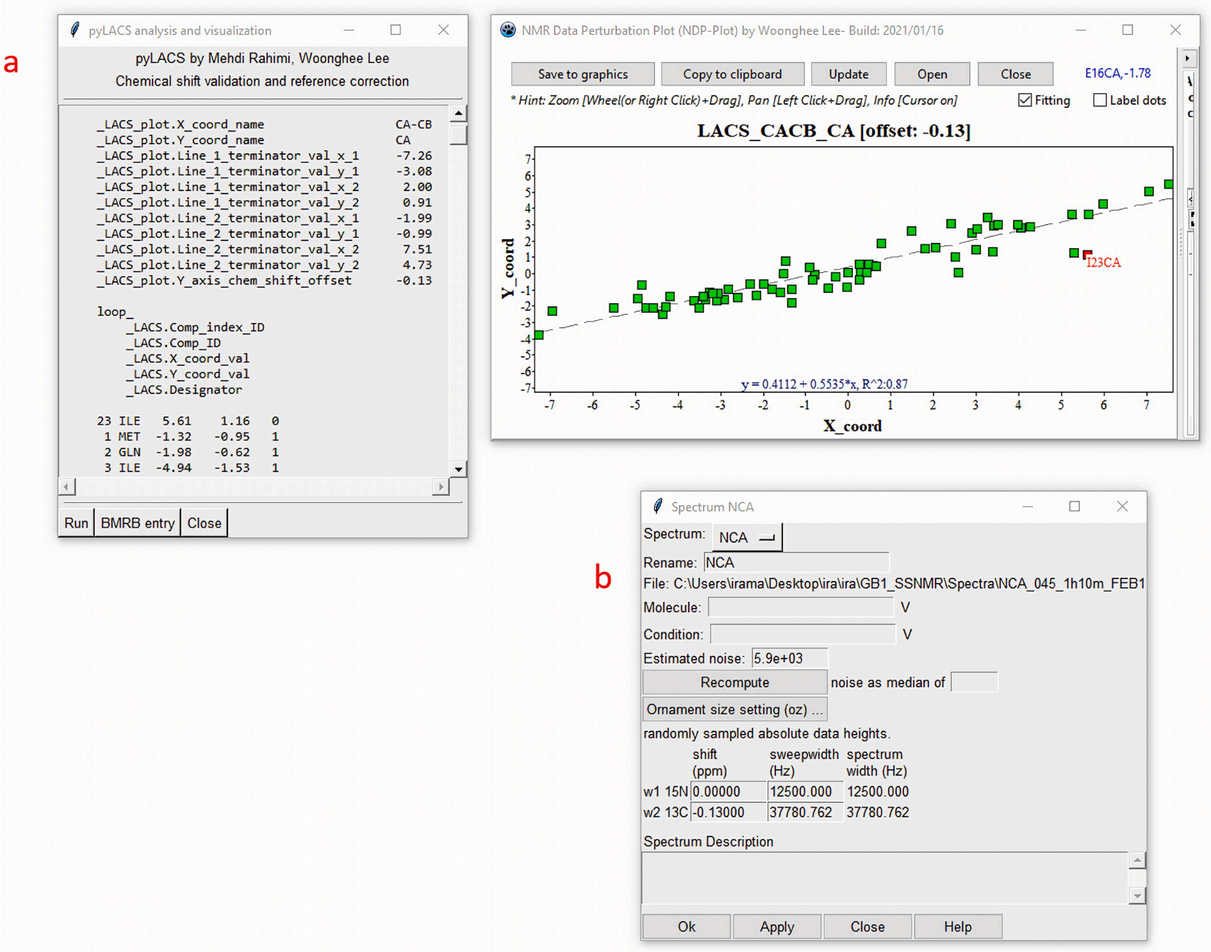


Fig. 3. Example of a *CA-CB Reference View*. Experimental ^{13}CA and ^{13}CB peaks (shown vertically on the right) are compared with the corresponding standard ^{13}C chemical shift ranges for all 20 amino acids. The vertical slide enables adjustment of the chemical shift window, and the horizontal slide bar allows adjustment of the proximity of the corresponding chemical shift ranges to the experimental peaks.

assignments predicted from the real-time probabilistic approaches implemented in *Reference Views*. Once probable assignments have been added to residue history, the sequence of predicted assignments can be compared with the protein sequence to verify their correctness. In addition to supporting full assignments, users can use the sub-sequence approach to analyze and compare NMR data collected before and after adding a ligand to detect binding sites. *Versatile Assigner* is functionally similar *PLUQin-POKY* (two-letter-code *pq*); however, its advantages for users include no dependency on an internet connection and faster calculation of amino acid probabilities. *Versatile Assigner* bases its predictions on selected ^{13}C/^{15}N resonances as well as connected ^{1}H resonances. Unlike *PLUQin-POKY*, *Versatile Assigner* is capable of “linking predictions” from multiple residues in order to assign sub-sequences. By finding two or more consecutive sequential predictions, the user can


a
b
pyLACS analysis and visualization
Run
BMRB entry
Close
NMR Data Perturbation Plot (NDP-Plot) by Woonghee Lee- Build: 2021/01/16
Save to graphics
Copy to clipboard
Update
Open
Close
E16CA,-1.78
* Hint: Zoom [Wheel(or Right Click)+Drag], Pan [Left Click+Drag], Info [Cursor on]
☑ Fitting
□ Label dots
LACS_CACB_CA [offset: -0.13]
Y_coord
7 6 5 4 3 2 1 0 -1 -2 -3 -4 -5 -6 -7
X_coord
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7
I23CA
y = 0.4112 + 0.5535*x, R^2:0.87
Spectrum NCA
Recompute
Ornament size setting (oz) ...
Ok
Apply
Close
Help

pyLACS by Mehdi Rahimi, Woonghee Lee

Chemical shift validation and reference correction

```
_LACS_plot.X_coord_name                     CA-CB
_LACS_plot.Y_coord_name                     CA
_LACS_plot.Line_1_terminator_val_x_1        -7.26
_LACS_plot.Line_1_terminator_val_y_1        -3.08
_LACS_plot.Line_1_terminator_val_x_2         2.00
_LACS_plot.Line_1_terminator_val_y_2         0.91
_LACS_plot.Line_2_terminator_val_x_1        -1.99
_LACS_plot.Line_2_terminator_val_y_1        -0.99
_LACS_plot.Line_2_terminator_val_x_2         7.51
_LACS_plot.Line_2_terminator_val_y_2         4.73
_LACS_plot.Y_axis_chem_shift_offset         -0.13

loop_
    _LACS.Comp_index_ID
    _LACS.Comp_ID
    _LACS.X_coord_val
    _LACS.Y_coord_val
    _LACS.Designator

23 ILE   5.61    1.16   0
 1 MET  -1.32   -0.95   1
 2 GLN  -1.98   -0.62   1
 3 ILE  -4.94   -1.53   1
```

