Rapid Quantification of Protein Secondary Structure Composition from a Single Unassigned 1D ^{13}C NMR Spectrum

Haote Li[†,1], Marcus D. Tuttle[†,1], Kurt W. Zilm[1,*], Victor S. Batista[1,*]

[1]Department of Chemistry, Yale University, New Haven, CT 06520.

[†,*] These authors contributed equally to this work.

Correspondence to K.W.Z. or V.S.B.

Abstract

The function of a protein is predicated upon its three dimensional fold. Representing its complex structure as a series of repeating secondary structural elements is one of the most useful ways by which we study, characterize, and visualize a protein. Consequently, experimental methods that quantify the secondary structure content allow us to connect a protein's structure to its function. Here, we introduce an automated gradient descent-based method we refer to as Secondary Structure Distribution by NMR that allows for rapid quantification of the protein secondary structure composition of a protein from a single, 1D ^{13}C NMR spectrum without chemical shift assignments. The analysis of nearly 900 proteins with known structure and chemical shifts demonstrates the capabilities of our approach. We show that these results rival alternative techniques such as FT-IR and circular dichroism that are commonly used to estimate secondary structure compositions. The resulting method requires only the primary sequence of the protein and its referenced ^{13}C NMR spectrum. Each residue is modeled in an ensemble of secondary structures with percentage contributions from random coil, α-helix, and β-sheet secondary structures obtained by minimizing the difference between a simulated and experimental 1D ^{13}C NMR spectrum. The capabilities of the method are demonstrated as applied to samples at natural abundance or enriched in ^{13}C, acquired by

either solution or solid-state NMR, and even on low magnetic field benchtop NMR spectrometers. This approach allows for rapid characterization of protein secondary structure across traditionally challenging to characterize states including liquid-liquid phase-separated, membrane-bound, or aggregated states.

Introduction

The relationship between protein structure and biological function is central to structural biology. One of the most useful organizing principles for communicating and classifying the complex structures of proteins is that of secondary structure elements. Our understanding of protein structure is greatly aided by visualization of three-dimensional arrangements of protein domains adopting canonical secondary structural forms. Therefore, a wide range of computational tools have been developed to predict, classify, and recognize secondary structure domains in protein structures(1). A variety of spectroscopic tools have also been developed to experimentally determine secondary structures and monitor changes in secondary structures induced by environmental conditions(2-4). For example, circular dichroism (CD) is routinely applied to determine changes in secondary structure elements of a protein upon ligand binding or melting.

A limitation of any optical spectroscopy-based secondary structure determination is the difficulty in comparing results from dilute aqueous phase samples with data from liquid protein droplets, condensed complex coacervates, and/or solids. At the same time, detection of changes in protein conformation within these states is of prime importance for understanding their role in biological function and disease. Unfortunately, routine characterization of those states is typically difficult for optical techniques due to light scattering (2, 5). Nuclear Magnetic Resonance (NMR) spectroscopy, on the other hand, is immune to such effects, and able to probe solution, gel and solid phases equally well (6, 7). This has allowed NMR to analyze

protein conformation and its response to environmental conditions across a variety of biologically relevant phases, including *in vivo (8)*.

NMR spectroscopy, especially ^{13}C NMR, is a powerful method for mapping secondary structure elements onto the protein primary sequence (9-11). ^{13}C NMR chemical shifts for CO, Cα, and Cβ carbons are largely a reflection of protein backbone torsion angles and thus their secondary structures (9, 10). For example, Figure 1A shows database(12) ^{13}C chemical shift distributions for alanine and isoleucine colored by their assigned secondary structures; α-helix as red, β-sheet as blue, and grey as random coil. In the case of alanine (Figure 1A, top), the peak of the respective secondary structure-dependent chemical shift distributions are well separated. In this case, any one of the backbone chemical shifts can lead to relatively confident identification of secondary structure for a single residue. Isoleucine, in contrast (Figure 1A, bottom), has less separated distributions. The CO and Cα atoms are significantly overlapped for β-sheet and random coil, while the Cβ shifts are instead overlapped between α-helix and random coil. Figure 1B plots the mode ^{13}Cα chemical shift for each amino acid color coded by secondary structure, showing that this shift is good at distinguishing α-helices for all amino acids.

Biological NMR studies have long utilized the relationship between chemical shift and secondary structure. The chemical shift index, for example, reports the presence of α-helical or β-sheet content according to the departure of an observed chemical shift of a residue relative to its random coil average(11, 13). Alternatively, tools such as TALOS-N use chemical shifts to directly predict the backbone dihedral angles which correspond to specific secondary structure elements (14). While powerful and information-rich, the use of NMR for this purpose is labor intensive, since sequential site-specific assignments of backbone chemical shifts are required. In many instances, however, the percentage composition by secondary structural type alone is sufficient for answering important questions. As a result, analytical methods such as CD are more routinely utilized for protein secondary structure composition determination in biophysical chemistry.

In this paper we introduce a method to determine the secondary structure percentage composition of a protein using the information content contained in a single, one-dimensional ^{13}C NMR spectrum. This method, called *secondary-structure distribution by NMR* (SSD-NMR), bypasses the need for chemical shift assignments. Instead, SSD-NMR uses the general relationship between ^{13}C shifts and secondary structure to fit each residue as an ensemble of secondary structures. The resulting output gives an overall secondary structure composition. We show that method works on data acquired with either solution or solid-state NMR and can be used on proteins across their differing states; from soluble monomers, to protein-rich liquids, protein aggregates, and crystals. The method presented here differs from prior NMR approaches to evaluating secondary structure from NMR spectra by employing tools commonly employed in deep learning algorithms, along with validation by application to a large curated collection of NMR chemical shifts and structures. Using gradient descent optimization, the SSD-NMR algorithm reproduces with high fidelity the percentage secondary structure content of nearly 900 proteins from the protein data bank. These include soluble proteins that are accessible to optical-based methods as well as those in condensed phases. This algorithm is automated and simple to apply to proteins that are isotopically enriched in ^{13}C, including data acquired on low-field benchtop NMR spectrometers. We also demonstrate practical approaches to apply this method to proteins without isotopic enrichment on instrumentation that is available in most shared NMR resource centers with standard experiments.

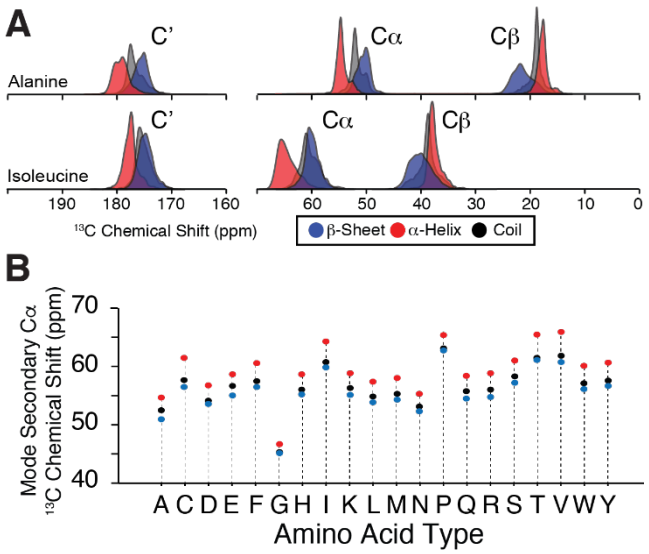

Figure 1: Protein backbone ^{13}C chemical shifts depend strongly on secondary structure. (A) Distributions of database assignments of the $^{13}C\alpha$,^{13}CO, and $^{13}C\beta$ chemical shifts for alanine as a function of secondary structure (top) which have excellent dispersion in the secondary chemical shift for each carbon. (bottom) The same distributions for isoleucine show that it has much less dispersion in the secondary shifts, most notably between β-sheet and random coil. (B) The mode $^{13}C\alpha$ chemical shift for each amino acid type colored by secondary structure: Red = α-helix, Grey = Coil, Blue = β-Sheet.

2. Methods

2.1 Generating Simulated ^{13}C NMR Spectra from Protein Secondary-Structure Content

We simulate a Secondary-Structure Content Simulated Spectra (SCSS) ^{13}C NMR spectrum of a protein by initializing peaks for each atom in a protein as an ensemble of three secondary structures: α-helix, β-sheet, and random coil. In this work we have chosen to use the STRIDE(15) algorithm to derive residue specific secondary structure classifications from structural data. For simplicity, we reduced the secondary structures defined by STRIDE to three types by combining: (1) 3-10 Helix, Pi-Helix, and α-Helix into α-helix; (2) Extended Conformation, Isolated Bridge, and β-Sheet into β-sheet; (3) Turn, Coil and other unidentified secondary structures into random-coil. We assign the mode backbone chemical shift for each atom given the residue type and secondary structure using the values reported by Fritzsching *et al. (12)* (Fig. 1B, Fig S1). The chemical shift of sidechain atoms beyond Cβ do not depend strongly on backbone dihedral angles and are thus initialized as a single peak at their respective mode chemical shifts irrespective of secondary structure. These peaks are initialized using the Lorentzian lineshape profile:

$$G_{Lorentzian}(x, \mu, \sigma) = \frac{\sigma}{2\pi\left((x-\mu)^2+\left(\frac{\sigma}{2}\right)^2\right)}, [1]$$

where μ is the center of the line shape and σ is the width, evaluated at point x.

