Journal of Magnetic Resonance 361 (2024) 107664

Contents lists available at ScienceDirect

Journal of Magnetic Resonance

journal homepage: www.elsevier.com/locate/jmr

ASAP: An automatic sequential assignment program for congested multidimensional solid state NMR spectra

Bo Chen

Department of Physics, University of Central Florida, Orlando 32816, USA

ARTICLE INFO

Keywords:
Sequential assignment
NMR
Multidimensional spectroscopy

ABSTRACT

Accurate signal assignments can be challenging for congested solid-state NMR (ssNMR) spectra. We describe an automatic sequential assignment program (ASAP) to partially overcome this challenge. ASAP takes three input files: the residue type assignments (RTAs) determined from the better-resolved NCACX spectrum, the full peak list of the NCOCX spectrum, and the protein sequence. It integrates our auto-residue type assignment strategy (ARTIST) with the Monte Carlo simulated annealing (MCSA) algorithm to overcome the hurdle for accurate signal assignments caused by incomplete side-chain resonances and spectral congestion. Combined, ASAP demonstrates robust performance and accelerates signal assignments of large proteins (>200 residues) that lack crystalline order.

1. Introduction

Solid state NMR (ssNMR) is the ideal structural biology technique to characterize insoluble biomolecular aggregates that lack the perfect structural order [1–7]. The assignment of resonance signals, called chemical shifts (CSs), to specific sites in the molecule, is the pre-requisite to extract site-specific structural information. Recent advancements in spectroscopic techniques lead to samples of increasing sizes solved by ssNMR, for example, the 41 kDa DsbA/DsbB [8], or the 72 kDa tryptophan synthase [9]. We note the samples used in these works are either micro or nanocrystalline quality, which produce spectra of extraordinarily sharp lines (0.3 ppm or better). Assignments of such well-resolved spectra can be readily completed manually or by auto-assignment programs such as FLYA [10] or ssPINE [11]. However, in most ssNMR studies of non-crystalline samples, the linewidth can be twice or triple of that observed with crystalline samples, which can lead to serious spectral congestion. Hence, it is challenging to make signal assignments with large proteins (≥150 residues) that lack crystalline order, which seriously cripples the application of ssNMR.

The typical workflow of ssNMR studies starts with the acquisition of multidimensional spectra, which disperse congested resonance signals and reveal intra and inter-residue correlations. The fundamentals are depicted in Fig. 1A and B, in terms of the ^{13}C detected 3D NMR experiments. Briefly, after the polarization signals on nuclei X evolve for a period at their respective CSs (1st CS labeling period), they are transferred to nearby Y nuclei to evolve at the CSs of Y sites (2nd CS labeling period). Then the polarizations are transferred nearby Z sites for direct detection. Thus, similar CSs of Z nuclei are dispersed along their distinct CSs of X and Y nuclei to achieve higher resolution. As shown in Fig. 1B, the NCACX experiment channels the polarization from amide nitrogen (X) to c-alpha (Y), and then to other carbons (Zs) in the same residue, to disperses the CSs of carbons (Zs) along the CSs of c-alpha (Y) and amide nitrogen (X), gathering the intra-residue correlation. Meanwhile, the inter-residue correlation is revealed by NCOCX by the polarization transfer from amide nitrogen (X) to carboxylic carbon (Y) and other carbons (Zs) in the preceding residue. 3D or 4D experiments such as CANCO/CANCX or CONCA/CONCX[12] may further improve the resolution, shown in Fig. 1B. They can provide critical additional information to facilitate signal assignments and disperse congested signals. These experiments work particularly well for crystalline or polycrystalline samples [12]. However, for typical ssNMR samples that lack structural order and with broad linewidth, [13,14] fewer residues show up in these spectra due to the weaker signal-to-noise ratio (SNR) of double heteronuclear polarization transfer.

After acquiring the spectra, the first step of signal assignments is to group resonances from the same residue together, and identify their residue types, referred to as the residue type assignments (RTAs). Then the polarization transfer pathways encoded in RTAs in different spectra are matched to the connections specified by the protein sequence, to complete the sequential assignment. Various auto-assignment programs or strategies exist, [11,15–37] which usually determine RTAs by the characteristic CSs of amino acids [38]. However, this can be quite

E-mail address: Bo.chen@ucf.edu.

https://doi.org/10.1016/j.jmr.2024.107664
Received 13 December 2023; Received in revised form 19 March 2024; Accepted 20 March 2024
Available online 21 March 2024
1090-7807/Published by Elsevier Inc.

challenging for congested ssNMR spectral with broad resonances.

Despite the help of multidimensional NMR experiments, signal broadening (>0.5 ppm) due to anisotropic interactions in non-crystalline solid samples still leads to poor spectral resolution, even with advanced magic angle spinning and decoupling pulse sequences [39]. Because the CSs dispersion of c-alpha is 2–3 times wider than those of the carboxylic sites, frequently the 3D NCOCX spectrum of a sample may become too congested for accurate RTAs, even when its 3D NCACX still displays sufficient spectral resolution. An example is shown in Fig. 1D and E by the 2D planes extracted from the 3D NCACX and NCOCX spectra of the tubular assembly formed by the 237-residue Rous sarcoma virus (RSV) capsid protein (CA) [13]. In addition, limited sidechain resonances makes the determination of RTAs by characteristic CS patterns unreliable. While more sidechain resonances may be induced by longer mixing time, it also incurs extra line broadening, and additional sidechain resonances in the 30–40 ppm regions will also exacerbate the signal congestion. As it provides indispensable inter-residue correlations, assignment of the over-congested NCOCX spectra usually becomes the bottleneck for a ssNMR project.

When reliable RTAs from different spectra can be obtained, to accelerate the sequential assignment, Tycko's group created the MCAssign program to automatically determine their sequential allocations [32,37]. It utilizes the Monte Carlo simulated Annealing (MCSA) algorithm to randomly shuffle RTAs to match their polarization transfer pathways with the protein sequence. Given the same set of input RTAs, the program often finds different sequential allocations with comparable final scores. To differentiate them, based on the MCAssign program, Hong's group developed a variant called NSGA-II [36]. It utilizes the non-dominated sorting genetic algorithm with an additional bias that increases the weight of RTAs forming good connections with their neighbors. We refer to both methods as the standard MCSA, as they employ the same MCSA process to determine the sequential allocations of given RTAs. These methods greatly accelerated the sequential assignment for ssNMR projects.

However, both MCAssign and NSGA-II demand accurate RTAs from all spectra. For large proteins, [13,14] while it may be possible to determine RTAs fairly quickly and accurately in NCACX, the NCOCX spectra are usually too congested to make accurate RTAs. Moreover, the performance of MCAssign quickly deteriorates with ambiguous RTAs, even with decent spectral resolution (~0.6 ppm full width half maximum (FWHM)) for proteins approaching 150 residues [40]. The RTAs have to be carefully revised repetitively to maximize the number

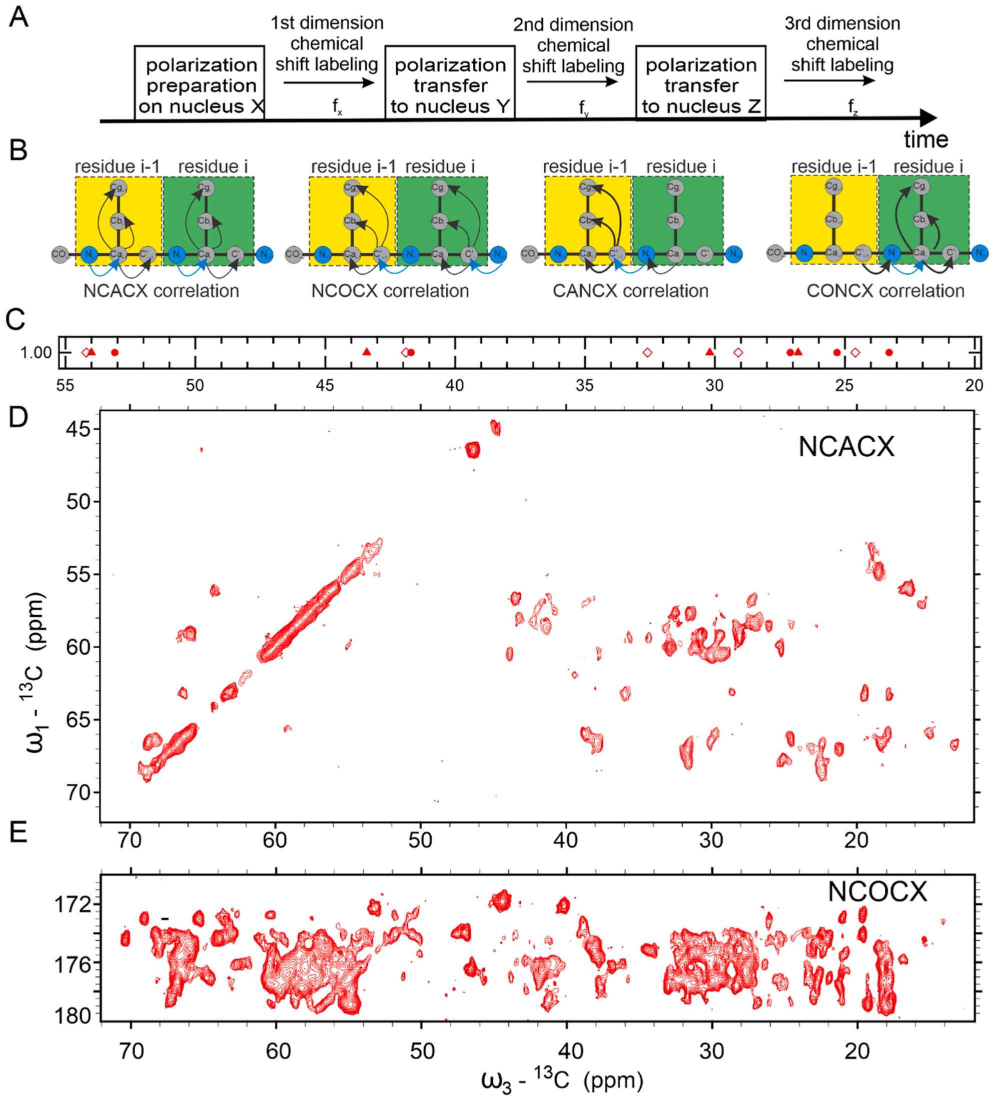


Fig. 1. 3D experiments setup and their resolution disparity. (A). Schematic 3D pulse sequence setup. (B). Polarization transfer pathways of 3D NCACX, NCOCX, CANCO/CX and CONCA/CX. (C) Illustration of the RTA complexity caused by coincidental alignment of resonances from different residues. Signals from a L (circles), R (triangles), and K (diamonds) are plotted together. (D) The most congested 2D plane extracted from 3D NCACX and NCOCX (E) of the tubular assembly of uniform ^{13}C, ^{15}N labeled RSV CA.

of sequentially assigned residues, which is quite challenging and problematic with a congested NCOCX spectrum for multiple reasons that will be discussed in this work. The assignment process can still take years, even for a well-trained researcher with the assistance of these state-of-the-art auto-assignment programs. Hence, ssNMR usually becomes the last resort for structural characterization of large proteins.

Here we introduce an auto-sequential assignment program (ASAP) that overcomes some of the challenges limiting the capability of ssNMR. ASAP integrates an innovative Automatic Residue Type Identification STrategy (ARTIST) with MCSA. It only needs the protein sequence, the peak list from the NCOCX spectrum, and the RTAs from the NCACX spectrum of a sample, which is more likely to be determined with confidence and less ambiguity due to its higher spectral resolution than NCOCX, to enable thorough sampling of all possible configurations of resonances that maximize the signal assignments. With ASAP, proteins of 250 residues with broad NMR lines that do not provide useful information from higher dimensional spectral such as CANCX and CONCX spectra can be assigned in days. It demonstrates superior effectiveness, accuracy, robustness against ambiguous RTAs, tested by multiple proteins [13,41,42]. ASAP is designed to work with the ^{13}C detected NCACX and NCOCX spectra, but it can be modified to assign other spectra with different polarization transfer pathways and detection methods [43–46]. The program is coded in python, [47] which allows easy revisions and improvements by other users.

2. Methods

2.1. Introduction of standard MCSA algorithm and its limitations

To understand the strength of ASAP, the limitations of the standard MCSA should be analyzed first. Accurate RTAs from multiple spectra are required by MCAssign and NSGA-II, including NCACX and NCOCX [11,40]. The RTAs in the NCACX and NCOCX spectra are identified as the residue type associated with each c-alpha and carboxylic site in the polarization transform pathway, respectively. Ambiguity in an RTA refer to the assignment of a group of signals as more than one possible residue types.

Despite their differences, both methods rely on the MCSA algorithm to sample millions of random allocations of RTAs in given spectra to determine their sequential allocations. Specifically, each MC attempt starts with the selection of a residue position randomly in the protein sequence. Assume that current MC attempt picks the residue position *kres*. Next, the program takes turns from the *k* input spectra to randomly select an RTA of matching residue type to replace the existing RTA occupying *kres* position from the same spectrum. For convenience of our discussion, assume that an RTA is selected from the NCACX spectrum. The program inspects the agreement of the ^{13}C CSs between the new RTA from NCACX, and those residing *kres* position from other spectra (NCOCX, CANCX and CONCX if they are available) to determine their compatibility. Then its ^{15}N resonance is also compared with that of the RTA from the NCOCX spectrum seated at residue position *kres-1,* to count the total number of good (n_g), bad (n_b) and edge (n_e) connections. An edge connection refers to the situation that a residue position is occupied by a null assignment. The same evaluation is repeated for the RTA from NCACX currently residing *kres* position. The change of the score between the old and new configurations is then computed:

$$S^i(\Delta n_g, \Delta n_b, \Delta n_e, \Delta n_u) = w_1^i \Delta n_g - w_2^i \Delta n_b - w_3^i \Delta n_e + w_4^i \Delta n_u \tag{1}$$

The superscript denotes that the parameters are pertinent to the *i*-th annealing step. Δn_g, Δn_b, Δn_e and Δn_u are the changes of connection numbers and used RTAs. Their coefficients w_j^i at annealing step *i* are set by:

$$w_j^i = w_{j0} + scale \times \frac{w_{jf} - w_{j0}}{nstep} i = scale \frac{w_{jf}}{nstep} i \tag{2}$$

here *nstep* is the total number of annealing steps in the entire MCSA process, with the annealing slope set by *scale*. Typically, their initial values w_{j0} are set to zero. The score determines the acceptance of the new configuration by the Metropolis criterion:

$$\exp\left(S^i(\Delta n_g, \Delta n_b, \Delta n_e, \Delta n_u)\right) \ge rand(0,1) \tag{3}$$

where $rand(0,1)$ is a random number between 0 and 1. The score function imparts a growing penalty as the annealing progresses, to guide the MCSA process towards the maximization of n_g and n_u, minimizing n_b and n_e. However, this setup may limit the efficiency, accuracy of the program, as well as the resilience against ambiguous RTAs in input, as analyzed below.