Spectrum: NCA

Rename: NCA

File: C:\Users\irama\Desktop\ira\ira\GB1_SSNMR\Spectra\NCA_045_1h10m_FEB1

Molecule: V

Condition: V

Estimated noise: 5.9e+03

noise as median of

randomly sampled absolute data heights.

	shift (ppm)	sweepwidth (Hz)	spectrum width (Hz)
w1 15N	0.00000	12500.000	12500.000
w2 13C	-0.13000	37780.762	37780.762

Spectrum Description

Fig. 4. *LACS* (Linear Analysis of Chemical Shifts) detects possible chemical shift referencing errors and suggests corrections. *a*) The *pyLACS* tool window (*left*) is accessed with the two-letter code *lv*. Each dimension can be inspected graphically (*right*) to investigate the fitted offset and outliers. This is a highly important tool when using *Versatile Assigner* and *I-PINE*, because offset errors greater than 0.25 ppm can cause significantly decreased assignment accuracy (see Supplementary Table 3). *b*) *Spectrum Setting* window in which the user can change spectrum names, compute noise levels, and enter shift offsets from *pyLACS*. The screen shot shows −0.130 ppm as the ^{13}C offset correction. The Spectrum Description area is a place where the user can leave notes about the spectrum.

narrow down their position in the overall peptide sequence.

2.6. Strip Plot enhancements

Strip Plot is a graphical interface in *POKY* used for choosing a peak in a 2D/3D/4D spectrum and extracting a two-dimensional slice containing that peak for viewing and analysis (Fig. 6). *Strip Plot* includes the following updated features: *a*) Up to three interactive *Strip Plots* can be viewed simultaneously. *b*) Entry tabs at the top of the window enable for quick manipulations of tolerance levels. *c*) Save and load functions capture settings and make it possible to reproduce working strips. *d*) Orthogonal planes can be viewed in order to check the validity of a signal. Tutorials on the function and use of strip plot, along with most other tools, can be found at our YouTube channel: https://www.youtube.com/c/LeeGroupatCUDenver.

3. Workflow creation for solution and Solid-State NMR

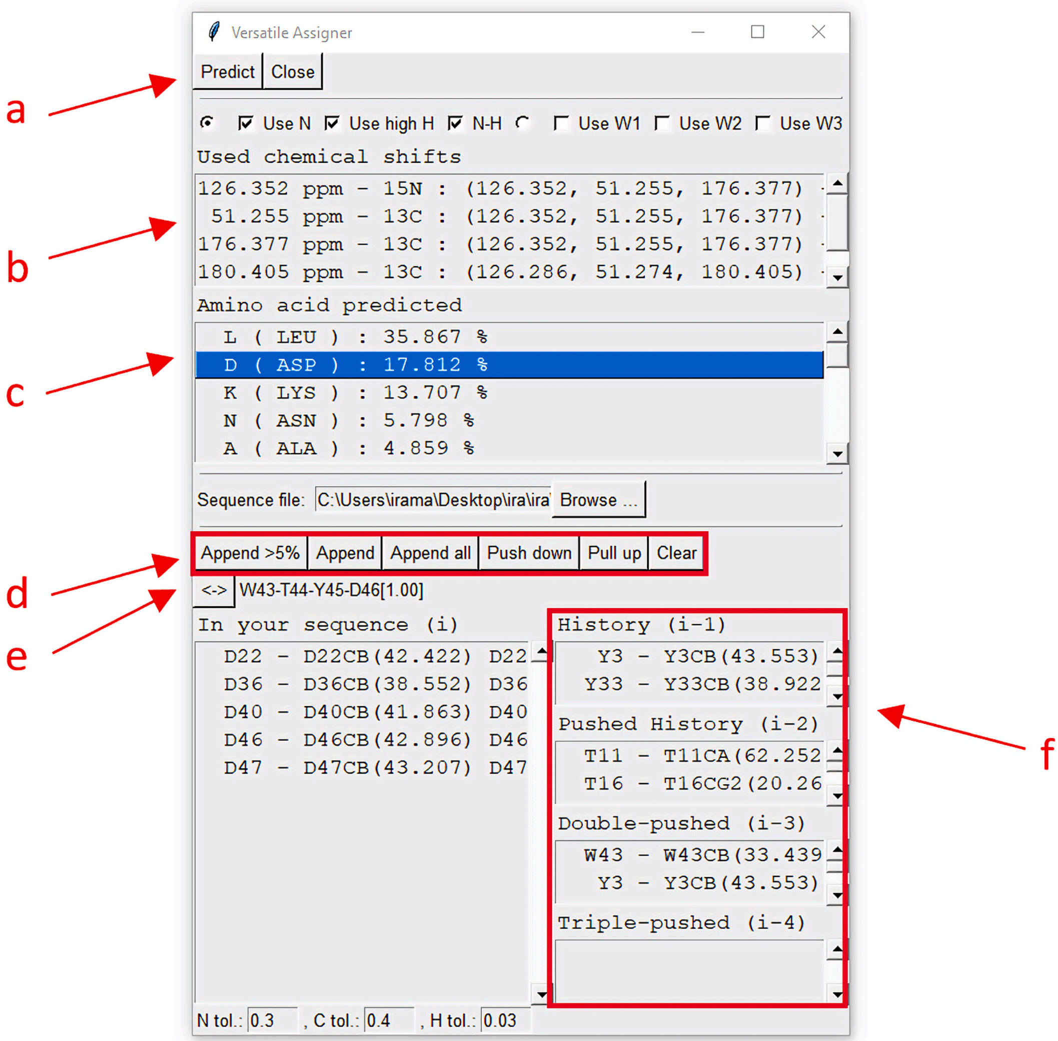
To determine the optimal strategy incorporated into the standardized workflows described here, we evaluated the performance of *Versatile Assigner* in analyzing solution and ss NMR data from a variety of proteins. The goal of this analysis was to determine which approaches led to improved assignment accuracy. The variables included walking direction and types of spectra. To evaluate its improvement, results from *Versatile Assigner* were compared with those from *PLUQin*.

