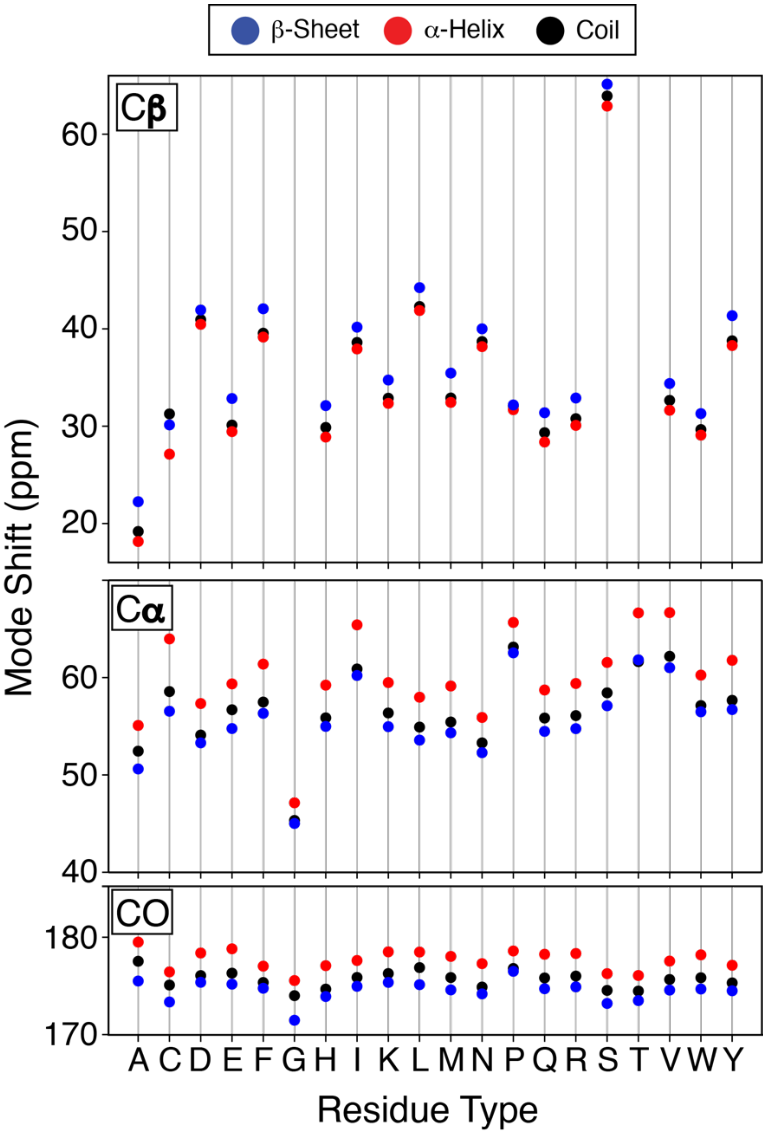
Supporting Information for:

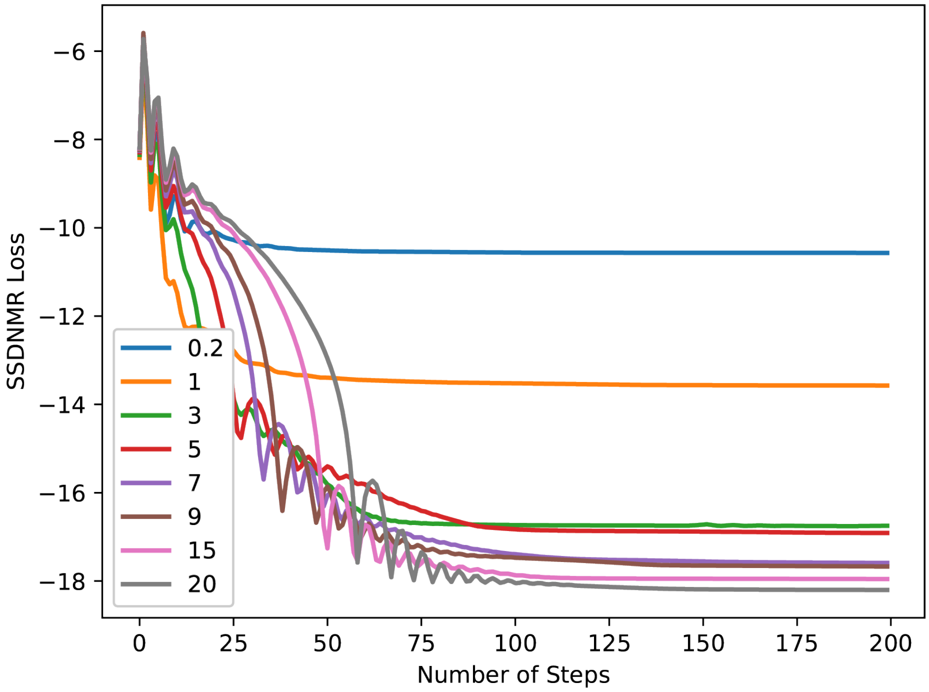
Rapid Quantification of Protein Secondary Structure Composition from a Single Unassigned 1D ^{13}C NMR Spectrum