A resonance signal in a 3D spectrum contains three coordinates in the frequency space: $\left(f_x, f_y, f_z\right)$, where f_x and f_y are the frequencies along the first two indirect dimensions, and f_z is the frequency along the direct detection dimension, shown in Fig. 1A. Assume that residue *kres* in a protein comprises N_{kres} carbon sites. The signal of its *i-th* carbon is designated by $\left(f_{axi}^{kres}, f_{ayi}^{kres}, f_{azi}^{kres}\right)$ in the 3D NCACX spectrum. The superscript *kres* denotes the residue that the signal is associated with. The first subscript *a* denotes that the signal is from NCACX. The last subscript *i* denotes that the resonance is from the *i-th* carbon along the directly detected dimension, with $i = 1, \ldots N_{kres}$. For convenience of discussion, we refer to the carboxylic carbon as $i = 1$, c-alpha as $i = 2$, and c-beta as $i = 3$, etc. Following this notation, the signal for the *i-th* carbon in residue *kres* in the 3D NCOCX spectrum is $\left(f_{bxi}^{kres}, f_{byi}^{kres}, f_{bzi}^{kres}\right)$, where the first subscript *b* denotes that the signal is from the NCOCX spectrum.

Obviously, when two different carbon sites *i and j* of residue *kres* both produce signals in NCACX, their indirect dimensions must align within their uncertainties:

$$\left|f_{axi}^{kres} - f_{axj}^{kres}\right| \le \sqrt{\left(\Delta f_{axi}^{kres}\right)^2 + \left(\Delta f_{axj}^{kres}\right)^2} \tag{4}$$

$$\left|f_{ayi}^{kres} - f_{ayj}^{kres}\right| \le \sqrt{\left(\Delta f_{ayi}^{kres}\right)^2 + \left(\Delta f_{ayj}^{kres}\right)^2} \tag{5}$$

here, $\left(\Delta f_{axi}^{kres}, \Delta f_{ayi}^{kres}, \Delta f_{azi}^{kres}\right)$ and $\left(\Delta f_{axj}^{kres}, \Delta f_{ayj}^{kres}, \Delta f_{azj}^{kres}\right)$ are their respective uncertainties, due to the non-zero resonance linewidth. In ssNMR spectra, this uncertainty is typically ∼ 1/2FWHM. Overlapping of two nearby resonances can shift their exact locations, and the shift is accounted for by their respective 1/2 FWHM. Hence, Eqs. 4 & 5 are still applicable to signals in the presence of overlapping. If more than 2 resonances overlap, it is possible that the shift of peak positions goes beyond their respective ½ FWHM.

Likewise, when these two sites produce signals in NCOCX, their indirect dimensions must align within their uncertainties:

$$\left|f_{bxi}^{kres} - f_{bxj}^{kres}\right| \le \sqrt{\left(\Delta f_{bxi}^{kres}\right)^2 + \left(\Delta f_{bxj}^{kres}\right)^2} \tag{6}$$

$$\left|f_{byi}^{kres} - f_{byj}^{kres}\right| \le \sqrt{\left(\Delta f_{byi}^{kres}\right)^2 + \left(\Delta f_{byj}^{kres}\right)^2} \tag{7}$$

Note that their x-coordinates f_{bxi}^{k} and f_{bxj}^{k} in NCOCX are the ^{15}N frequencies of the next residue *kres + 1* in the protein, as shown by Fig. 1B. Hence, the inter-residue correlation requires the match of ^{15}N frequencies:

$$\left|f_{ax}^{kres+1} - f_{bx}^{kres}\right| \le \sqrt{\left(\Delta f_{ax}^{kres+1}\right)^2 + \left(\Delta f_{bx}^{kres}\right)^2} \tag{8}$$

We dropped the last subscript in Eq. (8), as the frequencies of indirect dimensions are shared by all carbon sites in the same residue, in either NCACX or NCOCX. Meanwhile, if a specific carbon site *i* in residue *kres* produces resonances in both NCACX and NCOCX, the frequencies along the direct detection dimensions should match:

$$\left|f^{kres}_{azi} - f^{kres}_{bzi}\right| \le \sqrt{\left(\Delta f^{kres}_{azi}\right)^2 + \left(\Delta f^{kres}_{bzi}\right)^2} \quad (9)$$

Moreover, its frequency along the second indirect dimension in 3D NCOCX, should match with the carboxylic carbon's frequency of the same residue detected along the direct detection dimension in 3D NCACX, which is the z-coordinate of carbon site $i = 1$ according to our site notation convention:

$$\left|f^{kres}_{byi} - f^{kres}_{az1}\right| \le \sqrt{\left(\Delta f^{kres}_{az1}\right)^2 + \left(\Delta f^{kres}_{byi}\right)^2} \quad (10)$$

The above equations are the necessary conditions for resonances associated with the same site in the same residues. Usually, Eqs. (4–7) are used together with the characteristic CSs of side-chains to identify RTAs in NCOCX and NCACX. With good spectral resolution, the polarization mixing period could be extended to induce resonances of more side-chain carbons, so the residue type can be distinguished with a greater confidence. However, this strategy is not always applicable to ssNMR spectra with broad resonances (FWHM > 0.5 ppm). Extending the polarization mixing time also induces extra line-broadening, which will exacerbate the signal congestion. Additionally, a longer mixing time will not necessarily produce resonances from more side-chain sites, due to variations of local disorder or dynamics. The in-commutable Hamiltonians of interactions employed to mediate the polarization transfer may also limit the intensity of polarization transferred to a distanced site, referred to as the dipolar truncation problem [39,48,49]. Consequently, it is usually inevitable to end up with incomplete side-chain resonances. As the protein size increases, the narrower CS dispersion along the carboxylic dimension will lead to more congested signals in the NCOCX spectrum, which will incur coincidental alignments of resonances from different residues that satisfy Eqs. (6) and (7). This is why auto-assignment programs depending on the characteristic CSs of side-chain would encounter difficulty with congested ssNMR spectra, [11,25,50,51] as there are too many possibilities to isolate the resonances into different sets of RTAs, while all satisfy the alignment requirement of frequencies along their indirect dimensions.

To illustrate this challenge, Fig. 1C plots the random coil CSs of aliphatic sites of a K, L and R with coincidental alignment of their frequencies along indirect dimensions. Real scenarios would probably be more challenging with incomplete side-chain signals and overlapping resonances. Even with well resolved signals, they could be grouped into multiple different K, L, and R residues, and maybe also D, E, N, or Q assignments. Therefore, when multiple congested regions are present, the total variations of possible RTAs in a NCOCX spectrum could be numerous. Meanwhile, the usage of individual resonances must be tracked to ensure that they are used within the degeneracy values. This is what we refer to as the signal entanglement issue. When using MCAssign or NGSA-II to determine sequential assignments, sorting congested signals in the NCOCX spectrum into different RTA combinations is not incorporated into the random sampling of the MCSA algorithm, which obviously adds to the difficulty to achieve accurate sequential assignment.

Moreover, there are three different types of local minima that can cause erroneous sequential allocations for MCAssign or NGSA-II, even with accurate RTAs from all spectra. First of all, coincidental match of ^{15}N resonances of two residues in either or both of their signals in NCACX and NCOCX spectra would allow one residue position to be occupied by the RTAs of the other residue, but the reverse is not applicable. We refer to this scenario as the type 1 local minimum. Specifically, let's assume that the ^{15}N frequencies of of residue *kres* in NCACX and NCOCX are very close to those of residue *jres,* so that the corresponding RTAs of *kres* satisfy Eq. (8) with the neighbors of *jres* and can be allocated to residue position *jres* by a MC move. Meanwhile, at least one of the RTAs of residue *jres* cannot be allocated to residue *kres* due to their different ^{15}N frequencies. Thus, a MC move to misallocate RTAs of *kres* will probably be allowed, due to the positive score according to Eq. (1). If it happens, at least one of the displaced RTAs of *jres* will not find a position that satisfy Eq. (8), and decreases the total number of n_g. The misplaced RTAs will be corrected eventually, if the system undergoes thorough sampling. Unfortunately, there is no clear instructions for thorough sampling or even how to optimize the annealing setup. As we will show, insufficient sampling will trap RTAs in such kind of type 1 local minima.

Furthermore, when signals of two residues in both spectra possess sufficiently close ^{15}N resonances, each can occupy the other residue position and satisfy Eq. (8) equally well with all neighbors. Such erroneous allocations will not decrease the total number of n_g, and can never be eliminated by the MCSA algorithm, due to the equal scores. There is no mechanism in NGSA-II or MCAssign to detect their likely presence. If their secondary shifts differ significantly, it will lead to a different secondary structure. We refer to this scenario as the type 2 local minimum. If their secondary shifts are sufficiently close, such a swap will not result in a different structure prediction. We refer to this scenario as the type 3 local minimum, which can be safely ignored.

We note that these local minima may also be created by mistaken RTAs as well, the possibility of which cannot be excluded with a congested NCOCX spectrum.

As we will show, ASAP utilizes ARTIST to incorporate the sampling of all possible combinations of RTAs in the congested NCOCX spectrum into the MCSA algorithm, in addition to the sampling of all their possible sequential allocations, to achieve the optimized signal assignments result. ASAP also provides intelligent guidance to optimize the annealing setup for thorough sampling and eradicate erroneous assignments caused by type 1 local minima. ASAP cannot differentiate erroneous assignments caused by type 2 local minima if only NCACX and NCOCX are provided, as they are indistinguishable to the MCSA algorithm. Instead, a list of RTAs implicated in type 2 local minima will be provided by ASAP. This knowledge can be useful to design additional experiments with selective labeled samples to suppress or remove their influence.

2.2. *Unravel the signal entanglement and suppress erroneous RTAs by coincidental alignments by ARTIST*

The flowchart of ASAP is shown by Fig. 2. The front end of ASAP is ARTIST. Its function is to group individual resonances in NCOCX into matched RTAs based on reference RTAs determined in the better resolved NCACX, exploiting the CS dispersion disparity between 3D NCACX and NCOCX spectra. As shown in Fig. 1D and E, owing to the larger c-alpha CS dispersion, even the most congested regions in NCACX still demonstrate reasonable resolution for reliable RTAs, in contrast to the seriously overlapping resonances in NCOCX. Hence, the RTAs determined in NCACX carry a much higher confidence with lower ambiguity.

ARTIST designs two additional tests for NCOCX signals beyond the typical alignment of indirect dimension frequencies. Combined, they leverage this superior resolution in NCACX to eliminate some of the erroneous RTAs caused by coincidental alignment of signals along their indirect dimensions, which would be inevitable if RTAs are determined based on the characteristic CSs of amino acids.

Specifically, given a reference RTA of residue *kres* in NCACX, with signals designated by $\left(f^{kres}_{axj}, f^{kres}_{ayj}, f^{kres}_{azj}\right)$, $j = 1, N_{kres}$. The first test by ARTIST is to make use of the CS of its carboxylic carbon f^{kres}_{az1}, if it is resolved. It searches through all resonances $\left(f_{bxi}, f_{byi}, f_{bzi}\right)$ in NCOCX, and select those with their second indirect dimension f_{byi} aligned to f^{kres}_{az1} according to Eq. (10). If f^{kres}_{az1} is not resolved due to spectral congestion in NCACX, this test will be skipped for this reference RTA. On the other hand, if its f^{kres}_{az1} is resolved, but no match is found among all NCOCX resonances, we declare that this reference RTA does not have a matched RTA in NCOCX.

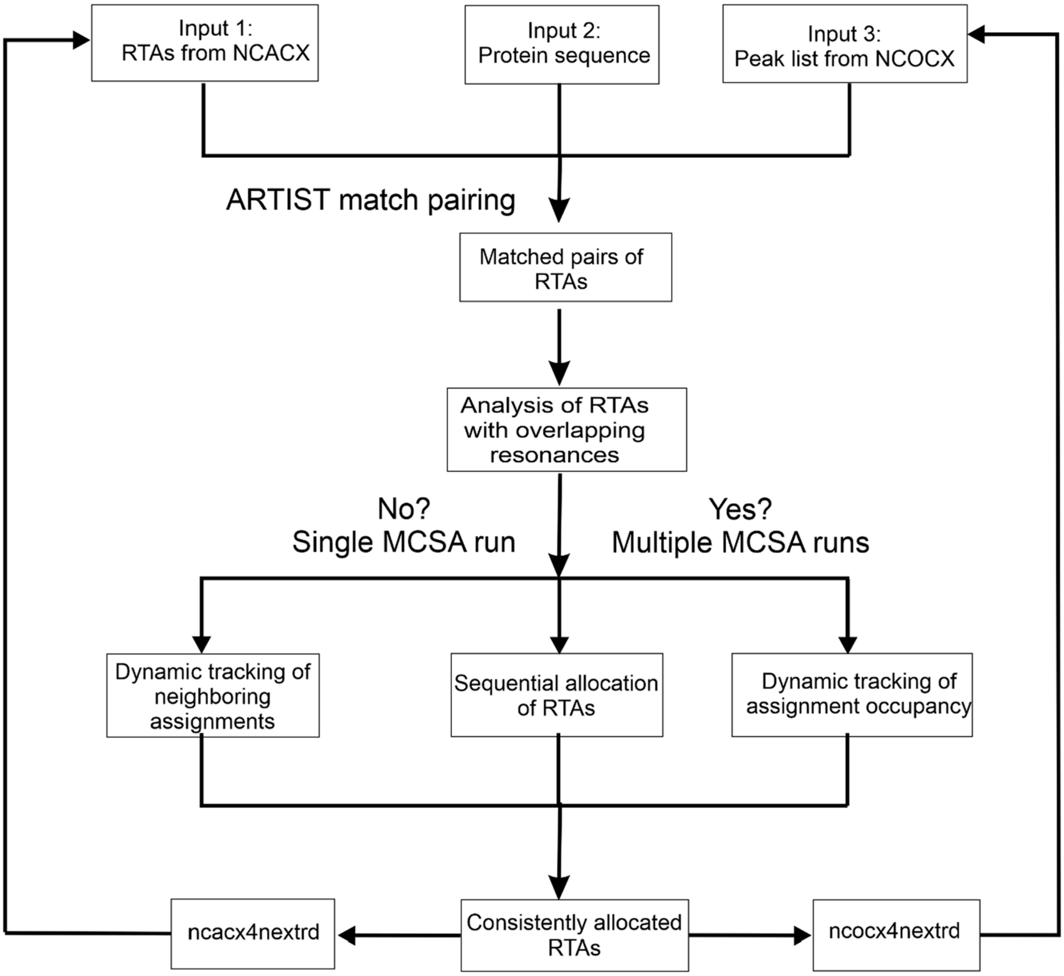

Input 1:
RTAs from NCACX
Input 2:
Protein sequence
Input 3:
Peak list from NCOCX
ARTIST match pairing
Matched pairs of RTAs
Analysis of RTAs with overlapping resonances
No?
Single MCSA run
Yes?
Multiple MCSA runs
Dynamic tracking of neighboring assignments
Sequential allocation of RTAs
Dynamic tracking of assignment occupancy
ncacx4nextrd
Consistently allocated RTAs
ncocx4nextrd

Fig. 2. The general workflow of the ASAP program.

