iPick: Multiprocessing Software for Integrated NMR Signal Detection and Validation

Mehdi Rahimi [1], Yeongjoon Lee [1], John L. Markley [2] and Woonghee Lee [1, *]

[1] Department of Chemistry, University of Colorado Denver, Denver, CO 80204 USA; mehdi.rahimi@ucdenver.edu, yeongjoon.lee@ucdenver.edu, woonghee.lee@ucdenver.edu
[2] National Magnetic Resonance Facility at Madison, Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706 USA; jmarkley@wisc.edu
[*] Correspondence: woonghee.lee@ucdenver.edu

Abstract: Peak picking is a critical step in biomolecular NMR spectroscopy. The program, *iPick*, presented here provides a scripting tool and a graphical user interface (GUI), which allow the user to perform interactive and intuitive peak picking and validation. The click-and-run GUI requires no computer programming skills, while the scripting tool can be used by more advanced users to customize the application. If used with a multi-core CPU, the multiprocessing feature of *iPick* reduces the processing time significantly by invoking parallel computing. The GUI is a plugin, compatible with the popular NMRFAM-SPARKY software package and its newly released successor, the POKY software. Features implemented in *iPick* include automated noise level detection and threshold setting, cross-validation against multiple spectra, and a method for quantifying peak reliability. The *iPick* software is cross-platform, open-source, and freely available from https://github.com/pokynmr/ipick.

Keywords: multidimensional NMR spectroscopy; signal recognition; peak validation; noise level calculation; graphical user interface; multiprocessing; NMRFAM-SPARKY; POKY; Python

1. Introduction

A wide range of biomolecular studies benefit from NMR spectroscopy. However, the interpretation of acquired data is often not straight-forward and requires complex and repetitive steps. An early critical step is signal detection, known as "peak picking", which builds the foundation for subsequent analysis of the NMR data [1]. Manual performance of this task with multi-dimensional spectra can be extremely difficult and time consuming, because each spectrum contains hundreds or thousands of peaks that must be investigated one-by-one by a visual search. Existing visual tools provided by software packages for analyzing NMR spectra, including NMRFAM-SPARKY [2] and its successor, POKY [3], NMRView [4], and CARA [5], detect individual signals, but their reliability is compromised by overlapping peaks, weak signals, and spectral artifacts.

The main objectives of peak picking are to establish a peak list that provides accurate information on peak parameters (position, height, width, shape, volume) while minimizing false-positive (noise or artifact) peaks [6]. The basic approach for peak peaking in biomolecular NMR spectra is searching for local maxima [7], and the prerequisite for avoiding many noise peaks is the selection of a noise level to be higher than the desired contour level. Clearly, the main obstacles to these tasks are limited digital resolution and low signal-to-noise ratio (SNR).

A number of software tools have been developed to overcome these problems and automate the process. For example, CYPICK [8] analyzes geometric properties of contour lines to identify peaks, and NMRNet [9] relies on deep learning to automate the peak picking task. The *iPick* method, described here, offers the option of using either automatically detected signal-to-noise ratio (SNR) or maximum contour levels as the basis of signal identification. Its use of the median value of randomly sampled data results in extremely rapid identification of the noise while identifying SNR maxima. Both the noise- and the contour-level approaches can be fine-tuned by the user.

Weak peaks close to noise level and peaks on the shoulder of a stronger peak in a spectrum are often evaluated in practice by comparison to corresponding signals from spectral regions in the same or different spectra. This enables the weeding out of false-positive peaks that lack support from

corresponding regions. Our group developed the APES plugin for NMRFAM-SPARKY, which automates this approach to peak picking [10]. However, APES only supports a limited set of experiments from liquid-state protein NMR spectroscopy. Therefore, we have included a more general resonance cross-validation module in *iPick*.

