

Extreme Non-Uniform Sampling for Protein NMR Dynamics Studies in Minimal Time

Gregory Jameson,^{1,2} Alexandar L. Hansen,³ Dawei Li,³ Lei Bruschweiler-Li³ and Rafael
Bruschweiler^{1,2,3,4*}

¹Department of Chemistry and Biochemistry, The Ohio State University, Columbus,
Ohio 43210, United States

²Biophysics Graduate Program, The Ohio State University, Columbus, Ohio 43210,
United States

³Campus Chemical Instrument Center, The Ohio State University, Columbus, Ohio
43210, United States

⁴Department of Biological Chemistry and Pharmacology, The Ohio State University,
Columbus, Ohio 43210, United States

*To whom correspondence should be addressed:

Rafael Bruschweiler, Ph.D.

Department of Chemistry and Biochemistry, Room 382 CBEC,

151 W. Woodward Ave, The Ohio State University, Columbus, Ohio 43210

E-mail: bruschweiler.1@osu.edu

ABSTRACT

NMR spectroscopy is an extraordinarily rich source of quantitative dynamics of proteins in solution using spin relaxation or Chemical Exchange Saturation Transfer (CEST) experiments. However, ^{15}N -CEST measurements require prolonged multidimensional, so-called pseudo-3D HSQC experiments where the pseudo dimension is a radio-frequency offset $\Delta\omega$ of a weak ^{15}N saturation field. Non-uniform sampling (NUS) approaches have the potential to significantly speed up these measurements, but they also carry the risk of introducing serious artifacts and the systematic optimization of non-uniform sampling schedules has remained elusive. It is demonstrated here how this challenge can be addressed by using fitted cross-peaks of a reference 2D HSQC experiment as footprints, which are subsequently used to reconstruct cross-peak amplitudes of a pseudo-3D dataset as a function of $\Delta\omega$ by a linear least-squares fit. It is shown for protein Im7 how the approach can yield highly accurate CEST profiles based on an absolutely minimally sampled (AMS) dataset allowing a speed-up of a factor 20 – 30. Spectrum-specific optimized non-uniform sampling (SONUS) schemes based on the Cramer-Rao lower bound metric were critical to achieve such a performance, revealing also more general properties of optimal sampling schedules. This is the first systematic exploration and optimization of NUS schedules for the dramatic speed-up of quantitative multidimensional NMR measurements that minimize unwanted errors.

KEYWORDS:

Protein dynamics, fast multidimensional NMR, CEST, non-uniform sampling, Cramer-Rao lower bound.

INTRODUCTION

NMR spectroscopy is a major source of experimental information about protein dynamics at atomic resolution on a broad range of timescales offering valuable insights about protein function. Common NMR experiments that provide such information include R_1 , R_2 , $R_{1\rho}$, and CPMG relaxation dispersion experiments¹⁻² as well as Chemical Exchange Saturation Transfer (CEST).³⁻⁶ In order to make a maximal number of protein resonances accessible to quantitative dynamics analysis, these experiments are performed in a pseudo-3D manner in the form of stacks of 2D HSQC-type spectra,⁷ where ^{15}N or ^{13}C nuclei are correlated with their directly attached ^1H protons resulting in unique cross-peaks, while an additional parameter is systematically varied along the “pseudo” 3rd dimension. Depending on the experiment, this dimension can correspond to a set of relaxation delays (R_1 and R_2 experiments), the effective radio-frequency (rf) field strength (CPMG, $R_{1\rho}$), or an rf spin-lock offset (CEST, $R_{1\rho}$).² Common to all these experiments is that the 2D HSQC-type spectra retain their resonances (cross-peaks) at the exactly same positions with identical lineshapes and only their amplitudes (or volumes) vary along the pseudo dimension. For some of these experiments the number of points probed along the pseudo dimension can be quite large. For example, for a ^{15}N -CEST experiment with moderately small rf saturation field strength ($\gamma B_1 \ll 25\text{Hz}$) it is common to measure 2D ^1H - ^{15}N HSQC spectra for 100 or more different rf offsets. This can lead to prolonged experiment times of the order of one or several days. For samples with good sensitivity, besides the number of points along the pseudo dimension, the measurement time is determined by the required number of increments N_1 along the indirect ω_1 dimension of each of the ^1H - ^{15}N HSQC-type spectra. For traditional Fourier transform NMR data processing, N_1 is typically around 128 (complex) increments to ensure a digital resolution that exceeds the natural linewidth of most cross-peaks.

The standard method used for pseudo-3D experiments measures each 2D HSQC-type plane separately as a function of the 3rd (pseudo) dimension and subjects it to 2D Fourier transform processing. The fact that cross-peak positions and shapes remain the same in each 2D plane and can be measured before beginning the full experiment is, however, generally not utilized to decrease measurement time or increase the accuracy of the final spectrum. In order to make the NMR time manageable, such information has

been used in a pseudo-4D CEST experiment.⁸ A recent approach (MERT NMR) utilizes such information by parametrizing cross-peak positions and shapes in a reference 2D HSQC plane and determines peak volumes along the pseudo dimension making cross-peaks in crowded regions better accessible to quantitative analysis.⁹

Over the recent past, non-uniform sampling (NUS) has gained significant traction in multidimensional NMR applications by measuring only a subset of t_1 evolution times followed by the reconstruction of the spectra using customized algorithms. For situations where sensitivity is not the limiting factor, sampling can be typically reduced to 50% or 25% of the total number of points amounting to two- to four-fold time savings. NUS implementation can take different forms, such as multi-dimensional decomposition (MDD),¹⁰ maximum entropy,¹¹ or compressed sensing.¹²⁻¹³

NUS invariably requires users to select a sampling schedule, i.e. the specific set of t_1 evolution time increments that are measured. The choice of the sampling schedule has important consequences for the final spectrum, since it directly affects the resolution, the sensitivity, and the possible appearance of artifacts.¹⁴⁻¹⁵ The total number of sampling schedules can be astronomically large, even if one limits oneself to “on-grid sampling” only, i.e. all sampled time points are integer multiples of the Nyquist dwell time $\Delta t_1 = 1/SW$, i.e. $t_1 = k \cdot \Delta t_1$, where SW is the spectral width along the indirect ω_1 dimension (in Hz) and k is an integer between 0 and N_1-1 . Specifically, the number of sampling schedules with n_1 increments chosen from a total of N_1 equidistantly spaced increments is

$$N_{NUS} = \binom{N_1}{n_1}. \text{ For } N_1 = 128 \text{ and } n_1 = 32 \text{ or } 64, N_{NUS} \text{ is } 1.5 \cdot 10^{30} \text{ and } 2.4 \cdot 10^{34},$$

respectively. The very large size of N_{NUS} has prevented the systematic exploration of sampling schedules and the best sampling schedule for a given experiment and sample remains unknown. Experience shows that the quality of a sampling schedule depends on many different factors, such as the number of cross-peaks, their positions in terms of chemical shifts, resonance linewidths, and amplitudes. Empirical rules have been developed for the generation of decent sampling schedules along with metrics that allow one to approximately assess their performance.^{14, 16-18} Randomization of increments has

been found useful to prevent a systematic violation of the Nyquist sampling theorem to minimize systematic artifacts, e.g., through spectral aliasing.¹⁵

Each sampling schedule represents a compromise between (i) a minimal number of increments to speed up data acquisition ($n_1 < N_1$), (ii) the use of increments with short t_1 delays that optimize sensitivity by allowing minimal transverse relaxation, and (iii) the use of long t_1 delays that enhance spectral resolution by disambiguating between cross-peaks with similar chemical shifts along ω_1 . Currently, the vast majority of sampling schedules is generated randomly in either a neutral manner (uniform random sampling¹⁹) or by introducing a bias that favors short t_1 delays over longer delays as implemented, for example, in exponential²⁰ or Poisson-gap sampling.²¹

We recently introduced the absolute minimal sampling (AMS) strategy to accurately reconstruct a spectrum with the absolute minimal number of increments.²² In the present work, we generalize AMS for the reconstruction of peak intensities in pseudo-3D experiments using extreme NUS for the purpose of spectral collection in a minimal amount of time without sacrificing accuracy with the NUS schedule spectrum-specifically optimized based on the Cramer-Rao lower bound metric. Similar to MERT NMR, the new method uses prior knowledge in the form of the 2D HSQC cross-peak footprints whose intensities are modulated along the pseudo dimension. Because the amplitude reconstruction is a linear-least squares problem, it can be solved by linear algebraic methods with high computational efficiency permitting a more systematic analysis of the effect of the sampling schedule on the accuracy of the spectral reconstruction than was previously possible. The new method, which is referred to as AMSi where “i” stands for intensity reconstruction, is demonstrated here for ¹⁵N-CEST where for the best predicted sampling schedule, a speed-up over the standard method of more than a factor of 20 - 30 is possible without introducing undesirable NUS artifacts.

