A Toolset for the Solid-State NMR-based 3D Structure Calculation of Proteins

Mehdi Rahimi[1], Yeongjoon Lee[1], Huong Nguyen[2,3], Abigail Chiu[1], and Woonghee Lee[1,] *

[1] Department of Chemistry, University of Colorado Denver, Denver, CO 80204, USA; mehdi.rahimi@ucdenver.edu, bioyj1012@gmail.com, abigail.chiu@ucdenver.edu, woonghee.lee@ucdenver.edu

[2] Computer Science Department, University of Wisconsin-Madison, Madison, WI 53706, USA

[3] URS Program, University of Wisconsin-Madison, Madison, WI 53706, USA; htnguyen23@wisc.edu

* Correspondence: woonghee.lee@ucdenver.edu

Abstract: Proteins are the building blocks of life. The shape of the protein determines its functionality. This understanding of the 3D structure of proteins has applications in study of diseases, medicine, body functions, and other aspects of life. Nuclear magnetic resonance (NMR) has been a powerful tool for researchers to get insight into the metabolome of cells, tissues, biofluids, secretions, and overall etiology of the disease state. Solid-state NMR (ssNMR) spectroscopy is used for samples that have low solubility in common NMR solvents. The use of ssNMR for 3D structure determination of proteins has been on the rise in the recent years especially for such samples. Still, one of the challenges that researchers face in this area is a shortage of easy and user-friendly computational aids. To address this, we are introducing our comprehensive software solution by automating every step of the process and essentially transforming the task into a few clicks of the mouse. The workflow for 3D structure determination has been simplified down to only a few procedures. Starting with selection of an ssNMR spectrum, user can receive its 3D structure along with an abundance of statistical information and validation tools using our software. We have tested this toolset to test the usefulness and user-friendliness with different data sets available on biological magnetic resonance bank (BMRB).

Availability: All the codes are freely available on our website: https://poky.clas.ucdenver.edu

Keywords: Solid-state NMR; Structure calculation; POKY; SPARKY; Automation; PONDEROSA-C/S; AUDASA;

1. Introduction

Functions and activities of biomolecules are known to be related to their structural properties. By understanding their structures, it allows for more innovative biomedical advances which have enriched human life. Some common applications from understanding the 3D structure of proteins include, but not limited to, the study of diseases, medicine, and body functions [1]. Mostly, the three-dimensional (3D) structures of proteins have been determined by one of these major techniques: Nuclear magnetic resonance (NMR) spectroscopy, X-ray crystallography, and cryogenic electron microscopy. Among these methods, solid-state NMR (ssNMR) spectroscopy is becoming a prominent approach since it can be used to discover the structure of both soluble and insoluble proteins with less size limitations. Moreover, the technique does not require crystallization which is often difficult or impossible, and is even applicable to proteins with large disordered regions [2, 3]. Not to mention, NMR has been used by researchers to investigate metabolome of cells, tissues, biofluids, secretions, and overall etiology of the disease state [4, 5].

Despite these advantages, there is still a limited selection of software tools for analyzing protein ssNMR spectra. To maximize the usability of ssNMR experiments, we introduce a comprehensive and easy-to-use computational aid for 3D structure determination using ssNMR spectra. We previously introduced the PONDEROSA-C/S (Peak-picking of Noe Data Enabled by Restriction of Shift Assignments-Client/Server) software package [6] for the structural determination of proteins using solution NMR data, which over the years has demonstrated outstanding performance and acceptance in the NMR community. Here, we extend the PONDEROSA-C/S package to offer the same straightforward approach for the ssNMR. This extension includes an updated graphical toolset and a new algorithm named AUDASA, based on our solution NMR algorithm, AUDANA [7]. Similar to the AUDANA, our new algorithm generates distance constraints to automate the assignment of ssNMR spectra, including but not limited to 2D-CC, 2D-CN, 2D-NN, 2D-CHHC, 3D-NCOCX, and 3D-NCACX. In fact, our tools support any set of experiments providing same experimental profile (e.g. C-C, C-N, N-N, H-H) with given examples. Then, iteratively it conducts simulated annealing to acquire the lowest energy structures followed by water refinement steps. AUDASA has been implemented into the PONDEROSA Server program and can be