2. Design and Implementation

2.1 Peak Picking

The *iPick* program offers two running modes: The *Basic Mode* (Figure 1A) and the *Advanced Mode* (Figure 1B). In the *Basic Mode*, the user only needs to select a spectrum and click the *Run iPick* button, which utilizes default parameters. Checking the *Import peaks* option causes the positions of the picked peaks to be displayed on the selected spectra. The *Advanced Mode* of *iPick* to enables the user to fine-tune the software to tackle more difficult data sets. In this mode, the user has control of each step of the program. Two backbone programs are available for the peak searching: UCSFtool and NMRGlue [11]. Positive or negative peaks (or both) can be chosen. All internal parameters are displayed and can be changed easily. Tips for the use of each tool are activated by hovering the mouse cursor over its name. A feature of the *Advanced Mode* is post-process automation: the user can specify the minimum Euclidean distance between two peaks to be considered as separate and the minimum drop factor below which the program assumes two local maxima to be part of a single peak.

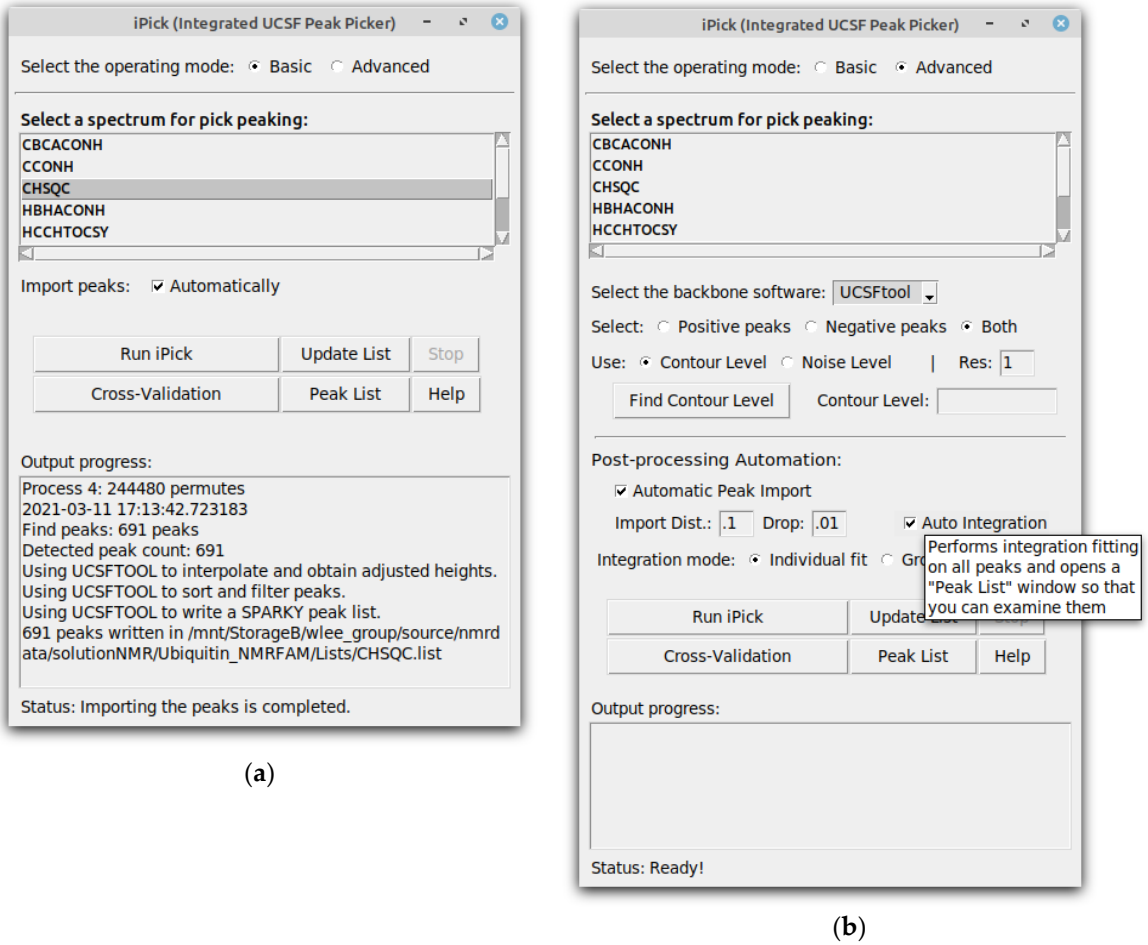


Figure 1. Screenshots of the *iPick* GUI's *Basic* and *Advanced Modes*. (**a**) The *Basic Mode* allows a simple one-click approach that automatically detects the best default options. (**b**) The *Advanced Mode* allows user customization of the parameters for fine-tuning. Tips providing information about the tool are activated by the mouse cursor. As an example, the result of the mouse hovering over *Auto Integration* is shown in the screen shot.