To generate the SCSS of a protein, we account for the contribution from the mode chemical shift of each atom from each residue and each secondary structure. In this way, a single backbone atom is initialized as a linear combination of three peaks, where each peak is centered at its respective mode chemical shift according to each secondary structure. The intensity from the backbone atoms $I_{backbone}$, at any ppm value x, contributed by the backbone atoms in the SCSS is given by the equation:

$$\tilde{I}_{backbone}(x) = \sum_{i,k,ss} G(x, \delta_{i,k}^{ss} + \Delta\delta_{i,k}^{ss} - \delta_r, \sigma) a_i^{ss}, [2]$$

where $G(x, \mu, \sigma)$ is the Lorentzian lineshape profile defined in equation 1. $\delta_{i,k}^{ss}$ is the mode chemical shift for atom k from amino acid i for each secondary structure (ss, from α-helix, β-sheet, and coil) and σ is a global linewidth parameter. Since gradient descent requires continuous derivatives, for expediency our algorithm fits a parameter b where $b^2 = \sigma$, which ensures that the linewidth is positive. $\Delta\delta_{i,k}^{ss}$ is a perturbation term for each initialized peak and are all initialized as 0. δ_r is a global self-referencing parameter which allows the SCSS to be uniformly shifted. This parameter can be initialized by the user if a referencing offset is known, fitted by our algorithm, or both. The a_i^{ss} values are the contribution from each secondary structure where

$$\sum_{ss} a_i^{ss} = a_i^{\alpha helix} + a_i^{\beta sheet} + a_i^{coil} = 1. \text{ [3]}$$

The intensity from the sidechain atoms $I_{sidechain}$, at any chemical shift x, in the SCSS for atom k from amino acid i is instead:

$$\tilde{I}_{sidechain}(x) = \sum_{i,k} G(x, v_{i,k} + \Delta v_{i,k} - \delta_r, \sigma), \text{ [4]}$$

where $v_{i,k}$ is the secondary structure-independent mode chemical shift for atom k from amino acid i. Here, σ is the same global linewidth parameter from Eq. 2. $\Delta v_{i,k}$ is the perturbation term for each peak, analogous to $\Delta\delta_{i,k}^{ss}$ from Eq. 2 and are initialized as 0. δ_r is the global re-referencing parameter from Eq. 2.

The SCSS of a protein, given its primary sequence, is then generated by summing over each atom (k) from each amino acid (i) in the protein. Each atom contributes the same total normalized intensity and has the same linewidth (σ). In this way, the spectrum depends solely on the secondary structure content of each residue (a_i^{ss}), the individual perturbations around the mode shift ($\Delta\delta_{i,k}^{ss}, \Delta v_{i,k}$), a global linewidth parameter (σ), and the optional global self-referencing parameter (δ_r). The final SCSS is given by summing the intensity from every backbone and sidechain atom in the protein. At any chemical shift x the intensity is given by:

$$\tilde{I}(x) = \tilde{I}_{backbone}(x) + \tilde{I}_{sidechain}(x) \text{ [5]}$$

2.2 Gradient Descent Loss Function

We optimize the values of the percentage secondary structure elements (a_i^{ss}) assigned to each residue by minimizing the following loss function which compares the jth point in the SCSS $\tilde{I}_j$ to the corresponding point in the experimental spectrum I_j:

$$L = \ln\left(\frac{1}{N}\sum_j^N\left(\tilde{I}_j - I_j\right)^2\right) + \sum_{i,k,ss}\left(\Delta\delta_{i,k}^{ss}\right)^2 + \sum_{i,k}\left(\Delta v_{i,k}\right)^2. \text{ [6]}$$

I_j is normalized such that $\sum_j I_j = 1$. The loss function introduced by Eq. (6) is composed of two main terms. The first is the penalty for the difference between the SCSS and the experimental spectrum. For this term, we chose to use the log of the squared difference because it will provide faster convergence speed when the sum of residuals between the two spectra are small. This can be shown by inspecting the derivative of the first term of the loss with respect to a_i^{ss}. The difference term in the denominator makes the gradient with respect to a_i^{ss} larger when the difference is small:

$$\frac{\partial L}{\partial a_i^{ss}} = \frac{\sum_{x,k} 2(\tilde{I}_x - I_x)G(x, \delta_{i,k}^{ss} + \Delta\delta_{i,k}^{ss} - \delta_r, \Omega)}{\sum_{j\prime}(\tilde{I}_{x\prime} - I_{x\prime})^2}, \text{ [7]}$$

The second term in the loss enforces a harmonic well on the offsets $\Delta\delta_{i,k}^{ss}$ and $\Delta v_{i,k}$ (Eqns 2,4). To minimize the loss, gradient descent optimization (GD) is used to adjust these offsets for each atom, the secondary structure contribution (a_i^{ss}) from each residue, the global linewidth (σ), and the optional global self-referencing parameter (δ_r) such that the next step is along the steepest descent along the loss function landscape. For each step in gradient descent, the parameters a_i^{ss} are updated such that the difference between the current step and the previous step, a_i^{ss}follows:

$$\Delta a_i^{ss} = -r\frac{\partial L}{\partial a_i^{ss}} \text{ [8]}$$

where r is the learning rate, which we set to 0.1. This process is repeated for a number of pre-determined steps (discussed below) using the ADAM optimizer(16) with default parameters β1 of 0.9, β2 of 0.99, ϵ of 10^{-8}, and a weight decay of 0, using PyTorch(17). The output provides the optimized parameters.

For each residue, the values of $a_i^{\alpha helix}$, $a_i^{\beta sheet}$ and a_i^{coil} are extracted. As we will explain in the discussion, many of the amino acids are near-indistinguishable during the fitting process based on their mode Cα, Cβ, and CO shifts (Fig S1). Moreover, the a_i^{ss} output cannot distinguish individual amino acids within the primary sequence. As such, any residue X in the sequence can be interchanged with residue $X + n$ in the sequence so long as it is the same amino acid type. Thus, instead of reporting the secondary structure of each residue, we report the overall secondary structure content by summing over all of the a_i^{ss}, yielding the proportion of each secondary structure, $P(ss)$ where ss is one of α-helix, β-sheet, or coil:

$$P(ss) = \frac{\sum_i a_i^{ss}}{\sum_{i'} \sum_{ss'} a_{i'}^{ss'}} [9]$$

2.3 Optimization of Initialization and Hyperparameters for Gradient Descent

An important consideration for our GD based approach is the selection of hyperparameters for GD and the initial parameters for the SCSS discussed above (Individual: $\Delta\delta^{k,ss}$, Δv^k , and a_i^{ss}. Global: σ and δ_r). To obtain a set of hyperparameters that are generally applicable for the majority of protein NMR ^{13}C spectra, we studied the effect of these parameters on convergence criteria. We set initial values of 0 for all $\Delta\delta^{k,ss}$ and Δv^k , and 0.04 ppm for the global linewidth σ. We assume equal probabilities for each secondary structure by setting a_i^{ss} of 0.33 for α-helix, β-sheet, and coil for all residues. We ran each minimization for 100 steps. The self-referencing parameter δ_r is initialized as 0 unless there is an appropriate guess for the referencing offset. For instance, the ^{13}C referencing offset between the biological referencing standard DSS (assumed here) and TMS referencing standards is approximately -2.4 ppm(18), and one could set this for δ_r if experimental data was instead referenced to TMS(12, 18, 19).

In order to evaluate the efficacy of this set of hyperparameters for GD to converge across the range of values expected in experimental data, we extracted the value of the loss

function at each step of the GD for simulated experimental spectra of microcrystalline ubiquitin (BMRB: 25123)(20) which were generated with varying underlying linewidths ranging from 0.2 to 20 (Fig. S2). The value of the loss function converged for all linewidths less than 20 ppm within 100 steps of GD as demonstrated by the stability in the value of the loss function. However, convergence of GD is not sufficient to show that the fit parameters have converged on a set of values that are consistent with the underlying data.

To determine if GD with our loss function converges in this way, we compare the global fitted linewidth parameter σ_{fit} to the simulated σ_{sim}. This shows whether convergence of the loss function results in a linewidth that matches the value we generated the spectra with. To do this, we extracted σ_{fit} at each step of the GD and observed when oscillations in the fitted value ended and GD reached a stable value. Figure 2 shows σ_{fit} as a function of the number of training steps for each generated spectrum with fixed linewidth σ_{sim}. These results indicate that the model converges to a stable σ_{fit} value very close to the true value σ_{sim} within 100 steps when the underlying linewidths are below 20 ppm, which far exceeds the linewidths expected in NMR spectra of proteins. When the linewidth is greater than 20 ppm, a choice of a larger starting width parameter close to the linewidth leads to faster convergence (Fig S3). Additionally, when using the self-referencing parameter, we also observed that additional steps are required for convergence, and we recommend a minimum of 500 steps. In the cases discussed below, we ran GD for 100 steps as the experimental or simulated linewidths were ≤ 10 ppm and were not self re-referenced unless specified.

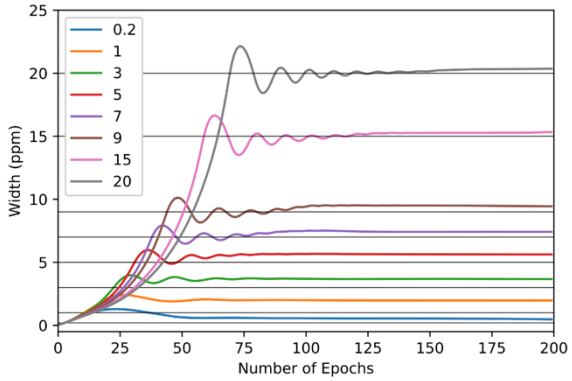

Figure 2: Convergence of Gradient Descent with Chosen Hyperparameters. Simulated experimental spectra were generated from ubiquitin with varying Lorentzian width ranging from 0.2 to 20 ppm. The Ω parameter was extracted at each step for each linewidth with different colored lines representing the SSD-NMR fitted results at each step. The black solid lines indicate the different underlying width parameters used. The set of hyperparameters (initial Ω=0.04, 100 steps, 0.1 learning rate) is sufficient for spectra with an underlying linewidth of ≤ 10 ppm.

2.4 Model Testing on Simulated Experimental Spectra from Databases

To validate our algorithm, we generated simulated experimental 1D ^{13}C spectra (SES) using the assigned chemical shifts from proteins found within the PACSY database(21) as described in the Experimental Section. PACSY is a SQL database connecting the assigned NMR chemical shifts in the Biological Magnetic Resonance Data Bank (BMRB)(22) to their applicable structures deposited in the Protein Databank (PDB)(23). We then curated the PACSY database using the following criteria: (1) the "Metal_ion" and "Paramagnetic" fields within the assembly heading must be filled in and must be either "." or "no" to exclude proteins with metal centers; (2) a protein must have ≥90% completeness for the assignment of backbone ^{13}C atoms; and (3) the protein sequence is at least 20 residues long. A total of 1839 proteins remained in the curated database following this criterion (BMRB and PDB accession codes provided in the github repository) This filter eliminated some outliers, which are known to exist in the BMRB, as well as proteins with few chemical shift assignments. However, we acknowledge that this filter does not exclude all paramagnetic/ion-containing data due to mislabeling in the BMRB. Additionally, some entries deposited in the BMRB are known to have mis-referenced chemical shifts relative to the DSS standard(12, 24). As a result, we were also able to assess the performance of our self re-referencing parameter using these SES.