used by the PONDEROSA plugins in NMRFAM-SPARKY (two-letter-code *c3*) [8], and its successor, the POKY software suite (two-letter-code *c3*) [9]. In the POKY suite, the Poky Structure Builder accessible by two-letter-code *Pb* provides more advanced features. The two-letter codes are shortcuts that the user can type into the program to access the modules faster. They are also accessible from the menu buttons.

2. Methods and Materials

We developed an Integrative NMR ecosystem [10] for 3D structure calculation of the proteins in ssNMR, which is analogous to the solution NMR version we developed previously. This system, consisting of multiple standalone software and plugins, along with calculation servers, provide an effortless approach for the user for structure calculation. A schematic diagram of the complete protocol is shown in Figure 1.

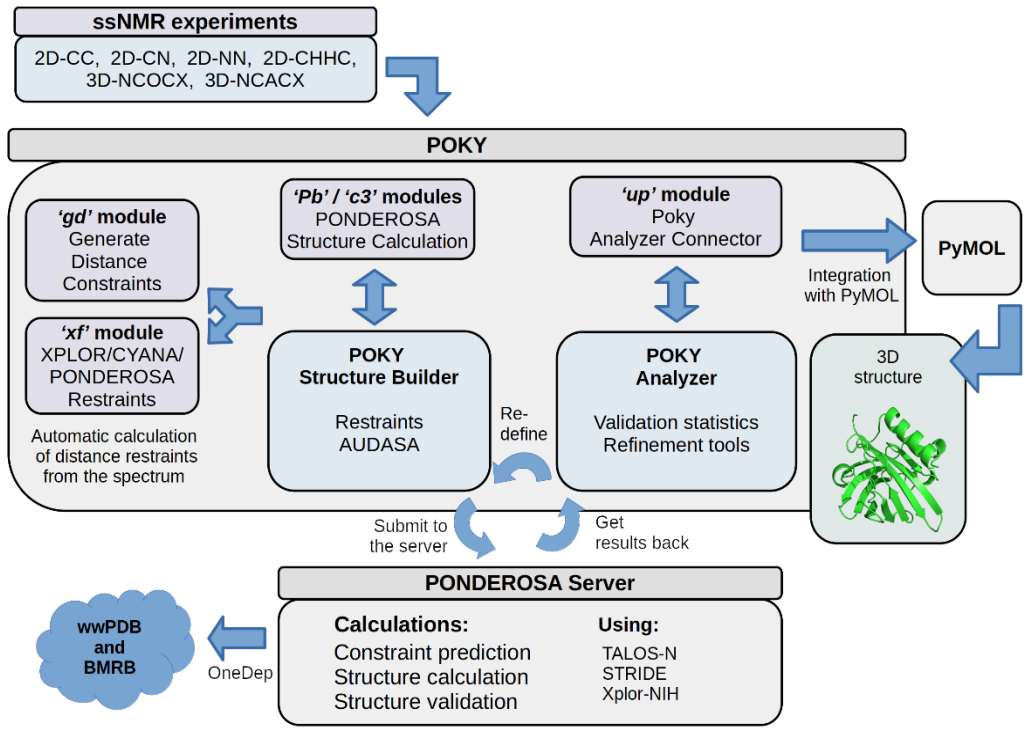

Figure 1. A schematic diagram of the system shows the steps that the user would take to obtain the 3D structure of a protein from ssNMR spectra.