2.2 Peak Reliability Score

Another important feature of the *Advanced Mode* is the *Auto Integration* option. By activating this option, *iPick* conducts an automated integration of all peaks, one-by-one and also by groups. Then, a customized *Peak List* window presents the fitted peak parameters (chemical shifts, peak height, fit height, fit volume, SNR, linewidth) and the *Reliability Score* (Figure 2). Double clicking on an entry in the peak list causes the corresponding peak to be centered in the *spectral view* for further investigation. The *Reliability Score* is calculated from a combination of the peak parameters. Peaks in the window can be sorted by height and *Reliability Score* to rapidly distinguish prominent from non-prominent peaks. To discard unreliable peaks, a user can set a threshold and click the *Remove* button. The threshold also changes interactively as the user selects a peak. Details regarding the *Reliability Score* are presented in *Supplementary Information*, including how a user can change the coefficients that formulate the *Reliability Score*.

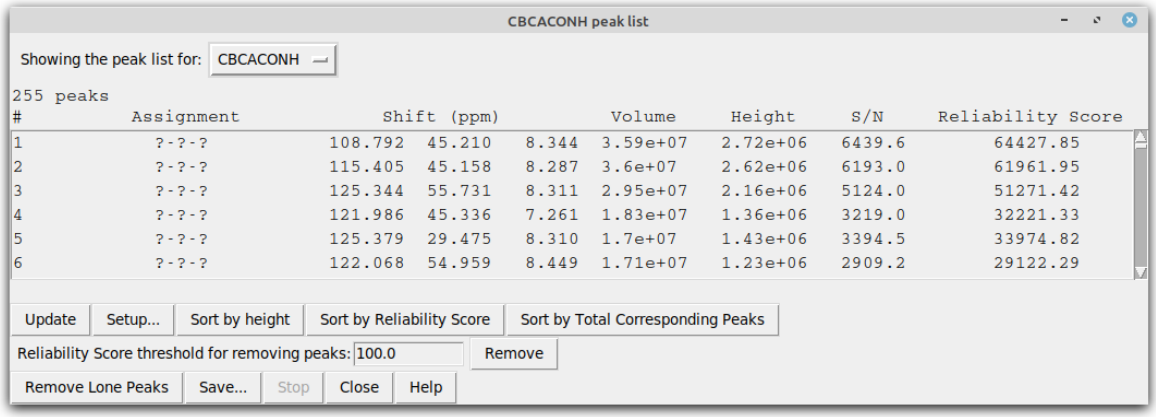


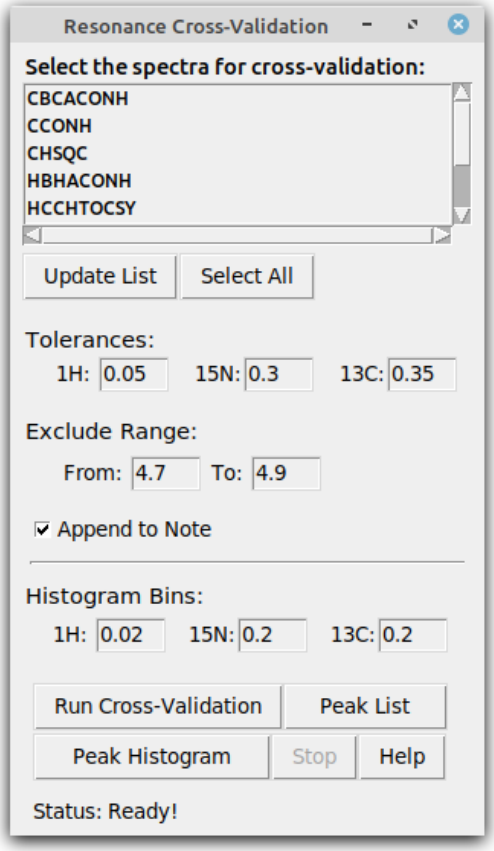
Figure 2. Example of a peak list generated by the *Auto Integration* option. The *Reliability Score* provides a measure of the probability that a peak is real and not noise or artifact.