Those resonances surviving the first test will be subjected to the second test, designed to address the incomplete side-chain profile in ssNMR spectra. We introduce a tunable parameter $N_{mandate}$ to demand the number of non-carboxylic carbons signals co-present in both NCACX and NCOCX spectra, counting from c-alpha ($j = 2$). The f_{bzi} of those resonances in NCOCX that passed the first test will be compared with f^{kres}_{azj} of every non-carboxylic carbon specified by $N_{mandate}$ in the reference RTA, according to Eq. (9). If the number of non-carboxylic carbons present in the reference RTA is less than $N_{mandate}$, only those present will be used for the second test. If any of the mandate non-carboxylic carbons in the reference RTA fails to find a match in the NCOCX signals surviving the first test, we declare that this reference RTA does not have a matched RTA in NCOCX.

Despite the uncertainty of side-chain resonances in ssNMR spectra, with an appropriate polarization mixing time, it is achievable that a limited number of sites in most of residues would contribute signals in both NCACX and NCOCX. Therefore, it is reasonable for $N_{mandate}$ to be small, so most of residues qualify for the second test. For instance, by setting it to 2, the second test would require the co-presence of two carbon sites (typically c-alpha and c-beta resonances) in both spectra, which is plausible without a long mixing time to incur extra line broadening. Hence, the second test further filters those resonances that passed the first test by coincidence, while at least partially overcomes the challenges caused by the incomplete side-chain resonances in ssNMR spectra.

Finally, all signals passing test 2 will be divided into different sets. In each set, each of the mandate sites in the reference RTA should have a matched resonance. Resonances in the same set will be subjected to the final test, according to Eqs. (6) and (7), to ensure their frequencies along the indirect dimensions are aligned. If test 3 is successful, signals in this set will be labeled as a matched NCOCX RTA and carry the same residue type ambiguity as the reference RTA in NCACX. If test 3 fails for all possible sets constructed by those resonances surviving test 2, we declare that this reference RTA dos not have a matched RTA in NCOCX.

During these three tests, peak uncertainties along each dimension are used to address the line broadening and possible resonance overlap, as mentioned earlier. Hence, the match pairing tests should retain their accuracy in the presence of peak shifts caused by overlaps between two resonances. However, we note that it is possible for more serious peak shift to take place by overlapping of more than two resonances, which will not be accounted for, unless the uncertainty value is further relaxed along the corresponding dimension associated with signal overlaps, or the peak positions should be determined by some reliable algorithm to deconvolute the overlaps.

We note that the ambiguity of the reference RTAs in NCACX will be transferred to their matched RTAs in NCOCX. This strategy partially avoids the negative effect caused by the resonance congestion in NCOCX. Ideally, if each RTA in NCACX only finds one unique RTA in NCOCX, the inferior spectral resolution of NCOCX then plays no ill-effect in the RTA determination, and NCOCX 100 % inherits the spectral resolution from NCACX. With a congested NCOCX spectrum, usually an RTA in NCACX finds multiple matched RTAs, reflecting the signal congestion in NCOCX. Nonetheless, the inflation of possible combinations of RTAs would be much higher if the NCOCX spectrum were assigned by purely the characteristic CSs of residues, as they will only check the alignment of resonances in NCOCX along their indirect dimensions according to Eqs. (6) and (7), which is the third test in ARTIST. In some auto assignment programs, [11,25,50,51] each possible RTA combinations is assigned with some probability to reflect the confidence of assignment, which depends on various factors, such as the number of side-chain resonances. Considering the unpredictable nature of side-chain resonances that depends on the local structural topology and dynamics in ssNMR, ARTIST does not discriminate any possibilities due to missing side-chain resonances and is more inclusive: individual resonances in a congested regions can be employed in multiple RTA combinations, and all possible RTAs are accounted for as long as they satisfy the three tests. Meanwhile, the criteria used by ARTIST are necessary conditions that the correct RTAs should satisfy, except for the assumption of $N_{mandate}$ that depends on the experimental setup. Therefore, the correct RTAs must be among the matched RTAs identified by ARTIST, which guarantees the inclusion of correct RTAs in a congested NCOCX spectrum, if the mandate sites of these residues all contribute signals in both spectra, and the peak positions are accurately determined.

ARTIST creates a registry to track the usage of individual resonances in construction of all possible matched RTA pairs between NCACX and NCOCX. The correct matched RTA should maximize the total number of n_g. It will be selected by the thorough sampling of MCSA in subsequent sequential assignment, as will be explained next. Consequently, ARTIST eliminates the workload to determine RTAs in congested NCOCX, and the accuracy of RTAs entirely depends on the assignment of the better resolved NCACX and the peak picking algorithm in NCOCX, which is probably more manageable and realistic.

The identification of matched RTAs in NCOCX enables the system to group each reference RTA in NCACX with each of its matched RTAs in NCOCX as individual matched RTA pairs. If no matched RTA is found in NCOCX, this RTA in NCACX is always paired with a null RTA.

After the identification of matched RTA pairs, ARTIST continues to identify those implicated to incur type 2 local minima, recorded in file *Overlap*. The program screens each matched RTA pair against every other pair of the same residue type. Their ^{15}N frequencies in the same spectrum are compared respectively according to Eqs. 4 and 6, with the right side of equations changed to the arithmetic means of uncertainties to ensure no possibility is excluded. Those matched pairs pass the test will be subjected to additional comparison of their secondary shifts. If their differences are larger than 0.5 ppm, these two matched RTA pairs are considered as potential suspects to cause type 2 local minima in the sequential assignment.

2.3. The advantage of allocation of matched RTA pairs by ASAP

The rearrangement of the input data as matched RTA pairs enables the sequential assignment by MCSA to adopt a slightly different design, to enhance the efficiency of sampling and suppress errors due to type 1 local minima.

Specifically, at each MC attempt, after a residue position *kres* is chosen by a random selection, an RTA of the same residue type as *kres* is selected randomly from NCACX. In addition, the program randomly picks one of its matched RTAs in NCOCX to be its matched pair. Together, they are to replace the existing matched RTA pair currently occupying *kres*. The program allows a percentage of the MC attempts to select a NCACX RTA to pair with a null assignment, or a pair of null assignments, to replace the existing matched RTA pairs at *kres*. It is a key strategy to remove erroneous allocated RTAs due to type 1 local minima, to maximize n_g and n_u.

Allocating RTAs in matched pairs brings several advantages. Firstly, it saves the computation to check the compatibility of RTAs from different spectra located at residue *kres*, improving the efficiency. More importantly, this pair-wise allocation of RTAs also increases the penalty to mistakenly allocate those RTAs with one coincidental match of ^{15}N resonances with the neighbors, implicated in type 1 local minima. In MCAssign or NGSA-II, as only one RTA from an individual spectrum is allocated at each MC attempt, if it carries a coincidental match of ^{15}N resonance with its neighbor, the move would probably be accepted according to Eq. (8). Moreover, an RTA with a coincidental match of ^{15}N resonances could be mixed with the correct RTA from another spectrum. By allocation of RTAs in matched pairs, it precludes such erroneous allocations if they have different ^{13}C resonances. Therefore, allocating RTAs in matched pairs eliminates a good fraction of misleading phase space associated with type 1 local minima and essentially "smoothens" the energy landscape for MCSA, compared to the rugged surface filled with false local minima in MCAssign or NGSA-II. As we will show, ASAP always eradicates bad assignments in its results. In contrast, bad assignments may survive in the standard MCSA assignments.

Fundamentally, allocation of RTAs in matched pairs improves the probability to find their correct sequential positions to $20/N_{res}$. In contrast, this probability scales with $\left(\frac{20}{N_{res}}\right)^k$ in MCAssign or NGSA-II, with k as the number of spectra, assuming a uniform amino acid composition in the protein. Combined, ASAP is expected to demonstrate improved efficiency and accuracy than the standard MCSA, which will be shown in our tests by three different proteins.

2.4. Optimization of the annealing setup to ensure thorough sampling

To ensure the accuracy of the MCSA algorithm, the data structure of individual resonances should be simultaneously optimized at two different levels. At the base level, the MCSA algorithm should explore all possible RTAs formed by individual signals. At the secondary level, the MCSA algorithm should explore all possible sequential allocations of RTAs.

MCAssign or NGSA-II does not incorporate the sampling of the phase space of the base level in the MCSA process. For well-resolved spectra, this is not an issue, since accurate RTAs can be obtained. Accurate signal assignments by MCSA would only require sampling of the phase space of the secondary data structure level. This is proved in our test with the SrtC. When spectra become congested, the possibility to group the same signals into different sets of RTAs quickly multiples, which leads to the proliferation of local minima at the base level of RTAs. Successful signal assignments by MCSA then demand sampling of the phase space of both data structure levels. When using MCAssign and NGSA-II, users have to manual revise RTAs, which is inefficient and susceptible to error. As explained earlier, the treatment of RTAs from different spectra as independent inputs creates extra local minima and additional phase space, which exacerbates the situation. Moreover, there is no intelligent instruction or guidance to ensure optimized annealing or thorough sampling. Even given correct and unambiguous RTAs, as we will show, for complicated systems like MLKL or RSV CA, MCAssign will face greater challenge to make correct assignments.

At the end of each MC move in ASAP, the change of scores caused by allocations of a new matched RTA pair is computed according to Eq. (1). If it is accepted, the system updates the registry that tracks the usage of the NCACX RTA in the matched pair, and individual NCOCX signals in the matched NCOCX RTA. Thus, the sampling of the phase space in both data structure levels is simultaneously incorporated in the MCSA process, and the signal entanglement issue is unwounded in the congested NCOCX spectrum. Therefore, given thorough sampling, the MCSA algorithm in ASAP will find the optimal combinations of signals into the RTA configurations that maximize n_g and n_u, minimizing n_e and n_b, the global minimum.

To achieve thorough sampling, the setup must have sufficient MC moves to allow reallocations of matched RTA pairs with coincidental matches of ^{15}N resonances forming good connections with both neighbors, which is essential to eradicate type 1 local minima to maximize n_g and n_u. The most likely trajectory is to replace it by a pair of null assignments. It can be shown that the least number of MC attempts at annealing step i for this to happen is:

$$N_a\left(n_g = 2\right) = \frac{3N_{res}^2}{10} e^{-\frac{S^i(-2,0,-2,-1)}{2}} \tag{11}$$

here N_{res} is the number of residues in the protein. $S^i(-2,0,-2,-1)$ corresponds to the score defined by Eq. (1). Meanwhile, the system should prohibit reallocations of RTAs forming one good connections with its neighbors towards the end of annealing. Similarly, it can be shown that the least number of MC attempts at annealing step i for this to happen is:

$$N_a\left(n_g = 1\right) = \frac{3N_{res}^2}{10} e^{-\frac{S^i(-1,0,0,-1)}{2}} \tag{12}$$

here $S^i(-1,0,0,-1)$ corresponds to the score defined by Eq. (1). Eqs. (11) and (12) can be used to guide the system setup for optimized annealing. Detailed derivations are described in the supporting information.

There is not a unique set of parameters to satisfy Eqs. (11) and (12), as we will demonstrate. But we would like to propose a general guideline, which will be validated in our results section. Firstly, the coefficient for bad connections w_2^i should outweigh all the rest. A single bad sequential assignment misplaces at least two pairs of RTAs from NCACX and NCOCX. Hence, we set its value to be 2.5 times of w_1^i, the coefficient for good connections. As we will show, it eradicates assignments forming bad connections. Meanwhile, w_1^i is the main positive drive to maximize the total sequential assignment, so we set it twice of w_4^i, the coefficient for used signals, which encourages the maximum usage of signals. The least important is the edge connections, which would be minimized naturally if most signals are sequentially allocated. Therefore, we set w_3^i as half of w_4^i. The penalty from w_3^i encourages continuously distributed good connections, so it is preferred to be at least nonzero.

To verify Eqs. (11) and (12), two matrices *neighbor_t* and *occupancy_step* are created to track the dynamic migration of RTAs. Specifically, *neighbor_t* and *occupancy_step* are a *npeak_nca* × *npeak_nca* × *nstep* and a *npeak_nca* × N_{res} × *nstep* 3D matrix. Here *npeak_nca* refers to the number of RTAs in NCACX. At the i-th annealing step, assume that a MC move is accepted to allocate an NCACX RTA j with its matched pair from NCOCX to residue *kres*. At this time, its neighboring residue positions *kres* ± 1 are occupied by matched RTA pairs with their NCACX RTAs identified as k and l, respectively. Here *j, k, l* are the indexes of RTAs in the input NCACX RTA file. Therefore, the correlation between the newly allocated matched RTA pair and its neighbors can be registered by adding 1 to the components at the l and k-th columns, j-th row and i-th stack in *neighbor_t*. Likewise, this change of occupancy at *kres* residue position can be registered by 1 increment to the entry at j-th column, *kres*-th row and i-th stack stack in *occupancy_step*. Thus, *neighbor_t* tracks the dynamic and correlated migration of RTAs, and *occupancy_step* records the dynamic residency of each residue position. To get a coarse-grained view, the total dynamic migration of each RTA and occupancy at each residue position along the annealing process are also computed by summing all columns, and recorded in matrices *instigator* and *occupancy_sum*. In case of parallel simulations, these matrices report their averaged values. Therefore, they provide a direct visual confirmation of Eqs. (11) and (12), which will be shown in our results section.

We note that the optimal setup optimizes the possibility to achieve thorough sampling but does not guarantee it. Practically, the direct conformation of thorough sampling is to conduct a series of sequential assignments, with optimized setup and progressively increasing n_t, the MC attempts per annealing step. As thorough sampling removes RTAs trapped in type 1 local minima, n_g and n_u continue to grow until thorough sampling is achieved. As we will show, for proteins ~ 150 residues with spectral quality comparable to MLKL or SrtC, it requires $n_t = 50$ million. For proteins of ~ 250 residues with spectral resolution comparable to the RSV CA, it requires $n_t = 500$ million. Hence, n_t increases with increasing spectral complexity, protein composition and size for a single ASAP or MCSA simulation. Alternatively, the search for the global minimum may be achieved via iterative parallel short simulations instead of a single long simulation.