2.1. POKY Suite

The POKY suite is a crucial software suite for multidimensional NMR spectroscopy and three-dimensional structure calculation of biomolecules using the acquired spectra. In the heart of our `ecosystem', multiple modules of the Poky suite facilitate the structure calculation process. One of these modules is the "Poky Structure Builder" as shown in Figure 1-a. This advanced module (two-letter-code *Pb*), along with the PONDEROSA Structure Calculation module (two-letter-code *c3*), provides the basic graphical user interface (GUI), prepares, and submits the data to the PONDEROSA Server. These modules pick peaks automatically by simply seeking local maxima above the contour threshold, however, the user can consider using the new *iPick* program (two-letter-code *iP*), which is also available in the POKY suite for further investigation on the peaks [11].

This preparation takes in the protein sequence, the chemical shift assignments, and the distance constraints or NMR thru-space data for the model. The next steps include a pre-processing, and finally, submission to the server by clicking the "Calculate" button. If the automated long-range cross peak assignment and restraint generation coupled with the structure calculation are required, the user can choose "AUDANA/AUDASA automation". The user can assign

non-sequential cross peaks in the thru-space NMR data and generate distance restraints using plugins described below. With sufficient number of restraints, the user can choose the "Constraints-only X" option to perform the simulated annealing with the given set of restraints without the automated cross peak assignments. Iterative restraint validation and calculation steps to generate violation-free converged structures are included in this option.

2.2. Distance Constraints

The distance constraints can be obtained independently from the structure calculation using two other developed modules: "Generate distance constraints" (two-letter-code *gd*) and "XPLOR/CYANA/PONDEROSA Restraints" (two-letter-code *xf*). The different file formats of XPLOR-NIH [12], CYANA [13], and PONDEROSA are supported. The "Generate distance constraints" can automatically calibrate upper limit distances for the distance restraints using peak assignments and intensities. Formulated on the selected approximation option (e.g. peak intensity is proportional to the distance to the power of minus three, r^{-3} in ssNMR [14]), the peak intensities are first used to generate temporary values that roughly correspond to each distance information. By default, interproton distances are calibrated between 2.4 Å and 5.5 Å, and intercarbon distances are calibrated between 4.4 Å and 7.7 Å. Then, the values are automatically calibrated based on the pre-set median distance. The user also has the option of modifying the upper limit (UPL) scaler percentage to avoid over-fitting to the restraints when folding the structure by the XPLOR-NIH program [15]. This plugin, along with the supported experiments, is shown in Figure 2-b.

The distance restraints can also be generated using another Poky plugin, "XPLOR/CYANA/PONDEROSA Restraints" (two-letter-code *xf*) as shown in Figure 2-c, or the Poky Structure Builder (two-letter-code *Pb*), which are updated with the support for ssNMR. These three plugins have different use cases. The "XPLOR/CYANA/PONDEROSA Restraints" plugin (two-letter-code *xf*) is a manual approach. The "Generate distance constraints" plugin (two-letter code: *gd*) is a semi-automated approach, and the Poky Structure Builder (two-letter-code *c3*) is a fully automated path to achieve this goal. The choice of the plugin is dependent on the data and workflow of the user.