Haote Li[†,1], MarcusD. Tuttle[†,1], Kurt W. Zilm[1], Victor S. Batista[1]

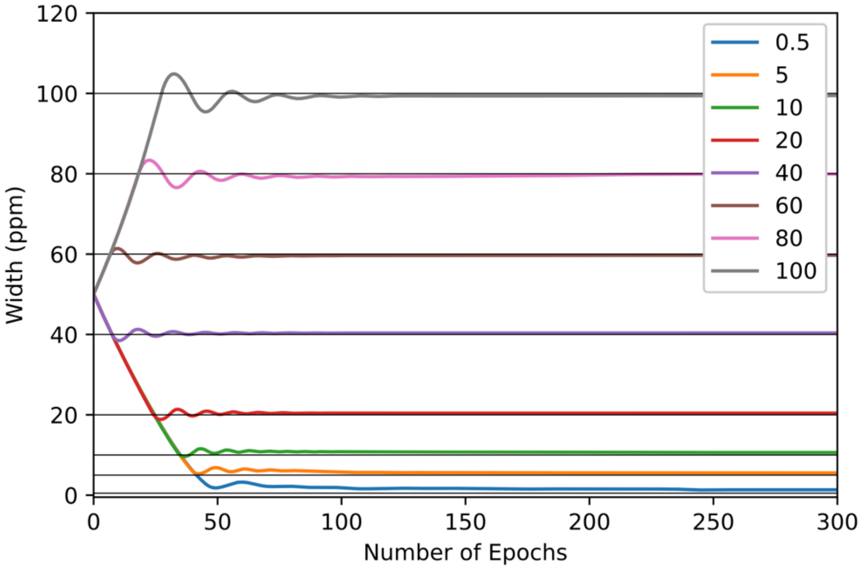
Table of Contents



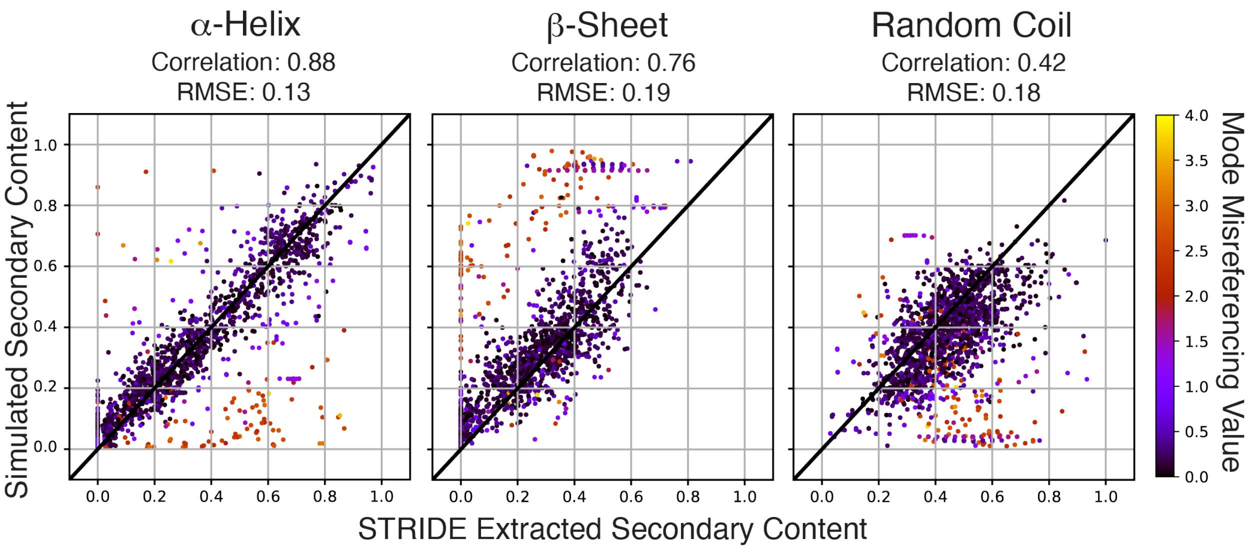
Supplemental Figure S1: The mode ^{13}C secondary structure-dependent chemical shifts for each backbone atom by amino acid. These values were reported by Fritzsching *et al.* Red colored dots are for α-helix, blue β-sheet, and black are random coil. (Top) Backbone Cα mode shifts. (Middle) Sidechain Cβ chemical shifts. (Bottom) Backbone CO shifts.