MATERIAL AND METHODS

AMSi theory

An NMR signal is modeled here as a weighted sum of decaying sinusoids.²³⁻²⁴ When a 2D HSQC has been Fourier transformed in the direct dimension, it may be expressed as:

$$s(t_1, \omega_2) = \sum_{j=1}^M a_j e^{i\Omega_j^{(1)}t_1 - R_{2,j}^{(1)}t_1} p_j(\omega_2 - \Omega_j^{(2)}) \quad (1)$$

$$\text{or} \quad \mathbf{s} = \mathbf{F}\mathbf{a} \quad \text{where} \quad F_{ij} = e^{i\omega_j^{(1)}t_{1,i} - R_{2,j}^{(1)}t_{1,i}} p_j(\Omega_i - \omega_j^{(2)})$$

In this equation, M is the total number of cross-peaks in the 2D HSQC footprint and a_j , Ω_j , and $R_{2,j}$ are the amplitude, Larmor frequency, and transverse relaxation rate of the j^{th} peak, respectively. Superscript values on frequencies and relaxation rates indicate the dimension for each of these values. $p_j(\omega)$ is the lineshape of the j^{th} peak along the processed direct dimension. The second equation is the discretized form of the first equation, where $\mathbf{F}$ is a matrix representing the footprint of the spectrum, column vector $\mathbf{a}$ contains the amplitude of each cross-peak as its elements, and the elements of column vector $\mathbf{s}$ represent the t_1 - and ω_2 -dependence of the signal. If the footprint matrix $\mathbf{F}$ is known, e.g. from fitting a reference 2D HSQC spectrum, then the cross-peak amplitudes $\mathbf{a}$ of any experiment with the identical footprint can be reconstructed according to:

$$\mathbf{a}^{fit} = (\text{Re}[\mathbf{F}^\dagger \mathbf{F}])^{-1} \text{Re}[\mathbf{F}^\dagger \mathbf{s}] \quad (2)$$

This reconstruction can be performed with no restriction on the sampling schedule, as long as the number of data points provided is greater than or equal to half the number of peaks M in the footprint (see Supporting Information).

Processing the direct dimension prior to fitting allows for several advantages, including the better separation of peaks along these dimensions and the application of common tools such as zero-filling, apodization, baseline correction, and water signal removal to be used prior to AMSi. However, the shape of the peaks along the Fourier transformed dimension must be carefully considered. The Voigt profile, which is a convolution of a Lorentzian and a Gaussian lineshape with unequal line widths,²⁵ has been previously used in lineshape analysis for apodized signals.²⁶ This lineshape performed best for AMSi reconstruction and is used throughout this work unless noted otherwise.

Larmor frequencies and lineshapes are most easily determined by direct peak fitting as illustrated in Figure 1 for the 2D ^{15}N - ^1H HSQC footprint spectrum of Im7

protein. It is important to ensure that all peaks or features are picked for the faithful representation of the original spectrum and that the peak locations are known with high accuracy. Uninteresting or artifact peaks may be disregarded further downstream, but must be included in the footprint. Accurate lineshapes along dimensions that will be Fourier transformed prior to AMSi are critically important. However, AMSi is less sensitive to errors in lineshapes along non-uniformly sampled dimensions.

The following steps summarize the method of AMSi data processing for an ^{15}N -CEST experiment. They can be directly adopted to other pseudo-3D relaxation experiments.

1. Collect a 2D HSQC reference spectrum
2. Perform peak fitting and determine a footprint (Larmor frequencies and lineshapes for all peaks)
- 3a. Optional: optimize non-uniform sampling schedule along indirect t_1 dimension
- 3b. Choose sampling schedule and collect the CEST experiment using NUS in the indirect dimension
4. Determine the footprint matrix from sampling schedule and footprint
5. Perform linear least-squares minimization as described above to extract intensities of each peak
6. Plot CEST profile and subject it to quantitative analysis (e.g. by ChemEx software)

Cramer-Rao lower bounds for *a priori* scoring of NUS schedules

The Cramer-Rao lower bound (CRLB) is a lower bound on the expected error of a parameter.²⁷ It is the inverse of the Fisher information, which is a way of determining the amount of information provided by an observable about a model parameter. When multiple parameters are being determined, the CRLB is a matrix that is related to the expected covariance matrix. As a result, each element in the diagonal places a lower bound on the expected uncertainty in a peak intensity, and the trace of the CRLB places a lower bound on the expected sum squared error for the estimated parameters.

The CRLB has been used previously in schedule analysis of model-based NMR.²⁸ In the context of NMR signals for which footprints are known, the CRLB is given in Eq. (3) (see SI for details).

$$CRLB = \frac{1}{\sigma^2} (\text{Re}[\mathbf{F}^\dagger \mathbf{F}])^{-1} \quad (3)$$

This value will be used as a metric to predict the scoring of NUS schedules for the purpose of spectrum-specific optimization with one modification: because additive random baseline noise is constant, the inverse variance term $1/\sigma^2$ is dropped. This scoring method was tested for accuracy through comparison to RMS error in amplitudes.

Sample Preparation

The DNA fragment encoding Im7 was PCR-amplified and subcloned into a pTBSG ligation independent cloning vector derivative (pTBSG1).²⁹ The resulting plasmid pTBSG1_Im7 was then transformed into *Escherichia coli* BL21(DE3) strain for protein overexpression. The expressed fusion protein contains a His₆-tag and a TEV protease cleavage site N-terminal to Im7. The overexpression was carried out as follows: a single colony was inoculated to 20 mL LB media under vigorous shaking of 250 rpm overnight at 37 °C, the overnight culture was then transferred into 1L M9 minimal media with 1g ¹⁵NH₄Cl and 5g D-glucose (or 4g D-glucose-¹³C₆ for ¹³C-labeled samples) as the sole nitrogen/carbon sources and incubated at 37°C under vigorous shaking. When OD₆₀₀ of the culture reached 0.8–1.0, isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to it to the final concentration of 0.5mM, and further incubated at 25°C under vigorous shaking for 18 hours. After overexpression, the cells were then pelleted by centrifugation, and lysed by EmulsiFlex-C5 homogenizer (AVESTIN, Inc.) The cell lysate was subjected to centrifugation at 20,000 xg for 20 minutes. The His₆-tagged Im7 protein in the supernatant was purified by a Ni-NTA agarose (QIAGEN) affinity column and mixed with tobacco etch virus (TEV) protease for His₆-tag cleavage. The final Im7 protein, which has three non-native residues (SNA) at its N-terminus, was recovered with a second Ni-NTA affinity column and was concentrated in 50 mM sodium phosphate buffer at pH 7.0.

NMR data collection and processing

^{15}N -CEST spectra were collected with a 25 Hz and 100 Hz saturation field on the colicin E7 immunity protein Im7 at 298 K on a Bruker AVANCE III 850 MHz spectrometer equipped with a cryogenically cooled TCI probe using 128 (complex) indirect time points. This data was processed using NMRPipe,³⁰ and the cross-peaks were assigned using assignments reported in the literature³¹ and confirmed by 3D HNCA and HNCOCA experiments. Kinetic and thermodynamic parameters were extracted from CEST profiles using ChemEx⁴ (<http://www.github.com/gbouvignies/chemex>).

For AMSi processing, the data was subsampled to $n_1 = 2, 3, 4, 6, 8, 12, 16$ complex time points. A footprint was determined from a high-resolution ^1H - ^{15}N HSQC and fit using an in-house program capable of Voigt lineshape fitting. Schedules were selected and AMSi was performed with an in-house program utilizing the nmrglue Python library.³² The root mean square error (RMSE) of the amplitudes was calculated according to:

$$RMSE = \sqrt{\frac{1}{M} \sum_{j=1}^M (a_j^{rec} - a_j^{true})^2} \quad (4)$$

In addition, the data was fit with ChemEx for validation of the method and compared with the fully sampled results.