For each protein, we generated a SES with signal-to-noise ratio (SNR) of 50 using the method described in the Experimental Section, using a Lorentzian line shape with a full-width-half-maximum (FWHM) of 1.0 ppm. These simulations used only the Cα, Cβ, and CO carbon shifts since many entries lack the additional side-chain chemical shift assignments. The results from this study showed a remarkable correlation between the predicted and the STRIDE extracted secondary structure (Fig 3A, Fig S4). We obtained a correlation of 0.97 for α-helix, 0.91 for β-sheet, and 0.74 for random-coil for well-referenced proteins (<0.2 ppm mode predicted referencing offset, 891 of the 1839 proteins, list in github repository). The correlation of α-helical content is the highest, and rivals the correlation observed for CD measurements (25). Our approach is less reliable at distinguishing between β-sheet and random

coil secondary structure elements. This is also observed with CD (25). Expectedly, the correlation decreases with increasing chemical shift referencing errors (Fig S6). An additional test using the re-referencing parameter δ_r (run for 500 steps for convergence) results in a better RMSE for proteins with large reference offsets (>2.7 ppm) but does not outperform the results for well-referenced data (<0.2 ppm) without inclusion of the self-referencing parameter δ_r (Fig S7).

For comparison, we simulated the spectra for the set of 891 well-referenced proteins from above with a range of simulated linewidths to test the robustness of our algorithm with respect to linewidth (Fig. 3B, 3C). The observed trend is that with narrow (<0.4 ppm) or increasing linewidths (>2 ppm) (Fig 3B), the algorithm tends to have a larger RMSE, most notably for random coil. However, the predicted secondary structure percentage composition remains relatively stable for α-helix and β-sheet over a range of simulated linewidths from 0.4 ppm to 9 ppm, with added RMSEs of only 0.01 for α-helix and 0.05 for β-sheet. Similarly, using a fixed linewidth of 1 ppm, we tested the effect of signal-to-noise by varying the SNR from 1 to 50 (Fig 3C). The prediction accuracy of α-helix and β-sheet content stabilizes at the same value for SNR as low as 2. We also repeated this simulation using an alternative SNR based upon the most intense point present in the spectrum (described in the methods). This reports on the SNR of the envelope of the spectrum, instead of the SNR of an individual peak. These results (Figure S5) show that SSD-NMR stabilizes even when the SNR of the envelope is as low as 5 ppm.

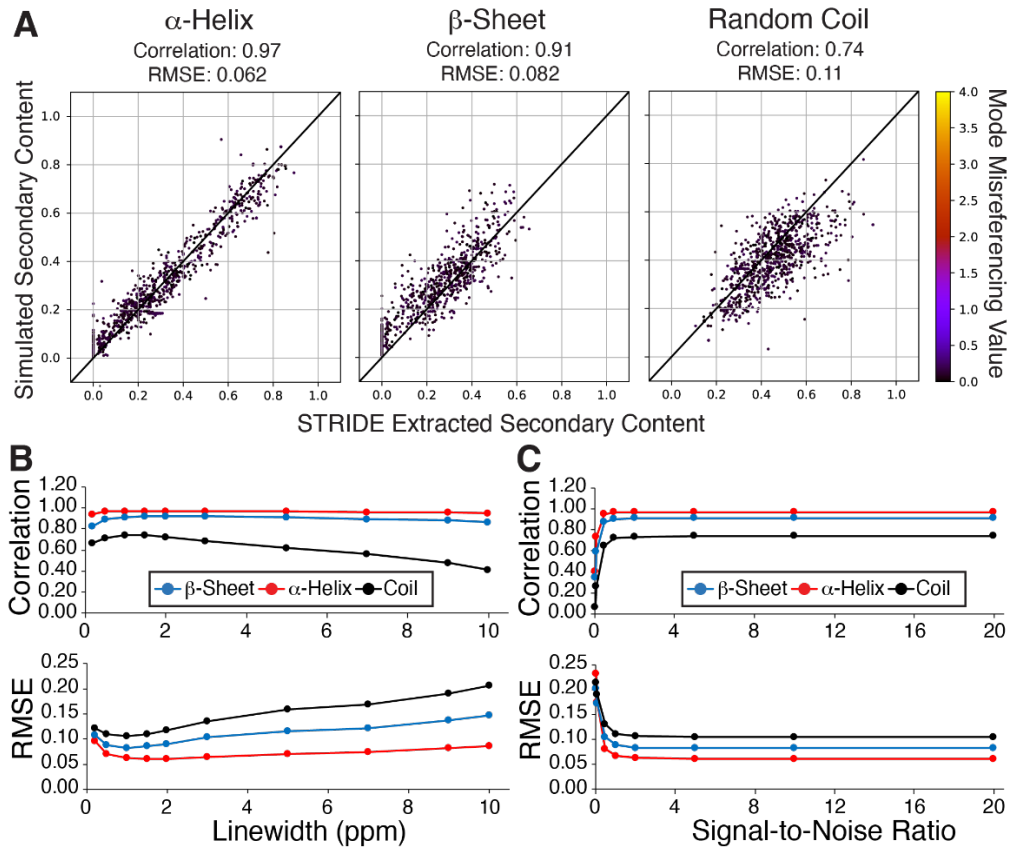

Figure 3: Secondary Structure Distribution Prediction from Simulated Spectra from Database Chemical Shifts. (A) Correlation between SSD-NMR secondary structure (vertical axis) and the secondary structure reported by STRIDE (horizontal axis) for each secondary structure type from simulated experimental spectra of 891 proteins. Correlation coefficients of 0.74 for random coil, 0.91 for β-sheet and 0.97 for α-helix show the accuracy of SSD-NMR. (B) The correlation and RMSE obtained by varying linewidth parameters in the simulated spectrum. While the performance decreases with increasing linewidth, the algorithm remains relatively constant for α-helix and beta-sheet predictions for linewidths that are less than 9 ppm. (C) The correlation and RMSE resulted by using different SNR_{sim} ratios demonstrate the reliability of this approach for signal to noise ratios as low as 2.

2.5 Comparison of SSD-NMR to TALOS-N

We further evaluated the performance of our model by comparing its results to those obtained when site specific chemical shift assignments are available. To do this, we ran TALOS-N(14), a widely used program for predicting secondary structure from assigned chemical shifts, on a subset of the 891 well-referenced proteins used above. From the BMRB website, 332 NMR-STAR files with complete backbone chemical shift assignments were downloaded. Figure 4 demonstrates the correlation between the resulting subset of 332 well-referenced proteins and the corresponding STRIDE extracted secondary content. TALOS-N predicts the probability (Q_L = coil, Q_H = α-helix, and Q_E = β-sheet) of each residue being in each of the three secondary structures. Since Q_L, Q_H, and Q_E represent a distribution for each residue, we report the extracted secondary structure content from TALOS-N by taking the sum over the continuous probabilities. TALOS-N shows excellent agreement to STRIDE, with correlations of 0.99, 0.97, and 0.92 for α-helix, β-sheet, and random coil, respectively. We compared the performance of SSD-NMR to the 332 proteins (Figure S8) and then compared them to the TALOS-N output, with TALOS-N RMSEs being only 2.2%, 2.9%, and 4.6% better for α-helix, β-sheet, and random coil, respectively (Fig. 3). Furthermore, Figure 5 shows the correlation of SSD-NMR to the TALOS-N results from Figure 4. The correlation between SSD-NMR and TALOS-N is higher for all secondary structure types than the correlation between SSD-NMR and STRIDE, as shown in Figure 3.

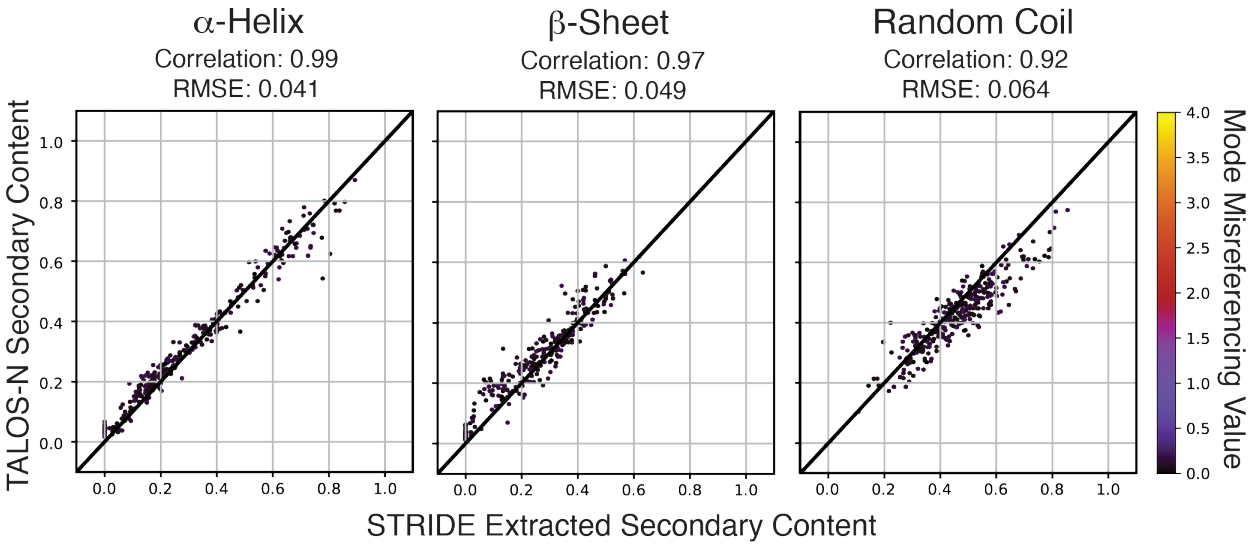

Figure 4: Secondary Structure Distribution Prediction from TALOS-N compared to STRIDE Extracted Secondary Structure. Results for 332 well-referenced proteins extracted from TALOS-N. The agreement between TALOS-N and STRIDE demonstrates the strong relationship between NMR chemical shifts and protein secondary structure.