2.5. Iterative ASAP simulations for accelerated convergence towards the global minimum and eliminations of local minima

The signal assignments result by a single MCSA simulation without thorough sampling is certain to contain some mistakenly assigned RTAs even in the absence of explicitly bad connections, due to the existence of type 1 local minima. Meanwhile, the actual trajectories of MCSA simulations differ due to the random shuffling of RTAs in the annealing process. If n_t is insufficient for thorough sampling, but most residues still find their correct RTAs, and the RTAs trapped in local minima are minorities. Among parallel MCSA simulations, consistently allocated RTAs probably correspond those assigned correctly, and the positions of erroneous allocated RTAs should vary. Therefore, instead of a long simulation, correct assignment can be obtained by iterative rounds of MCSA simulations with a less-than-optimal n_t, each round comprising multiple parallel simulations. The positions of consistently allocated RTAs will be fixed in subsequent simulations. As will be shown in our results section, accelerated convergence towards the global minimum can be achieved.

It is difficult to provide a universal parametrization for this approach. It depends on the details of the system, including the protein sequence composition and spectral resolution. A general rule of thumb is, a successful iterative MCSA simulation should keep removing RTAs trapped in type 1 local minima, thus more correctly assigned RTAs are expected in each subsequent round, and eventually plateau. On the other hand, if n_t is too low and the local minima are too populated, some of the mistakenly assigned RTAs will coincidentally occupy the same residue positions in all parallel simulations and be retained in subsequent iterations, which will lead to decreasing n_u and n_g. We will demonstrate examples in our results sections, both positive and negative. Additionally, only those NCACX RTA paired with a valid NCOCX RTA should be counted towards consistently allocated assignments before the last iteration, because RTAs involved in edge connections are more likely due to a coincidental match, and persist until the end of each simulation.

Type 1 local minima are more populated in MCAssign or NGSA-II, due to uncorrelated allocations of RTAs from different spectra. Iterative parallel simulations will face a greater challenge, as shown by our test with the level 2 ambiguity in the input data of the RSV CA by MCAssign.

Compared to MCAssign, NGSA-II executes many parallel sequential assignments by MCSA with an extra random weight factors to modulate the score function, so more diverse solutions with comparable number of good assignments can be identified. Essentially, this random variation surveys different trajectories towards the global minimum. Hence, NGSA-II is a variant of iterative MCAssign simulations with a controlled selection mechanism between iterative generations. As expected, the requirement of thorough sampling is relaxed, which exhibits improved performance when tested on small proteins (100 residue or less) [36]. With increasing protein sizes and spectral congestion, under sampling may still lead to erroneous results in iterative MCSA simulations, as we will show with tests by level 2 ambiguity in the RSV CA data. In addition, it demands accurate RTAs from all spectra and doesn't incorporate sampling of all possible RTAs in the congested NCOCX in the MCSA process. Hence, all performance comparisons are tested against the MCAssign program with a single simulation in this work.

2.6. Overall workflow of ASAP, input and output files

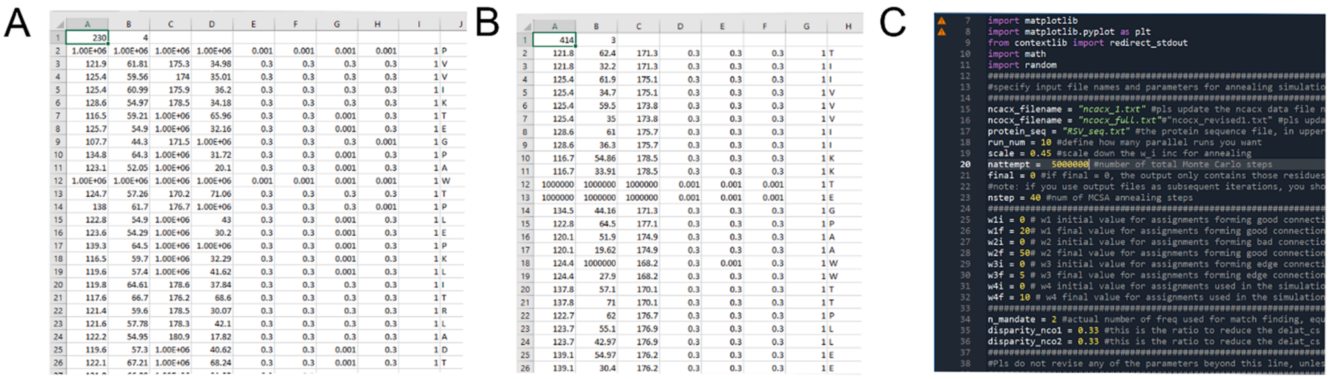
The ASAP program is coded in python, and requires three input files, shown in Fig. 2. Input 1 holds the RTAs in the 3D NCACX spectrum, input 2 is the protein sequence, and input 3 contains the peak list in the 3D NCOCX spectrum, specified by in the program by *NCACX_filename, protein_seq,* and *NCOCX_filename*, respectively, shown in Fig. 3C.

Input 1 and 3 adopt an identical format, as shown in Fig. 3A and B. The first row holds two numbers, separated by a tab or space. The first number denotes the number of RTAs in input 1 or individual resonances in input 3. The second number denotes the number of CSs per entry in input 1 or 3. The maximum allowed CSs is 7, which could specify up to 5 non-carboxylic carbon CSs. Starting from the second row, the CSs of each entry are listed in the order of their ^{15}N, c-alpha, carboxylic carbon, and any additional carbons. Their CS uncertainties are listed at corresponding columns after the last CS entry, estimated to be ~ ½ FWHM of the spectral linewidth. In input 3*,* if only three CSs of one resonance are listed per row, they should be listed in the order of f_{bxi},f_{bzi},f_{byi}, where the CS coordinate of the non-carboxylic site along the direct detected dimension is entered as the second column. For both files, if the CS of a particular site is unknown, that entry is filled by *1e6*, with its uncertainty set to *0.001*. The last two columns of each row are the signal degeneracy and its RTA. Here signal degeneracy refers to how many carbon sites actually contribute signals to this resonance. At the beginning of an ASAP simulation, the RTA column in input 3 is just a place holder. In contrast, this field in input 1 describes its RTA in upper-cased single letters. Ambiguous RTAs are accepted as consecutive upper-cased single letters representing each possible assignment.

If the sequential allocation is known (definitely assigned), the RTA column should be the upper-cased single letter for the residue type followed by its numeric position in both files. Definitely assigned RTAs and signals are exempted from match pairing by ARTIST, and their sequential positions are fixed in subsequent sequential assignment. Additionally, multiple resonances in NCOCX can be entered at a single row in input 3, as long as they share the indirect dimension frequencies and degeneracy values. Their common ^{15}N and carboxylic carbon frequencies should be entered only once at the same row at the first and third columns, with the CSs of additional non-carboxylic carbons listed after the carboxylic carbon entry. The program will parse them into individual resonances $\left(f_{bxi},f_{byi},f_{bzi}\right)$.

To perform an ASAP simulation, all input files should be located in the same data folder as the ASAP script. In addition to the three input files, users are expected to specify the following parameters, directly at the beginning section of the code, as shown in Fig. 3C:

1. *NCACX_filename:* the file name of input 1, the RTAs in the NCACX spectrum.
2. *NCOCX_filename:* the file name of input 3, the peak list in the NCOCX spectrum.
3. *protein_seq:* the file name of input 2, comprising upper-cased single letter abbreviations of the amino acid sequence.
4. *run_num*: the number of parallel sequential assignment simulations by ASAP.
5. *scale*: the variable *scale* in Eq. (2), to control the annealing slope.
6. *nattempt:* the MC attempts per annealing step in the MCSA simulation, corresponds to n_t in our discussion. For iterative simulations with parallel jobs, the recommended value is 5 million for proteins with 150——250 residues.
7. *final*: a flag to control which kind of sequentially allocated RTAs in parallel simulations will be labeled as definitely assigned signals in output files. If $final = 0$, only those NCACX RTAs paired with a valid NCOCX matched RTA will be counted towards consistently assigned signals. If $final = 1$, all consistently allocated RTAs will be treated as definitely assigned signals. For better accuracy, it should be set to zero for iterative simulations, and revised to 1 for the final iteration.
8. *nstep*: the number of total annealing steps in MCSA. The recommended value is 40.
9. *w1f to w4f:* correspond to variables w_{1f} *to* w_{4f} in Eq. (2), to control the penalty or bonus of n_g, n_b, n_e, and n_u. The recommended values for these coefficients are 20, 50, 10, and 5, respectively.
10. *N_mandate*: the number of mandatory non-carboxylic carbon sites in the reference RTA to find a match in the second test of ARTIST. It is recommended to be set to 2, so only the CSs of two non-carboxylic carbon sites (typically c-alpha and c-beta) in the reference NCACX RTA are required to be matched by signals in NCOCX. All results in this work use $N\_mandate = 2$.
11. *disparity_nco1*: a positive number ≤ 1. The multiplication of *disparity_nco1* with the listed uncertainty of the ^{15}N CS in input 3 sets the uncertainty Δf_{bxi}^{kres} in the final test according to Eq. 6. Usually, signals of the same site in the same residue in different spectra may exhibit some deviation, due to factors such as variations of sample conditions during spectra acquisition, field calibrations. This is captured by the uncertainty values listed in input 1 and 3, used for the first two tests in ARTIST according to Eqs. 9 and 10. However, when testing if signals belong to the same residue in the same spectrum, the alignment along their indirect dimensions should have a much tighter tolerance, no more than 0.2 ppm, or typically 0.1 ppm.
12. *disparity_nco2*: a positive number smaller ≤ 1. The multiplication of *disparity_nco2* with the listed uncertainty of the carboxylic carbon CS in input file 3 sets the uncertainty Δf_{byi}^{kres} in the final test according to Eq. 7. Just like *disparity_nco1*, it modifies the tolerance of alignment of the second indirect dimension of resonances belonging to the same residue in NCOCX.



	A	B	C	D	E	F	G	H	I	J
1	230	4								
2	1.00E+06	1.00E+06	1.00E+06	1.00E+06	0.001	0.001	0.001	0.001	1	P
3	121.9	61.81	175.3	34.98	0.3	0.3	0.3	0.3	1	V
4	125.4	59.56	174	35.01	0.3	0.3	0.3	0.3	1	V
5	125.4	60.99	175.9	36.2	0.3	0.3	0.3	0.3	1	I
6	128.6	54.97	178.5	34.18	0.3	0.3	0.3	0.3	1	K
7	116.5	59.21	1.00E+06	65.96	0.3	0.3	0.001	0.3	1	T
8	125.7	54.9	1.00E+06	32.16	0.3	0.3	0.001	0.3	1	E
9	107.7	44.3	171.5	1.00E+06	0.3	0.3	0.3	0.001	1	G
10	134.8	64.3	1.00E+06	31.72	0.3	0.3	0.001	0.3	1	P
11	123.1	52.05	1.00E+06	20.1	0.3	0.3	0.001	0.3	1	A
12	1.00E+06	1.00E+06	1.00E+06	1.00E+06	0.001	0.001	0.001	0.001	1	W
13	124.7	57.26	170.2	71.06	0.3	0.3	0.3	0.3	1	T
14	138	61.7	176.7	1.00E+06	0.3	0.3	0.3	0.001	1	P
15	122.8	54.9	1.00E+06	43	0.3	0.3	0.001	0.3	1	L
16	123.6	54.29	1.00E+06	30.2	0.3	0.3	0.001	0.3	1	E
17	139.3	64.5	1.00E+06	1.00E+06	0.3	0.3	0.001	0.3	1	P
18	116.5	59.7	1.00E+06	32.29	0.3	0.3	0.001	0.3	1	K
19	119.6	57.4	1.00E+06	41.62	0.3	0.3	0.001	0.3	1	L
20	119.8	64.61	178.6	37.84	0.3	0.3	0.3	0.3	1	I
21	117.6	66.7	176.2	68.6	0.3	0.3	0.3	0.3	1	T
22	121.4	59.6	178.5	30.07	0.3	0.3	0.3	0.3	1	R
23	121.6	57.78	178.3	42.1	0.3	0.3	0.3	0.3	1	L
24	122.2	54.95	180.9	17.82	0.3	0.3	0.3	0.3	1	A
25	119.6	57.3	1.00E+06	40.62	0.3	0.3	0.001	0.3	1	D
26	122.1	67.21	1.00E+06	68.24	0.3	0.3	0.001	0.3	1	T

	A	B	C	D	E	F	G	H
1	414	3						
2	121.8	62.4	171.3	0.3	0.3	0.3	1	T
3	121.8	32.2	171.3	0.3	0.3	0.3	1	I
4	125.4	61.9	175.1	0.3	0.3	0.3	1	I
5	125.4	34.7	175.1	0.3	0.3	0.3	1	V
6	125.4	59.5	173.8	0.3	0.3	0.3	1	V
7	125.4	35	173.8	0.3	0.3	0.3	1	V
8	128.6	61	175.7	0.3	0.3	0.3	1	I
9	128.6	36.3	175.7	0.3	0.3	0.3	1	I
10	116.7	54.86	178.5	0.3	0.3	0.3	1	K
11	116.7	33.91	178.5	0.3	0.3	0.3	1	K
12	1000000	1000000	1000000	0.001	0.001	0.001	1	T
13	1000000	1000000	1000000	0.001	0.001	0.001	1	E
14	134.5	44.16	171.3	0.3	0.3	0.3	1	G
15	122.8	64.5	177.1	0.3	0.3	0.3	1	P
16	120.1	51.9	174.9	0.3	0.3	0.3	1	A
17	120.1	19.62	174.9	0.3	0.3	0.3	1	A
18	124.4	1000000	168.2	0.3	0.001	0.3	1	W
19	124.4	27.9	168.2	0.3	0.3	0.3	1	W
20	137.8	57.1	170.1	0.3	0.3	0.3	1	T
21	137.8	71	170.1	0.3	0.3	0.3	1	T
22	122.7	62	176.7	0.3	0.3	0.3	1	P
23	123.7	55.1	176.9	0.3	0.3	0.3	1	L
24	123.7	42.97	176.9	0.3	0.3	0.3	1	L
25	139.1	54.97	176.2	0.3	0.3	0.3	1	E
26	139.1	30.4	176.2	0.3	0.3	0.3	1	E

```
import matplotlib
import matplotlib.pyplot as plt
from contextlib import redirect_stdout
import math
import random
################################################################
#specify input file names and parameters for annealing simulatio
################################################################
ncacx_filename = "ncacx_1.txt" #pls update the ncacx data file n
ncocx_filename = "ncocx_full.txt"#"ncocx_revised1.txt" #pls upda
protein_seq = "RSV_seq.txt" #the protein sequence file, in upper
run_num = 10 #define how many parallel runs you want
scale = 0.45 #scale down the w_i inc for annealing
nattempt =  5000000 #number of total Monte Carlo steps
final = 0 #if final = 0, the output only contains those residues
#note: if you use output files as subsequent iterations, you sho
nstep = 40 #num of MCSA annealing steps
################################################################
w1i = 0 # w1 initial value for assignments forming good connecti
w1f = 20# w1 final value for assignments forming good connection
w2i = 0 # w2 initial value for assignments forming bad connectio
w2f = 50# w2 final value for assignments forming good connection
w3i = 0 # w3 initial value for assignments forming edge connecti
w3f = 5 # w3 final value for assignments forming edge connection
w4i = 0 # w4 initial value for assignments used in the simulatio
w4f = 10 # w4 final value for assignments used in the simulation
################################################################
n_mandate = 2 #actual number of freq used for match finding, equ
disparity_nco1 = 0.33 #this is the ratio to reduce the delat_cs
disparity_nco2 = 0.33 #this is the ratio to reduce the delat_cs
################################################################
#Pls do not revise any of the parameters beyond this line, unles
```

Fig. 3. Input files for ASAP program. (A). A snapshot of input 1 file that supplies the RTAs in the 3D NCACX spectrum. (B). A snapshot of input 3 file that specifies the individual resonances in the 3D NCOCX spectrum. (C). A snapshot of the ASAP script where users specify the values of 12 parameters introduced in our discussion.