RESULTS

A fully sampled 2D ^{15}N - ^1H HSQC spectrum of Im7 was collected with 2048 direct time points and 128 indirect time points to extract the spectral footprint (all data points indicate complex points, unless noted otherwise). After standard spectral processing using apodization, zero-filling, and traditional Fourier transform and phasing, the resulting spectrum can be seen in Figure S6. The spectrum has generally a low degree of overlap, but has traces in the indirect dimension with up to 5 peaks making extreme time-

saving methods, such as SPEED³³, inapplicable. Peak picking found 102 peaks, 84 of which were successfully assigned to backbone amides covering all non-proline residues of Im7. Peak fitting was performed using Voigt profiles in both dimensions. The indirect dimension was found to be adequately reconstructed by a Lorentzian lineshape. However, the full Voigt profile was needed along the direct dimension for optimal reconstruction. This information was used to create a footprint of the spectrum for subsequent AMSi processing.

A fully sampled CEST spectrum of Im7 was collected with 128x2048 complex time points along the indirect and direct dimensions. Two B_1 fields were used (target field strengths 25 Hz and 100 Hz, with calibrated field values of 27.4 Hz and 110 Hz), and 116 B_1 offsets used for the 25 Hz experiment and 41 offsets used for the 100 Hz experiment. Four reference planes with no saturation transfer delay were taken throughout the experiment as reference. The 25 Hz experiment took 38.5 hours and the 100 Hz experiment 14.5 hours to collect (the 100 Hz experiment covered a wider range of offset values with fewer increments). Upon fitting with ChemEx, 7 peaks (I44, T45, E46, L53, I54, Y55, Y56) were determined to undergo exchange with a large change in chemical shift (>5.5 ppm), and 13 additional peaks were determined to undergo exchange with a smaller change in chemical shift.

We subsampled the data with $n_1 = 2, 3, 4, 6, 8, 12, 16$ increments using the CRLB-based prediction of the best sampling schedule and recreated the CEST data using AMSi. The $n_1 = 6$ point AMSi reconstruction can be seen in Figure 2 for an overlapped spectral region against the fully sampled spectrum and the fit of individual cross-peaks used for footprinting. It shows that the AMSi reconstructed spectrum represents a highly accurate depiction of both the original spectrum and the derived footprint. The amplitudes for the different numbers of n_1 increments are plotted as complete CEST profiles in Figure 3 for four representative residues I72, L53, E21, and Y55 in comparison with the fully sampled result. All features of the CEST profiles are well-reproduced, including an asymmetry of the main peak (Figure 3b) and the appearance of a minor CEST peak (Figures 3c,d) for all sampling schedules using as few increments as $n_1 = 2$ (magenta line). The same profiles are overlaid in Figure S1, with numerical values of select points listed in Table S1, demonstrating excellent agreement among the profiles

even for $n_1 = 2$. The only visually noticeable feature of smaller numbers of t_1 increments is the increase in noise, which reflects the shorter total acquisition time. For $n_1 = 2, 8, 16$ the data acquisition requires 50 minutes, 3.3 hours, and 6.6 hours vs. 53 hours for the fully-sampled reference spectrum.

Next, the CEST profiles were subjected to a fully quantitative chemical exchange analysis according to the same ChemEx analysis protocol. The number of peaks found to have significant chemical exchange for each value of n_1 is shown in Table 1. All peaks appearing to undergo chemical exchange in the NUS schedule were also found in the fully sampled spectrum, i.e. no artificial chemical exchange effect was introduced due to NUS reconstruction. When only the 7 peaks with a relatively large change in ^{15}N chemical shift difference > 5.5 ppm between ground and excited states are fit globally, there was no noticeable trend in error of each parameter as the number of time points increased (Table S2). Thus, the major benefit of collecting more data points is the improved recovery of exchange effects, especially for some of the most difficult peaks where the secondary CEST dip is close to the main peak (low $\Delta\omega$).

We were able to perform AMSi reconstruction and evaluation for all possible schedules with $n_1 = 2, 3$, and 4 indirect time points collected, with the requirement that the first point ($t_1 = 0$) always be included. We also performed AMSi reconstruction and evaluation for 1 million randomly selected schedules with $n_1 = 6, 8, 12$, and 16 time points. The total RMS error of all amplitudes for each schedule is rank-ordered and plotted in Figure 4. For all sets of schedules, except $n_1 = 2$, the majority of random schedules have similar error, with a small but non-negligible fraction of schedules performing significantly better or worse than others (left and right tails of curves). This change is so significant that an optimal choice of a schedule outperforms a median schedule with a 2 – 3 times larger number of indirect points taken. Similarly, a poorly chosen schedule will result in a performance comparable to a median schedule with 2 – 3 times fewer indirect points.

The best schedules as determined by true error in amplitude are given in Table 2. They show a clear preference for early time points and for relatively large gaps between sampled increments. The relative frequency of each time point occurring in the top and bottom 0.5% of the schedules is plotted in Figure 5 for $n_1 = 4, 8$, and 16. The best

schedules have a similar shape to an exponential distribution with a finite offset, similar to exponential sampling as has been previously suggested for standard NUS applications.^{11, 20, 34} The decay of this distribution becomes steeper when focusing on the absolute best set of schedules (see Figure S2). There is also an initial buildup of fractional occurrence in time points 1-3, with time point 4 being the most likely point chosen in all three displayed schedules, which is reflected in the set of best NUS schedules. For the 0.5% of the best sampling schedules, certain pairs of time points show distinct preferences and anti-preferences to be co-sampled (see SI Figure S3).

Although the CRLB-based scoring method does not identify the absolute best sampling schedules for each of the different n_1 values, it predicts very good sampling schedules as can be seen in Figure 4 where the best CRLB-based schedules (filled circles) are close to the left end. This is further corroborated in Figure 6 where for each of the schedules tested the RMS error is plotted against the CRLB trace. For all n_1 values (4, 8, 16), the relationship is funnel shaped where schedules with low predicted CRLB scores possess also low RMS errors. For all n_1 values (except $n_1 = 2$), the best schedule by CRLB score is within the top 5% of schedules by RMS error with the best scores obtained for larger n_1 values. As a demonstration of the accuracy of these predicted schedules, they were used for the AMSi CEST profiles depicted in Figure 3.

On the other end, schedules with poor CRLB scores have a much broader RMS error distribution, which is however inconsequential for practical applications. The consequences of the best, median, and worst sampling schedule are illustrated in Figure 7 for the reconstruction of CEST profiles of E25 and D35. As can be seen, the best sampling schedule reproduces the reference profile very accurately, whereas the median and worst schedules show shoulder effects, which may be misinterpreted as excited states.

Taken together, our results suggest that CRLB-based scoring of schedules is a reliable method for the generation of spectrum-specifically optimized NUS schedules (SONUS) avoiding schedules with high RMS errors that can produce misleading artifacts in CEST profiles. Especially for small n_1 values, which promise the largest gain in measurement time, the choice of a purely random schedule is *not* recommended. Instead, the CRLB-based scoring method is capable of correctly predicting very good schedules.

DISCUSSION

Non-uniform sampling permits the substantial shortening of the measurement time of multidimensional NMR spectra when the measurement is dictated by sampling of indirect time points rather than sensitivity. NUS is now routinely applied to many multidimensional NMR experiments, especially those that provide NMR resonance assignment information of proteins in solution.³⁵ By contrast, NUS is only rarely applied to pseudo-3D experiments, such as protein spin relaxation or CEST experiments, for the fully quantitative biophysical characterization of structural dynamics. Application of standard NUS to each HSQC plane is possible, but allows only a relatively modest speed-up. For ^{15}N - ^1H HSQC planes a reduction of the total number of t_1 increments to 64 or 32 may be possible resulting in a speed-up of a factor 2 – 4.

By definition, NUS requires a user to make a choice between many possible sampling schemes carrying the risk of spectral artifacts caused by the sampling scheme, rather than random noise. The sampling scheme must be good or, at least, acceptable, because after a NUS dataset has been collected, standard Fourier transform is not available any longer for the independent assessment of the accuracy of reconstructed spectrum.

Among the many pseudo-3D experiments used in protein NMR, CEST is one of the most challenging experiments as in many proteins the vast majority of cross-peaks do not show chemical exchange effects and those that do need to be identified with high accuracy. Hence, for any NUS-CEST scheme reliability is paramount whereby the number of both false positive and false negative CEST effects are to be kept extremely small and, if possible, at zero. NUS-CEST also provides a challenge for reconstruction as CEST planes invariably include a very wide range of peak amplitudes, including some near-zero amplitudes, that must be accurately reconstructed, while most other pseudo-3D experiments display a much more limited dynamic range.

