Journal of Biomolecular NMR (2022) 76:197–212
https://doi.org/10.1007/s10858-022-00405-0

ARTICLE

Characterization of conformational heterogeneity via higher-dimensionality, proton-detected solid-state NMR

Ekaterina Burakova[1,2] · Suresh K. Vasa[1,2] · Rasmus Linser[1,2]

Received: 17 May 2022 / Accepted: 7 September 2022 / Published online: 23 September 2022

Abstract

Site-specific heterogeneity of solid protein samples can be exploited as valuable information to answer biological questions ranging from thermodynamic properties determining fibril formation to protein folding and conformational stability upon stress. In particular, for proteins of increasing molecular weight, however, site-resolved assessment without residue-specific labeling is challenging using established methodology, which tends to rely on carbon-detected 2D correlations. Here we develop purely chemical-shift-based approaches for assessment of relative conformational heterogeneity that allows identification of each residue via four chemical-shift dimensions. High dimensionality diminishes the probability of peak overlap in the presence of multiple, heterogeneously broadened resonances. Utilizing backbone dihedral-angle reconstruction from individual contributions to the peak shape either via suitably adapted prediction routines or direct association with a relational database, the methods may in future studies afford assessment of site-specific heterogeneity of proteins without site-specific labeling.

Keywords Fast-MAS solid-state NMR · Proton detection · Sample heterogeneity · TALOS · PACSY

Introduction

Protein disorder plays a significant role in various cellular processes (Uversky 2013, 2018). Intrinsically disordered proteins (IDPs) and intrinsically disordered protein regions (IDRs), due to their high flexibility and high accessibility, are crucial elements of transcription factors (Sammak and Zinzalla 2015), voltage-dependent gating (Zhou et al. 2001; Kjaergaard and Kragelund 2017), protein phase separation (Turoverov et al. 2019), and many others. Intrinsic disorder gives proteins the ability to form low-affinity but highly specific complexes, which is important for regulatory pathways (Uversky 2013). Similarly, ensembles of partially folded intermediates can provide valuable insight into protein folding, mis- or refolding (Havlin and Tycko 2005a; Hu et al. 2009; Potapov et al. 2015). Whereas NMR has proven to be an invaluable tool to characterize the level of disorder in solution (Lindorff-Larsen et al. 2004; Nielsen and Mulder 2020), static disorder in the solid state, either related to the conformational ensemble of the protein in solution or as a biological property on its own, can be assessed by solid-state NMR spectroscopy (Siemer 2020). Upon aggregation, solidification, or even crystallization, disorder can be captured for part of the protein sequence, even if other parts transition into well-ordered structural elements. As a consequence, heterogeneity in the solid state is of significance for example for understanding the aggregation mechanisms of amyloids (Morris et al. 2012; Elkins et al. 2016; Xiang et al. 2017) as well as the formation principles of complex biological conjugates such as spider silk (Asakura et al. 2013a, b). "Arrested dynamic disorder" can be quantified in freeze-trapped solutions (Havlin and Tycko 2005b; Heise et al. 2005), which can capture the previous physiologically relevant conformational distribution. Disorder in the solid state may manifest itself in the presence of a low number of distinct forms, as it is sometimes the case for proteins in polymorphic amyloid assemblies (Paravastu et al. 2008; Tycko 2011, 2014; Amo et al. 2012; Jaroniec 2019). Alternatively, a distribution of numerous conformations can

✉ Rasmus Linser
rasmus.linser@tu-dortmund.de

[1] Department of Chemistry and Chemical Biology, Technical University Dortmund, Otto-Hahn-Str. 4a, 44227 Dortmund, Germany

[2] Department of Chemistry and Pharmacy, Ludwig-Maximilians-University Munich, Butenandtstr. 5-13, 81377 Munich, Germany

arise, as sometimes found for membrane protein preparations already at room temperature (Su and Hong 2011), for crystalline proteins upon freezing (Luo and Yu 2008; Linden et al. 2011; Siemer et al. 2012; Endapally et al. 2019), for folding intermediates (Chimon and Ishii 2005; Havlin and Tycko 2005b), or freeze-trapped IDRs (Hu and Tycko 2010), which in the extreme case can span a continuously sampled conformational space. With significant recent improvements in hardware, there is a growing interest in utilization of solid-state NMR assessment not only using standard MAS approaches but also—given the intrinsically low sensitivity of frozen solutions and the general hurdle of broad (and thus lower-height) peaks—via dynamic nuclear polarization (DNP) techniques (Siemer et al. 2012; Uluca et al. 2018; Jeon et al. 2019).

Different approaches toward quantitative characterization of the static disorder have been described, where either isotropic chemical shifts, CSA, or dipolar coupling correlations are used and translated into best-fit ensemble models via principle-component analysis of spectral features or via MD-derived ensembles *a-posteriori* weighted by comparison of in-silico with experimental data. Despite the enormous challenges on the way, e. g., the faithful simulation of isotropic chemical shifts of individual conformer contributions, the often-underdetermined relationship between experimental values and dihedral properties, as well as the resolution limitations occurring for the desired type of experiment for a more complex target protein, quantitative insights into secondary-structure polymorphism have been obtained in several studies by deconvolution of ^{13}C-detected NMR spectra. Some examples are an analysis of measured and simulated peaks in static 2D spin-diffusion experiments in model polymers (Asakura et al. 2001) or spider silk with isotope labeling by residue type (Kümmerlen et al. 1996), assessment of spider silk via DOQSY spectra (Beek et al. 2000), the conformation of peptide T in frozen glycerol/water solution using 2D ^{13}C–^{13}C exchange spectroscopy (Dios et al. 2004), double quantum/zero quantum- (DQ–ZQ) spectroscopy of the neurotensin peptide without cryoprotectant (Heise et al. 2005), a linear combination of “clean” SQ/SQ spectra or correlated anisotropic interactions of carbonyls for assessing the unfolding of HP35 (Havlin and Tycko 2005a; Hu et al. 2009; Hu and Tycko 2010), PCA analysis of melittin spectra interrupted upon folding (Hu et al. Oct. 2009), selectively labeled α-synuclein in frozen solution (Uluca et al. Apr. 2018), and others. One of the biggest hurdles in the presence of conformational ensembles is sufficient signal dispersion for downstream processing of NMR data for more in-depth analyses. To overcome it, selective amino acid labeling is often used (Tycko 2014), which, however, largely reduces the information content per sample and necessitates a high reproducibility of the sample preparation in case multiple sites are to be investigated.