Threshold contour levels were first set near the noise level by sight, checked in *POKY*, and then set directly in *iPick*. If *iPick* failed to complete its calculations within a reasonable period (e.g., 30 min), the process was stopped, and a new calculation was started with a raised threshold. *iPick* calculations were carried out repeatedly with higher or lower thresholds to achieve the expected number of peaks. The optimal input for *Versatile Assigner* was evaluated by choosing all or partial information from the available experiment types (CBCA(CO)NH and HNCACB for solution NMR; NCOCX and NCACX for solid state NMR). The prediction rank and percentage correctness for each residue were tabulated as a function of the input data used and whether the walk was forwards or backwards. The documented workflows take users through novel strategies of assignment. They provide a comprehensive guide for inexperienced users and suggest optimal pathways for experienced users. These workflows, which incorporate many of the new tools created, enable rapid and accurate spectral analysis.

4. Results

4.1. Analysis of Versatile Assigner

Versatile Assigner was tested with solution state CBCA(CO)NH and HNCACB data from the protein ubiquitin with and without N—H matching (having N and H dimension assignments agree in residue assignment). It was assumed that any residue assignment predicted to be



◉ ☑ Use N ☑ Use high H ☑ N-H ○ ☐ Use W1 ☐ Use W2 ☐ Use W3

```
Used chemical shifts
126.352 ppm - 15N : (126.352, 51.255, 176.377)
 51.255 ppm - 13C : (126.352, 51.255, 176.377)
176.377 ppm - 13C : (126.352, 51.255, 176.377)
180.405 ppm - 13C : (126.286, 51.274, 180.405)
Amino acid predicted
  L ( LEU ) : 35.867 %
  D ( ASP ) : 17.812 %
  K ( LYS ) : 13.707 %
  N ( ASN ) : 5.798 %
  A ( ALA ) : 4.859 %
```

Sequence file: C:\Users\irama\Desktop\ira\ira

W43-T44-Y45-D46[1.00]

```
In your sequence (i)            History (i-1)
  D22 - D22CB(42.422) D22         Y3 - Y3CB(43.553)
  D36 - D36CB(38.552) D36        Y33 - Y33CB(38.922
  D40 - D40CB(41.863) D40       Pushed History (i-2)
  D46 - D46CB(42.896) D46        T11 - T11CA(62.252
  D47 - D47CB(43.207) D47        T16 - T16CG2(20.26
                                Double-pushed (i-3)
                                 W43 - W43CB(33.439
                                  Y3 - Y3CB(43.553)
                                Triple-pushed (i-4)
```

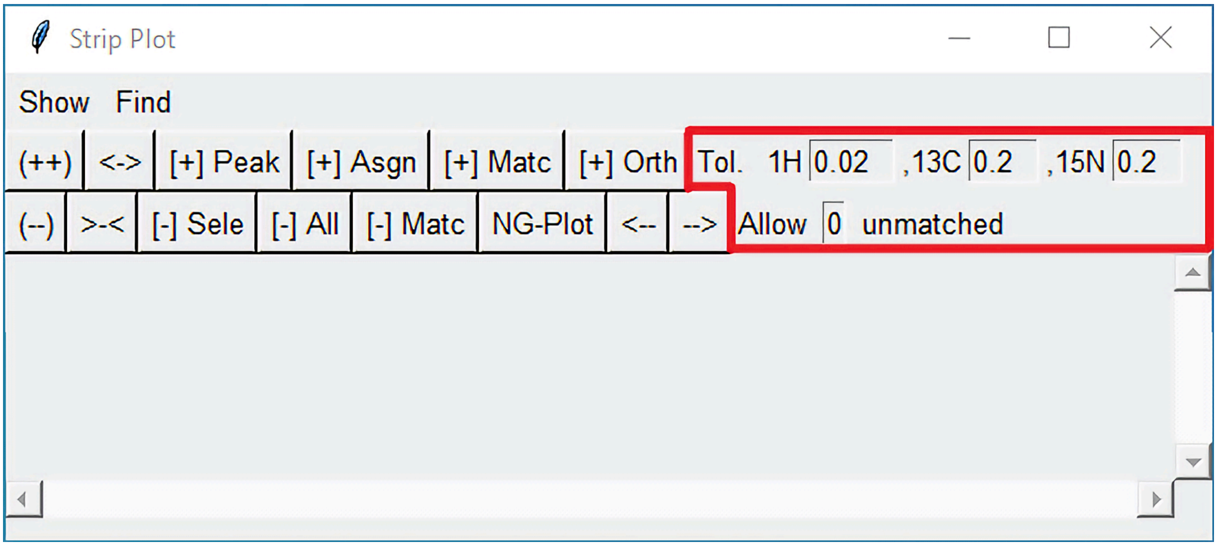
N tol.: 0.3 , C tol.: 0.4 , H tol.: 0.03

Fig. 5. *Versatile Assigner* is used to predict residue assignments, to validate assignments, and to designate sub-sequences. The power of *Versatile Assigner* comes from its diversity of features: The user can control which resonance information to use in predictions **(a)**. The user can view the chemical shift information **(b)** used to predict the amino acid types and their probabilities. The append button **(d)** adds the chosen residue (Asp46) to the growing subsequence. The "<->" button **(e)** specifies the direction in which the protein sequence is to be read. Forward, backward, and both directions are available. Once proteins have been entered into the history section, possible sub-sequences of the protein and the probability of each sub-sequence being the correct match for the residues investigated are shown here. This section does not give information on the probability that the selected residues assignments are correct, it only displays if a sub-sequence in the protein matches the assignments already chosen based on probability calculations. The history section **(f)** shows residue-specific assignment predictions discarded (pushed) by the user. The sections show residue positions relative to that of the residue currently under consideration (1). Up to five residues at a time can be investigated for sub-sequence matches.

above 5 % (the approximate value if all were equally likely) should be considered as possible until tested by further information. The tool predicted correct assignments at levels above 5 % ~97 % of the time, dropping slightly to ~ 94 % when using N—H matched information in HNCACB. The chosen walking direction had no significant impact on assignment prediction.