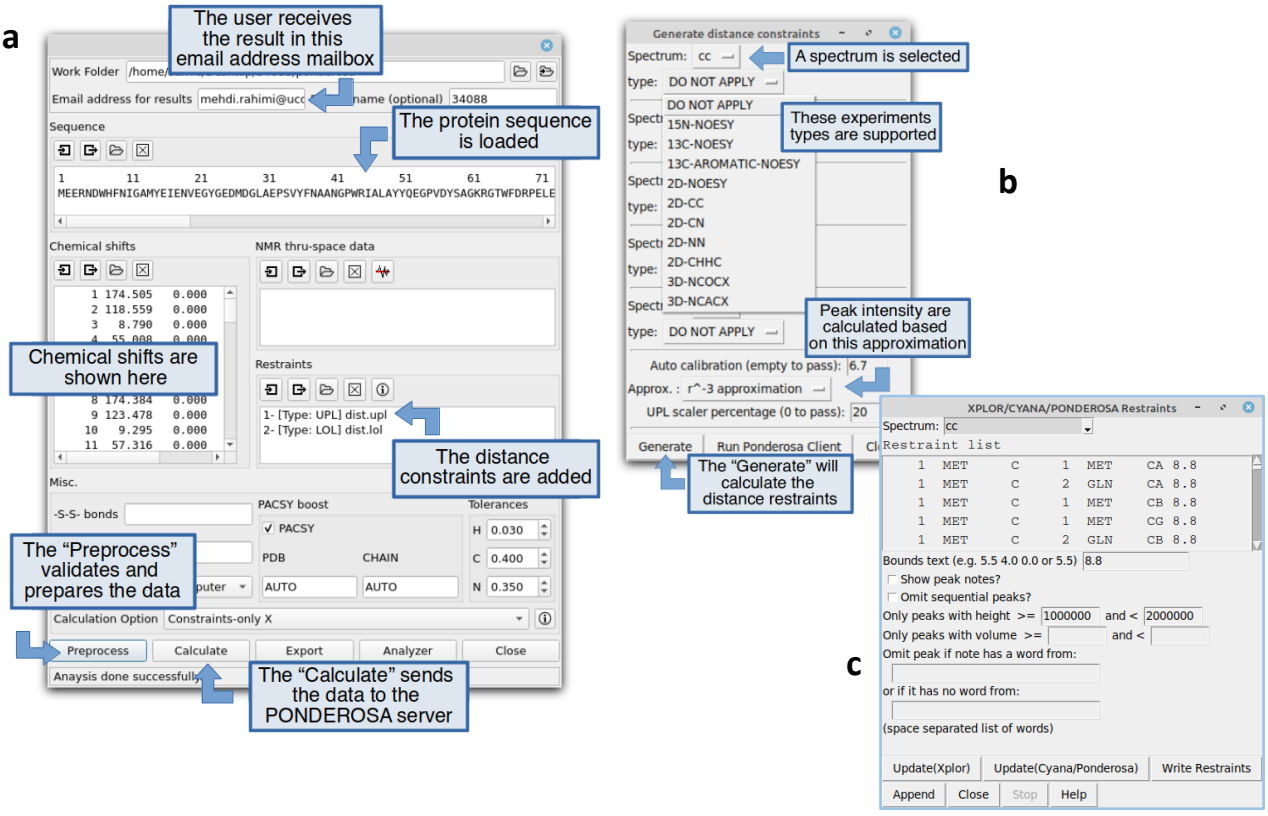

Figure 2. Two different programs to generate distance restraints. (a) The Poky Structure Builder (two-letter-code *Pb*). The distance restraints along with the protein sequence and the chemical shift assignment are loaded. The "Calculate" button sends the data to PONDEROSA Server for the automated 3D structure determination. (b) The "Generate distance constraints" plugin (two-letter-code *gd*) can be run from the POKY suite to automatically generate the distance restraints using assigned peaks. The list of supported experiments is shown in the drop-down menu. (c) The XPLOR/CYANA/PONDEROSA Restraints module is run by the two-letter-code *xf* and is updated to support the DIANA UPL format as well. This provides a more manual approach to the problem of generating the restraints.

2.3. Poky Structure Builder

The Poky Structure Builder sends the distance restraints along with the protein sequence and the chemical shift assignments to the PONDEROSA Server. A cluster of servers processes the user's submission for determining the distance and angle constraints, calculating the 3D structure, as well as estimating the quality of the structure. The PONDEROSA server uses multiple software including XPLOR-NIH [15], TALOS-N [16], and STRIDE [17] for calculation.