In this work, to enable the readout of multiple heterogeneously broadened peaks (from different residues) at once, we explore two purely chemical-shift-based approaches for residue-specific evaluation of the dihedral angle distribution in conjunction with higher-dimensionality spectra. The first one relies on exploiting chemical-shift based dihedral-angle predictions, most importantly via TALOS-N (Shen and Bax 2013). A complementary approach is a direct chemical shift database comparison (Fig. 1). For a possible utilization for future biological questions, we specifically include chemical shifts of backbone H^N, N^H, and C^α nuclei as well as C^β (i.e. four dimensions) to achieve a chemical shift dispersion as large as possible, which gives access to peak features within the higher-dimensionality shift correlations. To be able to interpret the resulting conformational ensembles in terms

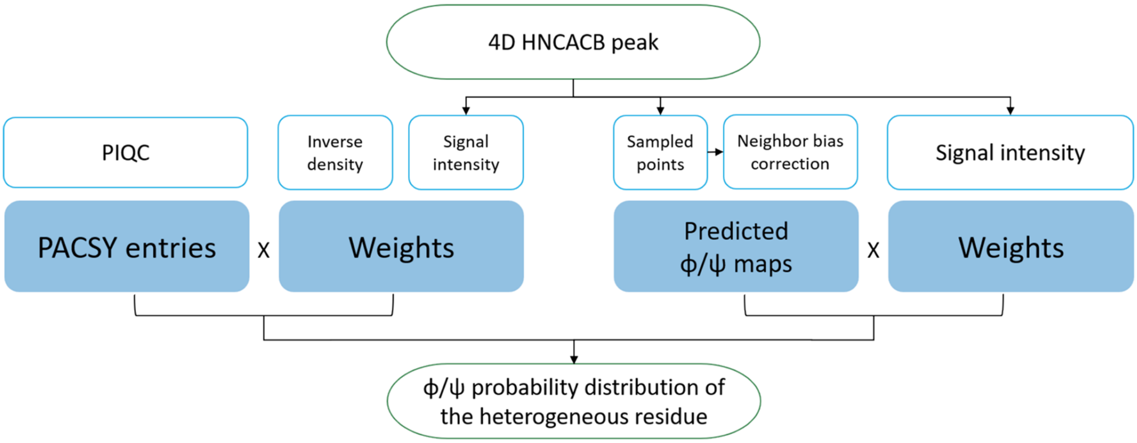

Fig. 1 Flowchart of the two developed approaches to study peak shapes in higher-dimensionality solid-state NMR spectra of heterogeneous proteins via chemical-shift patterns. Left: PACSY data base-derived reconstruction of conformational ensembles, right: prediction-based assessment of backbone dihedral-angle distributions

of a degree of conformational heterogeneity, we suggest a series of scores, which we apply to the outcomes of both reconstruction frameworks.

Materials and methods

Uniformly labelled (^{13}C, ^{15} N)-GGAGG pentapeptide was purchased from Thermo Fischer. An inhomogeneous sample was prepared by dissolving the peptide in 1 mL of ddH$_2$O, flash-freezing in liquid nitrogen, and drying in a vacuum chamber at 0.01 bar. Spectra of GGAGG were recorded on a Bruker Avance 800 MHz NMR spectrometer in a 1.3 mm MAS rotor. The rotor was filled by overnight centrifugation of the sample and spun at 40 kHz at a temperature of 10 °C. A 4D hCBCANH spectrum (Xiang et al. 2016) of GGAGG was acquired non-uniformly with 5% sampling density (19,208 points). A Poisson-Gap schedule and hmsIST as the reconstruction method were chosen based on previous work (Burakova et al. 2020). The DREAM (Verel et al. 2001) scheme was used to achieve C$^{\alpha}$–C$^{\beta}$ magnetization transfer. Digital resolution in the indirect dimensions was 127.7 Hz for nitrogen and 251.0 Hz in both carbon dimensions. The spectrum was recorded in seven blocks of eight scans to allow for manual field correction in between. Each block was recorded for ca. 2.5 days. The 2D correlation spectra were referenced to an external DSS standard as described in Aeschbacher et al. (2012). The 4D hCBCANH was indirectly referenced by superimposing the 2D projections with the 2D DREAM correlation. For details on acquisition and processing parameters see Table S1. Apodization had no significant effect on the GGAGG peak shapes and widths in comparison to the inhomogeneous broadening (see Fig. S1). Spectral processing was done using NMRPipe software (Delaglio et al. 1995).

The CSV-formatted PACSY database (version from Dec, 28 2020) (Lee et al. 2012), cleansed by methodology presented in Fritzsching et al. (2016), was analyzed and visualized using Python and Python-based packages including NumPy (Harris et al. 2020), NMRglue (Helmus and Jaroniec 2013), Pandas (McKinney et al. 2010) and others (Hunter 2007; Waskom et al. 2017; Fundamental Algorithms for Scientific Computing in Python et al. 2020).

Results

Model heterogeneous sample

We developed our approach on a short model heterogeneous sample of u-(^{13}C, ^{15} N)-GGAGG pentapeptide, checked for purity by mass spectrometry and analytical HPLC. We introduced a permanent conformational disorder by first flash-freezing in water and then freeze-drying to 10 mbar, resulting in a glass. Being minimally restrained by steric properties, the G-A-G peptide bonds in this sample can be assumed to represent an extreme example of dihedral-angle variability. In this sample, peak broadening can be assumed to derive almost exclusively from conformational heterogeneity, as long-range modulation of the chemical shift by ring-current effects were specifically avoided. Contributions from intermolecular contacts to the heterogeneous line shape, beyond the backbone dihedral angles of interest, can be assumed to be largely limited to the H and N shifts (due to differential H-bonding interactions). These are an acceptable compromise for developing the demonstrated algorithms, as in future applications in frozen solutions these would be severely reduced and because carbon shifts are less susceptible to intermolecular contacts. (See a discussion of this limitation below.)