The goals for ss NMR included those for solution NMR (determining which walking direction and which spectral data are optimal) but also included determining the average number of residues needed for a sub-sequence hit. From the MAS-ss NMR data for the protein GB1 (Frericks-Schmidt et al. 2007), *Versatile Assigner* predicted correct assignments at levels above 5 % ~96 % of the time when using NCACX data both with and without N—H matching; the level dropped slightly to ~ 95 % when NCOCX data were used. While both NCACX and NCOCX analyses gave correct assignments at high levels, it is worth mentioning that there was an average increase in prediction values of 9 % when N—H matched information was not used. This does not necessarily guarantee that not using N—H matching is optimal, but it is something to consider when given the option to choose. Walking direction showed no significant effect on prediction values, both directions giving an average of three residues needed for a sub-sequence hit.

Versatile Assigner performed better than *PLUQin* in assigning ubiquitin data. The tools rank predictions in order of certainty, given their parameters. *Versatile Assigner* gave correct assignments as the first ranked prediction ~ 66.66 % of the time, whereas *PLUQin* did so ~ 20.83 % of the time. The accuracy of *Versatile Assigner* increased markedly when looking at correct assignments ranked in the top one (~78.79 %), top two (~87.50 %), top three (~91.67 %) and top four (100 %) results, whereas *PLUQin* scored far lower prediction accuracy in each of the top one (~27.23 %), top two (~48.49 %), top three (54.55 %), and top four (~66.67 %) results. The only advantage *PLUQin* has over *Versatile Assigner* is its ability to derive secondary structural information and to analyze many peaks at once. See Supplementary Materials for detailed results for *Versatile Assigner* and *PLUQin*.



Tol. 1H 0.02 ,13C 0.2 ,15N 0.2

Allow 0 unmatched

Fig. 6. *Strip Plot* allows users to quickly access and work with multiple areas of a spectrum. Up to three *Strip Plot* windows can be viewed simultaneously. The tool assists with linking, validating, and assigning peaks. The highlighted area shows a feature of *Strip Plot* that enables the adjustment of chemical shift tolerances for matching ^{1}H, ^{13}C, and ^{15}N signals in different spectra and the specification of the number of peaks may remain unmatched.

4.2. Workflows

Many of the steps in solution NMR spectroscopy of proteins aspects are (semi-)automated. Users can start the workflow in *POKY* (Fig. 7), beginning with the *APES* program (two-letter code *ae*) for automated peak picking. *APES* shows the user a variety of signals in the spectrum and asks the user to assign them as peaks or noise. Once this process is complete, users should check their peak lists (two-letter code *lt*) and manually verify that there are no noise peaks. To do so, users can check peaks by data height until they find the largest peak that they consider noise and delete all peaks with lower intensity. This produces a highly accurate set of picked peaks, which can be further validated through assignment. *Strip Plot* can now be used to move through the spectra in order to check for noise peaks or missing peaks. Users can use *Strip Plot* on all of their spectra, and the easily accessible tolerance settings makes connecting strips easy. *Strip Plot* can help users visualize sections of the same spectrum, and corresponding sections of other spectra, at the same time. The ability to *Save/Load* strip plots further reduces the difficulty of