The AMSi strategy introduced here promises to meet this challenge as it permits the very reliable reconstruction of pseudo-3D CEST spectra with n_1 in the single digits and some instances as small as $n_1 = 2$ increments. This reduction in n_1 allows an NMR time speed up from over 2 days to under an hour.

For AMSi to be successful, the following conditions need to be met and key steps need to be followed. First, the reference HSQC spectrum needs to be footprinted accurately by decomposing it into a sum of cross-peaks with Voigt profiles. For this purpose, a high-quality 2D HSQC spectrum has to be collected first and a non-linear least squares fitting problem has to be solved to obtain the resulting footprints. Since such a 2D ^{15}N - ^1H HSQC is typically one of the first NMR spectra recorded for any given protein system, it does not add to the total NMR time of the project. This step needs to be done only once and the cross-peak footprints can subsequently be applied to any other pseudo-3D experiment based on the same HSQC spectrum (R_1 , R_2 , CPMG, CEST, DOSY) provided that the HSQC closely matches the reference plane of the pseudo-3D experiment.

Next, a NUS schedule needs to be chosen that ensures good performance. For many NUS applications with a relatively large number of increments ($n_1 > 32$) of increments, a randomized selection usually works just fine. However, for the kind of extreme speed-up described here with n_1 in the single digits, a randomly selected schedule has a non-negligible chance to have poor performance with major consequences on the quality of the resulting CEST profiles. We show that the Cramer-Rao lower bound (CRLB) metric permits the spectrum-specific optimization of sampling schedule performance allowing one to choose a very good to excellent schedule. Because the spectral reconstruction can be expressed in terms of linear algebra with good computational efficiency, an unprecedented large number of schedules can be tested and the one with the best CRLB metric can then be used for the actual experiment.

The analysis of a large number of sampling schedules performed here allowed, for the first time, an independent assessment of commonly used NUS strategies, such as exponential and Poisson gap sampling. Statistically, the best sampling schedules identified here sample early t_1 increments significantly more frequently than later time points in order to optimize the sensitivity. The frequency drops approximately exponentially toward a plateau value, which approaches zero for small n_1 . It is interesting to note that certain t_1 increments are much more frequently found than others. For example, $t_1 = 4\Delta t_1$, $8\Delta t_1$, $16\Delta t_1$, $25\Delta t_1$ are increments that have a significantly higher chance to be found in some of the best performing schedules than other increments,

whereas some increments, such as $t_1 = \Delta t_1$, $17\Delta t_1$, are clearly unfavorable. We expect that these “special” increments directly depend on the actual cross-peak distribution and are not transferrable between spectra.

The theoretical minimal number of increments is determined by:

$$2N_2n_1 \geq M \quad (5)$$

where M is the total number of footprinted cross-peaks, including spurious peaks, and N_2 is the number of direct time points collected. As there is essentially no time cost when increasing N_2 , this requirement can be simplified to $n_1 \geq 1$ (see Supporting Information). This assumes a sufficiently high signal-to-noise, which in practice may not be fulfilled. Such effects are reflected in higher CEST baseline noise levels (Figure 3) and the lower recovery (by software such as ChemEx) of small CEST effects, especially shoulder peaks of the main CEST peak (Table 1). The fitting errors of CEST-derived thermodynamic and kinetic parameters obtained using ChemEx (Table S2) do not follow any trend with the number of peaks fitted. In most cases, the error is less than 5%, which is larger than the errors in the traditionally constructed CEST profiles supplied by ChemEx (1.2% for k_{ex} , 0.76% for p_b).

Poor footprinting or poor schedules may have a detrimental effect on CEST profiles. Beyond an increase in noise due to sampling low signal-to-noise time points, a poor schedule can also cause instability in the shoulder of the main CEST peak (Figure 7). Even more notable are the effects of using a poor footprint, often due to inaccurate R_2 or lineshape in the direct dimension. The most commonly observed artifacts were “shadow features”, which is the effect of one peak subsuming a part of another peak within the same trace. For example, when Lorentzian lineshapes are used for the spectra processed above for the best schedule, systematic artifacts such as those seen in Figure S4 occur. Fortunately, these shadow peaks have consistent behavior, causing the appearance that the main peak either oversaturated or failed to fully saturate, and causing a secondary peak to appear in the profile either as a rise or dip, respectively. In addition, footprints can be improved or reconstructed after an experiment has been run in case that it is discovered that the original footprint is inadequate.

It is noteworthy that the majority of schedules perform remarkably similarly (Figure 4), especially for larger n_1 , with only significant changes in error occurring in the

best and worst few percent. However, for all sets of schedules (except $n_1 = 2$) the best two schedules only differ by at most 3% error, i.e. there is only limited benefit by identifying the very best schedule among $\sim 10^6$ candidate schedules. It is noted that the baseline noise in the best schedule does not scale with the square root of the number of points (Figure S5) and, instead, performs significantly better than would be expected from square root scaling. This suggests that the substantial speed-up of AMSi afforded over traditionally sampled experiments can also be leveraged for the reduction of noise by using the extra time to collect more scans.

The CRLB-based method for predicting the best schedule presented in this work is highly suitable for the optimization of NUS schedules, but there still exists some room for improvement in prediction of the absolute best schedule and, less importantly, in prediction of poor schedules. For those who intend to search for the absolute best schedule, we note that the time points of top schedules significantly differ from each other (Table S3) and there is only little correlation between chosen points in optimal schedules (Figure S3). This suggests that “greedy” search methods are unlikely to converge to the global optimum.

AMSi shares features with other NUS methods in the literature. It is algorithmically similar to MDD if peak positions and lineshapes were separately supplied and held constant, and it is the final result of AMS when restraining both frequencies and relaxation constants. Most notably, it is conceptually similar to the pseudo-4D method by Long et. al.,⁸ from which it differs by its use of varying lineshape for the direct dimension, and its removal of the option for small corrections in frequency in order to optimize speed and interpretability of the process. AMSi is also similar to MERT, but AMSi uses an interferogram instead of full time-domain data and samples the indirect time-domain non-uniformly to achieve large time reductions.

In this work, we restricted ourselves to the analysis of CEST data as this pseudo-3D experiment has the greatest potential for time savings through NUS. However, AMSi is capable of being used with any pseudo-3D experiment where peaks do not drift or change shape between planes, including R_1 , $R_{1\rho}$, DEST, CPMG relaxation experiments, and DOSY. Although these experiments typically collect fewer 2D planes, and thus are shorter experiments with less absolute time to be saved by NUS, we expect AMSi to

achieve the same 20-30 fold time reduction without compromising the quality of the data. It can also be combined with multi-frequency saturation methods that improve CEST speed by reducing the number of planes taken, such as cos-CEST and D-CEST.³⁶⁻³⁸

CONCLUSIONS

AMSi provides a general framework for the speed-up of CEST or any other pseudo-3D experiment. We anticipate this method to be especially useful for the rapid and yet accurate dynamic screening of cohorts of protein samples for alternative conformational states under different conditions (free vs. ligand bound, wild-type vs. mutants, variable temperature, etc.) and the simultaneous determination of R_1 and R_2 relaxation parameters for model-free analysis.³⁹ It will also be useful for the analysis of CEST spectra with very low B_1 fields that would typically require a week or longer by traditional sampling. The CRLB-based scoring method is designed to be performed with an AMSi experiment, but may be also applicable to other NUS experiments, and the identification of trends displayed by good schedules can be used for future schedule generators.

ASSOCIATED CONTENT

Additional information showing the method used for automated CEST fitting, along with a discussion of the linear least squares approach used to obtain Eq. (2) and a derivation of Eq. (3), is available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bruschweiler.1@osu.edu

Notes

The authors declare no competing financial interest.

ACKNOWLEDGEMENTS

We thank Dr. Lewis E. Kay for providing the plasmid of Im7 and AkzoNobel for the silica nanoparticles and Ms. Xinyao Xiang for confirmation of the resonance assignments of Im7. This work was supported by the NSF (award MCB-1715505) and the National Institutes of Health (grant R01 GM 066041). All NMR experiments were performed at the Campus Chemical Instrument Center NMR facility at the Ohio State University. Computations were performed at the Ohio Supercomputer Center (OSC).