As opposed to the previous works, we intended to utilize correlated chemical-shift data from as many nuclei as possible, expanding cross-polarization-based NMR experiments to four dimensions, which was put into practice via a 4D hCBCANH correlation. (See *Materials and Methods* and Table S1 for experimental details.) When applied to more complex proteins in the focus of biological questions in future studies, the 4D experiments will increase dispersion of the signals without individual site-specific labeling. Dispersion is a main bottleneck for the residue-resolved assessment of sample heterogeneity in case of severe peak broadening when many residues bear isotope labels. As expected, whereas the anti-correlated chemical-shift distributions of C$^{\alpha}$ and C$^{\beta}$ (see carbon–carbon correlation in Fig. 2B) resemble what would be expected for a distribution of different dihedral angles based on the statistical data (Lee et al. 2012), the H and N distributions can be assumed to be more strongly influenced by homogeneous line broadening and heterogeneous contributions independent of dihedral angles. Therefore, in this preparation, these dimensions predominantly represent the purpose of chemical-shift dispersion. In DNP assessment of flash-frozen preparations (where intermolecular protein contacts are avoided due to an excess of solvent), however, H and N shifts may serve as a more faithful reporter on secondary structure as well.

Conformational analysis based on predictions of TALOS-N

TALOS-N (Shen and Bax 2013) is a recent and widely-used program for predicting protein backbone dihedral angles from successive NMR chemical shifts that relies on an artificial neural network. The neural network ((φ, ψ)-ANN) is derived from proteins for which crystal structures are available together with their nearly complete backbone chemical shifts. The task of identifying the most likely (φ, ψ)

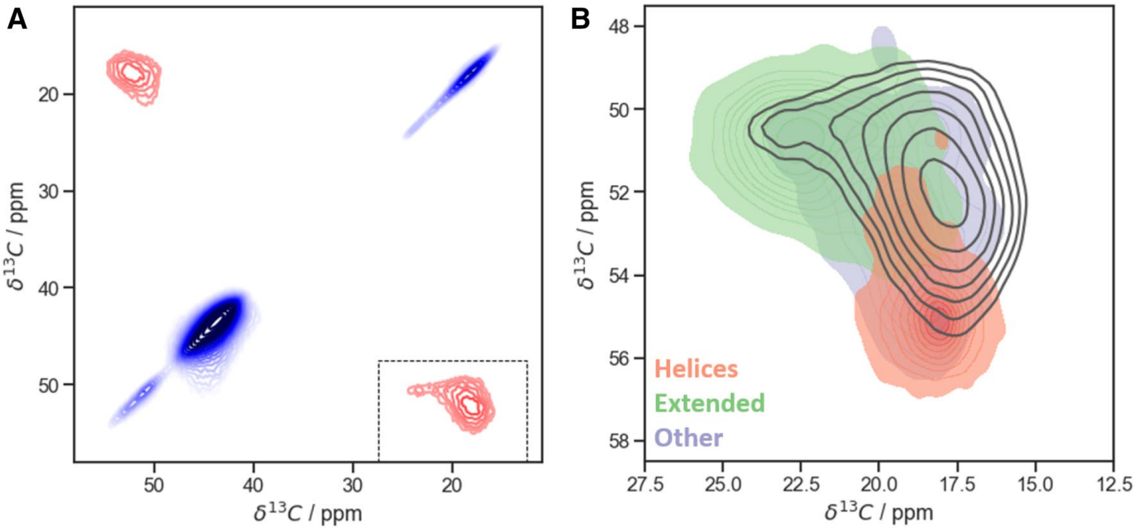

Fig. 2 ^{13}C–^{13}C 2D DREAM correlation of an inhomogeneous sample of GGAGG pentapeptide after freeze drying. **A** Full spectrum (line broadening coefficient LB = 20 Hz). **B** Overlay of $^{13}C^{\alpha}/^{13}C^{\beta}$ Ala cross-peak (black contours, with exponential line broadening of 150 Hz) with expected chemical-shift regions adopted by different kinds of secondary structure. These entries are color-coded by their secondary-structure class according to the STRIDE classification (Frishman and Argos 1995) with simplification: class "helices" includes alpha-, 3–10 and Pi-helices (H, G and I); "extended" class includes entries classified as E; other structures include the remaining T, B and b classes. Contours start from 4% of absolute intensity and increase with a factor of 1.2. Compare Fig. S4 for generation of secondary-structural color shades. Random-coil chemical shifts result from fast averaging of different conformations in solution and have been omitted here

combination for a given set of isotropic chemical shifts is closely related to the analysis desired here, where individual elements of the heterogeneously broadened peak need to be analyzed individually and an integrated distribution of angles be produced. In TALOS, the initial prediction of (φ, ψ) angle probability distribution is done with (φ, ψ)-ANN based on the chemical shift input for five consecutive residues. Then, the algorithm selects those 25 heptamer fragments with the best matching geometry and chemical shifts of the central residue to classify the result by quality and, if possible, find the most likely combination of (φ, ψ) angles.

Given the explicit expectation of non-standard secondary structural properties, our TALOS-N based approach to static conformational disorder (see Fig. 2, right) utilizes only the 18 × 18 grid (φ, ψ) distribution obtained from (φ, ψ)-ANN (324 φ/ψ combinations). No other output data (the most likely (φ, ψ) combination, secondary structure propensity, prediction of side-chain rotamers) is used. To analyze the entire, heterogeneously broadened peak, the volume occupied by signal intensity above a given threshold (here: 20% of peak maximum intensity, signal-to-noise ratio of 15) was probed for intensity at discrete grid points of the chemical-shift space obtained via regular sampling intervals. Sampling resolution needs to be sufficient to differentiate the areas expected for different secondary structural properties (compare Fig. 3A). For the alanine cross-peak in GGAGG we used a spacing of 0.4, 1.0, 1.5, and 1.5 ppm in the $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and ^{15}N dimensions, respectively, corresponding to 1407 (4 + 1)D grid points (four frequency dimensions plus one dimension for intensity). The expected homogeneous linewidths (at 700–750 MHz and 40 kHz MAS) are on the order of 360, 20, and 80 Hz (0.5, 0.3, 0.5 ppm) in ^{1}H, ^{15}N, and ^{13}C dimensions, respectively (Zhou et al. 2007; Linser et al. 2011). The inhomogeneously broadened lines of this sample (Fig. 3A), on the other hand, cover a shift range of several (~ 7–8) ppm in each dimension, which is in line with the expected large difference between helical and extended conformations present. As such, both the homogeneous contributions to the linewidth as well as the spacing of grid points are here sufficiently narrow in comparison to the extent of heterogeneous contributions but may have to be tightened in samples/residues with lesser extent of conformational heterogeneity.