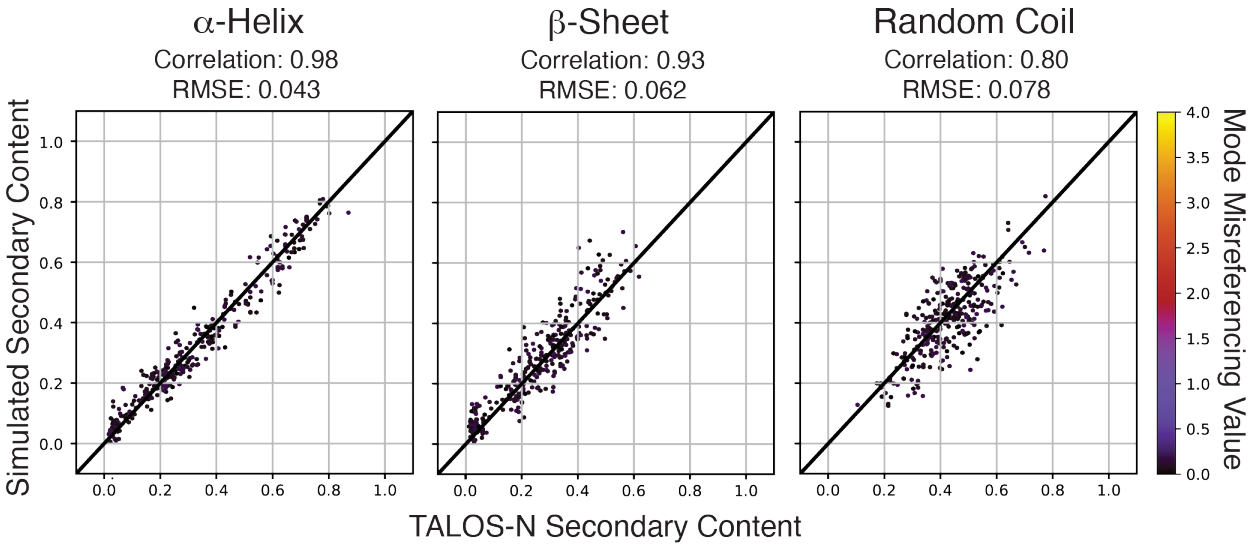

Figure 5: Secondary Structure Distribution Prediction from SSD NMR compared to TALOS-N Secondary Structure Distribution. Benchmarking SSD-NMR to TALOS-N on the

same subset of proteins as shown in Figure 4. The SSD-NMR shows better agreement with TALOS-N than with STRIDE (Figure 3).

2.6 Testing the Model's Robustness to Experimental Baseline Offsets

We generated an SES with a linewidth of 1 ppm and SNR of 50 and then applied a series of constant baseline offsets relative to the largest intensity in the spectrum (Figure S9). This was done to approximate a constant baseline offset in an experimental dataset. To do this, we first determined the maximum intensity, $Imax$, defined as $Imax=(I1,I2\ldots IN)$. We then shift the entire spectrum by the percent deviation, D, such that each intensity was adjusted by $D^{*}Imax$. For instance, If D is equal to 0.1, the spectrum of interest will be raised by 10% of its maximum intensity. By analyzing the result of the 891 well-referenced proteins, we determined that the algorithm produces optimum results when the baseline is flat with no deviation, (Figure S9). When $|D| < 0.05$, SSD-NMR produces excellent results with errors that are comparable to a case when there is no baseline offset. However, if D is around -0.07 or less, the correlations become negative or unstable. The loss of correlation is abrupt if D continues to take values less than -0.07. This is not observed when D takes values higher than 0.07. This is due to the fact that negative intensities result in ill-defined normalization. To ameliorate this instability from negative intensities that may occur when experimental data is to be fit, we calculate the mean negative intensity of the spectrum and add that value to each datapoint prior to GD minimization as further described in the Experimental Methods section.

3. Results

3.1 SSD-NMR Testing using Solution NMR of Ubiquitin

Figure 6A shows a 1D ^{13}C NMR directly polarized spectrum acquired on uniformly ^{13}C, ^{15}N isotopically labeled Ubiquitin at 500 MHz ^{1}H frequency. The pulse sequence used to acquire this data was a single pulse excitation spectrum with a 5 second interscan delay, and is

a default ^{13}C detected experiment included on most spectrometers. We then input this spectrum into SSD-NMR along with the amino acid protein sequence for ubiquitin. We initialized the algorithm with a window ranging from 10 to 195 ppm, and an initial σ of 0.04 ppm. We then ran SSD-NMR for 100 steps with a learning rate of 0.1, as discussed above. The SSD-NMR algorithm best-fit SCSS predicted 39.1% random coil, 25.8% α-helix, and 35.1% β-sheet content, which was nearly identical to the secondary structure extracted by STRIDE from the solution NMR PDB file 1D3Z (39.5% coil, 25.0% helix, and 35.5% sheet, Figure 6C,6D). We also processed the Ubiquitin spectrum with different apodization to produce different apparent linewidths in the processed data. Fig. 6B compares no apodization (top) and with a 10 ppm Gaussian filter (bottom). In both cases, the observed convergence is consistent with the estimate of the set of hyperparameters chosen for gradient descent and the estimates for the secondary structure content lie within the RMSEs predicted for each linewidth from our testing.

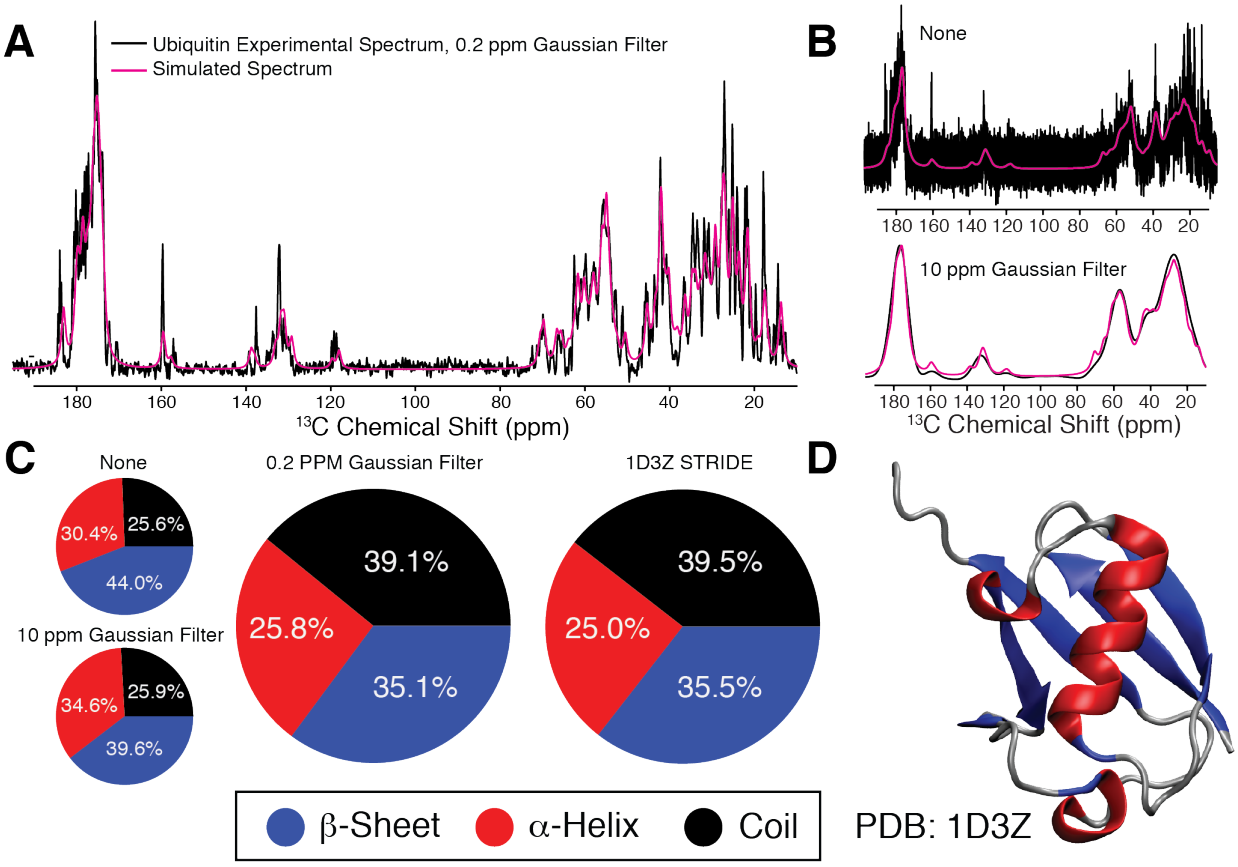

Figure 6 SSD NMR of Ubiquitin. (A) Experimental ^{13}C single pulse excitation spectrum (black) of uniformly ^{13}C, ^{15}N labeled Ubiquitin at 500 MHz ^{1}H Frequency, 128 scans with 0.2 ppm Gaussian line-broadening overlaid with SSD-NMR simulated spectrum (magenta). (B) Examples of the spectrum shown in from Fig.6A with no line broadening (top) and 10 ppm Gaussian line broadening (bottom) overlaid with the respective simulated spectrum. (C) Estimated secondary structure content of each spectrum compared to the secondary structure of Ubiquitin extracted via STRIDE from the solution NMR structure of Ubiquitin PDB: 1D3Z. (D) Structure of Ubiquitin (PDB: 1D3Z) colored via secondary structure.