2.3 Cross-Validation across Spectra

We have developed a resonance cross-validation module (Figure 3-5) that supplements *iPick* peak picking. Peaks in a given spectrum are validated in terms of the detection of expected corresponding peaks in other spectra. The user selects the spectra to be used for cross-validation and chooses the tolerance limits (in ppm) for peak correspondence. An area can be excluded for cross-validation, such as water signal region (usually around 4.8 ppm). The *Run Cross-validation* button executes the examination, and the results are tabulated in a *Peak List*, which displays the 2D frequencies of each peak in the spectrum to be validated and the number of corresponding peaks in each spectrum falling within the specified range in the other spectra used for validation (Figure 4). Peaks with no corresponding or supporting resonances in other spectra can be easily removed by clicking the *Remove Lone Peaks* button.

2.4 Parallelism

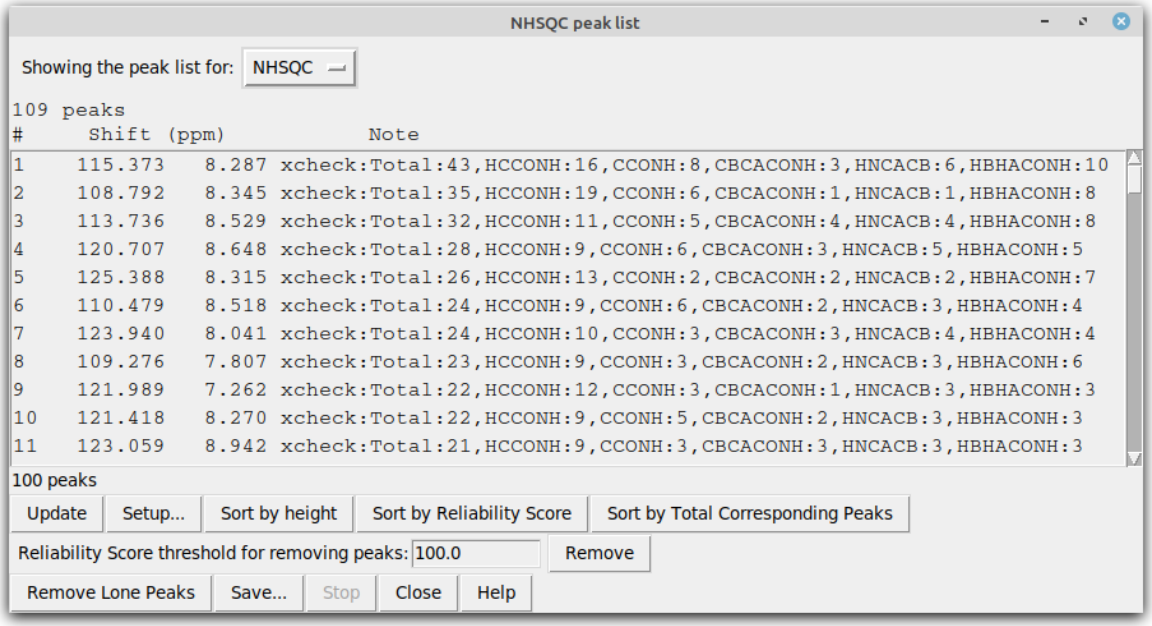
We developed *iPick* as a parallel algorithm to support modern CPUs with multiple processing cores when UCSFtool is chosen. Our benchmark tests have shown that the parallelism leads to much faster run times (Figure S5 in *Supplementary Information*).