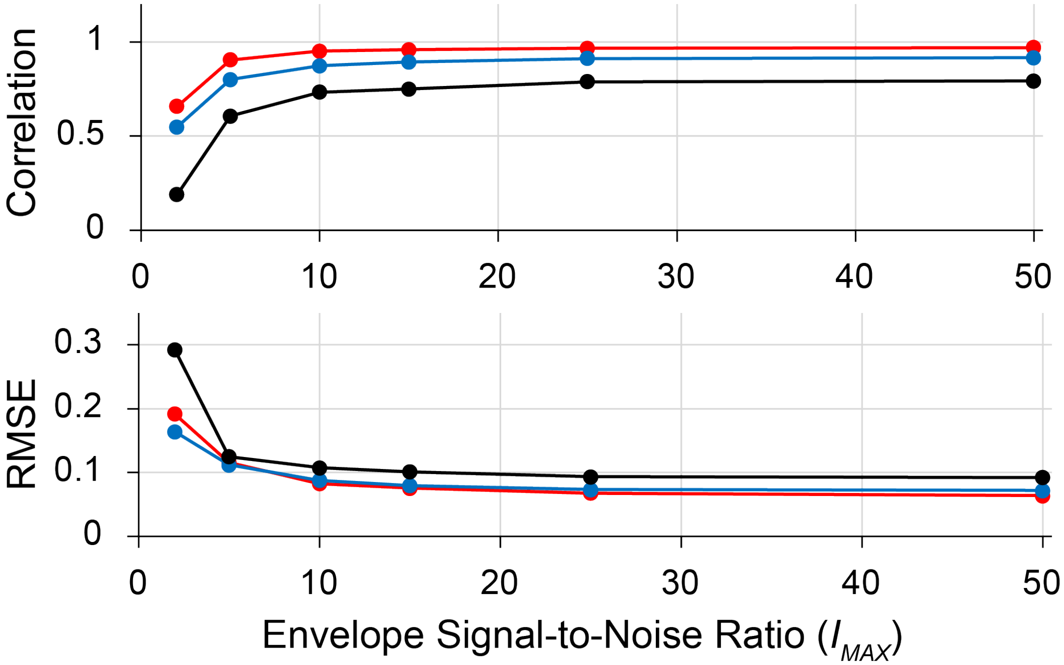
Supplemental Figure S2: Loss function output of simulated ubiquitin (BMRB 25123) as a function of GD steps. The value of the loss function at each step with simulated linewidths ranging from 0.2 to 20 ppm. This shows that the loss function converges in less than 100 steps with an initial σ at 0.4 ppm for linewidths less than 20 ppm.



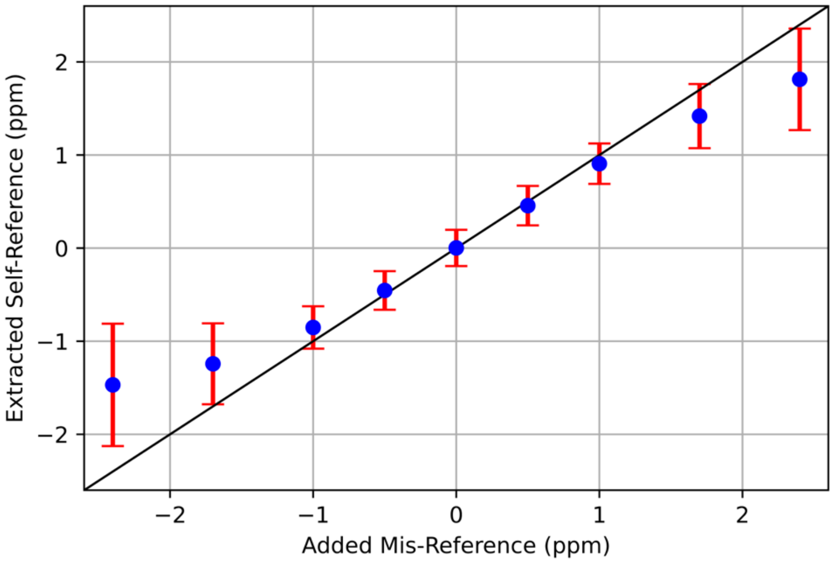
Supplemental Figure S3: BMRB 25123 convergence evaluation with an initial σ parameter set to 50 ppm. The result shows that the choice initial σ at 50 ppm accelerates the convergence when the SES is initialized at 60 ppm and 40 ppm than when σ is set to 0.04. Where is 0.04?