3.2 SSD-NMR using Solid-State NMR Spectroscopy

We further tested our SSD-NMR algorithm using solid-state NMR spectroscopy of the microcrystalline protein GB1. A directly polarized ^{13}C spectrum was acquired on uniformly ^{13}C,^{15}N isotopically labelled GB1 (Fig. 7A) at 800 MHz ^{1}H frequency. In this case, the pulse sequence was a single pulse excitation with a 110 μs spin echo to remove background signals. The spectrum was processed with 0.4 ppm Gaussian line broadening apodization. We initialized SSD-NMR as was done for the ubiquitin example. The predicted SCSS (Fig. 7A) indicated 18.5% random coil, 52.5% β-sheet, and 29.1% for α-helix. These results follow the secondary structure reported by Franks *et al.(26)* from the solid-state NMR chemical shift assignments (Fig. 7C) of 19.6% random coil, 51.8% β-sheet and 28.6% α-helix. However, they vary from the secondary structure extracted from STRIDE from the resulting ssNMR structure, PDB: 2QMT (30.4% coil, 42.9% sheet, and 26.8% helix). Nevertheless, these results fall within the predicted RMSEs from the database testing. Notably, SSD-NMR does not fit the first-order spinning side sideband from magic-angle spinning (Fig. 7A). This shows that SSD-NMR is robust to artifacts in regions of the spectrum not expected to have signals from protein, and that the constraint on the $\Delta\delta^{k,ss}$ and Δv^{k} parameters prevents SSD-NMR from overfitting the experimental data.

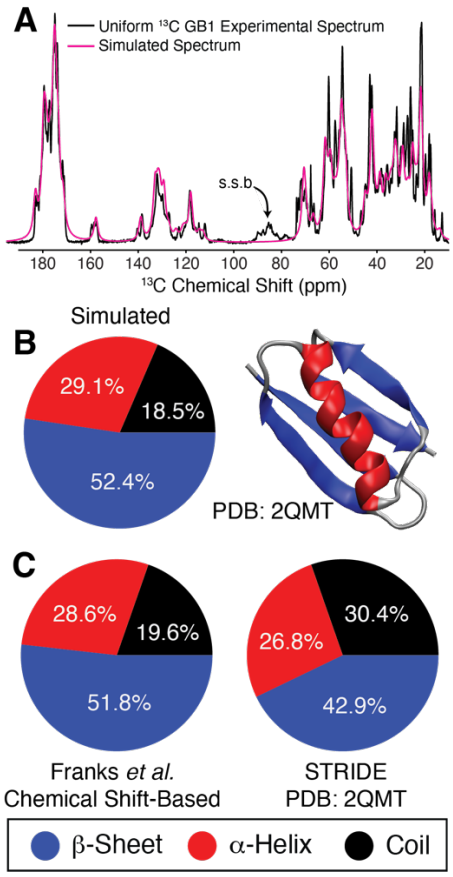

Figure 7: Secondary Structure Distribution of GB1 determined by Gradient Descent and Solid-State NMR. (A) Experimental Solid-State NMR spectrum (black) of uniformly ^{13}C, ^{15}N labeled GB1 ^{13}C acquired at 800 MHz ^{1}H frequency signal-averaged for 43 minutes overlaid with the simulated spectrum (magenta) derived from SSD-NMR. s.s.b. corresponds to the first order spinning side band from magic-angle spinning. (B) The secondary structure distribution found by this work derived from the simulated spectra in (A), compared to the 3D structure of GB1 from PDB: 2QMT colored by secondary structure: (blue) β-sheet (red) α-helix, (silver) coil. (C) Secondary structure distribution reported by Franks *et al.*(26) from chemical shift assignments (left) vs STRIDE classification extracted from PDB 2QMT (right).

3.3 Application of SSD-NMR to Natural Abundance Protein Samples

To assess the feasibility of using SSD-NMR on protein samples at natural abundance, we acquired a ^{13}C 1D NMR spectrum of 1 mM Hen Egg White Lysozyme at 600 MHz ^{1}H frequency doped with 20 mM CuEDTA and 2mM DSS. CuEDTA is a water-soluble paramagnetic doping agent commonly used in NMR spectroscopy to increase both the longitudinal and transverse relaxation rates of the samples (27). In practice, this permits dramatically shortened interscan delays which allows for rapid signal averaging at the cost of increased linewidths in the spectrum. We acquired 16,384 scans with an interscan delay of 1 s in 5 hours on a ^{1}H detect cryo-probe (Fig. 8A). The observed natural linewidths in the sample ranged from 20 to 40 Hz (0.1 to 0.2 ppm). The data was processed with 150 Hz of exponential apodization to bring the linewidths near the optimal width of 1.0 ppm found above. The SSD NMR results on this spectrum (Fig 6A) gave 28.4% coil, 26.4% β-sheet, and 45.2% α-helix (Fig. 8B). We then compared our result on lysozyme to the STRIDE extracted distribution from the X-Ray crystal structure (PDB:6LYZ) as well as those reported by literature from CD(28), Raman, and FT-IR (Fig. 8B). Results from SSD-NMR are consistent with the STRIDE extracted structure for α-helical content and follows the trend of increased β-sheet content predicted by other solution-based techniques.

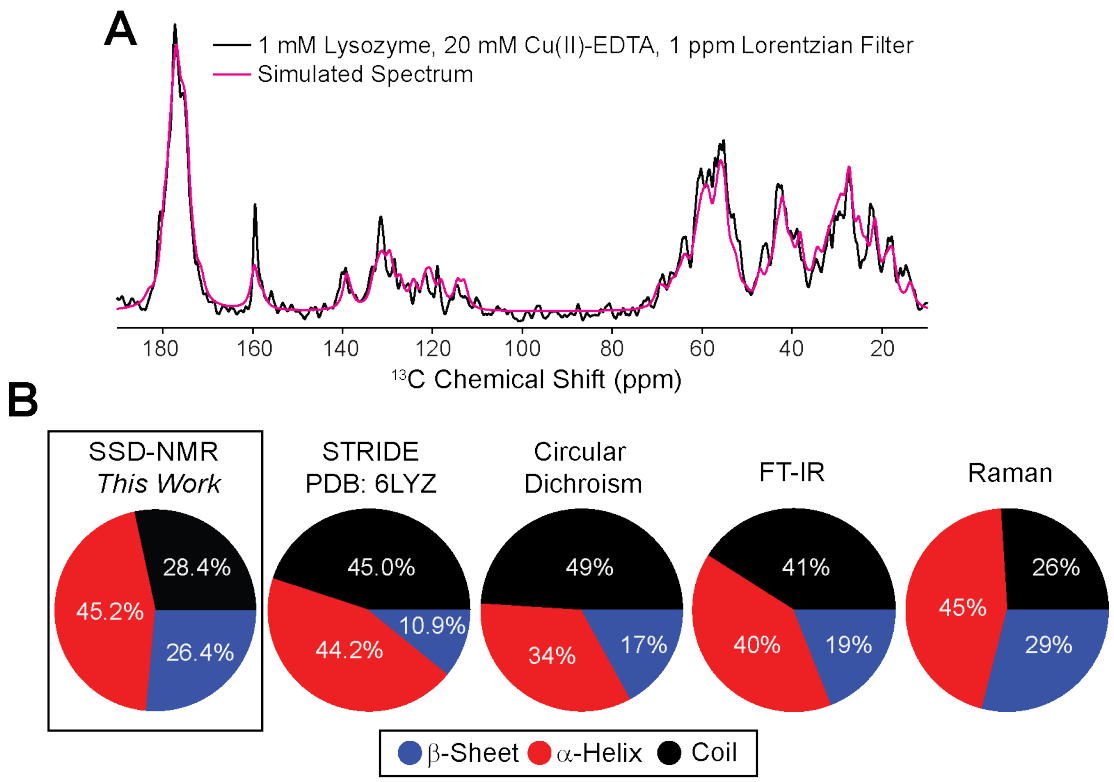

Figure 8: Secondary Structure Distribution of Hen Egg-White Lysozyme. (A) Experimental ^{13}C NMR Spectrum of natural abundance 1 mM Hen Egg White Lysozyme (black) doped with 20 mM Cu(II)-EDTA (5 hours acquisition) overlaid with the simulated spectrum (magenta). (B) Predicted secondary structure distributions for Lysozyme by different experimental methods. The SSD-NMR distribution predicted from (A) (left). The STRIDE reported secondary structure from the X-Ray structure 6LYZ (middle-left). Averaged secondary structure of Lysozyme by Circular Dichroism as reported by Greenfield(28) (middle). FT-IR distribution as reported by Dong *et al.* (29) (middle-right). Raman-determined distribution as reported by Di Foggia *et al.* (30) (right).

3.4 SSD-NMR at Low Magnetic Field

We also acquired a spectrum with the isotopically labeled ubiquitin sample studied above on a benchtop 60 MHz ^{1}H Frequency (1.4 T) NMR Spectrometer (Magritek Spinsolve 60 MHz). We acquired the default 1D CARBON+ WALTZ with 2048 scans, without NOE enhancement with an interscan delay of 3 s for a run time of 1.7 hours. This spectrum, shown in Figure 9, was then input into our SSD-NMR algorithm with the same initialization discussed above. The secondary structure content of 25.6% α-helix, 33.9% β-sheet, and 40.6% coil is very close to the 25.0%, 35.5%, and 39.5% percentages expected from the STRIDE classification, and is remarkably similar to the results shown in Figure 6, determined by SSD-NMR at 500 MHz ^{1}H field.

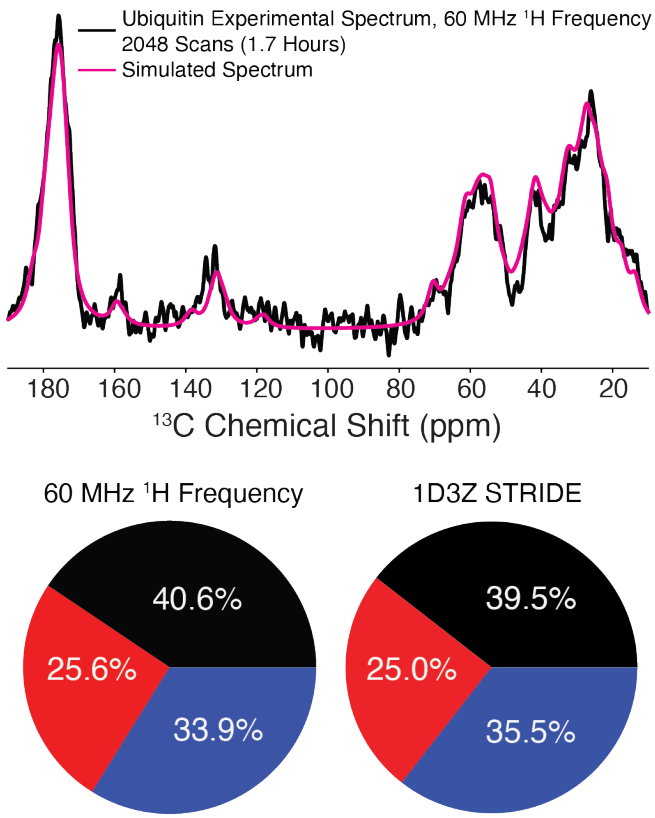

Figure 9: Secondary Structure Distribution of Ubiquitin at 60 MHz ^{1}H Frequency (1.4 T). Experimental ^{13}C single pulse excitation spectrum of uniformly ^{13}C, ^{15}N labeled Ubiquitin at 60 MHz ^{1}H Frequency, 2048 scans processed with 0.67 ppm Exponential line-broadening (black) overlaid with simulated spectrum from SSD-NMR (magenta). The SSD-NMR estimated secondary structure content (left) compared to the secondary structure of ubiquitin extracted via STRIDE from the solution NMR structure of Ubiquitin PDB: 1D3Z (right).