In TALOS-N, secondary-structural assessment is strongly improved by including additional residues before and after the residue of interest (ROI). However, in the analysis of heterogeneous peak shapes, it is close to impossible to decipher which individual peak sections of the neighbor peaks are connected with which peak elements of the ROI, in the sense that they stem from the same molecule (or at least similar conformations). (Additional inter-residual dimensions, in addition to the multiple intra-residual ones, as well as added magnetization transfer steps would be required, which is practically challenging.) Simplifying the chemical shifts of neighbor spins to one value (e. g., the global peak maximum position), on the other hand, would strongly bias

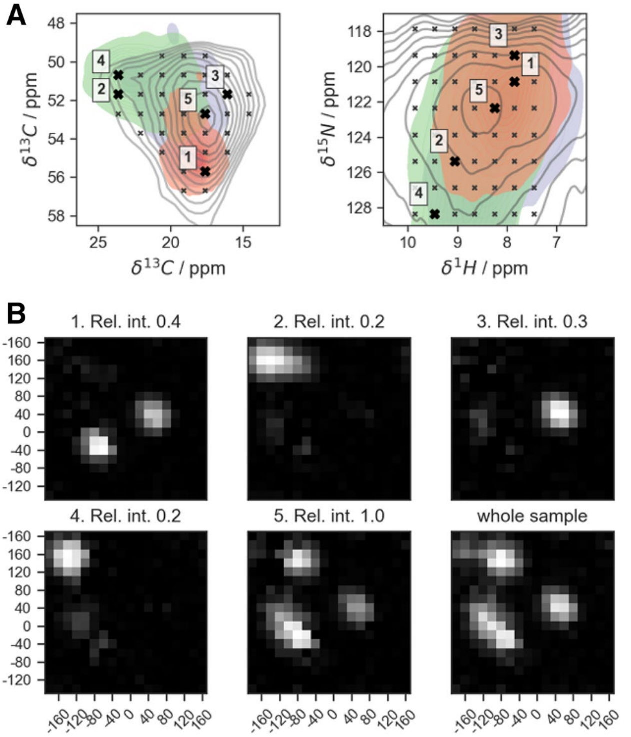

Fig. 3 Process and results of TALOS analysis of the Ala cross peak in an hCBCANH spectrum of freeze-dried GGAGG. **A** Grid points within the peak (black crosses) and the selection of five test cases ("Points 1–5", bold black crosses) over the orthogonal projections of the 4D peak (Cα/Cβ and H/N projection left and right, respectively) colored according to the chemical shifts expected for alanines (as occupied by 96% of the database entries, belonging to helical (red), extended structures (green), or other groups (purple)). Random coil ("C" class) entries have been left out—see caption to Fig. 2 and compare to Fig. S4. **B** TALOS predictions of the grid points defined in A (Points 1–5) for the extendend sequence, (G)GGAGG(G) (see main text), as well as the "whole sample". The latter is the summed-up φ/ψ distribution map of all the grid points of the peak shape, representing the overall conformational ensemble of alanine in the inhomogeneous (G)GGAGG(G) sample

the analysis of the ROI to the most populated conformation. Therefore, we introduced a step of translating the ROI propensity associated with each given shift combination (grid point) to neighbor residue shifts (see Fig. 2) before subjecting the shift combination to ANN. For this purpose, the five-dimensional vector $\mathbf{r_i}$ between each grid point i and the neighbor-corrected random-coil chemical shift was determined for the ROI, appropriately rescaled, and added to the neighbor's ($j \pm 1$ and $j \pm 2$) neighbor-corrected random-coil chemical shift values (Tamiola et al. 2010). Rescaling of $\mathbf{r_i}$ was done in nucleus- and residue-type-specific fashion that reflects the differential strength of shift modulation by secondary structure (Fig. S2), compared to the random-coil shifts, as determined using the conformation-specific centers of gravity for each nucleus and residue type (Fritzsching et al. 2016). For test purposes, we also artificially extended and translated the peptide chain by two glycine residues $j \pm 3$ at the termini before prediction, as the TALOS-N database search involves up to 7-mers to make a prediction. Fig. S3A and B show a comparison of TALOS-N predictions for the true (GGAGG) and extended sequence (GGGAGGG) for the same test coordinates. However, the differences were found to be negligible. The analysis in the following was anyways done on predictions for the extended sequence. Assessment of 1407 points by TALOS took about 5 h (running 3 subsets of points in parallel on 10 Intel® Core™ i7-8700 CPUs at 3.20 GHz, 64 bit).

The grid of individual samples from the heterogeneous 4D peak is shown (as a 2D projection) in Fig. 3A, overlaid there with the regions expected for different types of different secondary structure. (See Fig. S4 for the generation of such secondary-structural regions.) The prediction results for five exemplary grid points selected to represent different contributions to the volume of the overall heterogeneous peak, i.e., two samples from the helical region (Points 1 and 3), two samples from the strand-like region (Points 2 and 4), and the point of the maximum overall peak intensity (Point 5), are shown in Fig. 3B, panels 1–5. Final reconstruction of the conformational distribution represented by the *overall* heterogeneous peak D_k (the probability density at each of the 324 φ, ψ angle combinations k) is achieved by summing up the 1407 individual probability maps D_{ki}, weighted by the experimental intensity at each grid point I_i in the experimental 4D spectrum (Fig. 3B, last panel, "whole sample"):

$$D_k = \sum_{i=1}^{N} D_{ki} I_i \tag{1}$$

with D_k—probability density for each φ/ψ combination k on the Ramachandran map; I_i is the NMR intensity at the position i from the 4D peak volume ("grid point"), and N is the number of grid points covered by the peak (in this case, N = 1407). Note that we will use the variable D_k for the height (probability) at a given point (k) in Ramachandran space irrespective of what the respective map looks like in detail.