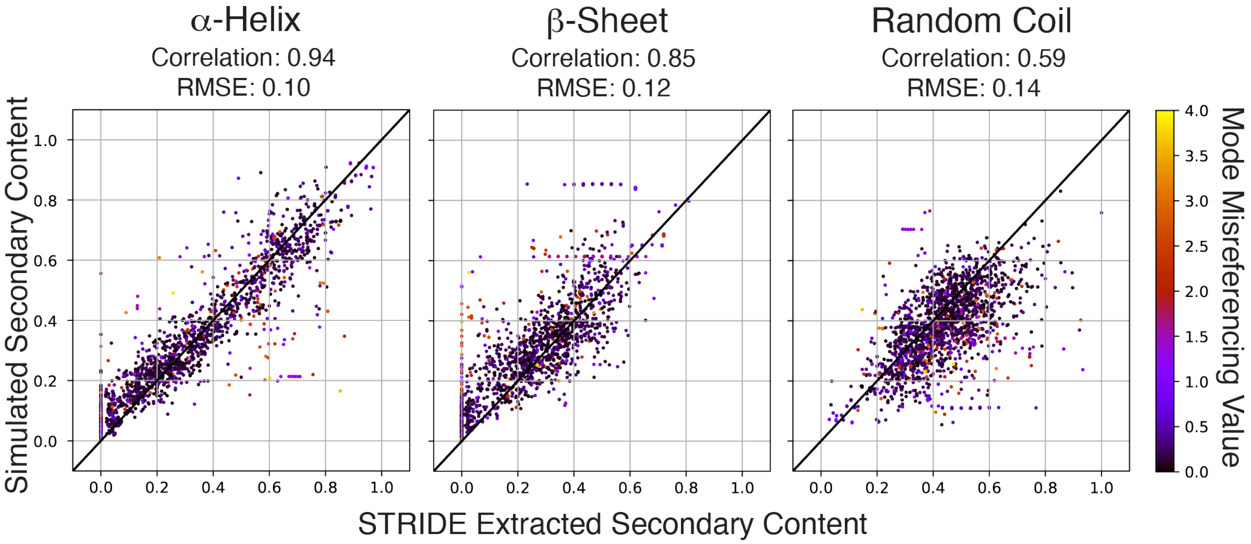
Supplemental Figure S4: The result from the entire filtered PACSY dataset. Correlation between SSD-NMR (vertical axis) and secondary structure content fit by STRIDE from their respective PDB file (horizontal axis). Sample of 1839 proteins with estimated absolute mode referencing errors was less than or equal to 4 ppm as reported by Fritzsching *et al.*. For each protein, its secondary structure content from its first conformer in the PDB are plotted from left to right in orders of Random Coil, β-sheet, and α-Helix. The method has correlation coefficients of 0.42, 0.76 and 0.88 to STRIDE classified percentages. This result emphasizes the need for accurate chemical shift referencing.



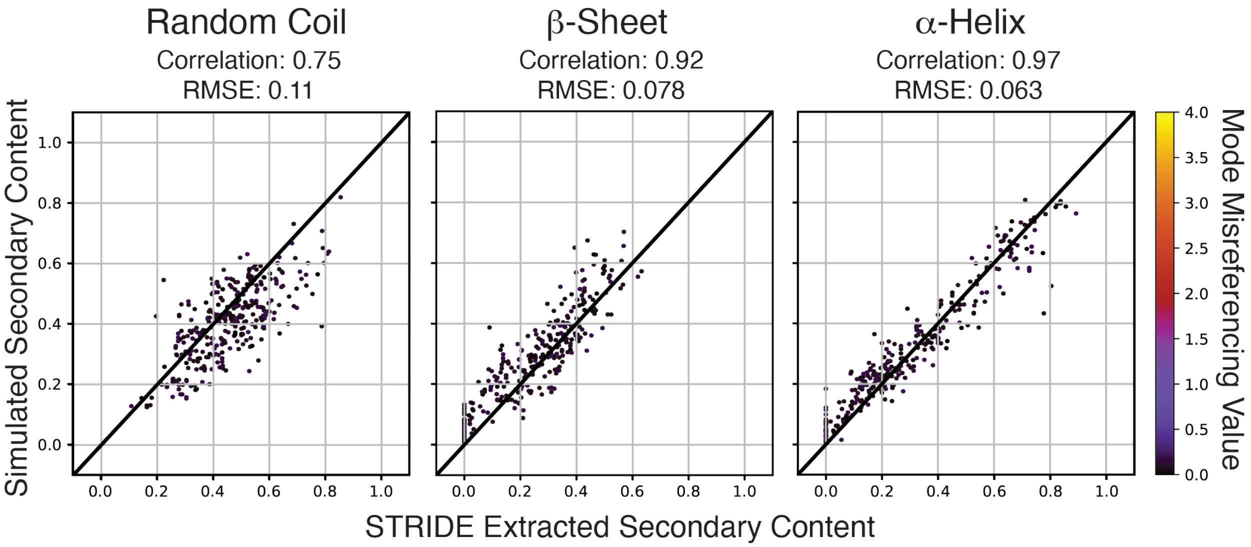
Supplemental Figure S5: Correlation and RMSE of SSD-NMR based on the SNR of the Envelope of the Spectrum. The correlation and RMSE of SSD-NMR of 891 proteins with SNR ranging from 2 to 50. This demonstrates that SSD-NMR is robust to the overall envelope of the spectrum, and not just the intensity of individual peaks.