As mentioned above, the AUDASA algorithm performs automated assignment for distance restraints, automated generation of torsion angle restraints and automated structure calculation. AUDASA accepts ssNMR spectra acquired with long mixing time as inputs along with the protein sequence and the chemical shift assignments. Distance restraints are generated and used to calculate the 3D structures in the iterative and automated manner like its cousin from solution NMR, AUDANA. It still requires fine-tuning with the real ssNMR spectra. However, the functionality with the tunable knobs has been developed to improve at the server side, and it does not require any reinstallation of the program at the user side in the future.

2.4. Obtaining the Result and Evaluation

The result is emailed to the user, which can be viewed and processed in the "Poky Analyzer" program (Figure 3-a), which is also part of the POKY suite. Alternatively, the user can utilize the Poky Analyzer Connector plugin (two-letter-code *up*) shown in Figure 3-d. Processing using Poky Analyzer includes the displays of NOE bar chart, Rama plot, and Contact map. Also, the "Distance Constraint Validator" and "Angle Constraint Validator" are two other parts of the program that help with identifying the violations in 3D structure modelling (Figure 3-b).

A PyMOL [18] script would be re-generated by the Poky Analyzer based on the user selection to visualize the distance restraints and facilitate further investigation of the 3D structure by typing *@p* in the PyMOL command line (Figure 3-c). This provides a seamless analytic environment. For example, because of this connection between the "Poky Analyzer" and PyMOL, the model violations can be easily tracked and analyzed. Once restraints are validated and finalized for the last step, the user can choose from "Final step w/ implicit", "Final step w/ explicit", "Final step w/ implicit (x2)", and "Final step w/ explicit (x2)" options. Basically, the options increase the number of structure calculations (one hundred without the x2 mark and two hundred with the x2 mark) and select twenty best structures based on pseudo-energy values (lower the better).

2.5. Workflow

A suggested workflow can be as follows: a) the user supplies the required data in the Poky Structure Builder program and send it to the server for the 3D calculation of the structure; b) the user receives the result of the calculation in an email; c) this result can be opened in the Poky Analyzer; d) the user can analyze the results using the *Distance Constraint Validator* of the Poky Analyzer; e) PyMOL can be started by a button in the plugin and the result is shown automatically; f) the user can further investigate the violations and if needed, repeat the process with corrected information.

This workflow creates the Integrative NMR platform ecosystem. The user can easily revise distance restraints and enhance the quality of structure by this re-iteration on the Integrative NMR platform.