REFERENCES

1. Mittermaier, A.; Kay, L. E. New tools provide new insights in NMR studies of protein dynamics. *Science*. **2006**, *312* (5771), 224-228.
2. Palmer, A. G., III. NMR characterization of the dynamics of biomacromolecules. *Chem. Rev.* **2004**, *104* (8), 3623-3640.
3. Fawzi, N. L.; Ying, J.; Ghirlando, R.; Torchia, D. A.; Clore, G. M. Atomic-resolution dynamics on the surface of amyloid- β protofibrils probed by solution NMR. *Nature*. **2011**, *480* (7376), 268-272.
4. Vallurupalli, P.; Bouvignies, G.; Kay, L. E. Studying "invisible" excited protein states in slow exchange with a major state conformation. *J. Am. Chem. Soc.* **2012**, *134* (19), 8148-8161.
5. Anthis, N. J.; Clore, G. M. Visualizing transient dark states by NMR spectroscopy. *Q Rev Biophys.* **2015**, *48* (1), 35-116.
6. Vallurupalli, P.; Sekhar, A.; Yuwen, T.; Kay, L. E. Probing conformational dynamics in biomolecules via chemical exchange saturation transfer: a primer. *J. Biomol. NMR*. **2017**, *67* (4), 243-271.
7. Cavanagh, J.; Fairbrother, W. J.; Palmer, A. G.; Rance, M.; Skelton, N. J., Experimental NMR Relaxation Methods. In *Protein NMR Spectroscopy (Second Edition)*, Academic Press: 2007; Vol. San Diego, pp 679-724.
8. Long, D.; Delaglio, F.; Sekhar, A.; Kay, L. E. Probing Invisible, Excited Protein States by Non-Uniformly Sampled Pseudo-4D CEST Spectroscopy. *Angew. Chem., Int. Ed.* **2015**, *54* (36), 10507-10511.
9. Matviychuk, Y.; Bostock, M. J.; Nietlispach, D.; Holland, D. J. Time-domain signal modelling in multidimensional NMR experiments for estimation of relaxation parameters. *J. Biomol. NMR*. **2019**, *73* (3-4), 93-104.
10. Orekhov, V. Y.; Ibraghimov, I. V.; Billeter, M. MUNIN: a new approach to multi-dimensional NMR spectra interpretation. *J. Biomol. NMR*. **2001**, *20* (1), 49-60.
11. Hoch, J. C.; Maciejewski, M. W.; Mobli, M.; Schuyler, A. D.; Stern, A. S. Nonuniform sampling and maximum entropy reconstruction in multidimensional NMR. *Acc. Chem. Res.* **2014**, *47* (2), 708-717.

12. Kazimierczuk, K.; Orekhov, V. Y. Accelerated NMR spectroscopy by using compressed sensing. *Angew. Chem., Int. Ed.* **2011**, *50* (24), 5556-5559.
13. Holland, D. J.; Bostock, M. J.; Gladden, L. F.; Nietlispach, D. Fast multidimensional NMR spectroscopy using compressed sensing. *Angew. Chem., Int. Ed.* **2011**, *50* (29), 6548-6551.
14. Hyberts, S. G.; Arthanari, H.; Wagner, G. Applications of non-uniform sampling and processing. *Top. Curr. Chem.* **2012**, *316*, 125-148.
15. Hyberts, S. G.; Arthanari, H.; Robson, S. A.; Wagner, G. Perspectives in magnetic resonance: NMR in the post-FFT era. *J. Magn. Reson.* **2014**, *241*, 60-73.
16. Aoto, P. C.; Fenwick, R. B.; Kroon, G. J.; Wright, P. E. Accurate scoring of non-uniform sampling schemes for quantitative NMR. *J. Magn. Reson.* **2014**, *246*, 31-35.
17. Mobli, M.; Maciejewski, M. W.; Schuyler, A. D.; Stern, A. S.; Hoch, J. C. Sparse sampling methods in multidimensional NMR. *Phys. Chem. Chem. Phys.* **2012**, *14* (31), 10835-10843.
18. Sidebottom, P. J. A new approach to the optimisation of non-uniform sampling schedules for use in the rapid acquisition of 2D NMR spectra of small molecules. *Magn. Reson. Chem.* **2016**, *54* (8), 689-694.
19. Mobli, M.; Stern, A. S.; Hoch, J. C. Spectral reconstruction methods in fast NMR: reduced dimensionality, random sampling and maximum entropy. *J. Magn. Reson.* **2006**, *182* (1), 96-105.
20. Barna, J. C. J.; Laue, E. D.; Mayger, M. R.; Skilling, J.; Worrall, S. J. P. Exponential Sampling, an Alternative Method for Sampling in Two-Dimensional Nmr Experiments. *J. Magn. Reson.* **1987**, *73* (1), 69-77.
21. Hyberts, S. G.; Takeuchi, K.; Wagner, G. Poisson-gap sampling and forward maximum entropy reconstruction for enhancing the resolution and sensitivity of protein NMR data. *J. Am. Chem. Soc.* **2010**, *132* (7), 2145-2147.
22. Hansen, A. L.; Brüscheweiler, R. Absolute Minimal Sampling in High-Dimensional NMR Spectroscopy. *Angew. Chem., Int. Ed.* **2016**, *55* (45), 14169-14172.
23. Levitt, M. H., NMR Spectroscopy. In *Spin dynamics: Basics of Nuclear Magnetic Resonance*, 2nd ed.; John Wiley & Sons: Chichester, England, 2008; pp 39-61.

24. Rule, G. S.; Hitchens, T. K., Practical Aspects of Acquiring NMR Spectra. In *Fundamentals of Protein NMR Spectroscopy*, Springer: Dordrecht, 2006; pp 29-64.
25. Armstrong, B. H. Spectrum Line Profiles - Voigt Function. *J. Quant. Spectrosc. Radiat. Transfer.* **1967**, 7 (1), 61-+.
26. Niklasson, M.; Otten, R.; Ahlner, A.; Andresen, C.; Schlagnitweit, J.; Petzold, K.; Lundstrom, P. Comprehensive analysis of NMR data using advanced line shape fitting. *J. Biomol. NMR.* **2017**, 69 (2), 93-99.
27. Cramér, H. *Mathematical Methods of Statistics*. Princeton University Press: Princeton, 1946.
28. Sward, J.; Elvander, F.; Jakobsson, A. Designing sampling schemes for multi-dimensional data. *Signal Processing.* **2018**, 150, 1-10.
29. Showalter, S. A.; Bruschweiler-Li, L.; Johnson, E.; Zhang, F.; Bruschweiler, R. Quantitative lid dynamics of MDM2 reveals differential ligand binding modes of the p53-binding cleft. *J. Am. Chem. Soc.* **2008**, 130 (20), 6472-6478.
30. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR.* **1995**, 6 (3), 277-293.
31. Whittaker, S. B.; Spence, G. R.; Gunter Grossmann, J.; Radford, S. E.; Moore, G. R. NMR analysis of the conformational properties of the trapped on-pathway folding intermediate of the bacterial immunity protein Im7. *J. Mol. Biol.* **2007**, 366 (3), 1001-1015.
32. Helmus, J. J.; Jaroniec, C. P. NmrGlue: an open source Python package for the analysis of multidimensional NMR data. *J. Biomol. NMR.* **2013**, 55 (4), 355-367.
33. Kupce, E.; Freeman, R. SPEED: single-point evaluation of the evolution dimension. *Magn. Reson. Chem.* **2007**, 45 (9), 711-713.
34. Palmer, M. R.; Suiter, C. L.; Henry, G. E.; Rovnyak, J.; Hoch, J. C.; Polenova, T.; Rovnyak, D. Sensitivity of nonuniform sampling NMR. *J. Phys. Chem. B.* **2015**, 119 (22), 6502-6515.
35. Kazimierczuk, K.; Orekhov, V. Non-uniform sampling: post-Fourier era of NMR data collection and processing. *Magn. Reson. Chem.* **2015**, 53 (11), 921-926.