As expected from the shift distribution of GGAGG in comparison with neighbor-corrected chemical shifts resulting for different secondary structures (Fig. 2B), for this extreme case of static disorder, the individual predictions sample almost the whole allowed Ramachandran space (Fig. 3). Points 2 and 4 correctly represent extended conformation with $\varphi = -155$ and $\psi = 140$, with some uncertainty for Point 2. Due to low relative intensity, these two points make only a minor contribution to the final result. Point 1, situated in the helical region of chemical shift distribution, correctly yields clear helical predictions. Interestingly, however, a mix of left- and right-handed helices is obtained.

For Point 3, which is located in one of the turn regions, a left-handed helix is confidently predicted (or, according to classification in Hutchinson and Thornton (1994), a type I’ turn). In the case of an alanine surrounded by glycines—residues without any C^{β}—both senses of winding are indeed possible and—given the enantiomeric character of the left-handed GGAGG helix relative to the right-handed one—would lead to similar chemical shift. Notably, overall, this sample contains more left-handed helical propensity than the conventional α-helical, according to TALOS-N (see panel “whole sample”). However, this is a special exception for this sample. (Accordingly, the PACSY database does not differentiate well between right- and left-handed helix, see below). Fig. S3 compares predictions for GGAGG with predictions at the same three test coordinates but for LLALL and LLLALLL as input and adjusting the shift translation by taking the Leu (instead of Gly) random coil values into account. In principle, this should give identical results since chemical shifts are equivalently translated, with the difference that the result is now in a purely *right-handed* helical conformation. This is indeed the case. Hence, in further analysis and for the general case, we omitted the sense of winding and considered *helical* contributions in a general sense. As such, the two different senses of helical properties derived from the shift combination at Points 1 and 3 were merged into single *helical* predictions by reflecting the right half of the resulting Ramachandran map onto the left one (point reflection about the 0, 0 coordinate, referred to as “folding” in the following, Fig. 4A).

Eventually, Point 5 yields a broader distribution of angles. It is associated with “dynamic” properties by TALOS-N and denotes overlapping contributions from turn ($\varphi = -80$, $\psi = 150$) and any of the helical conformations (the region around ($\varphi = -90$, $\psi = -40$) for alpha- and 3–10 helices and ($\varphi = 80$, $\psi = 10$) for the left-handed helix). For completely rigidified samples, averaging of chemical shifts to pure random-coil values does not occur. Fig. S5A shows the entries from the PACSY data base, with the coil entries removed, colored according to dihedral-angle combinations (rather than the STRIDE system). Most of the turn conformations, with dihedral-angle combinations diverging from helical or extended conformations, are located in the central region of the chemical-shift space, similar to the coil shifts in solution. Whereas, within the turn conformations, the most extreme helical and sheet shifts are not adopted and certain trends are still obvious, the association between shift and angle combinations is much less clear than between the major classes of secondary structure (compare Figs S5C, E, and F, as well as Figs. S6 and S7). Hence, even though the occurrence of “neither-E-nor-H” cases in the data base is much lower than the more faithfully predictable E and H shifts, the ambiguity for predictions of intermediate shift combinations constitutes a well-known shortcoming of shift-based secondary structure prediction. Equally importantly, every point in frequency space can be thought of as reflecting contributions from individual molecules with their specific backbone dihedral angles. However, intensity in central positions will always occur also as an artifact from overlap of individual peak “shoulders”, to an extent dependent on the level of homogeneous contributions to the linewidth and limited digital resolution. Hence, in addition to the turn residues resonating with exactly these shift combinations (Fig. S5C and F), shoulders from the more helical and extended conformations (Fig. S5B and D) strongly blend into the contributions from the “turn” dihedral angles in this area. In this respect, the inhomogeneous prediction results for central shifts are not entirely wrong, as indeed they agree with the presence of a mixture. Hence, whereas the “mixed” prediction outcome for central shift combinations may have distortive character (when in reality a narrow distribution around one or more of the turn torsion angle combinations would be correct), a reasonably representative outcome can be expected for mixtures mainly comprised of different populations of the more prototypical (helical/extended) conformations. (In these cases, if not absent, central chemical shifts are derived from homogeneous line broadening.)

Quantification of heterogeneity

Since there is no standard for the quantitative level of site-specific sample heterogeneity yet, here we explored different approaches to possibly represent the degree of heterogeneity in a fast and at least qualitatively reliable fashion. For this purpose, we considered eight distributions that represent a variety of possible scenarios: In addition to the "Points 1–5″ and the overall inhomogeneous peak shape (“whole sample”) considered above, two additional points that represent the “purest” homogeneous cases (expected helical and strand chemical shifts) are included in the comparison for convenience. (TALOS predictions for these pure helical and extended-structure cases were generated for the 6th residue of a Leu 10-mer sequence with chemical shifts being set to the expected values of either helix or strand (Fritzsching et al. 2016).) The eight scenarios are ordered tentatively from pure to mixed conformational content; however, it is clear that different measures (see approaches of quantification below) would be sensitive to specific features of the distribution. It may be useful to stress that, as a notion, the *degree of heterogeneity* makes sense only for the integral or “summed up” Ramachandran maps addressing the *whole* volume of a solid-state NMR peak (panel “whole sample” in Fig. 3B and the corresponding panel in Fig. 4A) and shall not be applied to the predictions from individual grid points of a heterogeneous peak (which were described in the previous section). In this section, we use the individual