3. Results and Conclusion

3.1. Results

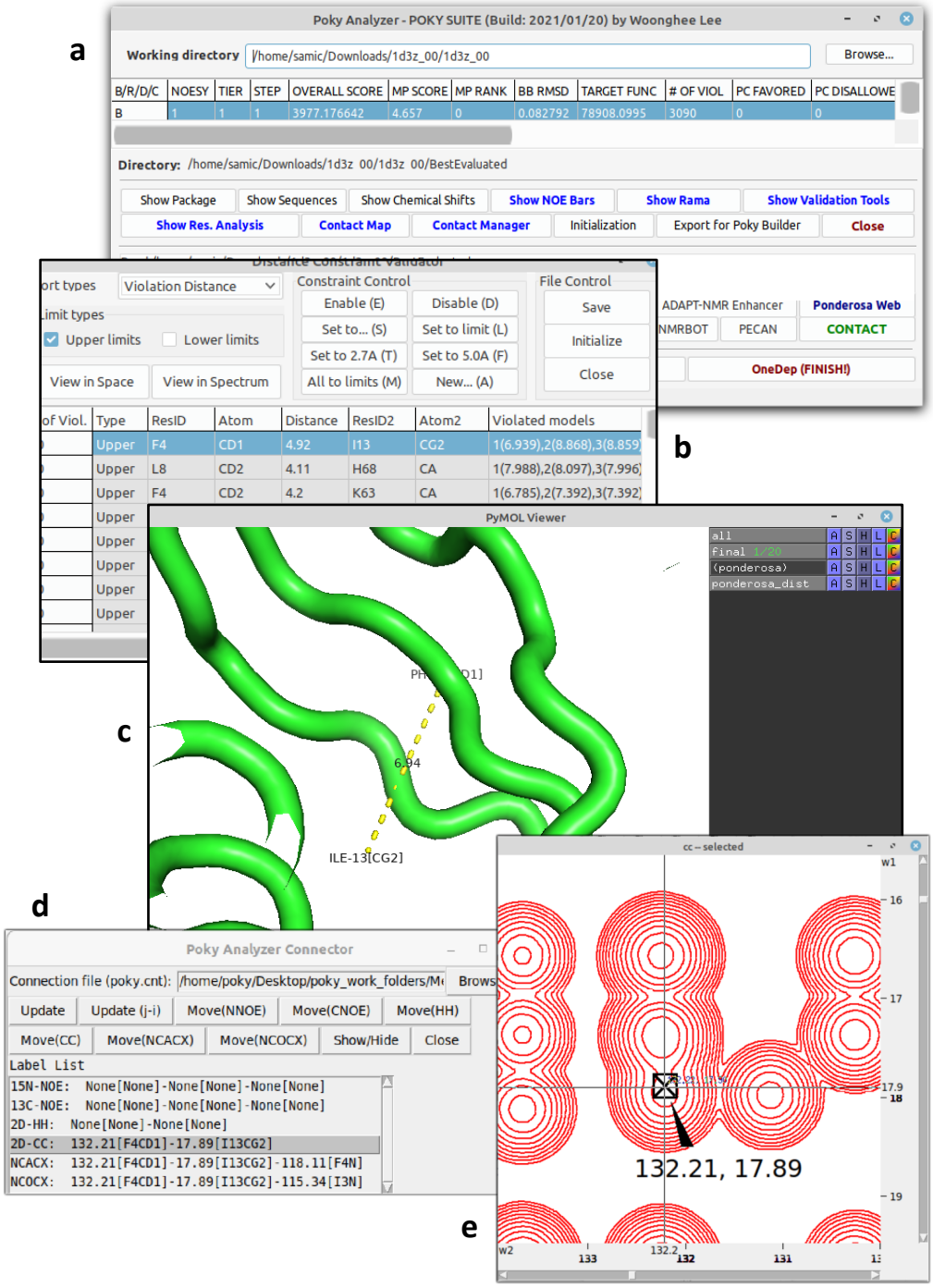

Figure 3. (a) Poky Analyzer provides validation tools and other information to analyze the results. (b) Distance Constraint Validator can be used to sort by violation distance. (c) the result is shown in PyMOL. (d) Poky Analyzer Connector plugin (two-letter-code up), can be called from POKY to investigate the result as well. (e) the result is shown on the spectrum. The visualized CC spectrum is simulated for the testing.

Because there are only a limited number of proteins with the ssNMR data for the structure, we simulated ssNMR spectra by the in-house ssNMR simulator program when developing the toolset. Seven proteins were selected from different SCOP classes and sequence sizes [19] ranging from 55 to 162 (Table S1; Figure S2-8). A scale factor of +20% was used in these experiments. This setting was named as “UPL Scaler percentage” in the plugin and it increases the values of the upper limits by the specified percentage. After the development and debugging, we also tested if we could reproduce the similar structures to the PDB-deposited structures with these proteins by a few mouse clicks [20]. We used the generated distance restraints from the simulated 2D CC spectra (two-letter-code *gd*) as inputs to the Poky

Structure Builder (two-letter-code *Pb*). NMR-STAR 3.1 files [21] were used to fill the sequence and chemical shift columns in the Poky Structure Builder. The "Constraints-only X" option was used with the inputs, and we clicked the "Preprocess" and "Calculate" buttons. We performed the pairwise alignment with the resulting structures and the PDB-deposited structures. The backbone RMSD (root-mean-square-deviation) were ranging between 1.084 Å and 1.895 Å (see the *Supplemental Information* for details).