36. Leninger, M.; Marsiglia, W. M.; Jerschow, A.; Traaseth, N. J. Multiple frequency saturation pulses reduce CEST acquisition time for quantifying conformational exchange in biomolecules. *J. Biomol. NMR*. **2018**, *71* (1), 19-30.
37. Yuwen, T.; Bouvignies, G.; Kay, L. E. Exploring methods to expedite the recording of CEST datasets using selective pulse excitation. *J. Magn. Reson.* **2018**, *292*, 1-7.
38. Yuwen, T.; Kay, L. E.; Bouvignies, G. Dramatic Decrease in CEST Measurement Times Using Multi-Site Excitation. *Chemphyschem.* **2018**, *19* (14), 1707-1710.
39. Gu, Y.; Hansen, A. L.; Peng, Y.; Bruschweiler, R. Rapid Determination of Fast Protein Dynamics from NMR Chemical Exchange Saturation Transfer Data. *Angew. Chem., Int. Ed.* **2016**, *55* (9), 3117-3119.

Tables and Figures with Captions

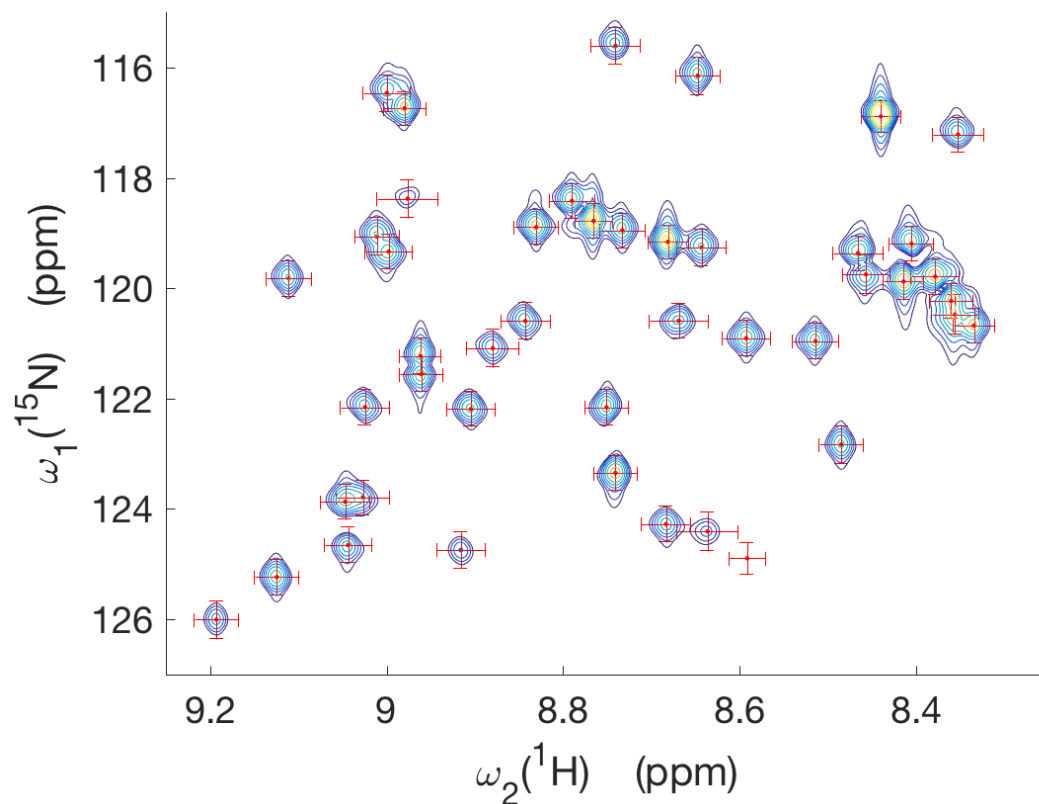


Figure 1. Section of 2D ^{15}N - ^1H HSQC “footprint” spectrum of Im7 protein (blue-to-yellow contours) with fitted footprints (red). A footprint of cross-peak, which is independent of its intensity (volume), captures the location and lineshape, including linewidth. The crosshairs (red) represent “footprints”, i.e. peak positions (centers) and effective R_2 relaxation parameters (cross-hair widths) along each dimension. AMSi uses parametrized footprints to reconstruct peak volumes in pseudo-3D experiments.

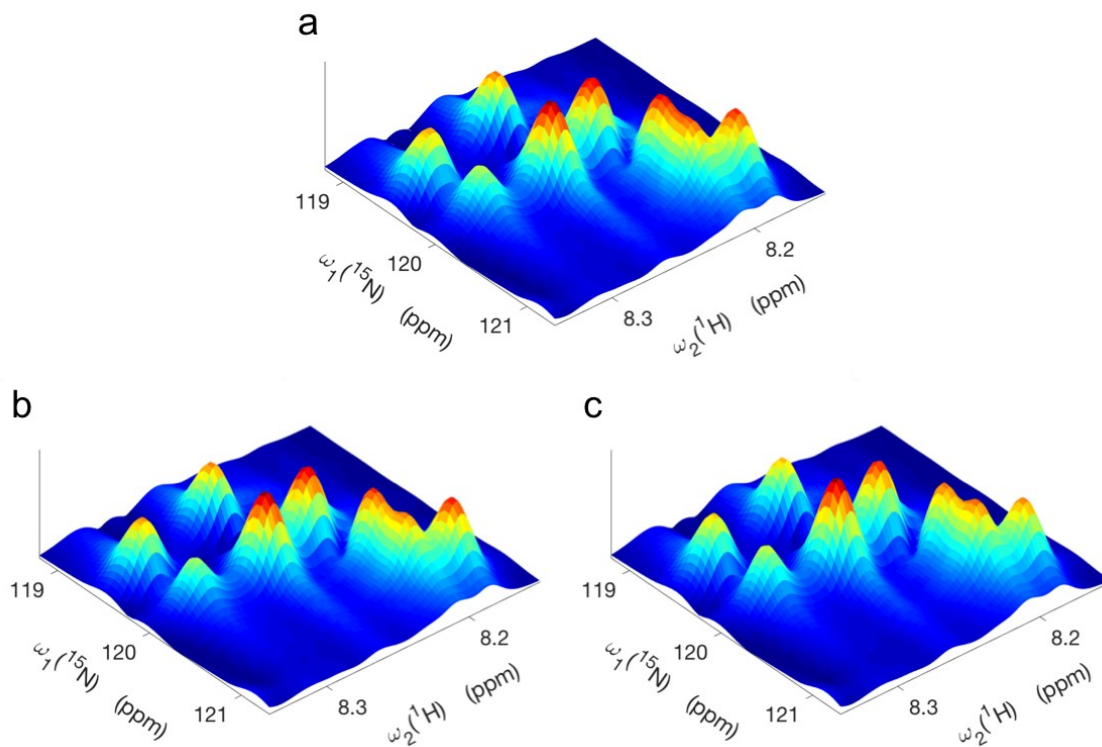


Figure 2. Portion of the (a) original Im7 spectrum, (b) fitted spectrum generated as a superposition of individual fitted cross-peaks using Voigt profiles, and (c) reconstructed spectrum via AMSi using 4.7% sampling along indirect t_1 dimension. The spectra are displayed as 3D surface plots to allow for a detailed visual comparison of peak amplitudes.