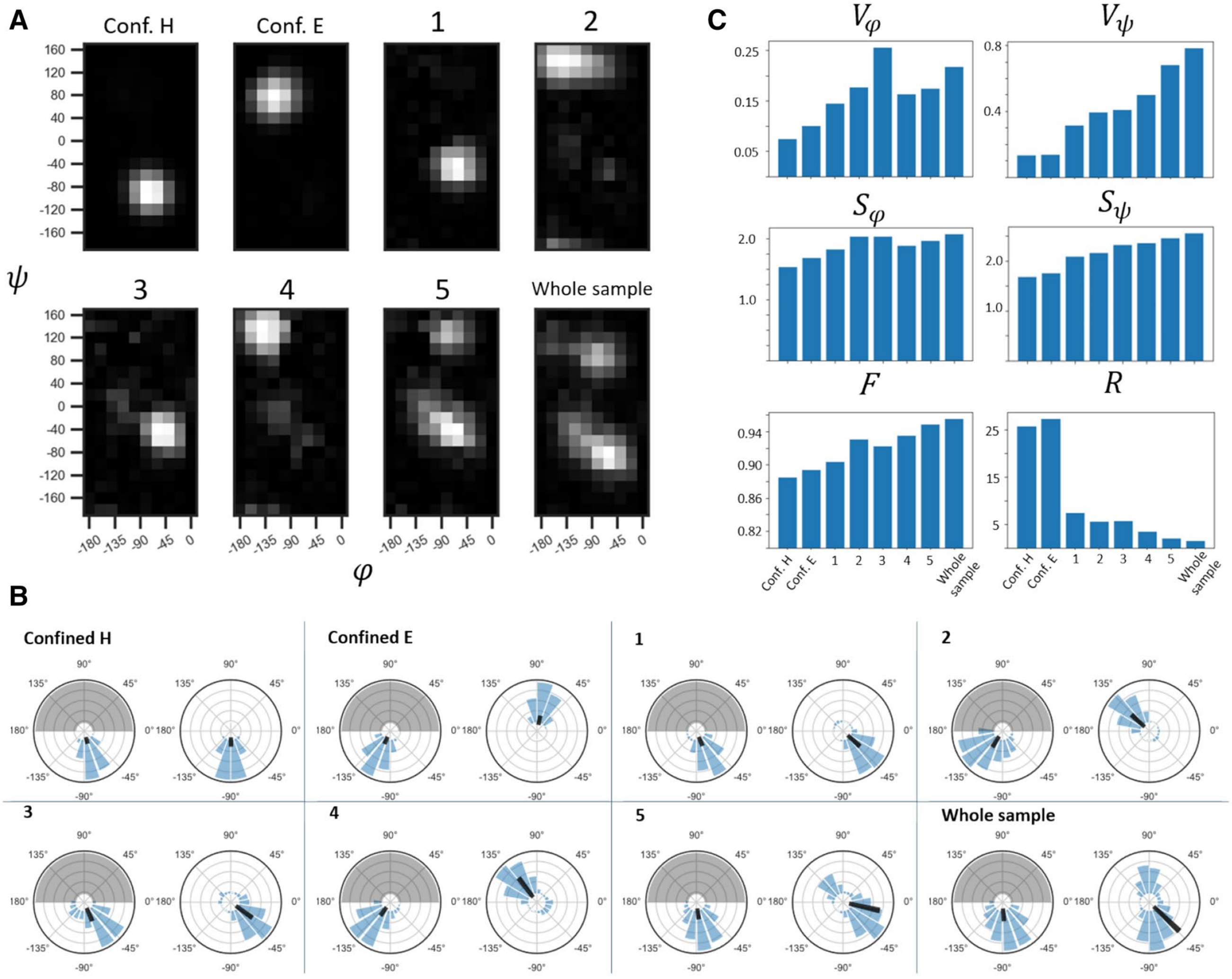

Fig. 4 Exploration of various methods of quantification of heterogeneity in solid-state NMR samples, applied to TALOS-based reconstruction of conformational distributions. **A** Folded Ramachandran maps of the test coordinates. Panel "Whole sample" corresponds to the weighted sum of predictions over the whole Ala peak of heterogeneous GGAGG. For generation of pure secondary structure, predictions were made of the 5th Leu in a Leu$_{10}$ chain with the corresponding expected chemical-shifts values (taken from (Fritzsching et al. 2016)). Grayscale is normalized from 0 (black) to 1 (white, maximum value). **B** Ramachandran maps from A) in polar coordinates. In each pair, the left plot corresponds to φ and the right one to ψ distributions. Gray area denotes the non-valid φ region for the calculations due to folding (see main text for details). Black vectors point into the *mean direction*, their length is set here to represent the circular variance, not the length of the resulting vector for the distribution. **C** Representation of different measures of heterogeneity (circular variance *V*, entropy *S*, flatness *F*, and secondary-structure ratio *R*) for the maps shown in A) as bar plots. See text for details

Ramachandran maps for demonstration purposes only, as if they were obtained for separate, homogeneous peaks, since all calculations remain exactly the same.

In order to characterise (φ, ψ) distributions quantitatively, at first, methods of circular statistics were applied. Each pixel in the Ramachandran map represents the (weighted) incidence of a specific angle combination. When generating a circular average over an angular distribution, one imagines an averaging of vectors in 2D space with an individual direction (given by the angle) and length (its weight within the ensemble). The (φ, ψ) distribution of the heterogeneous peak comprises of 180 vectors $\vec{v}_k$ (one for each (φ, ψ) combination *k*) with direction given by the pixel's angles φ and ψ and magnitude given by the pixel's relative intensity $\frac{D_k}{\sum_{k=1}^{180} D_k}$. Knowing the magnitude of an individual vector $|\vec{v}|_k$ and its angle(s), it can be dissected into vector components $|\overrightarrow{v}|\sin\theta$ and $|\overrightarrow{v}|\cos\theta$ using trigonometric relations (θ being ϕ or ψ angles). The average vector $\vec{v}$ and its magnitude, $|\vec{v}|$, can then be constructed via Pythagoras from the average vector components. In the case of only similar angles being populated in Ramachandran space, the average vector $\vec{v}$ is long (in the extreme case has a magnitude $|\vec{v}|$ of 1, as the sum of relative intensities is 1), whereas with a large variety of angles being populated, the vector sum or average has a short length (in the extreme case 0). The *circular variance V* (with higher values for broader distributions and vice versa) is reflected by $1-|\vec{v}|$:

$$V_\theta = 1 - |\vec{v}|_\theta = 1 - \sqrt{\left(\overline{|\vec{v}|_k \sin\theta_k}\right)^2 + \left(\overline{|\vec{v}|_k \cos\theta_k}\right)^2}$$
$$= 1 - \sqrt{\left(\frac{\sum_{k=1}^{180} sin\theta_k \cdot D_k}{\sum_{k=1}^{180} D_k}\right)^2 + \left(\frac{\sum_{k=1}^{180} cos\theta_k \cdot D_k}{\sum_{k=1}^{180} D_k}\right)^2} \quad (2)$$

V_θ thus ranges from zero to one, where lower values correspond to concise distributions. Note that in our approach, the distibutions have a period of only 180° (π) for the φ dimension due to the above folding of left-handed into right-winded structures (see section *Conformational analysis based on predictions of TALOS-N*). The Ramachandran plots (Fig. 4A) can be visualized using polar coordinates (Fig. 4B), in which the distribution of vectors (blue histograms) and the average property (black) are visualized. Note that the magnitude of the black arrow was chosen to reflect the variance ($1-|\vec{v}|$), not the resulting vector's magnitude. The rising trend of greater "inhomogeneity" in the individual panels (conf. H, conf. E, Points 1–5), which is apparent from both V_ϕ and V_ψ, shows the uncertainty within the prediction of individual chemical shift combinations: In contrast to the clear (easy-to-predict) shift combinations, an increasingly large *V* is found for Points 3–5, i. e., when shifts do not adhere to the standard values expected for helical or extended structures. This is consistent with the above observation that shifts in central regions are inherently associated with a broader (φ, ψ) distribution on their own.