105

106 **Figure 3.** Screenshot of the main window of the *Resonance Cross-Validation* module. The user selects
107 the spectra to be used, the tolerances for peak matches, and a signal exclusion range to be used in the
108 cross-validation. Buttons initiate the run and display the results as the *Peak List* or *Peak Histogram*.

109



110

111 **Figure 4.** Screenshot of the cross-validation results in the peak list. Each peak has a *Note* section that
112 shows the number of corresponding peaks in other spectra. Peaks with the fewest correspondences
113 are more likely to be false-positives.

The *Peak Histogram* button in the *Cross-Validation* module causes the computation and graphical display of histograms of correlated peak resonances in two or more spectra selected by the user (Figure 5). The user can use the *Show the selected peaks* button from the spectral view to display the positions of one or more peaks on the histogram. See the *Supplementary Information* for details.

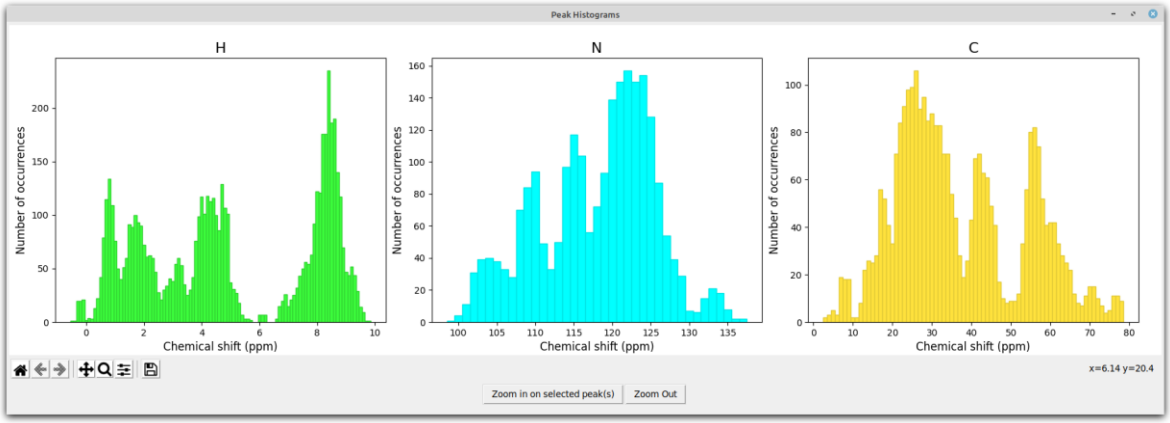


Figure 5. Example of a *Peak Histogram* display.

2.5 Integrative NMR Platform

We have built *iPick* into NMRFAM-SPARKY [2] and its successor, the POKY software [3] as a plugin as part of our development of an integrative NMR platform for biomolecular NMR research. The *iPick* GUI is accessible from the Extensions/Peak menu of the latest NMRFAM-SPARKY or POKY (two-letter-code *iP*). Also, it can be loaded as a module in the *Python Shell* of the NMRFAM-SPARKY or POKY (two-letter-code *py*). Either *ipick_gui_sparky.py* for the GUI or *iPick.py* for the command line interface (CLI) can be loaded by typing "*import ipick_gui_sparky*" or "*import iPick*". The CLI approach allows a user to write and run custom Python scripts using functions provided by *iPick*.

3. Results

The *iPick* program is written in Python with the Tkinter library. This program detects peaks in multidimensional NMR spectra. It supports multi-processing, a feature of modern multi-core CPUs, for achieving maximum performance. Automatic noise level determination helps in measuring the accurate SNRs. The *Basic Mode* of operation features an intuitive GUI for use by non-specialists in the field. The *Advanced Mode* of operation allows expert users to customize each step. Validation of the results is aided by modules for automated integration fitting and determination of a *Reliability Score*. A cross-validation tool, which finds corresponding peaks in multiple spectra, provides the means of weeding out peaks that lack expected correspondences. In combination, these tools provide robust platform for picking and validating peaks. The automated tools presented here, along with various fine-tuning options, can be efficiently integrated into a researcher's workflow to successfully expedite the overall process. An example of such a workflow is presented in the *Supplemental Information*.