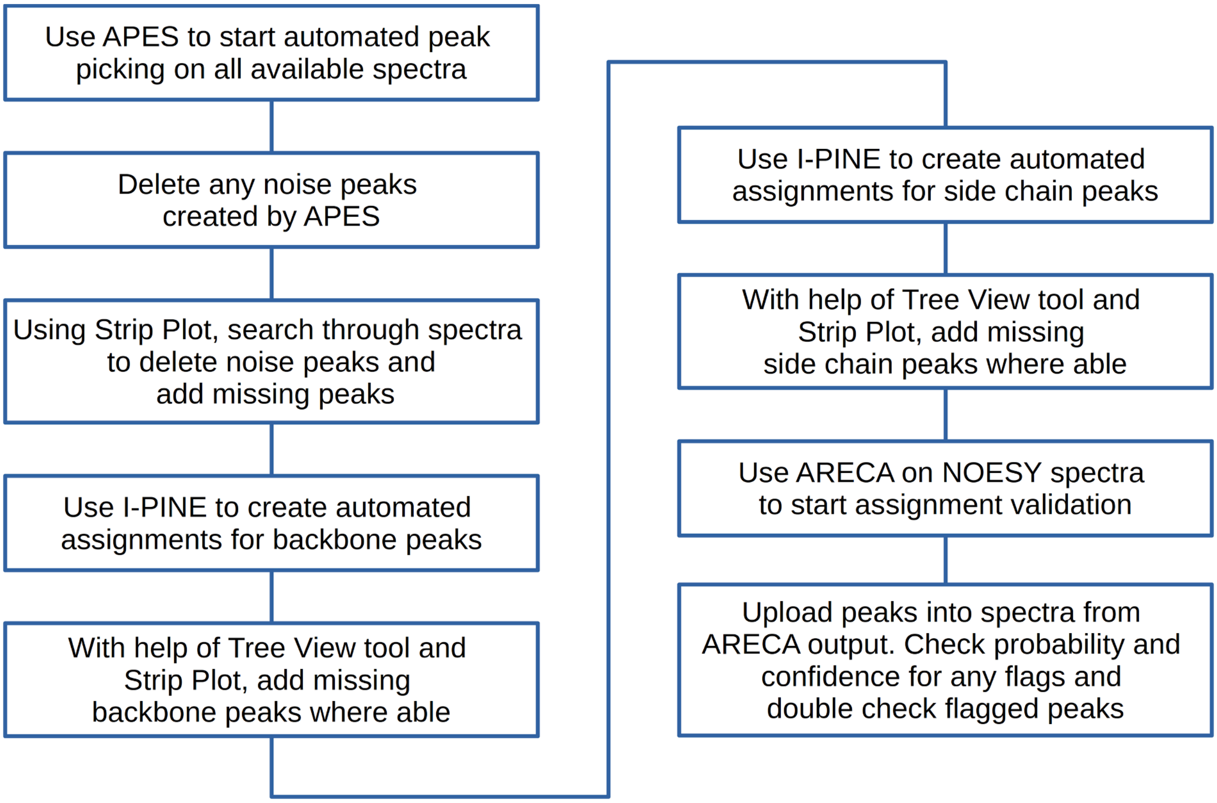

Use APES to start automated peak picking on all available spectra
Delete any noise peaks created by APES
Using Strip Plot, search through spectra to delete noise peaks and add missing peaks
Use I-PINE to create automated assignments for backbone peaks
With help of Tree View tool and Strip Plot, add missing backbone peaks where able
Use I-PINE to create automated assignments for side chain peaks
With help of Tree View tool and Strip Plot, add missing side chain peaks where able
Use ARECA on NOESY spectra to start assignment validation
Upload peaks into spectra from ARECA output. Check probability and confidence for any flags and double check flagged peaks

Fig. 7. Workflow for the assignment of solution state NMR and ^{1}H-detected solid-state NMR spectra of proteins with tools from the *POKY* suite. The workflow guides inexperienced users through the NMR assignment process and offers experienced spectroscopists an optimized path for faster NMR assignments.

working with many spectra at once.

We suggest that users set tolerances for solution NMR peaks at around 0.2 ppm for carbon and nitrogen dimensions, and 0.02 ppm for hydrogen dimensions. These values are commonly large enough to find peaks that should be linked, but not too large to include incorrect links. These tolerance values are the default ones in *POKY*. Once all peaks have been identified, *I-PINE* is the automation algorithm used to determine assignments and yield secondary structural information. The *I-PINE* webserver can be accessed using the *PINE-SPARKY.2* plugin (two-letter-code *ep*) in *POKY*. Peaks created by *I-PINE* will be color-coded, allowing users to easily see which peaks may need verification before accepting. I-PINE can be used for both backbone and sidechain assignments, making solution NMR spectra nearly entirely automated. After the assignment of each section (backbone and sidechain) users should use the *Tree View* tool (two-letter code *tv*) to check that all assignments have been made. If the user finds atoms with unassigned signals, *Strip Plot* can be used to create a peak and/or an assignment for it. Once users have created all of their desired assignments, *ARECA* (Assessment of the REliability of Chemical shift Assignments) (Dashti et al., 2016) can be used to cross-validate peaks with NOESY spectra. To do this, *ARECA* uses a truth model based on expected probabilities of NOESY contacts between intra- and inter-residue protons. By referencing these expected probabilities, *ARECA* can validate whether or not the experimental NOESY peak agrees with the assignment. Once validation is complete, the solution NMR spectra should be fully assigned. While the process is almost completely automated, it is important that users verify that individual tasks are completed accurately.

^{1}H-detected ss NMR gives a peak pattern similar to those of solution NMR spectra. Recently, the Veglia group demonstrated the ability of I-PINE to assign all backbone chemical shifts from GB1 ^{1}H-detected ss NMR data (Gopinath et al. 2022). The ^{1}H-detected ss NMR experiments therefore can use the solution NMR workflow (Fig. 7).

The analysis of other types of ss NMR data from proteins requires longer workflows incorporating manual steps (Fig. 8) and may require additional analysis through supplemental workflows (Fig. 9). We suggest that *iPick* (two-letter code *iP*) be used for peak picking ss NMR spectra. The *iPick* GUI in *POKY* comes with two modes: *Basic* and *Advanced*. The *Basic* mode simplifies the peak picking process to a click of one button. The *Advanced* mode gives fine-tuning options for ultimate control of the process. The *Reliability Score* can guide the user to distinguish strong reliable peaks from weak peaks or noise peaks. Another peak picking tool, *Restricted Peak Picking* (two-letter-code *kr*), can used in areas of low peak density, but *iPick* is highly recommended for 3D spectra.

The “*Find/Add peak*” cursor mode can yield accurately picked peaks on any 2D plane. This is useful for 2D spectra but becomes exceedingly tedious in a 3D spectrum as users need to use the cursor on each individual 2D slice of the spectrum. To circumvent this problem, the new tool *Transfer/Assign Peaks* enables the compression of peak-picked 3D spectra peaks to 2D. With this, users can compress a 3D NCOCX/NCACX spectrum into a 2D NCO/NCA spectrum. Users must remember to specify in the settings of *Strip Plot* which spectrum was called up. The *Save/Load* function of *Strip Plot* is particularly useful when analyzing complicated ss NMR spectra. If users have a C—C correlation spectrum collected with a short mixing time, by focusing on intra-atom correlations they can use it to determine whether a peak occurs at the same location as a peak missing in their NCOCX/NCACX spectrum. The peaks present in the C—C spectrum but not in the main spectrum can be transferred in. If only a few peaks need to be transferred, it is faster to use the POKY option (copy and paste peaks and/or assignments) than