In practice, users only need to update the values for parameters listed in 1 to 7, keeping the rest as recommended. *Disparity_nco1* and *disparity_nco2* may be revised according to the spectral quality or users' preferred rigor.

Before initiating an ASAP simulation, users are advised to use python script *Plot_survivability* based on Eqs. (11) and (12) to set parameters *scale* and *nattempt*, demonstrated in Fig. 4D. Alternatively, they can refer their project parameters (protein size and amino acid composition, spectral quality) to the examples in this work to set these parameters properly. Optimized values of *scale* and *nattempt* should maximize the number of effective annealing steps. In simple words, it should maximize the annealing steps with *nattempt* $\geq N_a(n_g = 2)$. Meanwhile, there should be a few annealing steps with *nattempt* $\leq N_a(n_g = 1)$.

The general workflow of ASAP is described by Fig. 2. The program first uses ARTST to find the matched RTAs in NCOCX for each reference RTA in NCACX. The results are summarized in the file *NCO_MatchSummary*, as shown in Fig. S1, with details listed in file *NCO_MatchDetail*. The number of matched RTAs for each RTA in NCACX is stored in file *knmatch_rd*, which is used to generate the plot shown in Fig. 4A by python script *Plot_nmatch*.

After match pairing, ARTIST identifies those RTA matched pairs implicated in type 2 local minima, and records the number in file *Overlap*. Users can inspect the content, as plotted in Fig. S2.

Next, the program proceeds to determine the sequential allocations of these matched RTA pairs by the MCSA algorithm as described above. If the protein is small and ambiguity in RTAs is low, most of the RTAs finds a unique matched RTA in NCOCX, with low values in *Overlap*, users may try a single MCSA simulation with a high *nattempt*. However, most of ssNMR projects probably need iterative parallel simulations due to the less-than-ideal spectral quality. The progress of the ASAP simulation is recorded in file *runrecord*. When each parallel simulation ends, the final values of n_g, n_b, n_e, n_u are recorded in file *runsummary*. Each parallel simulation also updates the RTA columns of RTAs that are successfully assigned, reported in *NCACXbknum* and *NCOCXbknum*, in the same format as the input files, where *num* is the job index in the parallel simulations. When all parallel simulations are completed, the program identifies those RTAs being consistently allocated and generate another pair of output files *NCACX4nextrd* and *NCOCX4nextrd*. In these files, only the RTA columns of those consistently assigned RTAs are revised to their allocated residue positions.

The consistently allocated RTAs of all parallel simulations can be plotted together with the number of matched RTAs for each NCACX RTA by python script *Plot_nmatch*, as shown in Fig. 4A. The progress of n_g, $n_{b,}n_e$, n_u along an MCSA simulation can be plotted by python script *Plot_numbers*, as shown in Fig. 4B. The dynamic occupancy at each residue position can be visualized by python script *Plot_occupancy_sum*, as shown in Fig. 4C.

After all parallel simulations end, if a new iteration should be performed, *protein_seq, NCACX4nextrd* and *NCOCX4nextrd* can be copied to a new folder, together with the ASAP script. The names of *NCACX4nextrd* and *NCOCX4nextrd* should be revised as the new input 1 and 3 to start the new iteration.

3. Results and discussion

3.1. Construction of input files

Three proteins are used to test the resilience of ASAP against ambiguous RTAs vs MCAssign: the 237-residue RSV CA tubular assembly, [13] and the catalytic domain of *Bacillus anthracis sortase* protein SrtC (147 residues) and the N-terminal domain of the mixed-lineage kinas domain-like protein MLKL (166 residues) [40].

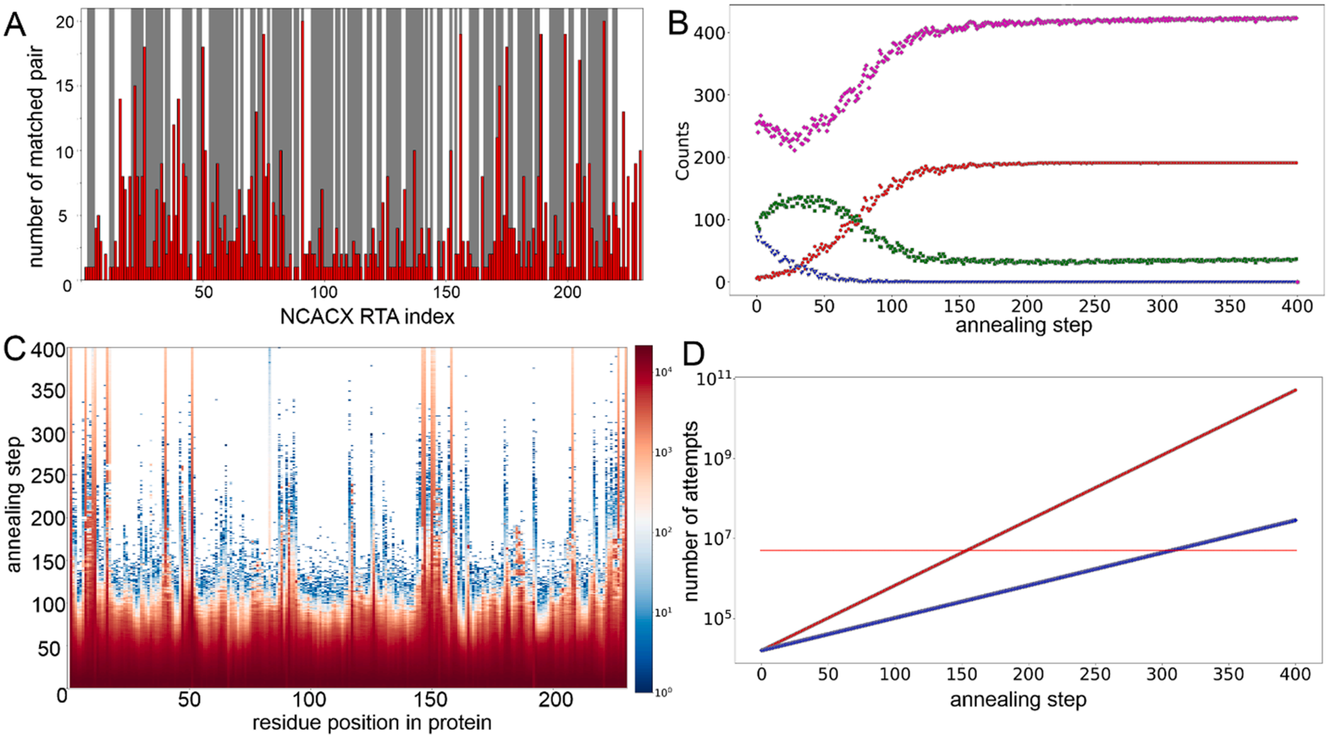


Fig. 4. Optimization of the ASAP setup. (A) Number of matched RTAs identified by ARTIST in NCOCX for each RTA in NCACX (red filled bars). The dark shaded positions are RTAs consistently allocated to the same residue positions in all parallel simulations in run 6 in Table 1. (B) The progress of the total number of good (red circles), bad (blue triangles), edge connections (green squares) and used RTAs (purple diamonds) along the annealing process. (C). The number of re-allocated occupancies at each residue positions along the annealing process. The scale bar to the right denotes the re-allocation frequency. (D). The minimum number of MC attempts to guarantee at least one successful move to remove a matched RTA pair forming two good connections (red tilted line) or one good connection (blue tilted line), described by Eqs. (11) and (12) respectively. The thin red horizontal line is the number of MC attempts at each annealing steps. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Specifically, the NCACX and NCOCX spectra of the RSV CA were acquired at the 900 MHz field at National High Magnetic Field Laboratory, with 50 ms and 150 ms DARR mixing [52]. They exhibit ~ 0.6–0.75 ppm linewidth along the directly detected ^{13}C dimension, and 0.8–1 ppm along the ^{15}N dimension, which justifies our setting of 0.3 ppm CS uncertainties. 230 RTAs were determined by multiple samples of different isotopic labeling patterns, including the uniform ^{13}C,^{15}N labeled sample, sparsely ^{13}C labeled samples by 1–1,3-^{13}C and 2-^{13}C glycerol, and R and L selective residue ^{13}C,^{15}N labeled samples [13]. The frequencies of individual resonances were determined by Poky in each spectrum [53]. They are used to construct input 1 and 3 for ASAP simulations. The RTA columns of input 1 for ASAP hold their experimentally determined RTAs, while the corresponding field in input 3 is merely a place holder. Meanwhile, MCAssign simulations use the experimentally determined RTAs in all input files.

The input files for tests on the 147-residue SrtC and 166-residue MLKL are prepared based on their assignments deposited at Biological Magnetic Resonance Bank (BMRB). The structure of these proteins were solved by Robson *et al.* (PDB 2LN7, BMRB 18152) [41] and Su *et al.* (PDB 2MSV, BMRB 25135) [42]. They are smaller than the RSV CA, but still considerably larger than the peptide samples for ssNMR. Specifically, the input 1 file of SrtC contains 119 RTAs, for residues 7–21, 23–37, 43–53, 55–64, 75–114, 116–121, 125–131, and 133–147. The input 3 file of SrtC contains the signals from 116 residues, for residues 7–20, 22–36, 39, 43–52, 54–64, 75–113, 115–120, 125–130, 133–146. The input 1 file of MLKL contains 148 RTAs, for residues 13–52, 54–65, 67–106, 109–134, 136–137, and 139–166. The input 3 file of MLKL contains signals from 148 residues, for residues 13–51, 53–64, 66–106, 108–133, 135–136, 139–165. Hence, they are a good representation of typical ssNMR data with signals missing from certain stretches. Furthermore, the SrtC structure comprises mostly β-stand structure, and that of the MLKL is dominated by α-helices. Therefore, they represent proteins with different secondary structure compositions. For both samples, we assume that their spectral linewidth is comparable to our RSV CA sample ~ 0.6 ppm, with the uncertainties ~ 0.3 ppm. Only the CSs of ^{15}N, carboxylic carbon, c-alpha and c-beta are used to construct input files for ASAP simulations following the format specification as described. The RTA columns of input 1 hold their experimentally determined RTAs, while the corresponding field in input 3 is merely a place holder. Meanwhile, MCAssign simulations use the experimentally determined RTAs in all input files.

To test the resilience of ASAP simulations against ambiguous RTAs, we adopt two levels of ambiguity in RTAs. At the first level, we adopt the same type of ambiguity described by Tycko in his work [40]: E or Q are assigned as EQ, W or H are assigned as HW, D or N are assigned as DN, and F or Y are assigned as FY. Together, it amounts to 21.7 %, 32.8 %, and 34.5 % RTAs in input 1 to be ambiguous for the RSV CA, SrtC and MLKL, respectively. At the second level, all D, N, E and Q residues are labeled as DENQ in the RTA columns.

3.2. Input data analysis and match pairing by ARTIST

We first test the match pairing function of ARTIST with the RSV CA data. The number of matched RTAs identified in NCOCX for each NCACX RTA is plotted in Fig. 4A. Briefly, 20 RTAs in NCACX fail to find any match in NCOCX, they are colored yellow in *NCACX_1CompareNCO* in supporting materials. They are residues missing mandatory resonances in NCOCX, or those with CS deviations slightly beyond the specified uncertainty that fail the first two tests, highlighted in red. They reflect the robotic (or rigorous) aspect of ARTIST to identify matched RTAs. In practice, human interventions can easily salvage such obvious outliers.

There are additional 30 RTAs in experimentally determined lists with fewer number of mandatory carbon signals in NCOCX than NCACX, or at least one of their CS deviations beyond the specified values, colored in orange in *NCACX_1CompareNCO*. They should not find a match in NCOCX. However, with *disparity_nco1* and *disparity_nco2* set to 1, ARTIST still identifies at least one match for these entries, by mixing signals in NCOCX that are experimentally assigned to other residues, but nonetheless satisfy all three tests in ARTIST. Experimentally, these signals are assigned based on a holistic evaluation of additional side-chain signals, either from the same spectrum of the uniform labeled sample, or from those of other samples. This additional information is not provided to ARTIST. It testifies the complexity induced by incomplete side-chains and signal entanglement in ssNMR, and the ability of ARTIST to account for all possible combinations. We also highlight four pairs of neighboring NCACX and NCOCX assignments with bad connections. They should not be sequentially assigned based on the large deviation of ^{15}N frequencies. Therefore, if we count the remaining consecutively connected experimental sequential assignments in input 1 and 3, the total number of n_g should be 150 residues. However, it may be possible to achieve a higher n_g by ASAP, as ARTIST is able to enumerate all possible combinations of signals in NCOCX for each RTA in NCACX. On the other hand, because simulations by MCAssign use actual experimentally determined RTAs from both spectra as inputs, this analysis suggests that the maximum total number of n_g should be ~ 452 for MCAssign simulations, since RTAs from NCACX and NCOCX are counted individually.

ARTIST also reports the number of matched RTA pairs that can be assigned interchangeably due to overlapping ^{15}N frequencies in file *Overlap*. The histogram of this statistics is plotted as the red bars in Fig. S2A. Altogether, 51 NCACX RTAs and their matched NCOCX RTAs find at least one matched RTA pair with overlapping ^{15}N resonances, with sufficiently different secondary shifts to potentially incur type 2 local minima. Among them, four matched RTA pairs are reported to have 5 sets of RTA pairs that can be assigned interchangeably, corresponding to A68, A84, A93 and A185.