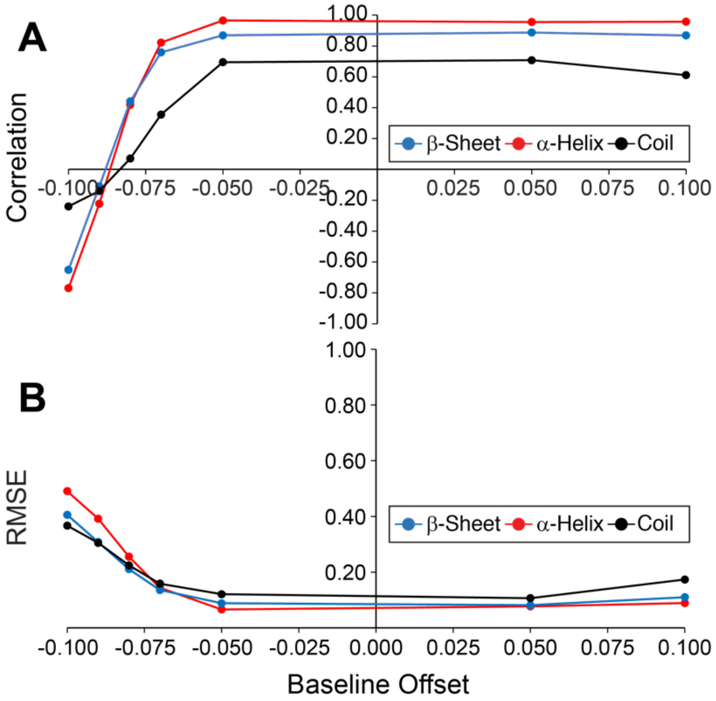
Supplemental Figure S6: The extracted result that measures the performance of the self-referencing parameter. For each of the data points above, a mis-reference is added to the 891 "well-referenced" proteins. Our algorithm, in each case, returns 891 self-reference parameters. For each distribution of extracted self-reference parameters, it is fitted to a standard Gaussian. The y-axis represents the means of the extracted Gaussians with the error bars representing the STDs. For 100 steps in each optimization, our algorithm shows better performance when the mis-reference is in between –0.5 and 0.5 ppm with errors around –0.2 ppm. In the tested range, the self-referencing parameter is consistent in predicting the sign of mis-referencing. For larger re-reference errors, the algorithm requires more steps for the self-referencing parameter to converge. The self-referencing parameter is suitable for detecting potential mis-referencing in spectrum preparation and processing. However, when the absolute value of mis-referencing is suspected to be greater than 1, more than 100 steps are necessary for getting accurate referencing results.