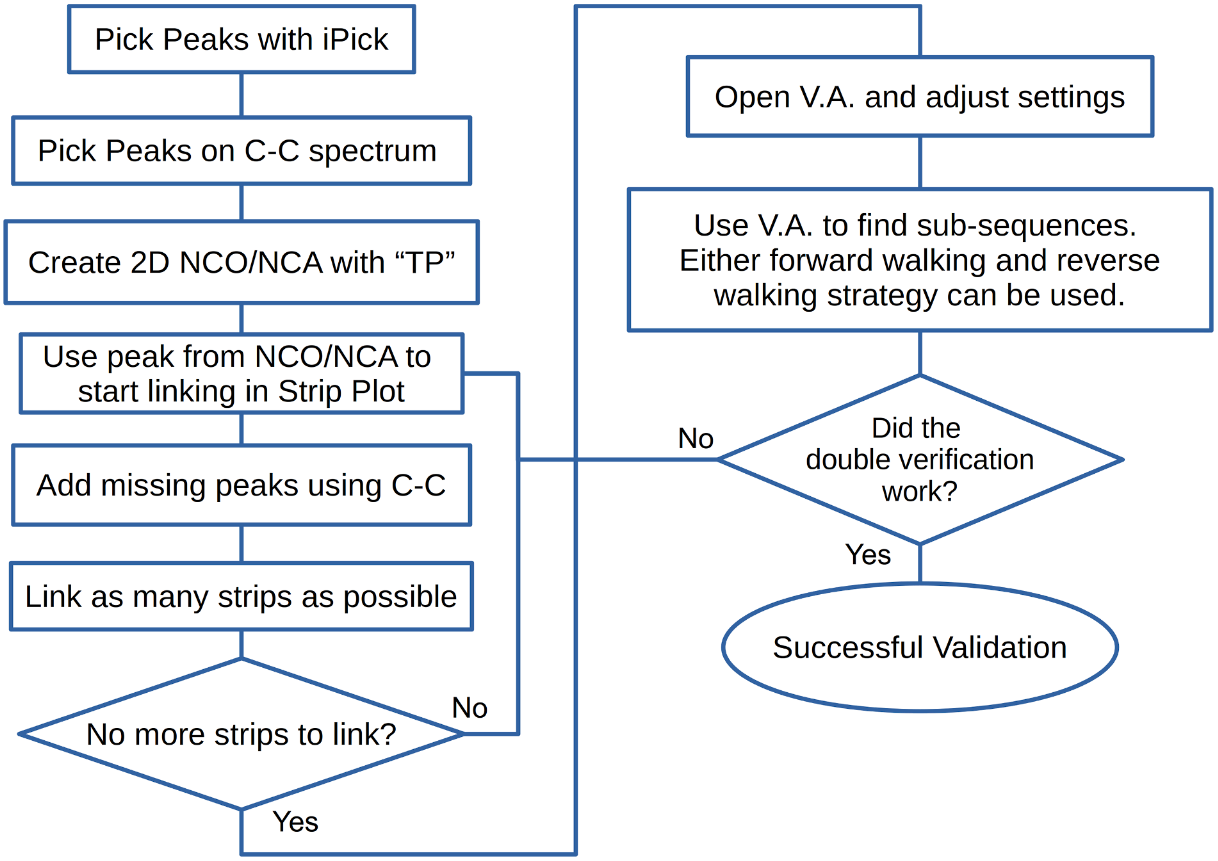

Pick Peaks with iPick
Pick Peaks on C-C spectrum
Create 2D NCO/NCA with “TP”
Use peak from NCO/NCA to start linking in Strip Plot
Add missing peaks using C-C
Link as many strips as possible
No more strips to link?
No
Yes
Open V.A. and adjust settings
Use V.A. to find sub-sequences. Either forward walking and reverse walking strategy can be used.
Did the double verification work?
No
Yes
Successful Validation

Fig. 8. Basic workflow for the analysis of ss NMR spectra of proteins. It walks users through the generation of peaks and linking of strips in *Strip Plot*, as well as validation and assignment. Additional steps may be required for peak validation and sub-sequencing; these are shown in Fig. 9. Abbreviation: V.A., *Versatile Assigner*; “TP”, two-letter-code for *Transfer/Assign Peaks* tool.