Furthermore, we investigated the effect of the additional tolerance to the upper limit distances after the intensity calibration. We chose two proteins with PDB IDs 1UBQ and 1DF3. We prepared four additional restraint files for each entry: +10%, +20%, +30% and +40%. For the 5 Å, the upper limit restraints in the original 0% version are 5.5 Å, 6.0 Å, 6.5 Å and 7.0 Å in +10%, +20%, +30% and +40% restraint files, respectively. We calculated the structures with these files including the original 0%, so that the newly calculated structure from the simulated annealing gave a slightly different number from the simulated 2D CC test above for 1DF3 (1.607 Å vs. 1.634 Å). As a result, +20% restraint files gave the best result among all for both proteins (Figure S9; Table S2-S3), thus we use 20% as the default value in the "Generate distance constraint" plugin (Figure 2-b).

After the benchmark on the synthetically prepared ssNMR data, we selected four BMRB-deposited entries that were determined by ssNMR, and which had restraint files available in the NMR Restraints Grid [22] database. These tests did not include the restraint generation step (two-letter-code *gd*) but the direct use of the Poky Structure Builder (two-letter-code *Pb*) with the distance restraints available from the database. The pairwise backbone RMSD between the resulting structures and the BMRB-deposited entries were ranging from 1.125 to 3.569 Å. More explanation about these tests can be found in the *Supplemental Information*.

Poky Analyzer and Poky Analyzer Connectors are crucial parts of this developed ecosystem. An abundance of statistical analysis and validation tools such as NOE bar charts, Rama plot, Residue Analysis, Contact Map, and others, are provided for the user. This facilitates the process of validation of the results. One important part of this process is evaluating the violations that has been occurred. From Poky Analyzer, the user can click a button to see the "Distance Constraint Validator" window and further investigate these violations. Another important component of Poky Analyzer is the connection that has been made to PyMOL. The user can easily select any violation, and by the click of a button, see the exact position in PyMOL. A script is automatically made and run in PyMOL behind-the-scenes so that a seamless connection to PyMOL is obtained. A list of useful commands in PyMOL is also shown to the user to better assist in interacting with PyMOL. For example, typing "@p" will update the model based on the selection on Poky Analyzer window. This harmonious connection between different parts of the system creates an uninterrupted workflow for the user.

3.2. Conclusion

In this work, we have developed multiple programs and plugins to provide an easy-to-use yet comprehensive integrative NMR approach for 3D structure determination for ssNMR samples. The workload is simplified down to a few clicks for the user, while all the processing happens on the PONDEROSA cluster of servers.

The process of structure determination has been tested through numerous experiments using synthetic ssNMR spectra as well as real ssNMR spectra. Detailed information about the results can be found in the *Supplemental Information*.

All the source codes and plugins are freely available on our GitHub page and the access to PONDEROSA servers is open to researchers without any cost. The programs and the plugins are integrated into the POKY software suite and the NMRFAM-SPARKY. The Poky Structure Builder is exclusively available only in the POKY software suite.

The latest version of the POKY suite can be downloaded from https://poky.clas.ucdenver.edu.

Supplementary Materials: The supplemental information is available online at www.com/xxx/s1

Also, we recorded tutorial videos for the benefit of the user. The videos are accessible at https://poky.clas.ucdenver.edu/ponderosa-videos

Author Contributions: Conceptualization, W.L.; methodology, H.N. and W.L.; software, M.R. and W.L.; validation, M.R., Y.L., A.C. and W.L.; writing—original draft, M.R.; writing—review and editing, Y.L., H.N., A.C. and W.L.; supervision, W.L.; funding acquisition, W.L.

All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation [DBI-2051595 and DBI-1902076 to W.L.] and the startup support from the University of Colorado Denver.