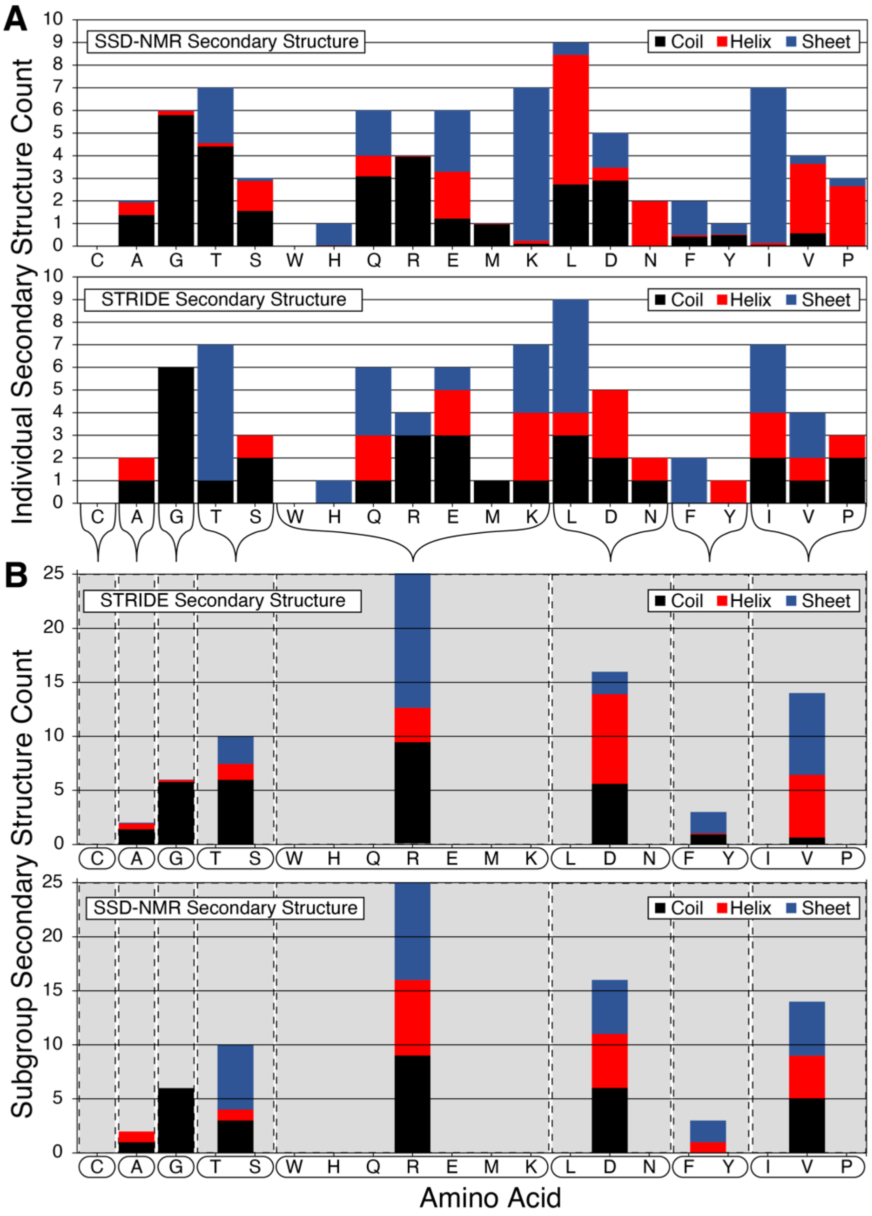
Supplemental Figure S7: The result from the entire filtered PACSY dataset with included self-re-referencing via SSD-NMR. Correlation between SSD-NMR with self-referencing (vertical axis) and secondary structure content fit by STRIDE from their respective PDB file (horizontal axis). In this case, we ran GD for 500 steps to allow better convergence on the same 1839 proteins as Figure 3. For each protein, its secondary structure content from its first conformer in the PDB are plotted from left to right in orders of Random Coil, β-sheet, and α-Helix. The method has correlation coefficients of 0.59, 0.85 and 0.94 to STRIDE classified percentages, demonstrating a marked increase in correlation compared to the case when self-referencing is not included (Fig. S6), demonstrating the utility of our re-referencing parameter.