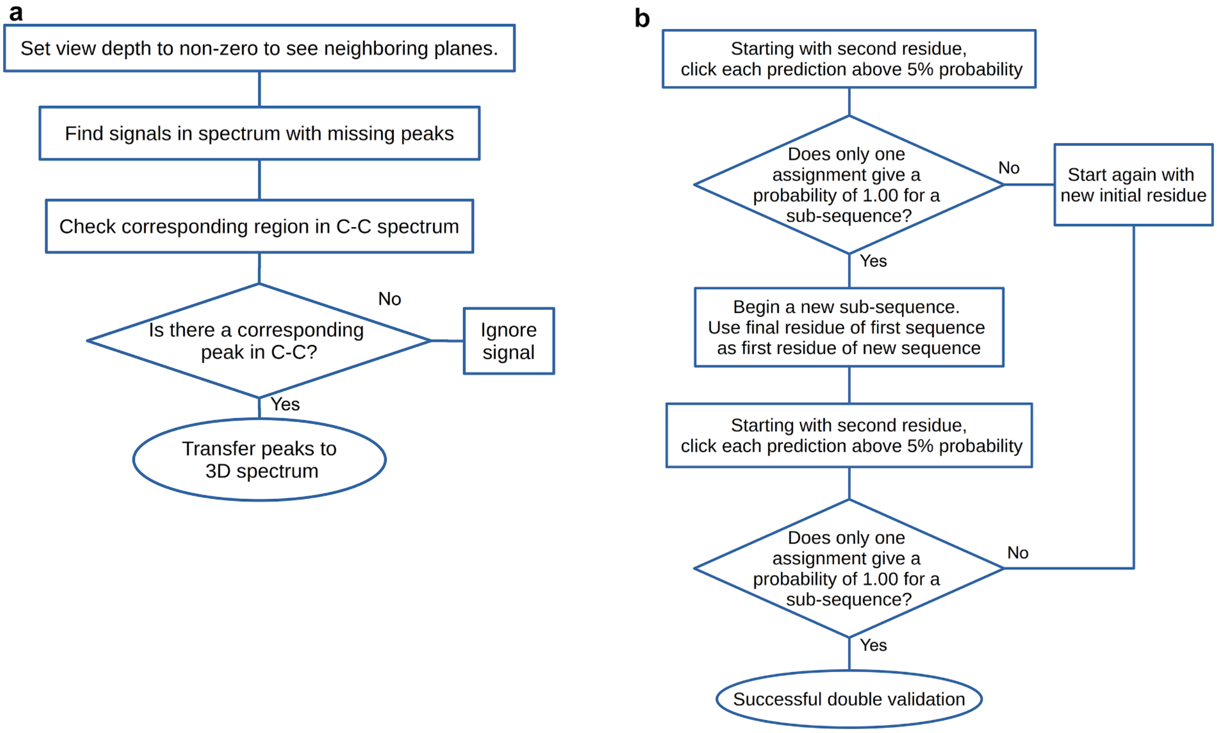

a
Set view depth to non-zero to see neighboring planes.
Find signals in spectrum with missing peaks
Check corresponding region in C-C spectrum
Is there a corresponding peak in C-C?
No
Ignore signal
Yes
Transfer peaks to 3D spectrum
b
Starting with second residue, click each prediction above 5% probability
Does only one assignment give a probability of 1.00 for a sub-sequence?
No
Start again with new initial residue
Yes
Begin a new sub-sequence. Use final residue of first sequence as first residue of new sequence
Starting with second residue, click each prediction above 5% probability
Does only one assignment give a probability of 1.00 for a sub-sequence?
No
Yes
Successful double validation

Fig. 9. Optional workflows for solid state NMR. These workflows may be needed only occasionally. (*Left*) Workflow used in checking and validating missing peaks with information from a C—C spectrum. The method is used to determine whether signals not assigned as a peak by *iPick* are valid peaks. (*Right*) Double validation by the Linking Method. This is an important step for validation of a sub-sequence as it reduces error significantly.

the *Transfer/Assign Peaks* module. We suggest using *Versatile Assigner* in the final assignment process, because it enables the validation of residue assignments. *Strip Plot* makes it easy to move through spectra one residue at a time. Users can obtain the probability of an assignment by selecting the predicted carbon peaks for a residue from the NCOCX/NCACX spectrum and clicking the “predict” button. While all carbons of a single residue can be used, we suggest that CA and CB be selected, because the addition of other carbons improves the prediction only marginally whereas misassignment of carbon signals can invalidate the residue prediction. *Versatile Assigner* can predict the assignment probability of up to five residues at once. By using these limited assignment options, one can narrow down the possible sequential assignment probabilities. Users can be confident of their assignments when only one possibility remains. The Double Validation method (Fig. 9 *Right*) can be used to further validate the results. To do this, users can validate a series of consecutive sub-sequence, each using the last residue of the previous sub-sequence as its first residue. If each sub-sequence is validated to be in agreement with the protein sequence, then the user can be highly confident in the validity of the entire assignment.

The output data can be made available in NMR-STAR 3.1/3.2 format (two-letter-code *es* or *bg*) ready for deposition in the BMRB archive.

5. Discussion

Solid-State NMR data are often difficult for users to work with owing to large line widths and low signal-to-noise ratios. These issues with data collection cause the loss of many peaks because they lie under the noise or because they are hidden under or merged with other peaks. The aim of computational analysis is to overcome these problems. The recommended approach is to open the *Reference Views* to the region of the CA and CB signals and to start classifying them by residue type as completely as possible. These results enable *Versatile Assigne*r to carry out sub-sequencing far faster than if all carbon signals were uncharacterized. Additionally, ss NMR spectra contain peaks from intra-residue interactions. These peaks should be ignored during assignment analysis; therefore, building assignments from CA and CB signals of individual residues is a good strategy.

Reference Views, *Versatile Assigner*, *LACS*, and the host of supporting tools in *POKY* offer users a wide range of options in spectral analysis. *Versatile Assigner*, which is shown here to yield more accurate and complete assignments than its predecessor (*PLUQin*) offers the advantages of greater speed because it does not need to connect to a webserver. The workflows proposed here are designed to support both inexperienced and experienced users. They provide step-by-step protocol for novices or suggestions for veteran spectroscopists who have their own workflow.

Funding

This work was supported by the National Science Foundation [grant number DBI-2051595 to W.L.; old number DBI-1902076], the University of Colorado Denver [speedtype number 61193205 to W.L.], the National Institute of Health [grant number P41GM103399 to J.L.M.], and the University of Wisconsin-Madison (URS Scholars Program to I.M.).

CRediT authorship contribution statement

Ira Manthey: Methodology, Validation, Writing – original draft, Writing – review & editing. **Marco Tonelli:** Methodology, Validation, Writing – review & editing. **Lawrence Clos II:** Methodology, Writing – review & editing. **Mehdi Rahimi:** Software, Writing – review & editing. **John L. Markley:** Writing – review & editing, Funding acquisition. **Woonghee Lee:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Authors thank Prof. Tatyana Polenova (U. Delaware) for providing ssNMR data for the microcrystalline agglutinin from *Oscillatoria agardhii* (OAA) and Prof. Chad Rienstra (U. Wisconsin) for the GB1 ssNMR data sets. Because the OAA assignments were incomplete during our active development, we used the data only to test and improve our ssNMR workflow.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yjsbx.2022.100073.

References

Berman, H., Henrick, K., Nakamura, H., Markley, J.L., 2007. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. 35, D301-303. https://doi.org/10.1093/nar/gkl971.

Dashti, H., Tonelli, M., Lee, W., Westler, W.M., Cornilescu, G., Ulrich, E.L., Markley, J.L., 2016. Probabilistic validation of protein NMR chemical shift assignments. J. Biomol. NMR 64, 17–25. https://doi.org/10.1007/s10858-015-0007-8.

Fritzsching, K.J., Yang, Y., Schmidt-Rohr, K., Hong, M., 2013. Practical use of chemical shift databases for protein solid-state NMR: 2D chemical shift maps and amino-acid assignment with secondary-structure information. J. Biomol. NMR 56 (2), 155–167.