3.3. Validating the optimization strategy for annealing setup

Following ARTIST's match finding, ASAP proceeds to perform sequential allocation by the MCSA algorithm. Our first task is to validate Eqs. (11) and (12) to optimize the MCSA setup. Following our guidelines, we set $n_t = 5 \times 10^6$, $w_{1f} = 10$, $w_{2f} = 25$, $w_{3f} = 2.5$, and $w_{4f} = 5$, with $scale = 1.0$. Only c-alpha and c-beta resonances of the NCOCX spectrum of the uniform ^{13}C, ^{15}N labeled RSV CA sample are included in input 3. For better resolution to track the migration of matched RTA pairs during the annealing process, a single MCSA simulation is performed with $nstep = 400$. The progress of n_g, n_b, n_e and n_u is plotted in Fig. 4B as red circles, blue triangles, green squares, and purple diamonds, respectively. N_g smoothly increases to 191 after about 150 annealing steps. It surpasses the estimated lower bound of 150 good connections, proving the power of ARTIST to find all possible matched RTAs in the congested NCOCX for given reference RTAs in NCACX, and the ability of MCSA to unwind resonance entanglements to maximize n_g. Meanwhile, n_b drops to zero quickly, and n_e stabilizes after the first 150 annealing steps. The fluctuation of n_e gradually damps down to ± 1 beyond annealing step 350.

To visualize the annealing progress, *occupancy_sum* is plotted in Fig. 4C. At the initial stage, every residue position exhibits high counts of re-allocations in red. As the annealing progresses, the penalty to relocate RTAs with good connections gradually increases, so those matched RTA pairs making good connections with both neighbors permanently reside at their positions, manifested as the fading red to blue and eventually to blank, at about ~ 150 steps. It agrees well with the predicted values $n_t = N_a(n_g = 2) = 153$ according to Eq. 11, shown in Fig. 4D. Beyond this step, RTAs forming only one good connection with their neighbors can still be re-allocated, but with decreasing frequency. This is evident from the gradually fading blue intensity between annealing steps 300–350, also consistent with the $n_t = N_a(n_g = 1) = 307$ prediction by Eq. 12, shown in Fig. 4D. There are some residue positions being frequently accessed until the end of annealing. These are positions that cannot find a good assignment.

In summary, this test validates Eqs. (11) and (12) as a crude approximation to optimize the annealing setup for MCSA. In addition, the re-allocation frequency of each RTA can also be visualized by script *Plot_instigator*, shown in Fig. S3. It shows which RTAs are seated properly. At each annealing step, the re-occupancy frequency at each residue position in the protein can be plotted by script *Plot_occupancy_step*, as shown by Fig. S4A and B for annealing step 5 and 400, respectively. Likewise, RTAs allocated to the preceding and next residue positions (indexed by x coordinates) to the specified RTAs (indexed by the y coordinate) can be visualized by plotting file *neighbor* by script *Plot_neighbor*, as shown by Fig. S4C and D for annealing step 5 and 400, respectively. They may be helpful to identify NCACX RTAs that need revisions.

3.4. Validating thorough sampling and differentiation of the global minimum and local minima

To validate the guideline for thorough sampling, we conduct four series of ASAP runs with the same input data of the RSV CA tubular assembly in the previous section, comprising run 1–4, 5–8, 10–11 and 12, respectively, shown in Table 1. All simulations are performed with $w_{1f} = 20$, $w_{2f} = 50$, $w_{3f} = 5$, $w_{4f} = 10$, $nstep = 40$, with n_t increased from 0.5×10^6 to 500×10^6 to probe the progress of assigned RTAs approaching thorough sampling. For different runs among each of the first three series, the *scale* values are adjusted to modulate $n_t = N_a(n_g = 2)$ and $n_t = N_a(n_g = 1)$, to vary the effective annealing steps. The $N_a(n_g = 2)$ and $n_t = N_a(n_g = 1)$ values are computed according to Eqs. (11) and (12). The results reported in the table are the average of 10 simulations, to average the randomness associated with MC simulations, except the last run, as it took 48 h to complete a single simulation.

As shown in Table 1, in each series, the effective annealing range expands as the *scale* value decreases, indicated by $n_t = N_a(n_g = 2)$ and $n_t = N_a(n_g = 1)$. Accompanied with this trend, n_g and n_e improves. The same trend is observed for runs in different series with similar annealing ranges but larger n_t. The maximum n_g approaches a plateau ~ 191, suggesting simulations in run series 3 is close to thorough sampling, while series 1 and 2 are under sampled. We note the number of consistently assigned RTAs continues to increase from series 1 to 3. Within a series, simulations with more effective annealing steps produce more consistently assigned RTAs. However, we note there are some exceptions. Run 5 in series 2 produces fewer consistently assigned RTAs, due to the threshold $n_t = N_a(n_g = 1)$ that goes beyond the last annealing step. Furthermore, run 1 in series 1 with the most effective annealing steps does not have the most consistently assigned RTAs, due to the severity of under sampling. We will show in the next section that most of the consistently assigned RTAs are allocated to their correct positions (experimental assigned positions).

Therefore, to determine if thorough sampling is achieved, users should execute ASAP simulations with increasing n_t to maximize the annealing steps $n_t \geq N_a(n_g = 2)$, until the total number of n_g stops improving. To maximize RTAs allocated at their global minimum positions, the threshold step $n_t = N_a(n_g = 1)$ should be sufficiently distant from the last annealing step to maximize the effective annealing. Note that the simulations in the previous section also achieved $n_g = 191$. It shows that thorough sampling can be achieved by different parameter setups.

3.5. Alternative strategies to achieve thorough sampling by iterative ASAP simulations

A single ASAP simulation with $n_t = 5.0 \times 10^6$ and $nstep = 40$ takes ~ 29 min on a desktop with 12th gen Intel i5-12500. Therefore, considerable time is needed to achieve thorough sampling for a large protein. Converting the program to Fortran or C will probably reduce the time by hundreds of folds, according to our experience simulating the self-assembly of the HIV CA by MCSA [54–57]. Alternatively, thorough sampling can be achieved by iterative ASAP simulations as discussed earlier.

To prove this, iterative ASAP runs are performed with the identical setup as run set 6 in Table 1 with the same RSV CA data, results shown in Table 2. As iteration progresses, n_g and n_e improve quickly. At the end of the third iteration, they produce comparable total number of n_g and n_e to those obtained by ASAP with the maximum n_t in Table 1. The number of consistently allocated RTAs saturates at ~ 183. Moreover, the number of correctly assigned RTAs quickly grows from 113 to 144, approaching the maximum possible value. Hence, iterative ASAP runs are more favorable than a single ASAP runs with a large n_t, if done correctly.

The success of iterative ASAP is to strike a balance between increasing the number of parallel simulations in each iteration and the decreasing sampling n_t. A less-than-optimal sampling would probably produce some erroneous sequential allocations in each simulation. However, if sufficient number of parallel simulations are performed, a mistaken assignment due to a specific local minimum will not be repeated in all parallel simulations, so consistently assigned RTAs should correspond to those allocated correctly. However, as the protein size and spectral congestion increases, local minima also proliferate. It is possible for all parallel simulations experience one or more identical

Table 2
Thorough sampling achieved in 3 iterations, with $n_t = 5.0 \times 10^6$, $scale = 0.45$. Each run set are performed with 10 parallel simulations, with *disparity_nco1* and *disparity_nco2* set to 1.

Iteration	Mean n_g	STD n_g	Mean n_e	STD n_e	Consistently assigned	Correctly assigned
1	188.8	1.48	33.7	2.11	133	113
2	189.7	0.48	34.8	1.03	182	144
3	190	0	35	1.56	183	144

Table 1
ASAP sequential assignments with different setups to test thorough sampling. All simulations end with $n_b = 0$. Each run set are performed with 10 identical simulations to average the randomness, except for run 12 with one simulation. All simulations are performed with *disparity_nco1* and *disparity_nco2* set to 1.

Run index	n_t*(millions)*	*scale*	$n_t = N_a(n_g = 2)$	$n_t = N_a(n_g = 1)$	Mean n_g	Mean n_e	Consistently assigned
1	0.5	0.25	18	37	183.2	37.6	97
2	0.5	0.45	10	20	179.6	43.7	92
3	0.5	0.6	8	15	179.5	45.3	105
4	0.5	1.0	5	9	177.0	49.5	90
5	5	0.25	31	61	188.8	31.9	97
6	5	0.45	17	34	187.2	36.2	136
7	5	0.6	13	26	186.9	37.7	133
8	5	1.0	8	15	185.8	40.9	130
10	50	0.6	18	36	190.0	30.5	171
11	50	1.0	11	22	190.1	36.3	171
12	500	1.0	14	28	191	35	NAN

local minima. Retaining such erroneously assigned RTAs for subsequent iterations would lead to decreasing total number of n_g. This observation would alert users to increase either n_t, or the number of parallel simulations in each iteration. This does happen if we use *NCACX4nextrd* and *NCOCX4nextrd* of run 1 in Table 1 with $n_t = 0.5 \times 10^6$ as input files to seed subsequent ASAP simulations, where n_g decreases to 181 in the second round.

3.6. Robustness against ambiguous assignments in input

Previously, Tycko demonstrated that ambiguity in RTAs was particularly detrimental for correct sequential assignments with MCAssign, even for proteins of ∼ 150 residues [40]. In our discussion, we claimed that the matched RTA pairs by ARTIST can greatly suppress local minima. To prove this, simulations are performed by MCAssign [37] and ASAP to test the effect of ambiguous RTAs, tabulated in Table 3.

The performance of ASAP against ambiguity in RTAs is first tested by the NCACX RTAs and the NCOCX peak list of the 147-residue SrtC [41]. As shown by Table 3, run 1 is first performed to assign SrtC with unambiguous RTAs in input 1, as a baseline reference. To compare with the results obtained by MCAssign, run 2 is performed with level 1 ambiguity, the same as in reference [40]. As shown in Table 3, all residues in the input files for SrtC are assigned correctly and consistently. This is shown by Fig. 5A, where the ASAP allocated positions for those consistently assigned RTAs are plotted against the experimentally assigned positions. In contrast, 2 were assigned mistakenly by MCAssign even in the absence of any ambiguous inputs [40].

Similarly, run 3 and 4 are ASAP simulations with the unambiguous and ambiguous RTAs in input 1 for MLKL, with 142 RTAs always consistently assigned, as shown in Table 3. Among them, 7 and 15 pairs of RTAs are placed to different positions than their experimentally assigned positions, for simulations with unambiguous and ambiguous inputs respectively. Specifically, with unambiguous input, the 7 erroneously assigned RTAs are those due to overlapping ^{15}N resonances implicated in type 2 local minima, falling onto the grey bar highlighted positions in Fig. 5B. This is an improvement in contrast to the 22 mistaken assignments by MCAssign [40]. In the presence ambiguity, 13 out of the 15 misplaced RTAs are those implicated in type 2 local minima. The other 2 are those implicated in type 3 local minima (K78 vs K122, L128 vs L81). As shown in Fig. 5B, we note that adjacent residue positions to these erroneous assignments are correctly assigned, indicated by their distribution along the diagonal direction. It agrees with our discussion that type 2 or 3 local minima are caused by RTAs with completely interchangeable sequential allocations, so neighboring assignments are not disrupted.

Next, the performance of ASAP is compared against MCAssign with the RSV CA data. We note that MCAssign program counts the RTAs in NCACX and NCOCX individually, so n_g in its results corresponds roughly to twice of that in ASAP simulations. When $n_t = 5 \times 10^6$, for MCAssign simulations with level 1 ambiguity in input 1, the number of consistently and correctly assigned RTAs decreases ∼ 1/3 of those with unambiguous RTAs, shown by run 5 and 6 in Table 3. When the ambiguity of RTAs is increased to level 2, the consistently and correctly assigned RTAs decreased further, shown by run 7. It implies the system is seriously under sampled to remove the populated type 1 local minima caused by ambiguous RTAs. When n_t is increased to 50×10^6, the number of consistently and correctly assigned RTAs are both greatly improved for simulations with unambiguous or level 1 ambiguity, shown by run 8 and 9 in Table 3. It indicates that sufficient sampling can effectively overcome the local minima caused by level 1 ambiguity. However, the consistently and correctly assigned RTAs stay roughly unchanged, even with more sampling for simulations with level 2 ambiguity, shown by run 10, suggesting that the system is still under sampled to overcome the local minima with MCAssign. In addition, according to our analysis of input experimental RTAs in the previous section, the maximum n_g for MCAssign should be ∼ 452, or ∼ 226 NCACX RTA.

In comparison, ASAP demonstrates a stronger resilience against both levels of ambiguity. When $n_t = 5 \times 10^6$, the decrease of consistently and correctly assigned RTAs with both levels of ambiguity (run 11 and 12 in Table 3) is much milder compared to results in the absence of ambiguity (run 6 in Table 1). Moreover, when n_t is increased to 50×10^6, results for simulations with both levels of ambiguity are greatly improved, shown by run 13 and 14. Recall that our analysis of input data for ASAP suggests that the maximum number n_g is only ∼ 150.

To inspect the consistently allocated RTAs, Fig. 5C plots the sequentially allocated residue positions of consistently allocated RTAs against their experimentally assigned positions. Specifically, in the presence of level 2 ambiguity in input 1, there are 25 distinct RTAs shifted from their experimentally determined positions. Among them, 18 are implicated in type 2 local minima, falling in the grey bar highlighted regions. Out of the remaining 7 shifted RTAs, 6 are signals forming type 3 local minima with the experimentally assigned values, listed in Table S1. The only exception is the T17 that is assigned to residue position 227, seated alone with both neighbors occupied by null assignments. Similar to our observations in Fig. 5B for MLKL results, adjacent residue positions to these mistaken assignments are correctly assigned, with signals distributed along the diagonal directions in Fig. 5C. They confirm that these mistaken assignments are caused by those RTAs with entirely overlapping ^{15}N resonances. Thus, to simplify our discussions, we will stop listing the IDs of these mistaken assignments.

Note that while there are nearly always bad connections in all MCAssign results, they are eradicated in all ASAP simulations. This is due to the higher penalty to allocate matched RTA pairs with bad connections. We can use *Plot_occupancy_sum* to visualize the ill-effect caused

Table 3
Comparison of resilience against ambiguous RTAs between MCAssign and ASAP. Each run set comprise10 identical simulations to average the randomness, with *disparity_nco1* and *disparity_nco2* set to 1.