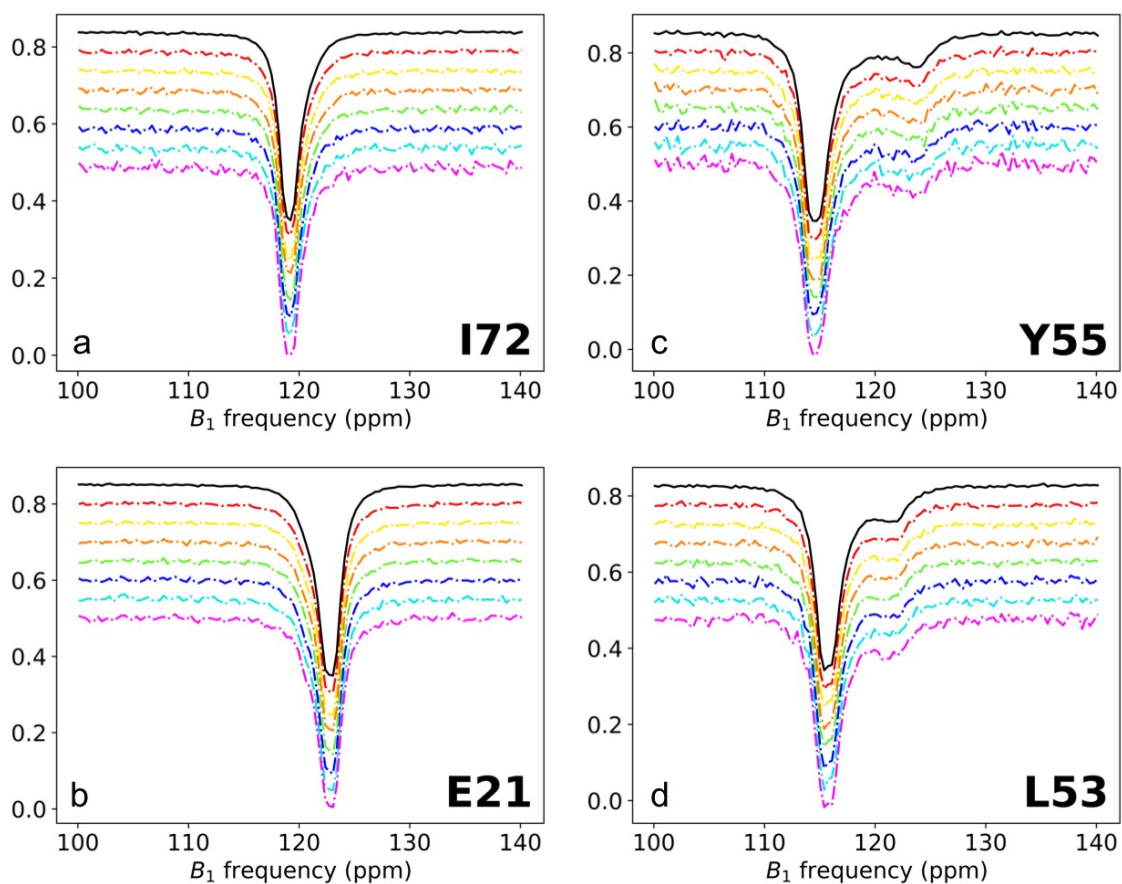


Figure 3. ^{15}N -CEST profiles of 4 cross-peaks of Im7 using both fully sampled and traditionally processed data (black, solid line), and NUS data with AMSi processing using 2 (magenta), 3 (cyan), 4 (blue), 6 (green), 8 (orange), 12 (yellow), and 16 (red) indirect complex t_1 time points for each of the 116 CEST offset frequency. NUS schedules were chosen using the best predicted schedule of a total of 10^6 randomly chosen schedules according to the CRLB trace metric, except for $N_1 = 2, 3, 4$ where all schedules were systematically enumerated and tested. The same data without vertical displacement are depicted in Figure S1.

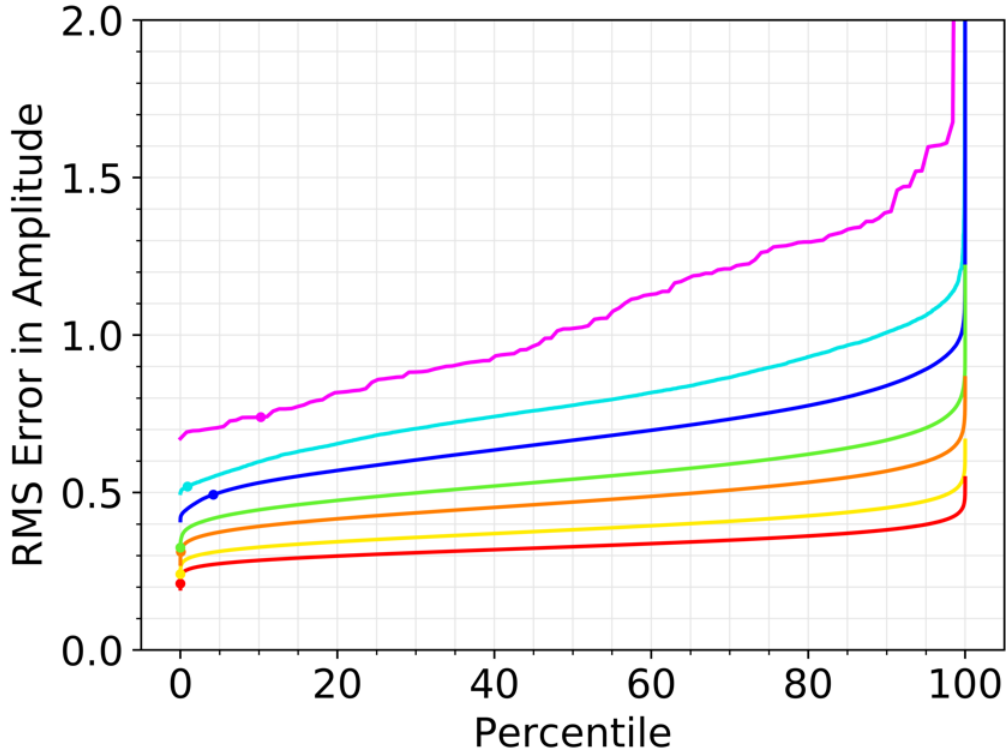


Figure 4. Rank-ordered RMS errors of CEST amplitudes for different values of N_1 . The fully sampled ($N_1 = 128$) and standard 2D FT processed dataset was used as reference for the determination of errors. Displayed curves belong to $N_1 = 2$ (magenta), 3 (cyan), 4 (blue), 6 (green), 8 (orange), 12 (yellow), and 16 (red). $N_1 = 2, 3$ and 4 data sets include all possible schedules, while the other data sets represent 10^6 randomly chosen schedules. Large dots indicate the best predicted schedules. Vertical and horizontal axes have been restricted to better display the majority of schedules. Typical peak amplitudes vary between 30 and 50.

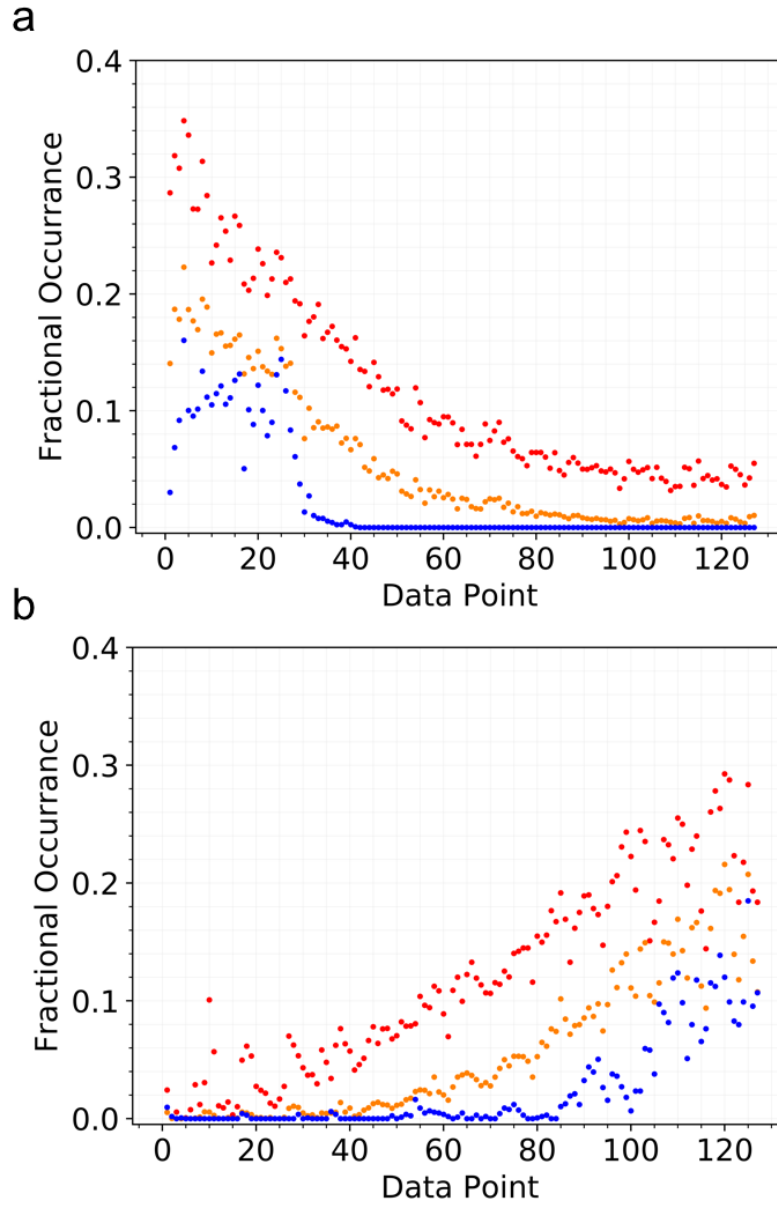


Figure 5. Fractional occurrence (relative frequency) of each data point in the best (a) and worst (b) NUS schedules. Best and worst schedules are the top and bottom 0.5% of all schedules tested as determined by RMS error in amplitude (with the fully sampled and standard 2D FT processed data taken as reference). Points displayed belong to schedules with $N_1 = 4$ (blue), 8 (orange), and 16 (red) indirect t_1 time points.