4. Experimental Methods

4.1 Simulated Experimental Spectrum

Simulated experimental spectra were generated from database chemical shift assignments of proteins with known structure. To approximate experimental data, the SES are composed of peaks with a single global linewidth which are centered at each assigned chemical shift from a protein of interest. The intensity of any point, $I(x)$, in an SES is defined as:

$$I(x) = \sum_i G(x, \delta_i, \sigma) \text{ [10]}$$

Where $G(x, \mu, \sigma)$ is the Lorentzian from Eq. 1, evaluated at point x (i.e. chemical shift). δ_i corresponds to the i^{th} assigned chemical shift from the database from the protein of interest. Therefore $I(x)$ is the sum of the contribution from every assigned chemical shift in a protein of interest using the lineshape profile $G(x, \mu, \sigma)$ evaluated at point x. Unless specified differently, the tests above were performed with a σ of 1.0 ppm.

4.2 Simulated Gaussian Noise

To add simulated noise to a simulated experimental spectrum as would be present in an experimental spectrum, we insert random noise with a defined signal-to-noise ratio SNR_{sim} based on the intensity of an individual peak in the spectrum. This random noise is added to the original intensities, $I(x)$ obtained from Eq.10. This gives a new intensity, $I_{SNR}(x)$ at the jth point in the spectrum, with a user specified SNR_{sim}, defined by:

$$I_{SNR}(x_j) = I(x) + I_0 \left(\frac{\in_j}{SNR_{sim}}\right) \text{ [11]}$$

The noise amplitude $\in_j$ is randomly drawn from a Gaussian distribution with a mean of 0 and width 1. I_0 is the peak amplitude of a single line contributing to the SES envelope from Eq. 10. Unless specified otherwise, a SNR_{sim} of 50 was used.

For comparison to experimental data, it is more convenient to consider the signal to noise of the spectral envelope, especially if an individual peak for a single resonance is not

readily discernible. To this end we also alternatively added noise to an SES based on the maximum amplitude of the envelope of the entire spectrum (I_{MAX}). The intensity of the at the jth point of the spectrum $I_{SNRe}(x_j)$ is defined as:

$$I_{SNRe}(x_j) = I(x) + I_{MAX}\left(\frac{\epsilon_j}{SNR_{sim}}\right) [12]$$

The noise amplitude ϵ_j is randomly drawn from a Gaussian distribution with a mean of 0 and width 1.

4.3 Sample Preparation

Lyophilized powder of egg white lysozyme was purchased from Millipore-Sigma and dissolved into 99% D_2O /1% H_2O purchased from Cambridge Isotopes Laboratory with 2 mM sodium trimethylsilylpropanesulfonate (DSS) as a referencing standard. For samples with a relaxation dopant, CuEDTA (Sigma) was added to the solution to bring the concentration to 20 mM. A uniformly ^{13}C,^{15}N labelled ubiquitin standard sample from Cambridge Isotopes Laboratory was used for solution NMR experiments. Uniformly ^{13}C,^{15}N GB1 was prepared as previously reported(26) and packed into a 2.5 mm zirconia solid-state NMR rotor.

4.4 NMR Spectroscopy

NMR spectra were acquired at 1.4 T (Magritek Spinsolve 60 MHz Carbon Spectrometer), 11.7 T, (Agilent DD2 500 MHz NMR Spectrometer, OneNMR probe with auto-tune-and-match), 14.1 T (Agilent DD2 600 MHz NMR Spectrometer, ^{13}C (H) cold probe with auto-tune-and-match), and 18.8 T (Agilent 800 MHz VNMRS NMR Spectrometer, custom built triple resonance ^{1}H,^{13}C,^{15}N ssNMR probe(31)). When magic angle spinning was used the spin rate was monitored and set by an Agilent MAS controller to 17,777 Hz with 2.5 mm rotor outer diameter zirconia rotors and custom Kel-f spacers for optimal RF homogeneity(31). Pulse widths for ssNMR experiments were 2.5 us for ^{1}H and 3.5 us for ^{13}C, with 100 kHz small phase incremental alternation (SPINAL-64) decoupling(32) applied during evolution and acquisition periods on ^{1}H. Solid-state experiments were run at 5° C which was controlled with a Ranque-

Hilsch vortex tube cooler (33). Solution NMR Experiments were performed at 25° C set by a Varian/Agilent temperature controller. Chemical shift referencing in the solution state for natural abundance sampled was referenced with the inclusion of 2 mM DSS and setting the DSS peak to 0 ppm. Isotopically enriched samples were referenced to the DSS scale using the D_2O lock frequency. Adamantane was used as an external chemical shift reference for ssNMR experiments with the downfield peak set to 40.48 ppm(18).

4.5 Experimental Spectrum Processing

1D spectra were converted, processed, and analyzed in either MestreNOVA or VNMRJ 3.2. The data was apodized with Gaussian or exponential line broadening and zero filled to double the number of points prior to Fourier transformation and phasing. Normal baseline adjustments used only a direct current offset adjustment, but when significant baseline roll was present, a 3rd order polynomial baseline correction was instead applied. The data was then exported in a x,y format with the chemical shift axis in referenced ppm and a unitless intensity axis. The spectrum is then indexed for a user definable ppm window. Unless otherwise specified, the window was set to cover 5 to 195 ppm for the results presented above. As discussed previously, excessive negative signals can lead to broadening in the simulated spectrum. To ameliorate this problem, we measure the number of negative points, N_{mn}, and raise the intensity of all points in the spectrum by the mean negative intensity, I_{mn}, defined as,

$$I_{mn} = \frac{1}{N_{mn}} \sum_{i, I_i \in I<0} I_i \text{ [12]}$$

Where the corrected intensity $I_{corr}(x)$ at any chemical shift value x is given by:

$$I_{corr}(x) = I(x) - I_{mn} \text{ [13]}$$

For experimental data shown above, the spectrum has been processed in this manner, with the index removed from $I_{corr}(x)$, giving $I(x)$ to represent the corrected intensity at chemical shift x.

Discussion

In this work we introduce SSD-NMR, a method which can extract the overall secondary structure content of a protein using a single experimental 1D ^{13}C NMR spectrum. To do this, we utilize the relationship between secondary structure and $^{13}C\alpha$,^{13}CO, and $^{13}C\beta$ chemical shifts to simulate an NMR spectrum of a protein based on its secondary structure content. We then then minimize the difference between this simulated spectrum and an experimental spectrum using gradient descent guided by a loss function we developed. The results from SSD-NMR are surprisingly robust and have a high correlation with proteins in the PACSY database, which covers a wide range of the 3D structures available in the PDB.

Given the high fidelity of the results obtained by SSD-NMR, we investigated how well the individual a_i^{ss} parameters for each residue type matched those of the data being fitted. These residue-specific distributions were notably less accurate than the overall distribution (Figures 6,7,8 and S10,S11,S12). It was also observed that while the overall SSD is stably reproduced for a range of SNR conditions, the SSD for individual amino acid types could vary quite markedly. While initially concerning, reaxaminaiton of the mode ^{13}C chemical shifts. identifies the source of this apparent discrepancy. As can be seen in in Figure S1, there are several classes of amino acids that have very similar sets of backbone secondary-structure dependent ^{13}C shifts. Their contributions to the spectrum are not entirely linearly independent, so it is not expected they will be readily distinguished from one another during the fitting procedure. Based on this observation, we examined whether residues that have similar shift sets could be put into subgroups that better match the secondary structures extracted by STRIDE. To do this, we formed 5 subgroups of amino acids: TS, WHQREMK, LDN, FY, and IVP. We kept C, A, and G separate as they have distinct patterns of backbone shifts and can be fit more independently. For Ubiquitin (Fig. S10), GB1 (Fig. S11), and Lysozyme (Fig. S12) the secondary structure distributions of these subgroups better match with those obtained from STRIDE. The GD procedure is then expected to interchange

secondary structure content between residues (via their a_i^{ss} parameters) with similar backbone chemical shifts, yet still arrive at a reliable overall SSD.

In addition to the similarities in sets of chemical shifts between different amino acid types, we also note that every residue of the same amino-acid type is initialized identically in our fitting procedure; they have the same mode shifts, the same a_i^{ss}, same $\Delta\delta^{k,ss}$, and the same linewidth σ. Because of this, the gradient at each step of the fitting for the same amino acid type will be identical, resulting in the same optimized parameters. This means that each residue of the same amino acid type is not fit independently. In the Secondary-Structure Content Simulated Spectrum (SCSS) then there are effectively 3 peaks for each backbone atom and 1 for each sidechain atom for each amino acid type. Because of this SSD-NMR works best on protein spectra where it only need fit the envelope and not each individual peak.

This helps to explain why SSD-NMR has surprising resilience to a wide range of experimental linewidths and why the optimal linewidth is near 1 ppm (Figure 3C). Our initial instinct was that higher-resolution data would yield the best results. However, because SSD-NMR is fitting the envelope of the spectrum, the best results are observed when the linewidth is narrow enough to resolve the differences in the mode shifts between secondary structure types (which are separated by ~2-5 ppm between β-sheet and α-helix, Figure S1), but wide enough to broaden out narrow resonances that are offset from the mode shifts. The SSD-NMR algorithm then focusses more on adjustment of the a_i^{ss} terms while keeping $\Delta\delta^{k,ss}$ and Δv^k small. We observed this in our testing – the distribution of extracted $\Delta\delta^{k,ss}$ values (Fig. S13) at 1 ppm linewidth resulted in $\Delta\delta^{k,ss}$ terms that were almost always less than 0.01 ppm; Under these conditions, SSD-NMR does not require $\Delta\delta^{k,ss}$ and Δv^k terms, but we have elected to keep them in the model to stabilize the algorithm should there be large intensity offsets from the mode shifts.