Run index	*Protein*	*Algorithm*	*n_t(millions)*	*scale*	*Ambiguity in input RTAs?*	Mean n_g	Mean n_e	Mean n_b	Consistently assigned	Correctly assigned
1	SrtC	ASAP	50	0.7	0	111	8	0	111	111
2	SrtC	ASAP	50	0.7	Level 1	111	8	0	111	111
3	MLKL	ASAP	50	0.65	0	142	7	0	142	135
4	MLKL	ASAP	50	0.65	Level 1	142	7	0	142	127
5	RSV CA	MCAssign	5	1.0	0	414.3	14.3	1.3	121	120
6	RSV CA	MCAssign	5	1.0	Level 1	402.4	25.9	1.1	79	78
7	RSV CA	MCAssign	5	1.0	Level 2	396	30.2	0.6	56	54
8	RSV CA	MCAssign	50	1.0	0	416.1	13.1	1.4	139	137
9	RSV CA	MCAssign	50	1.0	Level 1	416.7	12.8	1.4	138	136
10	RSV CA	MCAssign	50	1.0	Level 2	396	30.2	0.6	56	56
11	RSV CA	ASAP	5	0.45	Level 1	185.0	38.1	0	113	92
12	RSV CA	ASAP	5	0.45	Level 2	185.2	38.6	0	111	95
13	RSV CA	ASAP	50	0.6	Level 1	189.3	34.4	0	144	115
14	RSV CA	ASAP	50	0.6	Level 2	189	34.6	0	143	116

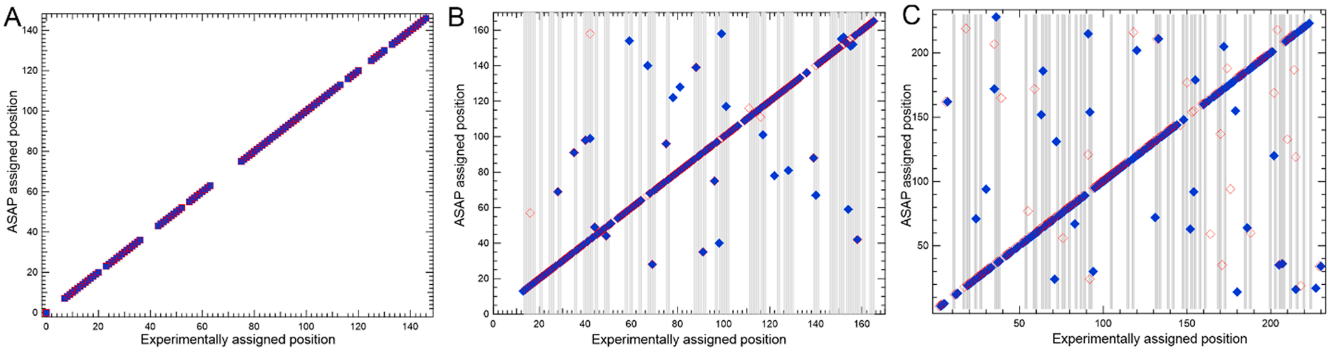


Fig. 5. Resilience against ambiguity in RTAs in ASAP simulations. Allocated positions of consistently assigned RTAs are plotted against their experimentally assigned positions. (A). Allocated positions for consistently assigned RTAs of SrtC data with zero ambiguity (run 1 in Table 3, red empty squares) vs level 1 ambiguity in input 1 (run 2 in Table 3, blue solid squares). (B). Allocated positions for consistently assigned RTAs of MLKL data with zero ambiguity (run 3 in Table 3, red empty diamonds) vs level 1 ambiguity in input 1 (run 4 in Table 3, blue solid diamonds). (C). Allocated positions for consistently assigned RTAs of the RSV CA with zero ambiguity (run 10 in Table 1, red empty diamonds) vs level 2 ambiguity in input 1 (run 14 in Table 3, blue solid diamonds). Grey bars are RTAs implicated with overlapping ^{15}N resonances to incur type 2 local minima. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

by ambiguous RTAs in input 1, which leads to more shuffling in the annealing process. They are shown by Fig. S5A for run 10 in Table 1 and Fig. S5B for its counterpart run 14 in Table 3.

In summary, our results confirmed that consistently allocated RTAs may be considered as those assigned to their global minimum locations if thorough sampling can be guaranteed, except for those implicated in type 2 and 3 local minima. To correct erroneous assignments caused by type 2 local minima, additional restrains are needed from experiments such as CANCX or CONCX spectra, if SNR allows. For proteins with highly repetitive stretches of sequences, frequency selective dipolar dephasing experiments can be performed with selective residue labeled samples to resolve the ambiguity [58]. Owing to the correlation established by ARTIST for input RTAs, ASAP demonstrates a stronger resilience against ambiguity in input RTAs.

3.7. Capability to work with the full NCOCX peak list

In all prior tests, only c-alpha and c-beta resonances in NCOCX spectra are included in input 3. This is justifiable when the NCOCX spectra are recorded with a short DARR mixing time. To demonstrate the full capability of ASAP, simulations are performed to test ASAP with input 3 comprising only c-alpha and c-beta resonances vs the full NCOCX peak list of the RSV CA acquired with 150 ms DARR mixing.

To make the test more challenging, the full NCOCX peak list contains more resonances than those obtained by Poky peak picking. Specifically, if the resonance of a non-mandatory side-chain site is present in NCACX but absent in the NCOCX spectrum, a fictitious resonance (f_{bx},f_{by},f_{bz}) is added to input 3, with f_{bz} obtained by randomly shifting ($\leq \Delta f_{az}$) its f_{az} frequency in NCACX, and f_{bx} and f_{by} as the common indirect frequencies identified for the c-alpha and c-beta in NCOCX. Thus, the number of individual signals in input 3 increases from 414 to 627. Because only c-alpha and c-beta resonances are used to identify matched RTAs by ARTIST, it maximizes the interferences from extra side-chain resonances in NCOCX. All runs are performed with 10 parallel simulations to average out the randomness factors.

In addition, all tests in prior sections are performed with *disparity_nco1* and *disparity_nco2* set to 1, so the full uncertainty values listed in input 3 are used by ARTIST to check resonances alignment along their indirect dimensions. To help cope with the full NCOCX peak list, tests are also performed with both parameters set to 0.33. It requires the frequencies of the indirect dimensions to match within 0.1 ppm in NCOCX, for them to pass the final test in ARTIST. This is the standard used in manual assignment, and should help to disqualify interference of resonances belonging to different residues.

The results are shown in Table 4 and Fig. 6. As shown in Fig. 6A, introduction of extra side-chain peaks in input 3 does lead to slightly inflated matched RTAs, due to the coincidental matches to mandatory resonances by other side-chain carbons. However, with *disparity_nco1* and *disparity_nco2* set to 0.33, when all RTAs in NCACX are unambiguously determined, the results are not impacted by the extra side-chain peaks, indicated by similar n_g, n_e and consistently assigned RTAs, as shown by the statistics of run 1 vs run 2 in Table 4. They also find comparable correctly assigned RTAs among the consistently assigned RTAs, shown in Fig. 6D. Majority of those shifted assignments are those implicated in type 2 local minima, indicated by the grey bars. Those fall in the blank regions are RTAs with overlapping ^{15}N and ^{13}C resonances, implicated in type 3 local minima. Again, due to the interchangeability of their sequential allocations, we note that assignments of adjacent residues are not affected, as their signals nearly always fall along the diagonal in Fig. 6D. Similar patterns are observed in the following two tests.

To investigate ASAP's ability to resist ambiguity in input 1 with the full NCOCX peak list, additional tests are performed with level 1 ambiguity in input 1, with input 3 comprising just c-alpha and c-beta resonances (run 5), or the full peak list from NCOCX (run 6), respectively. Level 1 ambiguity further increases the coincidentally matched RTAs

Table 4
Capability of ASAP to work with the full NCOCX peak list. All runs comprise 10 parallel simulations to average randomness, with n_t = 5 *million*, *scale* = *0.45*.

Run index	*signalsinNCOCX*	*disparity_nco1 disparity_nco2*	*Ambiguity in input RTAs*	Mean n_g	Mean n_e	Consistently assigned	Correctly assigned
1	c-alpha and c-beta	0.33	0	188.0	34.0	142	116
2	Full peak list	0.33	0	188.6	35.2	141	121
3	Full peak list	1.0	0	186.5	37.0	123	94
4	Full peak list with absolute R and L	0.33	0	190.2	33.6	156	133
5	c-alpha and c-beta	0.33	Level 1	185.1	38.0	122	108
6	Full peak list	0.33	Level 1	186.4	36.6	126	95
7	Full peak list	1.0	Level 1	186.3	37.3	114	91
8	Full peak list with absolute R and L	0.33	Level 1	188.3	35.0	137	119

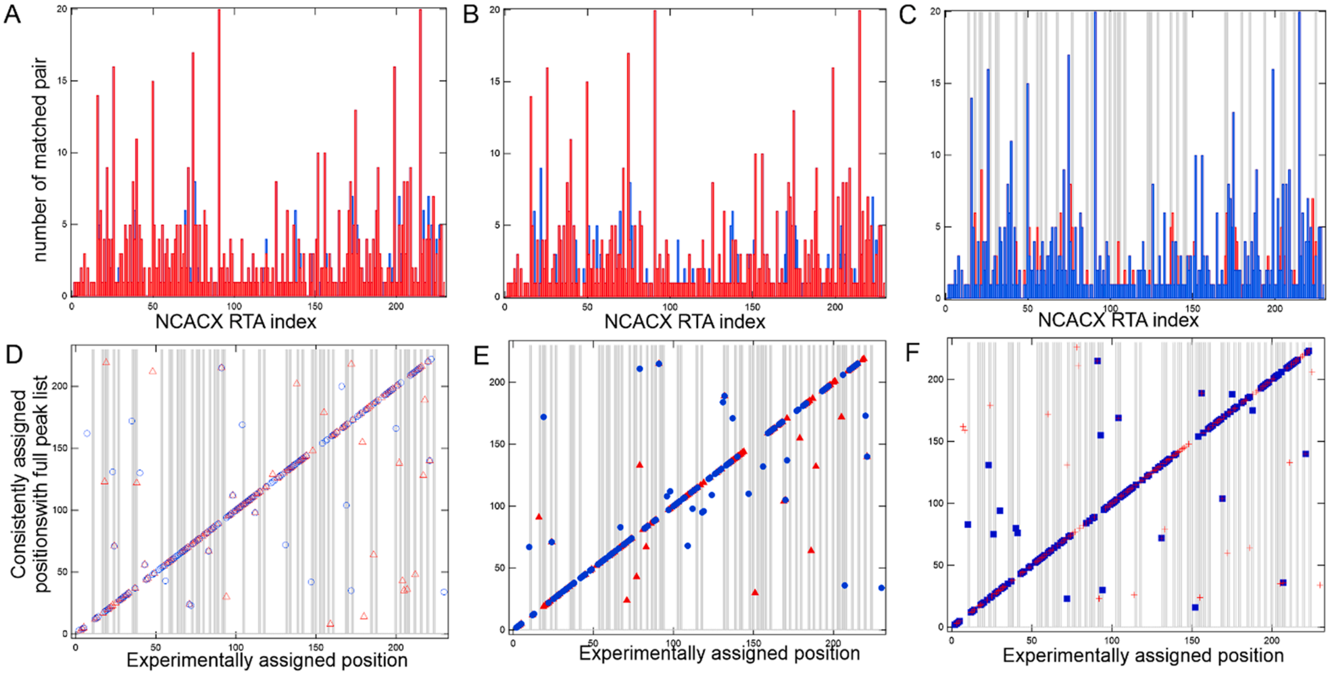


Fig. 6. Sequential assignment tests of ASAP with the full NCOCX peak list. (A) to (C) are the number of matched NCOCX RTAs identified for each NCACX RTA. (A) Input 3 comprising only c-alpha and c-beta resonances in input 3 (red) vs the full NCOCX peak list (blue), with no ambiguity in input 1. (B). Input 3 comprising the full NCOCX peak list and input 1 comprising zero (red) vs level 1 ambiguity (blue). (C). Input 3 comprising the full NCOCX peak list and input 1 with level 1 ambiguity. The red bars are for signals of R and L residues entered as their actual positions in NCOCX, and the blue bars are for their positions fixed to the observed frequencies in NCACX. The grey bars highlight the R and L residues in input 1. (D) to (F) are the allocated positions of consistently assigned RTAs plotted against their experimentally assigned positions. (D). Input 3 comprising only c-alpha and c-beta resonances (red empty triangles, run 1 in Table 4) vs the full NCOCX peak list (blue empty circles, run 2 in Table 4), with no ambiguity in input 1. (E). Input 3 comprising only c-alpha and c-beta resonances (red filled triangles, run 5 in Table 4), vs the full NCOCX peak list (blue filled circles, run 6 in Table 4), with level 1 ambiguity in input 1. (F). Input 3 comprising the full NCOCX peak list with R and L residues fixed to those identified in NCACX, with zero ambiguity (red cross, run 4 in Table 4) vs level 1 ambiguity in input 1 (blue filled squares, run 8 in Table 4). In (D) to (F), the grey bars highlight the RTAs implicated in type 2 local minima. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

identified in NCOCX, shown in Fig. 6B. The quality of sequential assignments for both run 5 and 6 also is decreased, shown in Table 4. Specifically, compared to run 5, run 6 produces slightly fewer consistently and correctly assigned RTAs. As shown in Fig. 6E, the consistently assigned RTAs but shifted from their experimentally determined positions are mostly those implicated in type 2 local minima, falling in the grey bar highlighted positions. Additionally, there are 5 pairs corresponding to those implicated in type 3 local minima, outside the grey bar highlighted regions for type 2 local minima.

Simulations are also performed with full uncertainty values in input 3 with *disparity_nco1* and *disparity_nco2* set to 1, with the full NCOCX peak list, for zero ambiguity (run 3) and level 1 ambiguity in input 1 (run 7). They produce fewer consistently assigned RTAs compared to simulations with *disparity_nco1* and *disparity_nco2* set to 0.33, proving the interference adding extra side-chain resonances, and the ability of the program to suppress this interference by a tighter tolerance in the final test of ARTIST. Moreover, we note that run 3, 6 and 7 produce comparable number of correct assigned RTAs. It suggests the mistakenly assigned RTAs are predominantly due to the side-chain interferences, not the ambiguity in input 1.

With the full NCOCX peak list, in addition to the standard approach of increasing n_t or iterative simulations, the results of ASAP can be further improved by strengthening the correlations of resonances in input 1 and 3. Specifically, in experiments, selective residue labeled samples can help us unambiguously correlate signals from the same residue across spectra, as shown in our prior work [13]. Run 4 and 8 are performed, with zero or level 1 ambiguity in input 1 respectively. In both runs, we simulate this situation that all signals of R and L residues can be unambiguously resolved and correlated in both spectra, by imposing identical c-alpha and c-beta frequencies for f_{az} and f_{bz} for these residues in input 1 and input 3, with their uncertainty set to 0.001 ppm. As shown by Fig. 6C, it suppresses coincidentally matched RTAs in NCOCX not only for these residues, but also for 10 other residues, as the side-chain resonances of these R and L residues could be used to match the mandatory carbons of other residues. Indeed, both runs produce improved results, with better n_g, n_e and consistently assigned RTAs. More importantly, the number of correctly allocated RTAs are greatly improved, by nearly 30 compared to their reference runs (run 3 and run 5). Hence, if a residue exhibits distinct and well-resolved CS signals in both NCACX and NCOCX spectra, such as A, S, T, I, P, V or G, we should impose identical f_{az} and f_{bz} for their c-alpha and c-beta resonances in input 1 and input 3, and remove non mandatory side-chain resonances in input 3 to prevent coincidental matching to other residues. This strategy will further improve the performance of ASAP.