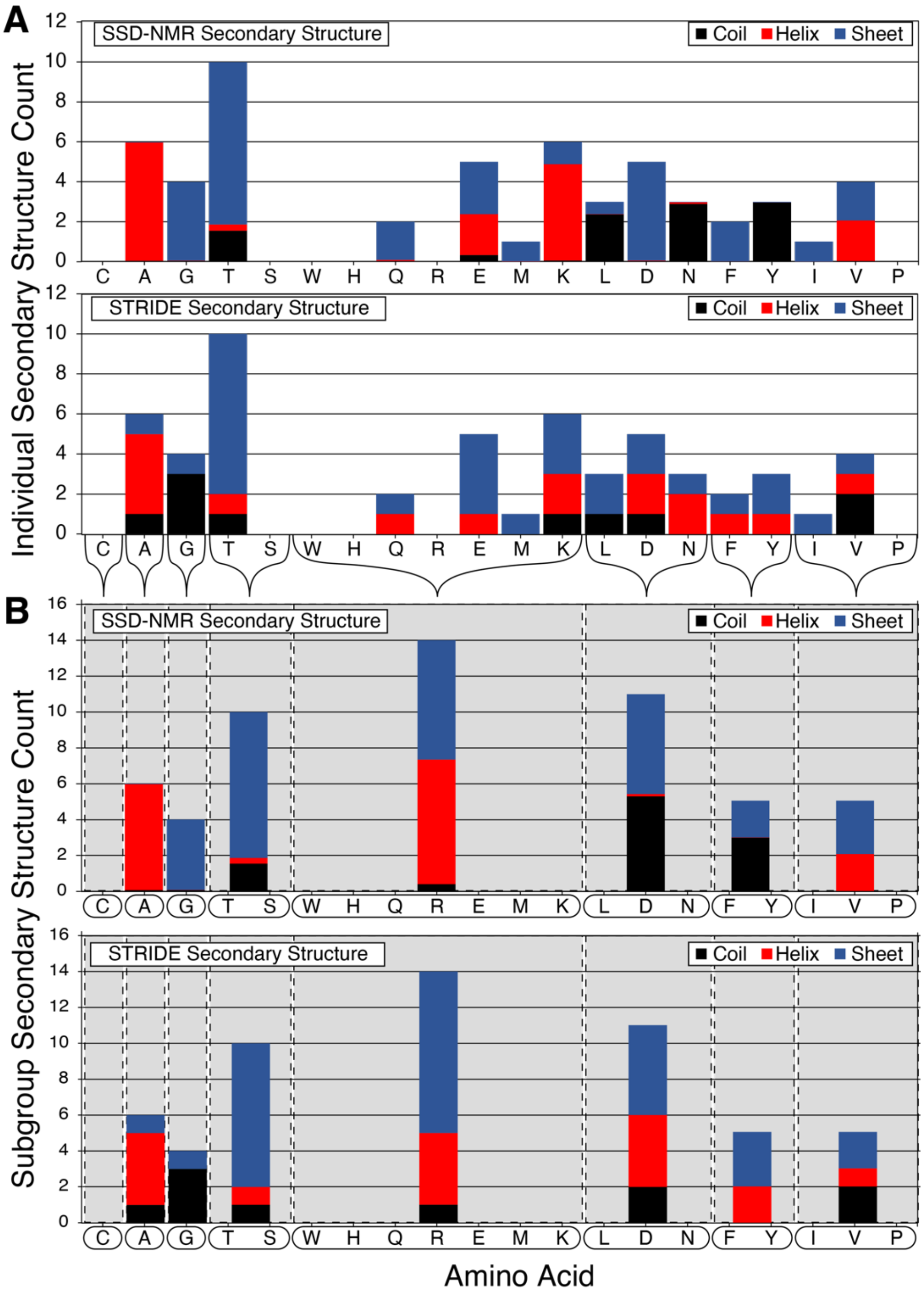
Supplemental Figure S8: Secondary Structure Distribution Prediction from SSD NMR compared to STRIDE Secondary Structure Distribution using proteins that passed the TALOS-N benchmarking. A subset of 332 proteins from Figure 3 showing that the subset of proteins that were passed through TALOS-N had similar correlation to the full data-set.



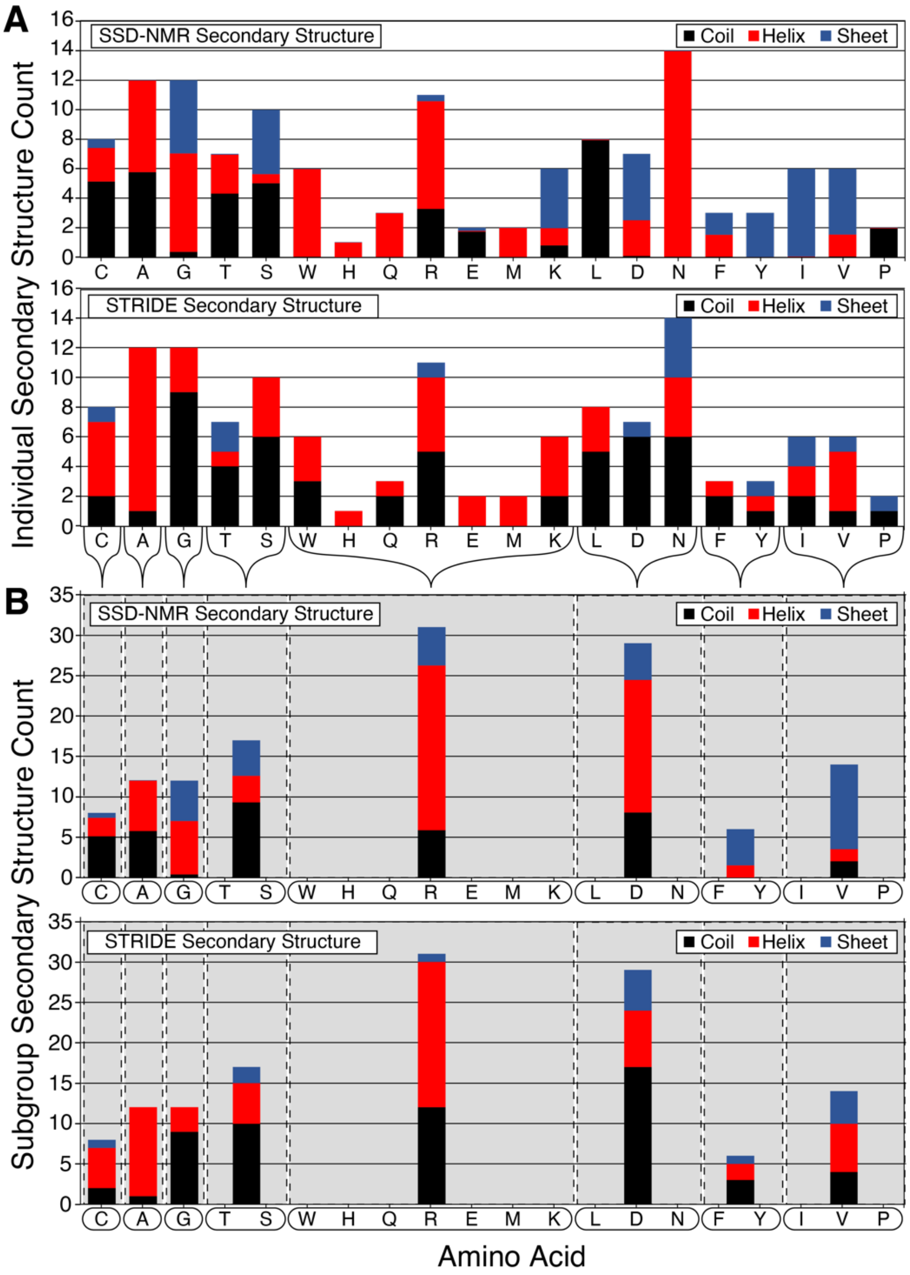
Supplemental Figure S9: Uniform baseline offsets lower correlation and RMSE. To each of the 891 proteins from Fig. 3, a uniform offset was added to each point to either lift or lower the entire spectrum. The offset (vertical axis) is the percentage to its maximum intensity in the no-offset spectrum. For small baseline offsets, the RMSE is very stable, but moderate (greater than -0.05 or -5%) negative offsets cause instabilities in the algorithm due to ill-posed normalization.