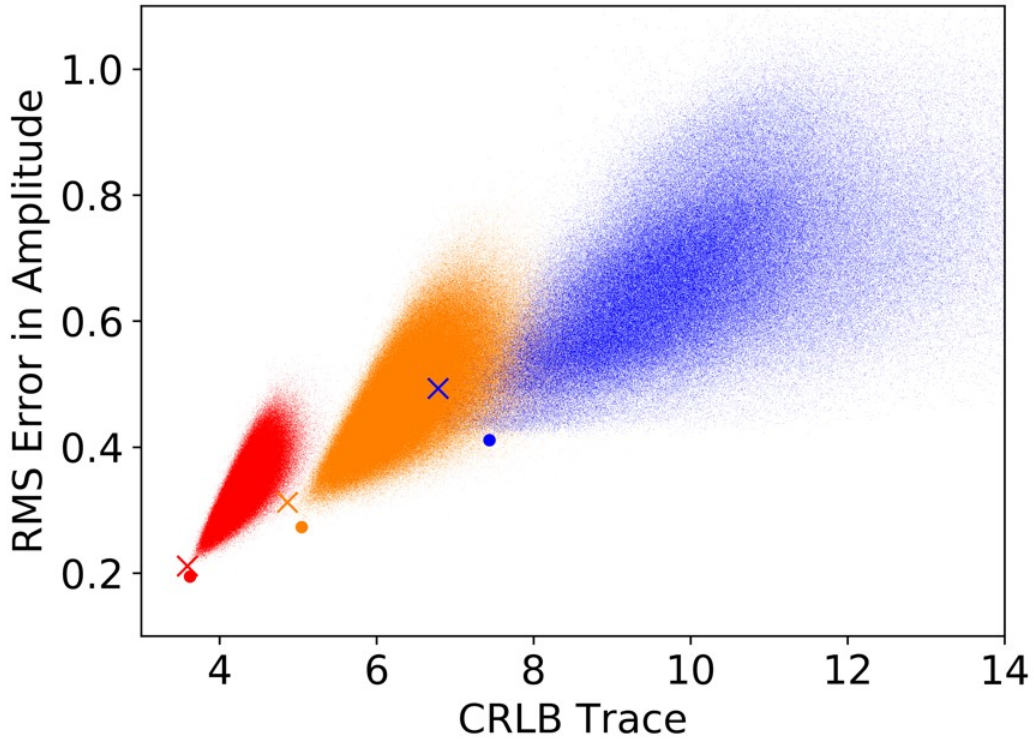


Figure 6. RMS error in amplitude plotted against the CRLB trace metric for large number of schedules (up to 10^6) using $N_1 = 4$ (blue), 8 (orange), and 16 (red) (with the fully sampled and traditionally processed amplitudes taken to be the standard for determination of error). The $N_1 = 4$ data set includes all 333,375 possible schedules, whereas other data sets have 10^6 randomly selected schedules plotted. Large dots indicate the schedule with lowest true error for each N_1 value and “x” symbols indicate the best scoring schedule for each N_1 value according to the CRLB trace. Vertical axis has been restricted to allow better comparison of schedules.

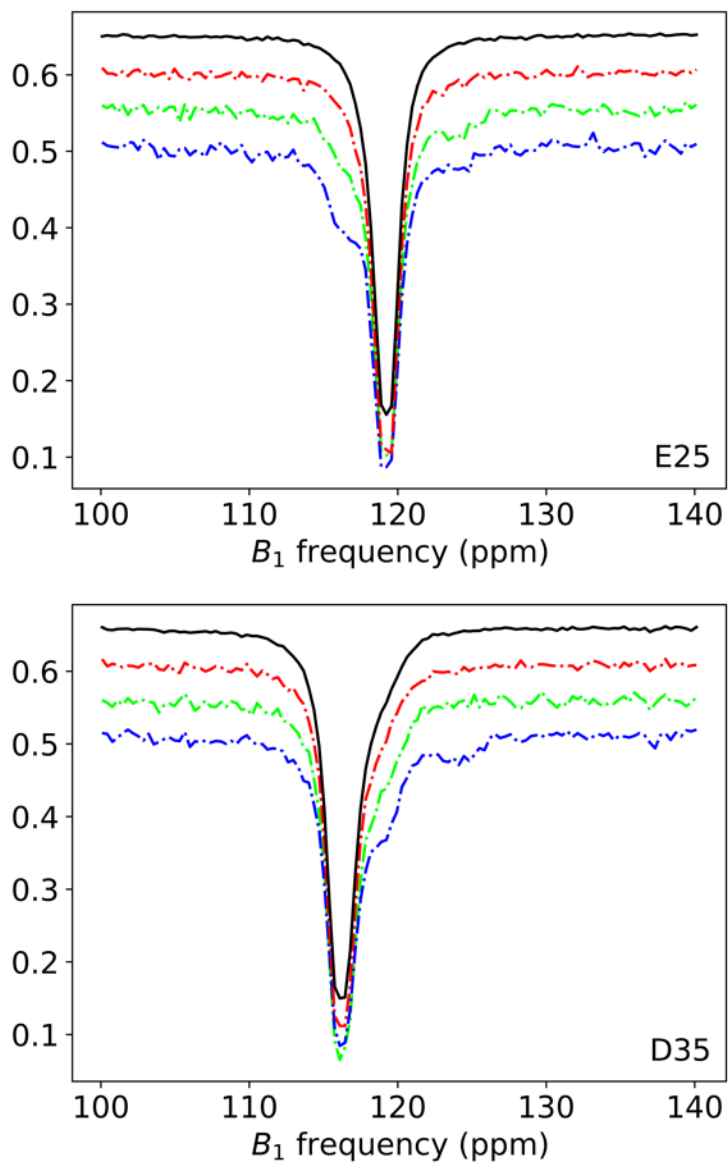


Figure 7. Demonstration of the effect of poor sampling schedules on the corruption of ^{15}N -CEST profiles. The panels belong to Im7 spectra of (a) residue Glu25 and (b) residue Asp35 using both fully sampled and traditionally processed data (black, solid line) and NUS data with AMSi processing using $N_1=6$ indirect complex time points using the best (red), median (green), and worst (blue) schedules (according to their errors with respect to the full sampling) (see also Table 2).

Table 1. Number of peaks that undergo chemical exchange detected for different values of N_1 .

N_1	128 ^a	2	3	4	6	8	12	16
No. of peaks with $\Delta\omega$ > 5.5 ppm	7	5	6	6	6	7	6	7
No. of peaks with $\Delta\omega$ < 5.5 ppm	13	2	3	7	8	10	11	12

^a Processed by traditional FT

Table 2. Absolute best and worst schedules for different values of N_1 based on RMS error in amplitude. Each number corresponds to the evolution delay t_1 expressed as integer multiple of the dwell time Δt_1 (by default, $t_1 = 0$ is included in each schedule).

N_1	best NUS schedule	worst NUS schedule
2^b	0, 13	0, 90
3^b	0, 6, 24	0, 60, 122
4^b	0, 4, 9, 14	0, 86, 90, 91
6	0, 9, 11, 20, 24, 26	0, 53, 56, 107, 110, 111
8	0, 4, 8, 12, 16, 22, 25	0, 110, 111, 119, 123, 124, 125, 126
12	0, 3, 8, 11, 12, 13, 17, 18, 21, 24, 31, 102	0, 89, 91, 98, 105, 115, 118, 121, 122, 125, 126, 127
16	0, 1, 3, 6, 7, 9, 10, 11, 12, 13, 16, 17, 32, 34, 35, 56	0, 38, 67, 77, 85, 87, 95, 99, 104, 109, 113, 114, 117, 118, 120, 121

^b For schedules with $N_1 \leq 4$, all possible schedules (with the first time point always taken) were analyzed.

TOC Figure