As detailed in the *Supplementary Information*, the performance of the software was evaluated through analysis of spectra from multiple 2D and 3D NMR experiments with ubiquitin protein labeled uniformly with ^{13}C and ^{15}N. To test the performance of the program with solid-state NMR data, we used spectra from multiple 2D and 3D experiments with GB1 protein labeled uniformly with ^{13}C and ^{15}N.

4. Availability and Future Directions

The *iPick* software is cross-platform, open-source, and freely available from https://github.com/pokynmr/ipick under the BSD 2-Clause License. The code is compatible with both Python 2 and 3, which also are free and open-source. *iPick* comes pre-installed in recent versions of the popular NMRFAM-SPARKY software package

(http://pine.nmrfam.wisc.edu/download_packages.html), and the POKY software package (https://poky.clas.ucdenver.edu), and no other installation step is necessary. The NMRFAM-SPARKY software package is available to use for subscribers of NMRbox.org [12] and SBGrid Consortium [13]. Updated algorithms and GUIs will be merged into the master branch of the iPick GitHub repository and also included in the NMRFAM-SPARKY and POKY. Developers will interact with users on NMR POKY/SPARKY user group (https://groups.google.com/g/nmr-sparky) and additional functionalities with bug fixes will be suggested there.

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

Funding: This work was supported by the National Science Foundation (Grant No. DBI-2051595 & DBI-1902076) and the University of Colorado Denver.

Acknowledgements: We are grateful to Dr. Marco Tonelli and Prof. Chad Rienstra (University of Wisconsin-Madison) for providing their NMR data benchmarked in this development.

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

References

1. Cheng Y, Gao X, Liang F. Bayesian peak picking for NMR spectra. Genomics, proteomics & bioinformatics. 2014;12(1):39-47.
2. Lee W, Tonelli M, Markley JL. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics. 2015;31(8):1325-7.
3. Lee W, Rahimi M, Lee Y, Chiu A. POKY: a software suite for multidimensional NMR and 3D structure calculation of biomolecules. Bioinformatics. 2021. doi: 10.1093/bioinformatics/btab180.
4. Johnson BA, Blevins RA. NMR View: A computer program for the visualization and analysis of NMR data. Journal of biomolecular NMR. 1994;4(5):603-14.
5. Keller R, Wuthrich K. Computer-aided resonance assignment (CARA). Verl Goldau Cantina Switz. 2004.
6. Smith AA. INFOS: spectrum fitting software for NMR analysis. Journal of biomolecular NMR. 2017;67(2):77-94.
7. Lee W, Cornilescu G, Dashti H, Eghbalnia HR, Tonelli M, Westler WM, et al. Integrative NMR for biomolecular research. Journal of biomolecular NMR. 2016;64(4):307-32.
8. Würz JM, Güntert P. Peak picking multidimensional NMR spectra with the contour geometry based algorithm CYPICK. Journal of biomolecular NMR. 2017;67(1):63-76.
9. Klukowski P, Augoff M, Zięba M, Drwal M, Gonczarek A, Walczak MJ. NMRNet: a deep learning approach to automated peak picking of protein NMR spectra. Bioinformatics. 2018;34(15):2590-7.
10. Shin J, Lee W, Lee W. Structural proteomics by NMR spectroscopy. Expert review of proteomics. 2008;5(4):589-601.
11. Helmus JJ, Jaroniec CP. Nmrglue: an open source Python package for the analysis of multidimensional NMR data. Journal of biomolecular NMR. 2013;55(4):355-67.
12. Maciejewski MW, Schuyler AD, Gryk MR, Moraru II, Romero PR, Ulrich EL, et al. NMRbox: a resource for biomolecular NMR computation. Biophysical journal. 2017;112(8):1529-34.
13. Morin A, Eisenbraun B, Key J, Sanschagrin PC, Timony MA, Ottaviano M, et al. Cutting edge: Collaboration gets the most out of software. elife. 2013;2:e01456.