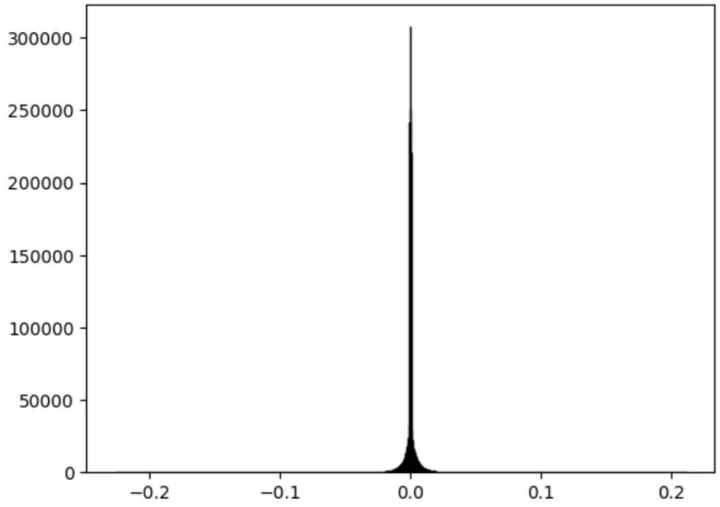
Supplemental Figure S10: Amino Acid and Subgroup-Specific Secondary Structure Distributions for Ubiquitin . (A) The count of each residue type in each secondary using the a_i^{ss} from SSD-NMR (top) or STRIDE (bottom). The individual distributions are not as accurate as the overall distribution from Figure 6. By combining the individual amino acids into the subgroups (B) C, A, G, TS, WHQREMK, LDN, FY, and IVP, the agreement is better between SSD-NMR (top) and STRIDE (bottom).



Supplemental Figure S11: Amino Acid and Subgroup-Specific Secondary Structure Distributions for GB1 . (A) The count of each residue type in each secondary using the a_i^{ss} from SSD-NMR (top) or STRIDE (bottom). The individual distributions are not as accurate as the overall distribution from Figure 6. By combining the individual amino acids into the subgroups (B) C, A, G, TS, WHQREMK, LDN, FY, and IVP, the agreement is better between SSD-NMR (top) and STRIDE (bottom).



Supplemental Figure S12: Amino Acid and Subgroup-Specific Secondary Structure Distributions for Lysozyme: (Top) The count of each residue type in each secondary using the a_i^{ss} from SSD-NMR (left) or STRIDE (right). The individual distributions are not as accurate as the overall distribution from Figure 8. By combining the individual amino acids into the subgroups (bottom) C, A, G, TS, WHQREMK, LDN, FY, and IVP, the agreement is better between SSD-NMR (left) and STRIDE (right).



Supplemental Figure S13: Distribution of Extracted $\Delta\delta^{k,ss}$ and Δv^{k} from Figure 3: Distribution of $\Delta\delta^{k,ss}$ values from every protein in Figure 3 at 1 ppm linewidths. In almost every case, the values of $\Delta\delta^{k,ss}$ are less than the experimental error of the measurement.