The combination of these observations also resolves why SSD-NMR is able to provide meaningful results when the peak envelope SNR is as low as 5. We were initially suspicious of this result – a SNR of 5 for the envelope is below what would be required for typical NMR

analysis of a protein. However, the intensity of the envelope of a spectrum is dependent on the total amount of ^{13}C in the sample, which is a combination of both the SNR of each peak and the number of peaks present. This means that, for the same amount of experimental acquisition time, SSD-NMR should provide similar quality results for the same mass of protein, regardless of the size of the protein, as they will have the same total number of ^{13}C atoms in the sample tube. Because of this, SSD-NMR should work for proteins much larger than is typically considered for NMR studies. For solution NMR, this means SSD-NMR can be applied even to proteins where the lines are significantly broadened by long correlation times. In cases where the protein is not soluble, magic-angle spinning solid-state NMR, which has no inherent size limitation, may be able to provide spectra for SSD-NMR on proteins of any size. In fact, results on larger proteins may provide results that are more consistent with STRIDE, as the impact of individual peaks with larger perturbations from the mode shift will be minimized.

One challenge with developing new computational tools is how to estimate the error of the predictions. While this is often accomplished by reporting error terms like those derived from a non-linear regression model, they reflect the model's prediction variance, not necessarily the error from the underlying population. In our case, because we are trying to match chemical shift-based results to those extracted from a PDB file by STRIDE, database testing (Figures 3 and 5) will best reflect our knowledge of the uncertainty in SSD-NMR. These tests show RMSEs of 6.2% for α-helix, 8.2% for β-sheet, and 11% for coil from the STRIDE secondary structure content. These errors are highly competitive with those reported by other comparable techniques such as CD(25, 28) and FT-IR(3, 34). Even in cases where SSD-NMR predictions differ from STRIDE assignments, our method produces a consistent outcome for a protein in a specific conformation at reasonable SNR (Figure 3). This can be used to track conformational changes in a protein, as distinct conformations of the same protein will result in different predictions by SSD-NMR. In our lab, we are now routinely using SSD-NMR on liquid-liquid phase separated proteins to track such changes.

SSD-NMR has a remarkably high correlation to distributions created from TALOS-N output for proteins in the PACSY database. This is particularly noteworthy as TALOS-N requires site specific chemical shift assignments and uses joint information from the chemical shift assignment as well as those of neighboring residues. In fact, it is likely that SSD-NMR's accuracy is near the maximum that could be expected without including residue-residue connections in the fitting process. This is because classification of protein secondary structure relies not only on the backbone torsion angle of individual residues, but also on those of their neighboring residues and on patterns of hydrogen bonding. Our prediction correlation is higher for α-helix in part because the hydrogen bonding network in a helix leads to distinctive torsion angles, giving rise to the separable helical mode chemical shift (Fig. 1). In comparison, β-sheet residues occupy a much larger region of Ramachandran space (35) and may require more information than torsion angles alone to be differentiated from turns and random coils.

Chemical shifts are also representative of the ensemble average of the protein conformation on the NMR timescale. Because of this, in some cases, the observed shifts may not correspond to the singular structure reflected in a static PDB file. This explanation is corroborated by the lower RMSE for SSD-NMR versus TALOS-N (Fig. 5) than versus STRIDE (Fig. 3). SSD-NMR and TALOS-N prediction are both based on measured chemical shifts and are biased by the same effects. Additionally, secondary structure classification is not well-defined during the transition from one secondary structure element to another (such as the transition from an α-helix to a random coil). This is a well-known phenomenon, as classification programs such as STRIDE and DSSP(36) have different thresholds that define secondary structure subtypes.

Our model was developed to be straightforward and to be generalizable to the type of data one expects from NMR spectra of proteins acquired on high-field instruments. However, there are alternative models, improvements, or loss functions that could be explored in future studies. For example, in consideration of the discussion above, we wanted to include nearest neighbor effects in our model. In addition to aiding in classifying secondary structure, it is

well-established that neighboring residues can induce identity-specific chemical shift perturbations to proximal residues (37, 38). The largest of these perturbations are for residues directly preceding a proline, which can move the Cα shift by more than 1 ppm (38). This effect is not accounted for in SSD-NMR and as such, proline-rich proteins are likely to have a poorer fit. However, the representation of all tripeptides in each secondary structure in the BMRB is not complete enough to account for these perturbations in our model. To address this, targeted experimental studies are needed to extend these statistics beyond the limitations of our current databases, and may provide an avenue for site-specific secondary structure assignment from 1D NMR spectra alone. Barring such experimental studies, chemical shift prediction of high-quality X-ray structures, similar to how TALOS-N generates its database (14), or high-level theory calculations may be able to provide sufficient data to fill this gap.

Additional spectra, such as a ^{1}H 1D, ^{15}N 1D, and/or multidimensional spectra, may also be incorporated into future iterations of SSD-NMR. Previous approaches for estimating protein secondary structure from NMR *in lieu* of complete chemical shift assignment have been reported using ^{1}H 1D or multidimensional spectra. In general, these methods focused on solution NMR and relied on manual analysis of individual peaks(39), acquisition of more challenging multidimensional experiments(40), or both(41). These methods thus generally required more manual interpretation of the data, are difficult to compare to solid-state NMR spectra on proteins in condensed phases, and required higher resolution and/or sensitivity than is needed for SSD-NMR. Further, ^{13}C shifts are more disperse and are less sensitive to environmental changes (such as solvent and pH) when compared to ^{1}H and ^{15}N shifts. These properties have made ^{13}C chemical shifts appealing targets for chemical shift prediction(42-44) and algorithmic development(45) and contribute to the robustness of SSD-NMR.

When acquiring experimental data for use in SSD-NMR, we recommend following experimental protocols designed to provide accurately referenced spectra with quantitative relative intensities. SSD-NMR assumes that data is referenced to the DSS standard, ideally by inclusion of DSS in solution samples (19) or external referencing using adamantane for solid

samples(18). While SSD-NMR is able to detect systemic referencing errors using our re-referencing term δ_r (Figure S6,S7), the performance is lower than if the experimental data was properly referenced. SSD-NMR further assumes that each carbon contributes equal intensity in the final spectrum. While SSD-NMR is robust to varied intensities, as evidenced by its high correlation even at low SNR, if the experimental spectrum is lacking signals or has non-uniform intensities then the accuracy will be lower. In this work we chose to acquire fully relaxed directly polarized Bloch decay spectra for the experimental results presented above. However, any quantitative ^{13}C spectrum(46, 47) should work similarly.

If the spectrum has obvious baseline distortions, that should also be addressed at the spectrometer prior to acquiring data. To this end, collecting data with a short spin echo can alleviate some baseline distortion. Digital baseline corrections can be applied during processing but should be applied judiciously to not distort relative intensities. In general, we expect our method to provide consistent results for referenced experimental NMR data with a flat baseline and linewidths between 0.4 and 9 ppm and with signal-to-noise ratios greater than 5 post-apodization. Should the observed linewidth be narrower than 1ppm, the peaks can be broadened during processing. Optimal performance is expected when the linewidths in the apodized spectrum are near 1 ppm and with a peak signal-to-noise ratio greater than 10. SSD-NMR's reliance on solution or solid-state NMR spectroscopy provide it unique advantages when compared to other biophysical techniques for characterizing secondary structure. It can be applied to the wide array of biological systems across all physiologically and disease relevant phases accessible to NMR spectroscopy – including soluble proteins, membrane proteins(48), microcrystalline proteins(26), protein complexes, protein aggregates(49-52), liquid-liquid phase separated coacervates(7, 53, 54), and even living cells(8). Additionally, ^{13}C isotope incorporation provides a structurally non-perturbing approach to labeling that can be used to selectively probe protein conformation in heterogeneous mixtures. For example, one can isotopically label a single protein of interest in a large complex. The resulting spectrum has signals from the labeled protein, while the rest of the complex is suppressed(6). The same

approach can be extended to probing specific domains in a single protein using intein segmental labeling. We envision that SSD-NMR and future iterations will greatly increase the throughput of such studies and be accessible to any researcher with access to an NMR spectrometer.

In summary, we have introduced a gradient-descent based approach, called *secondary-structure distribution by NMR* (SSD-NMR), to determine the secondary structure composition of a protein sample using a single 1D ^{13}C NMR spectrum and the sequence of the protein. SSD-NMR can accurately determine the secondary structure percentage composition of a protein as demonstrated on nearly 900 proteins from simulated experimental spectra. We further showed experimentally that this approach can be used by either solution or solid-state NMR and demonstrated practical approaches to apply it to natural abundance proteins. The model's robustness to signal-to-noise ratio and resolution extends it applicability to samples at low magnetic field, as demonstrated with isotopically labelled ubiquitin on a benchtop 60 MHz spectrometer. This method is especially useful for proteins found in condensed phase environments, such as those related to membranes, aggregated states, and liquid-liquid segregated phases.

Acknowledgements

This work was supported by the Kavli Institute for Neuroscience at Yale University Postdoctoral Fellowship to M.D.T.. K.W.Z. acknowledges support from the grants NIH R01 AG034924-06A1 and 5 R01 NS118796-03. V.S.B. acknowledges support from the grant NIH R01GM136815.

Data Availability

All code, reference accession codes, and the current implementation of SSD-NMR can be found at: https://github.com/haoteli/Secondary-Structure-Determination-NMR
The experimental data from the manuscript are included as examples of how to run the algorithm in the github respository.