Data Availability Statement: Publicly available data were analyzed in this study. This data can be found here: https://bmrb.io and https://restrainsgrid.bmrb.io. The POKY suite with the installation instruction is available from https://poky.clas.ucdenver.edu.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alberts, B., et al., *The shape and structure of proteins*, in *Molecular Biology of the Cell. 4th edition*. 2002, Garland Science.
2. Mandala, V.S., J.K. Williams, and M. Hong, *Structure and dynamics of membrane proteins from solid-state NMR.* Annual review of biophysics, 2018. **47**: p. 201-222.
3. Loquet, A., et al., *3D structure determination of amyloid fibrils using solid-state NMR spectroscopy.* Methods, 2018. **138**: p. 26-38.
4. Ranjan, R. and N. Sinha, *Nuclear magnetic resonance (NMR)–based metabolomics for cancer research.* NMR in Biomedicine, 2019. **32**(10): p. e3916.
5. Jayalakshmi, K., et al., *Solid state 13C NMR analysis of human gallstones from cancer and benign gall bladder diseases.* Solid state nuclear magnetic resonance, 2009. **36**(1): p. 60-65.
6. Lee, W., J.L. Stark, and J.L. Markley, *PONDEROSA-C/S: client–server based software package for automated protein 3D structure determination.* Journal of biomolecular NMR, 2014. **60**(2-3): p. 73-75.
7. Lee, W., et al., *The AUDANA algorithm for automated protein 3D structure determination from NMR NOE data.* Journal of biomolecular NMR, 2016. **65**(2): p. 51-57.
8. Lee, W., M. Tonelli, and J.L. Markley, *NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy.* Bioinformatics, 2015. **31**(8): p. 1325-1327.
9. Lee, W., et al., *POKY: a software suite for multidimensional NMR and 3D structure calculation of biomolecules.* Bioinformatics, 2021.
10. Lee, W., et al., *Integrative NMR for biomolecular research.* Journal of biomolecular NMR, 2016. **64**(4): p. 307-332.
11. Rahimi, M., et al., *iPick: Multiprocessing software for integrated NMR signal detection and validation.* Journal of Magnetic Resonance, 2021. **328**: p. 106995.
12. Schwieters, C.D., J.J. Kuszewski, and G.M. Clore, *Using Xplor–NIH for NMR molecular structure determination.* Progress in nuclear magnetic resonance spectroscopy, 2006. **48**(1): p. 47-62.
13. Güntert, P., *Automated NMR structure calculation with CYANA*, in *Protein NMR techniques*. 2004, Springer. p. 353-378.
14. Russell, R.W., et al., *Accuracy and precision of protein structures determined by magic angle spinning NMR spectroscopy: for some 'with a little help from a friend'.* Journal of biomolecular NMR, 2019. **73**(6): p. 333-346.
15. Schwieters, C.D., et al., *The Xplor-NIH NMR molecular structure determination package.* Journal of magnetic resonance, 2003. **160**(1): p. 65-73.
16. Shen, Y. and A. Bax, *Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks.* Journal of biomolecular NMR, 2013. **56**(3): p. 227-241.
17. Frishman, D. and P. Argos, *Knowledge–based protein secondary structure assignment.* Proteins: Structure, Function, and Bioinformatics, 1995. **23**(4): p. 566-579.

18. Schrodinger, L., *The PyMol molecular graphics system, version 2.0. Schrödinger, LLC, New York, NY*. 2017.
19. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* Journal of molecular biology, 1995. **247**(4): p. 536-540.
20. Berman, H., et al., *The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data.* Nucleic acids research, 2007. **35**(suppl_1): p. D301-D303.
21. Ulrich, E.L., et al., *NMR-STAR: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments.* Journal of biomolecular NMR, 2019. **73**(1): p. 5-9.
22. Meyer, P.A., et al., *SBGrid Databank.* Foundations of Crystallography, 2017. **73**: p. a264.