3.8. Limitations and incorporation of additional spectral assignments into ASAP

ASAP cannot differentiate those RTA pairs implicated in type 2 or 3 local minima that can be assigned interchangeably due to their similar ^{15}N resonances, with only NCACX and NCOCX spectra. They may cause fluctuations of consistently and correctly assigned RTAs in the simulations with the same or similar setup, shown by the differences between run 11 in Table 3 vs run 7 in Table 4. However, if their adjacent residues exhibit distinct resonances, incorporation of correlations revealed by CANCX or CONCX will eliminate these type 2 or 3 local minima. Current version of ASAP does not automate the inclusion of this correlation, and must be implemented manually by adjusting the amide ^{15}N CS uncertainties of corresponding RTAs in input 1 and 3 that are correlated in CANCX or CONCX, so that they are the only pair to satisfy Eq. 8. This process can be accelerated by ARTIST to find matched RTAs in these spectra against the reference NCACX RTA. The damage of these errors is probably localized, as we showed, assignments of adjacent residue are not perturbed by these errors.

The more challenging limitation is to accurately determine the peak positions in the presence of severe resonance overlap, as it directly impacts the accuracy of ARTIST and ASAP. While smaller *disparity_nco1* and *disparity_nco2* help to suppress coincidental matches, it weakens the ability to account for peak shifting caused by resonance overlap along the two indirect dimensions. Hence, separating congested resonances in NCOCX along the two indirect dimensions entirely depends on the accurate peak deconvolution algorithm of other programs such as Poky [53].

ASAP is designed for ^{13}C detected spectra NCACX and NCOCX. However, it can be easily adapted for inputs derived from other multidimensional spectra, or recorded with proton detection, by replacing corresponding entries along their indirect and direct dimensions.

3.9. Data availability

The ASAP source code with all input and output files are provided as the zip file in supporting materials for simulations in Table 4. Brief instructions are provided for the plotting scripts in supporting materials, as well as *NCACX_1CompareNCO* that tabulates the original RSV CA input files. Please contact Bo Chen at bo.chen@ucf.edu if you need additional help.

4. Conclusion

In conclusion, we demonstrate that ASAP is a robust sequential assignment program for congested multidimensional NMR spectra. Compared to other popular auto-assignment programs for ssNMR such as FLYA or ssPINE that require a plethora of multidimensional spectra, [10,11] some of which such as CANCX and CONCX may be challenging to obtain for noncrystalline samples, ASAP only need 3DNCACX and NCOCX spectra. It largely eliminates the laborious efforts in sequential assignments of congested NCOCX spectra, so long as the accuracy of peak picking in NCOCX and RTAs in the better resolved NCACX spectrum can be guaranteed. With ASAP, the sequential assignments of large proteins that lack premium spectral resolution for ssNMR can be reduced to days. It relieved the resolution cap for assignment capped by the more congested NCOCX spectrum to the NCACX spectrum.

CRediT authorship contribution statement

Bo Chen: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [Bo Chen reports financial support was provided by National Science Foundation. Bo Chen has patent pending to Patent declared by University of Central Florida. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper].

Acknowledgements

This work is supported by NSF award #1856055, UCF Mid-career award and seed grant. The spectra were acquired at National High Magnetic Field Laboratory, United States, supported by NSF/DMR #2128556 and the State of Florida, with the assistance of Dr. Zhehong Gan and Ivan Hung. The author sincerely acknowledge Dr. Robert Tycko's insights to help revise the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmr.2024.107664.

References

[1] M.D. Tuttle, et al., Solid-state NMR structure of a pathogenic fibril of full-length human alpha-synuclein, Nat. Struct. Mol. Biol. 23 (2016) 409–415.
[2] S.D. Cady, et al., Structure of the amantadine binding site of influenza M2 proton channels in lipid bilayers, Nature 463 (2010) 689–U127.
[3] A.W.P. Fitzpatrick, et al., Atomic structure and hierarchical assembly of a cross-β amyloid fibril, PNAS 110 (2013) 5468–5473.
[4] M.M. Lu, et al., Atomic-resolution structure of HIV-1 capsid tubes by magic-angle spinning NMR, Nat. Struct. Mol. Biol. 27 (2020) 863-+.
[5] D.T. Murray, et al., Structure of FUS protein fibrils and its relevance to self-assembly and phase Separation of low-complexity domains, Cell 171 (2017) 615-+.
[6] R. Rogawski, A.E. McDermott, New NMR tools for protein structure and function: spin tags for dynamic nuclear polarization solid state NMR, Arch. Biochem. Biophys. 628 (2017) 102–113.
[7] M.T. Colvin, et al., Atomic resolution structure of monomorphic Aβ<sub>42</sub> amyloid fibrils, J. Am. Chem. Soc. 138 (2016) 9663–9674.
[8] L.J. Sperling, et al., Solid-state NMR study of a 41 kDa membrane protein complex DsbA/DsbB, J. Phys. Chem. B 117 (2013) 6052–6060.
[9] A. Klein, et al., Atomic-resolution chemical characterization of (2x)72-kDa tryptophan synthase via four- and five-dimensional <SUP>1</SUP>H-detected solid-state NMR, PNAS 119 (2022).
[10] E. Schmidt, P. Güntert, A new algorithm for reliable and general NMR resonance assignment, J. Am. Chem. Soc. 134 (2012) 12817–12829.
[11] A.E. Lopez, A. Dwarasala, M. Rahimi, J.L. Markley, W. Lee, ssPINE/ssPINE-POKY: automated chemical shift assignment with an intuitive graphical user interface for solid-state NMR data from complex, Biophys. J . 122 (2023) 141A–A.
[12] W.T. Franks, K.D. Kloepper, B.J. Wylie, C.M. Rienstra, Four-dimensional heteronuclear correlation experiments for chemical shift assignment of solid proteins, J. Biomol. NMR 39 (2007) 107–131.
[13] J. Jeon, et al., Structural model of the tubular assembly of the rous sarcoma virus capsid protein, J. Am. Chem. Soc. 139 (2017) 2006–2013.
[14] T. Thames, et al., Curvature of the retroviral capsid assembly is modulated by a molecular switch, J. Phys. Chem. Lett. 12 (2021) 7768–7776.
[15] N.E.G. Buchler, E.R.P. Zuiderweg, H. Wang, R.A. Goldstein, Protein heteronuclear NMR assignments using mean-field simulated annealing, J. Magn. Reson. 125 (1997) 34–42.
[16] J.A. Lukin, A.P. Gove, S.N. Talukdar, C. Ho, Automated probabilistic method for assigning backbone resonances of (C-13, N-15)-labeled proteins, J. Biomol. NMR 9 (1997) 151–166.
[17] M. Leutner, et al., Automated backbone assignment of labeled proteins using the threshold accepting algorithm, J. Biomol. NMR 11 (1998) 31–43.
[18] T.K. Hitchens, J.A. Lukin, Y.P. Zhan, S.A. McCallum, G.S. Rule, MONTE: an automated monte carlo based approach to nuclear magnetic resonance assignment of proteins, J. Biomol. NMR 25 (2003) 1–9.
[19] J. Volk, T. Herrmann, K. Wuthrich, Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH, J. Biomol. NMR 41 (2008) 127–138.
[20] E. Schmidt, P. Guntert, A new algorithm for reliable and general NMR resonance assignment, J. Am. Chem. Soc. 134 (2012) 12817–12829.
[21] D.E. Zimmerman, et al., Automated analysis of protein NMR assignments using methods from artificial intelligence, J. Mol. Biol. 269 (1997) 592–610.
[22] K.B. Li, B.C. Sanctuary, Automated resonance assignment of proteins using heteronuclear 3D NMR.2. side chain and sequence-specific assignment, J. Chem. Inf. Comput. Sci. 37 (1997) 467–477.
[23] H.S. Atreya, S.C. Sahu, K.V.R. Chary, G. Govil, A tracked approach for automated NMR assignments in proteins (TATAPRO), J. Biomol. NMR 17 (2000) 125–136.
[24] M. Andrec, R.M. Levy, Protein sequential resonance assignments by combinatorial enumeration using C-13 alpha chemical shifts and their (i, i–1) sequential connectivities, J. Biomol. NMR 23 (2002) 263–270.
[25] B.E. Coggins, P. Zhou, PACES: protein sequential assignment by computer-assisted exhaustive search, J. Biomol. NMR 26 (2003) 93–111.
[26] J.T. Nielsen, N. Kulminskaya, M. Bjerring, N.C. Nielsen, Automated robust and accurate assignment of protein resonances for solid state NMR, J. Biomol. NMR 59 (2014) 119–134.
[27] H.N.B. Moseley, G. Sahota, G.T. Montelione, Assignment validation software suite for the evaluation and presentation of protein resonance assignment data, J. Biomol. NMR 28 (2004) 341–355.
[28] Y.S. Jung, M. Zweckstetter, Mars - robust automatic backbone assignment of proteins, J. Biomol. NMR 30 (2004) 11–23.

[29] J.Y. Wang, T.Z. Wang, E.R.P. Zuiderweg, G.M. Crippen, CASA: an efficient automated assignment of protein mainchain NMR data using an ordered tree search algorithm, J. Biomol. NMR 33 (2005) 261–279.
[30] G.M. Crippen, A. Rousaki, M. Revington, Y.B. Zhang, E.R.P. Zuiderweg, SAGA: rapid automatic mainchain NMR assignment for large proteins, J. Biomol. NMR 46 (2010) 281–298.
[31] E.R.P. Zuiderweg, I. Bagai, P. Rossi, E.B. Bertelsen, EZ-ASSIGN, a program for exhaustive NMR chemical shift assignments of large proteins from complete or incomplete triple-resonance data, J. Biomol. NMR 57 (2013) 179–191.
[32] R. Tycko, K.N. Hu, A Monte Carlo/simulated annealing algorithm for sequential resonance assignment in solid state NMR of uniformly labeled proteins with magic-angle spinning, J. Magn. Reson. 205 (2010) 304–314.
[33] L.J. Sperling, D.A. Berthold, T.L. Sasser, V. Jeisy-Scott, C.M. Rienstra, Assignment strategies for Large proteins by magic-angle spinning NMR: the 21-kDa disulfide-bond-forming enzyme DsbA, J. Mol. Biol. 399 (2010) 268–282.
[34] E. Schmidt, et al., Automated solid-state NMR resonance assignment of protein microcrystals and amyloids, J. Biomol. NMR 56 (2013) 243–254.
[35] J. Lapin, A.A. Nevzorov, Automated assignment of NMR spectra of macroscopically oriented proteins using simulated annealing, J. Magn. Reson. 293 (2018) 104–114.
[36] Y. Yang, K.J. Fritzsching, M. Hong, Resonance assignment of the NMR spectra of disordered proteins using a multi-objective non-dominated sorting genetic algorithm, J. Biomol. NMR 57 (2013) 281–296.
[37] K.N. Hu, W. Qiang, R. Tycko, A general Monte Carlo/simulated annealing algorithm for resonance assignment in NMR of uniformly labeled biopolymers, J. Biomol. NMR 50 (2011) 267–276.
[38] D.S. Wishart, et al., H-1, C-13 AND N-15 chemical-shift referencing in biomolecular NMR, J. Biomol. NMR 6 (1995) 135–140.
[39] Chen, B. *Fundamentals of Recoupling and Decoupling Techniques in Solid State NMR*, (AIP Publishing LLC, 2020).
[40] R. Tycko, On the problem of resonance assignments in solid state NMR of uniformly N-15, C-13-labeled proteins, J. Magn. Reson. 253 (2015) 166–172.
[41] S.A. Robson, A.W. Jacobitz, M.L. Phillips, R.T. Clubb, Solution structure of the sortase required for efficient production of infectious Bacillus anthracis spores, Biochemistry 51 (2012) 7953–7963.
[42] L.J. Su, et al., A plug release mechanism for membrane permeation by MLKL, Structure 22 (2014) 1489–1500.
[43] D.H. Zhou, et al., Proton-detected solid-state NMR spectroscopy of fully protonated proteins at 40 kHz magic-angle spinning, J. Am. Chem. Soc. 129 (2007) 11791–11801.
[44] V. Kurauskas, et al., Sensitive proton-detected solid-state NMR spectroscopy of large proteins with selective CH3 labelling: application to the 50S ribosome subunit, Chem. Commun. 52 (2016) 9558–9561.
[45] P. Fricke, et al., Backbone assignment of perdeuterated proteins by solid-state NMR using proton detection and ultrafast magic-angle spinning, Nat. Protoc. 12 (2017) 764–782.
[46] M. Cordova, P. Moutzouri, B.S. de Almeida, D. Torodii, L. Emsley, Pure isotropic proton NMR spectra in solids using deep learning, Angewandte Chemie-International Edition (2023).
[47] Python. 3.11.2 edn (Python Software Foundation, Python Language Reference, Version 3.11.2. Available at http://www.python.org, 2023).
[48] M.K. Pandey, Z. Qadri, R. Ramachandran, Understanding cross-polarization (CP) NMR experiments through dipolar truncation, J. Chem. Phys. 138 (2013).
[49] M.J. Bayro, et al., Dipolar truncation in magic-angle spinning NMR recoupling experiments, J. Chem. Phys. 130 (2009).
[50] K.J. Fritzsching, Y. Yang, K. Schmidt-Rohr, M. Hong, 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 (2013) 155–167.
[51] Y.J. Wang, O. Jardetzky, Probability-based protein secondary structure identification using combined NMR chemical-shift data, Protein Sci. 11 (2002) 852–861.
[52] K. Takegoshi, S. Nakamura, T. Terao, <SUP>13</SUP>C-<SUP>1</SUP>H dipolar-assisted rotational resonance in magic-angle spinning NMR, Chem. Phys. Lett. 344 (2001) 631–637.
[53] W. Lee, M. Rahimi, Y. Lee, A. Chiu, POKY: a software suite for multidimensional NMR and 3D structure calculation of biomolecules, Bioinformatics 37 (2021) 3041–3042.
[54] G.P. Zhao, et al., Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics, Nature 497 (2013) 643–646.
[55] X. Qiao, J. Jean, J. Weber, F.Q. Zhu, B. Chen, Mechanism of polymorphism and curvature of HIV capsid assemblies probed by 3D simulations with a novel coarse grain model, BBA-Gen. Subjects 1850 (2015) 2353–2367.
[56] B. Chen, R. Tycko, Simulated self-assembly of the HIV-1 capsid: protein shape and native contacts are sufficient for two-dimensional lattice formation, Biophys. J . 100 (2011) 3035–3044.
[57] X. Qiao, J. Jeon, J. Weber, F.Q. Zhu, B. Chen, Construction of a novel coarse grain model for simulations of HIV capsid assembly to capture the backbone structure and inter-domain motions in solution, Data Brief 5 (2015) 506–512.
[58] X.Y. Ding, F.Q. Fu, F. Tian, De novo resonance assignment of the transmembrane domain of LR11/SorLA in E. coli membranes, J. Magn. Reson. 310 (2020).