References

1. W. Pirovano, J. Heringa, "Protein Secondary Structure Prediction" in Data Mining Techniques for the Life Sciences, O. Carugo, F. Eisenhaber, Eds. (Humana Press, Totowa, NJ, 2010), 10.1007/978-1-60327-241-4_19, pp. 327-348.
2. B. Ranjbar, P. Gill, Circular Dichroism Techniques: Biomolecular and Nanostructural Analyses- A Review. *Chemical Biology & Drug Design* **74**, 101-120 (2009).
3. M. Fevzioglu, O. K. Ozturk, B. R. Hamaker, O. H. Campanella, Quantitative approach to study secondary structure of proteins by FT-IR spectroscopy, using a model wheat gluten system. *International Journal of Biological Macromolecules* **164**, 2753-2760 (2020).
4. J. Lippert, D. Tyminski, P. Desmeules, Determination of the secondary structure of proteins by laser Raman spectroscopy. *Journal of the American Chemical Society* **98**, 7075-7080 (1976).
5. L. S. Vermeer, A. Marquette, M. Schoup, D. Fenard, A. Galy, B. Bechinger, Simultaneous Analysis of Secondary Structure and Light Scattering from Circular Dichroism Titrations: Application to Vectofusin-1. *Scientific Reports* **6**, 39450 (2016).
6. P. C. A. van der Wel, New applications of solid-state NMR in structural biology. *Emerging Topics in Life Sciences* **2**, 57-67 (2018).
7. A. C. Murthy, N. L. Fawzi, The (un)structural biology of biomolecular liquid-liquid phase separation using NMR spectroscopy. *Journal of Biological Chemistry* **295**, 2375-2384 (2020).
8. S. Narasimhan, G. E. Folkers, M. Baldus, When Small becomes Too Big: Expanding the Use of In-Cell Solid-State NMR Spectroscopy. *ChemPlusChem* **85**, 760-768 (2020).
9. S. Spera, A. Bax, Empirical correlation between protein backbone conformation and C.alpha. and C.beta. 13C nuclear magnetic resonance chemical shifts. *Journal of the American Chemical Society* **113**, 5490-5492 (1991).
10. E. Oldfield, Chemical shifts and three-dimensional protein structures. *Journal of Biomolecular NMR* **5**, 217-225 (1995).
11. D. S. Wishart, B. D. Sykes, The 13C Chemical-Shift Index: A simple method for the identification of protein secondary structure using 13C chemical-shift data. *Journal of Biomolecular NMR* **4**, 171-180 (1994).
12. K. J. Fritzsching, M. Hong, K. Schmidt-Rohr, Conformationally selective multidimensional chemical shift ranges in proteins from a PACSY database purged using intrinsic quality criteria. *Journal of Biomolecular NMR* **64**, 115-130 (2016).
13. D. S. Wishart, B. D. Sykes, F. M. Richards, The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* **31**, 1647-1651 (1992).
14. Y. Shen, A. Bax, Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *Journal of Biomolecular NMR* **56**, 227-241 (2013).

15. M. Heinig, D. Frishman, STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic acids research* **32**, W500-W502 (2004).
16. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
17. A. Paszke *et al.* (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems*, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alche-Buc, E. Fox, R. Garnett (Curran Associates, Inc.).
18. C. R. Morcombe, K. W. Zilm, Chemical shift referencing in MAS solid state NMR. *Journal of Magnetic Resonance* **162**, 479-486 (2003).
19. R. K. Harris, E. D. Becker, S. M. Cabral de Menezes, P. Granger, R. E. Hoffman, K. W. Zilm, Further conventions for NMR shielding and chemical shifts (IUPAC Recommendations 2008). *Pure and Applied Chemistry* **80**, 59-84 (2008).
20. H. K. Fasshuber *et al.*, Structural heterogeneity in microcrystalline ubiquitin studied by solid-state NMR.
21. W. Lee, W. Yu, S. Kim, I. Chang, W. Lee, J. L. Markley, PACSY, a relational database management system for protein structure and chemical shift analysis. *Journal of biomolecular NMR* **54**, 169-179 (2012).
22. E. L. Ulrich *et al.*, BioMagResBank. *Nucleic Acids Research* **36**, D402-D408 (2008).
23. H. M. Berman *et al.*, The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).
24. B. Wang, Y. Wang, D. S. Wishart, A probabilistic approach for validating protein NMR chemical shift assignments. *Journal of Biomolecular NMR* **47**, 85-99 (2010).
25. A. Micsonai *et al.*, BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. *Nucleic acids research* **46**, W315-W322 (2018).
26. W. T. Franks *et al.*, Magic-angle spinning solid-state NMR spectroscopy of the β1 immunoglobulin binding domain of protein G (GB1): 15N and 13C chemical shift assignments and conformational analysis. *Journal of the American Chemical Society* **127**, 12291-12305 (2005).
27. N. P. Wickramasinghe, M. Kotecha, A. Samoson, J. Past, Y. Ishii, Sensitivity enhancement in 13C solid-state NMR of protein microcrystals by use of paramagnetic metal ions for optimizing 1H T1 relaxation. *Journal of Magnetic Resonance* **184**, 350-356 (2007).
28. N. J. Greenfield, Using circular dichroism spectra to estimate protein secondary structure. *Nature Protocols* **1**, 2876-2890 (2006).
29. A. Dong, P. Huang, W. S. Caughey, Protein secondary structures in water from second-derivative amide I infrared spectra. *Biochemistry* **29**, 3303-3308 (1990).
30. M. Di Foggia, S. Bonora, V. Tugnoli, DSC and Raman study on the effect of lysozyme and bovine serum albumin on phospholipids liposomes. *Journal of Thermal Analysis and Calorimetry* **111**, 1871-1880 (2013).
31. R. W. Martin, E. K. Paulson, K. W. Zilm, Design of a triple resonance magic angle sample spinning probe for high field solid state nuclear magnetic resonance. *Review of Scientific Instruments* **74**, 3045-3061 (2003).
32. G. Comellas, J. J. Lopez, A. J. Nieuwkoop, L. R. Lemkau, C. M. Rienstra, Straightforward, effective calibration of SPINAL-64 decoupling results in the enhancement of sensitivity

and resolution of biomolecular solid-state NMR. *Journal of Magnetic Resonance* **209**, 131-135 (2011).
33. R. W. Martin, K. W. Zilm, Variable temperature system using vortex tube cooling and fiber optic temperature measurement for low temperature magic angle spinning NMR. *Journal of Magnetic Resonance* **168**, 202-209 (2004).
34. J. De Meutter, E. Goormaghtigh, Evaluation of protein secondary structure from FTIR spectra improved after partial deuteration. *European Biophysics Journal* **50**, 613-628 (2021).
35. S. C. Lovell *et al.*, Structure validation by Cα geometry: φ,ψ and Cβ deviation. *Proteins: Structure, Function, and Bioinformatics* **50**, 437-450 (2003).
36. R. P. Joosten *et al.*, A series of PDB related databases for everyday needs. *Nucleic Acids Research* **39**, D411-D419 (2011).
37. Y. Wang, O. Jardetzky, Investigation of the Neighboring Residue Effects on Protein Chemical Shifts. *Journal of the American Chemical Society* **124**, 14075-14084 (2002).
38. S. Schwarzinger, G. J. A. Kroon, T. R. Foss, J. Chung, P. E. Wright, H. J. Dyson, Sequence-Dependent Correction of Random Coil NMR Chemical Shifts. *Journal of the American Chemical Society* **123**, 2970-2978 (2001).
39. D. S. Wishart, B. D. Sykes, F. M. Richards, Simple techniques for the quantification of protein secondary structure by 1H NMR spectroscopy. *FEBS Letters* **293**, 72-80 (1991).
40. Y. Wang, O. Jardetzky, Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein Sci* **11**, 852-861 (2002).
41. S. P. Mielke, V. V. Krishnan, Characterization of protein secondary structure from NMR chemical shifts. *Progress in Nuclear Magnetic Resonance Spectroscopy* **54**, 141-165 (2009).
42. B. Han, Y. Liu, S. W. Ginzinger, D. S. Wishart, SHIFTX2: significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR* **50**, 43-57 (2011).
43. Y. Shen, A. Bax, SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *Journal of Biomolecular NMR* **48**, 13-22 (2010).
44. J. Li, K. C. Bennett, Y. Liu, M. V. Martin, T. Head-Gordon, Accurate prediction of chemical shifts for aqueous protein structure on "Real World" data. *Chemical Science* **11**, 3180-3191 (2020).
45. Joseph M. Courtney *et al.*, Experimental Protein Structure Verification by Scoring with a Single, Unassigned NMR Spectrum. *Structure* **23**, 1958-1966 (2015).
46. R. L. Johnson, K. Schmidt-Rohr, Quantitative solid-state 13C NMR with signal enhancement by multiple cross polarization. *Journal of Magnetic Resonance* **239**, 44-49 (2014).
47. E. Caytan, G. S. Remaud, E. Tenailleau, S. Akoka, Precise and accurate quantitative 13C NMR with reduced experimental time. *Talanta* **71**, 1016-1021 (2007).
48. V. Ladizhansky, Applications of solid-state NMR to membrane proteins. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1865**, 1577-1586 (2017).
49. M. D. Tuttle *et al.*, Solid-state NMR structure of a pathogenic fibril of full-length human α-synuclein. *Nature Structural & Molecular Biology* **23**, 409-415 (2016).

50. M. Lee, U. Ghosh, K. R. Thurber, M. Kato, R. Tycko, Molecular structure and interactions within amyloid-like fibrils formed by a low-complexity protein sequence from FUS. *Nature Communications* **11**, 5735 (2020).
51. H. Van Melckebeke *et al.*, Atomic-Resolution Three-Dimensional Structure of HET-s(218−289) Amyloid Fibrils by Solid-State NMR Spectroscopy. *Journal of the American Chemical Society* **132**, 13765-13775 (2010).
52. J.-X. Lu, W. Qiang, W.-M. Yau, Charles D. Schwieters, Stephen C. Meredith, R. Tycko, Molecular Structure of β-Amyloid Fibrils in Alzheimer's Disease Brain Tissue. *Cell* **154**, 1257-1268 (2013).
53. P. Brady Jacob *et al.*, Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proceedings of the National Academy of Sciences* **114**, E8194-E8203 (2017).
54. M. A. Kostylev *et al.*, Liquid and Hydrogel Phases of PrPC Linked to Conformation Shifts and Triggered by Alzheimer's Amyloid-β Oligomers. *Molecular Cell* **72**, 426-443.e412 (2018).