Alternatively, the level of heterogeneity contained in broad (φ, ψ) angle distributions can be measured by *Shannon's entropy*. In statistics and information theory, the concept of entropy is widely used to quantify the amount of uncertainty in a given distribution of a random variable. Considering each φ/ψ combination *k* of the Ramachandran map as an independent state of an amino acid residue, with its intensity D_k representing its likelihood to be true/adopted, the entropy of a prediction would be calculated as follows:

$$S_\phi = -\sum_{k=1}^{10} D_k ln(D_k) \quad (3.1)$$

$$S_\psi = -\sum_{k=1}^{18} D_k ln(D_k) \quad (3.2)$$

$$S_{total} = -\sum_{k=1}^{180} D_k ln(D_k) \quad (3.3)$$

Entropy of a hypothetical case where only one state is populated equals zero; by contrast, it increases up to $S = \ln(180) \approx 5.19$ for the hypothetical case of a uniform distribution.* If the angle-specific contributions to the entropy are of interest, they can be calculated via projection of the Ramachandran probability map onto the individual axes and applying the above routine (then 0 ≤ k ≤ 10 for φ;$S_\varphi^{max} = \ln(10) \approx 2.3$ and 0 ≤ k ≤ 18 for ψ; $S_\psi^{max} = \ln(18) \approx 2.9$). When describing the full heterogeneous peak, the Ramachandran map D_k refers to the result from weigthed averaging over the heterogeneous chemical-shift pattern (see Eq. 1). For the heterogeneous GGAGG sample of this study, the total entropy S_{total} is 4.46, whereas entropy values of individual (one-dimensional) φ and ψ distributions amount to 2.07 and 2.56, respectively. Note that in Eq. 3.3, D_k (the probability for the φ/ψ *combination k* in the Ramachandran map) applies to the *folded* map with $k = \{1, \ldots, 180\}$. For single-angle entropies (Eq. 3.1 and 3.2, *k* bearing 10 or 18 values for φ and ψ, respectively), D_k refers to probabilities for *individual* φ or ψ values in one-dimensional Ramachandran maps. (Such projections are obtained by adding all those D_k values that are within the same column or row, respectively). It may be useful to correct for the level of ambiguity of the prediction for a well-defined event, which *excess entropy* results from simple subtraction of the entropy for a confined helix: $\Delta S = S - S^{conf.H}$. For the heterogeneous GGAGG sample, the overall ΔS amounts to 1.30, the highest-possible (but sterically challenging) value would be 2.02.

A simple approach to probe the level of homogeneity found in a distribution is the measure of *flatness*, which gives the relative abundance of the highest-probability event (normalized by the sum of overall occurrence of different events of the prediction):

$$F = \frac{\max(D_k)}{\sum_{k=0}^{180} (D_k)} \quad (4)$$

By definition, it is insensitive to the number of modes and rather characterizes how confined the distribution is overall (Fig. 4C). In addition, it may be interesting to consider the *ratio between the population of helical and extended regions* (*R*), as determined from the integral over relative densities in the typical areas of the Ramachandran plot.

$$R = \begin{cases} H/E \ \ if \ \ H > E; \\ E/H \ \ if \ \ H \leq E \end{cases} \quad (5)$$

where H and E are the integrals of the allowed regions in the folded φ / ψ maps. (Regions taken into account for H and E are depicted in Fig. S8.) Tables 1 and 2 and Figs. 4 and 5

Table 1 Heterogeneity parameters obtained for the folded Ramachandran maps predicted by TALOS-N for the local test scenarios as well as for the broad heterogeneous peak. Shown are the two reference cases (H and E), five individual coordinates from the Ala HNCACB peak ("Points 1–5"), and the cumulative peak volume in the GGAGG sample (bold, see Fig. 4A). The underlining in the sec. structure column denotes the excess of helical content

Scenario	Sec. struct	Circular variance *V*		Entropy *S*				Flatness *F*	*R*
		ϕ	ψ	ϕ	ψ	total	ΔS_{total}		
Conf. H	H	0.07	0.13	1.54	1.68	3.17	0.00	0.886	25.89
Conf. E	E	0.1	0.14	1.68	1.76	3.35	0.18	0.894	27.38
Point 1	H	0.15	0.32	1.82	2.10	3.68	0.51	0.904	7.53
Point 2	E	0.18	0.40	2.04	2.17	4.08	0.91	0.943	7.31
Point 3	H	0.26	0.41	2.04	2.33	3.99	0.82	0.923	5.80
Point 4	E	0.16	0.5	1.88	2.37	3.99	0.82	0.935	3.55
Point 5	H+E	0.18	0.68	1.96	2.47	4.24	1.07	0.949	1.99
Whole sample	**H+E**	**0.22**	**0.78**	**2.07**	**2.56**	**4.46**	**1.30**	**0.955**	**1.53**

Table 2 Quantitative analysis of Ramachandran maps obtained using the PACSY approach, focusing on chemical-shift combinations of confined helix and sheet, Points 1–5, and the full heterogeneous GGAGG peak (bold). Since for the PACSY approach in clean cases no population of incorrect secondary structure is produced, the *R* values tend to be infinity (division by 0) or very high, which hence represents a clean prediction. *N* stands for the number of PACS entries at the respective chemical-shift grid point. The underlining in the sec. structure column denotes the excess of helical content