Fritzsching, K.J., Hong, M., Schmidt-Rohr, K., 2016. Conformationally selective multidimensional chemical shift ranges in proteins from a PACSY database purged using intrinsic quality criteria. J Biomol NMR 64, 115–130. https://doi.org/10.1007/s10858-016-0013-5.

Hu, K.-N., Qiang, W., Tycko, R., 2011. A general Monte Carlo/simulated annealing algorithm for resonance assignment in NMR of uniformly labeled biopolymers. J Biomol NMR 50, 267–276. https://doi.org/10.1007/s10858-011-9517-1.

Lee, W., Markley, J.L., 2018. PINE-SPARKY.2 for automated NMR-based protein structure research. Bioinformatics 34, 1586–1588. https://doi.org/10.1093/bioinformatics/btx785.

Lee, W., Rahimi, M., Lee, Y., Chiu, A., 2021. POKY: a software suite for multidimensional NMR and 3D structure calculation of biomolecules. Bioinformatics 37, 3041–3042. https://doi.org/10.1093/bioinformatics/btab180.

Lee, W., Westler, W.M., Bahrami, A., Eghbalnia, H.R., Markley, J.L., 2009. PINE-SPARKY: graphical interface for evaluating automated probabilistic peak assignments in protein NMR spectroscopy. Bioinformatics 25, 2085–2087. https://doi.org/10.1093/bioinformatics/btp345.

Lee, W., Yu, W., Kim, S., Chang, I., Lee, W., Markley, J.L., 2012. PACSY, a relational database management system for protein structure and chemical shift analysis. J. Biomol. NMR 54 (2), 169–179.

Lee, W., Stark, J.L., Markley, J.L., 2014. PONDEROSA-C/S: client–server based software package for automated protein 3D structure determination. J. Biomol. NMR 60 (2-3), 73–75.

Lee, W., Tonelli, M., Markley, J.L., 2015. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics 31, 1325–1327. https://doi.org/10.1093/bioinformatics/btu830.

Lee, W., Cornilescu, G., Dashti, H., Eghbalnia, H.R., Tonelli, M., Westler, W.M., Butcher, S.E., Henzler-Wildman, K.A., Markley, J.L., 2016a. Integrative NMR for biomolecular research. J Biomol NMR 64, 307–332. https://doi.org/10.1007/s10858-016-0029-x.

Lee, W., Petit, C.M., Cornilescu, G., Stark, J.L., Markley, J.L., 2016b. The AUDANA algorithm for automated protein 3D structure determination from NMR NOE data. J. Biomol. NMR 65, 51–57. https://doi.org/10.1007/s10858-016-0036-y.

Lee, W., Bahrami, A., Dashti, H.T., Eghbalnia, H.R., Tonelli, M., Westler, W.M., Markley, J.L., 2019. I-PINE web server: an integrative probabilistic NMR assignment system for proteins. J. Biomol. NMR 73, 213–222. https://doi.org/10.1007/s10858-019-00255-3.

Marassi, F.M., Opella, S.J., 2000. A solid-state NMR index of helical membrane protein structure and topology. J. Magn. Reson. 144, 150–155. https://doi.org/10.1006/jmre.2000.2035.

Markley, J.L., Ulrich, E.L., Berman, H.M., Henrick, K., Nakamura, H., 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. https://doi.org/10.1007/s10858-008-9221-y.

Moseley, H.N.B., Sperling, L.J., Rienstra, C.M., 2010. Automated protein resonance assignments of magic angle spinning solid-state NMR spectra of β1 immunoglobulin binding domain of protein G (GB1). J. Biomol. NMR 48, 123–128. https://doi.org/10.1007/s10858-010-9448-2.

Opella, S.J., Marassi, F.M., 2004. Structure determination of membrane proteins by NMR spectroscopy. Chem. Rev. 104, 3587–3606. https://doi.org/10.1021/cr0304121.

Polenova, T., Gupta, R., Goldbourt, A., 2015. Magic angle spinning NMR spectroscopy: A versatile technique for structural and dynamic analysis of solid-phase systems. Anal. Chem. 87, 5458–5469. https://doi.org/10.1021/ac504288u.

Rahimi, M., Lee, Y., Markley, J.L., Lee, W., 2021. iPick: Multiprocessing software for integrated NMR signal detection and validation. J. Magn. Reson. 328, 106995 https://doi.org/10.1016/j.jmr.2021.106995.

Schmidt, E., Gath, J., Habenstein, B., Ravotti, F., Székely, K., Huber, M., Buchner, L., Böckmann, A., Meier, B.H., Güntert, P., 2013. Automated solid-state NMR resonance assignment of protein microcrystals and amyloids. J Biomol NMR 56, 243–254. https://doi.org/10.1007/s10858-013-9742-x.

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J.M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K.K., Lemak, A., Ignatchenko, A., Arrowsmith, C.H., Szyperski, T., Montelione, G.T., Baker, D., Bax, A., 2008. Consistent blind protein structure generation from NMR chemical shift data. Proc. Natl. Acad. Sci. U.S.A. 105, 4685–4690. https://doi.org/10.1073/pnas.0800256105.

Shin, J., Lee, W., Lee, W., 2008. Structural proteomics by NMR spectroscopy. Expert Review of Proteomics 5, 589–601. https://doi.org/10.1586/14789450.5.4.589.

Wang, J., Denny, J., Tian, C., Kim, S., Mo, Y., Kovacs, F., Song, Z., Nishimura, K., Gan, Z., Fu, R., Quine, J.R., Cross, T.A., 2000. Imaging membrane protein helical wheels. J. Magn. Reson. 144, 162–167. https://doi.org/10.1006/jmre.2000.2037.

Wang, L., Eghbalnia, H., Bahrami, A., Markley, J.L., 2005. Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. J. Biomol. NMR 32, 13–22. https://doi.org/10.1007/s10858-005-1717-0.

Weber, D.K., Wang, S., Markley, J.L., Veglia, G., Lee, W., 2020. PISA-SPARKY: an interactive SPARKY plugin to analyze oriented solid-state NMR spectra of helical membrane proteins. Bioinformatics 36, 2915–2916. https://doi.org/10.1093/bioinformatics/btaa019.