Scenario	Sec. struct	N	Circular variance V		Entropy *S*				Flatness *F*	*R*
			ϕ	ψ	ϕ	ψ	Total	ΔS_{total}		
Conf. H	H	2030	0.04	0.06	0.96	1.16	2.02	0.00	0.666	115.9
Conf. E	E	105	0.12	0.08	1.70	1.55	3.04	1.02	0.911	inf
Point 1	H	1303	0.02	0.04	0.88	0.99	1.83	− 0.19	0.645	419.6
Point 2	E	60	0.09	0.09	1.51	1.45	2.61	0.59	0.786	inf
Point 3	H	6	0.28	0.22	1.36	1.09	1.36	− 0.66	0.671	0.72
Point 4	E	20	0.06	0.07	1.31	1.0.37	2.26	0.24	0.790	inf
Point 5	H+E	422	0.17	0.74	1.64	2.45	3.71	1.70	0.860	1.54
Whole sample	**H+E**	**13,565**	**0.17**	**0.80**	**1.78**	**2.36**	**3.94**	**1.92**	**0.869**	**1.52**

show *R* for various cases. For the whole sample, *R* amounts to ~ 1.53, with a slight excess of helical properties.

Conformational analysis based on predictions of DANGLE

As an alternative approach to neural-network based tranformation of chemical-shift combinations into dihedral angle space, the DANGLE algorithm has been suggested (Cheung et al. 2010). DANGLE uses Bayesian inference-based methodology and was set up in 2009 for improving predictions over the TALOS-N predecessor TALOS + (Shen et al. 2009) at the time. We subjected the 1407 chemical-shift combinations covering the heterogeneous 4D peak derived above also to DANGLE, using exactly the same procedure as described in the framework of TALOS predictions. Quantitative evaluation of the predictions of the exemplary grid points described above (confined H, confined E, and Points 1–5) as well as the summed prediction for the entire heterogeneous peak was also done as described above (Fig. S9 and Table S3). A series of DANGLE test runs where—as an alternative—the secondary-structure propensity of the middle residue (Ala) was *not* propagated to its neighbors (all four Gly) yielded identical φ/ψ maps for all grid Points 1–5. With translation of secondary-structural propensity to neighbors (determining the deviation between the grid point shift combination to the random-coil shift combination and calculation of neighbor shift combinations from their random-coil shifts by adding the same difference as found for the ROI, as described above), however, results were obtained that are very similar to the TALOS approach. In particular, the expected angular properties are faithfully reproduced for confined H, confined E, and Points 1, 2, 4, and 5. A deviation is found only for grid point 3, which represents a rather sparsely populated area in chemical-shift space and also fails to yield faithful predictions in the approach based on direct data base correlations (see below). Generally, however, the individual predictions seem more discrete, and lower

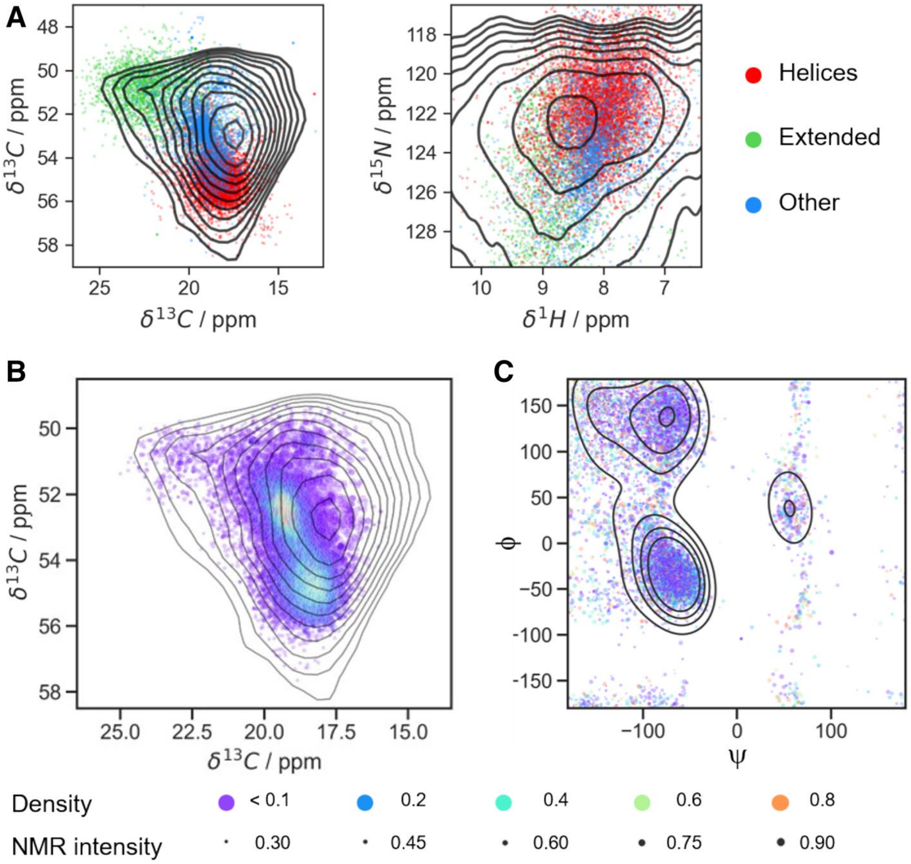

Fig. 5 Selection of points and results of PACSY-based assessment of conformational heterogeneity in the GGAGG sample. **A** Representations of entries from the cleansed PACSY database for Ala, color-coded by their secondary-structure class according to the STRIDE classification (Frishman and Argos 1995) with simplification: class "helices" includes alpha-, 3–10 and Pi-helices (H, G and I); the "extended" class includes entries classified as E; other structures include the remaining T, B and b classes; random coil entries were excluded from the presentation. Left: $^{13}C^{\alpha}$ / $^{13}C^{\beta}$ and right: $^{1}H/^{15}N$ projections of the data base and 4D hCBCANH spectrum. Contours start from 4% of absolute intensity and increase with a factor of 1.2. **B** Points from the data bank that belong to the 4D volume of the Ala cross-peak in the hCBCANH spectrum; Gray contours depict a bivariate weighted kernel density estimate: The weight of each point is a product of interpolated intensity of the 4D peak and the inverse point density in the PACSY database in the 4D chemical shift space. (Colors, as denoted in the bottom, refer to original (non-inverted) data bank entry density.) Contours start at 15% density and succeed with a factor of 1.1. **C** The same entries plotted in φ/ψ space

variability in the prediction results from individual chemical-shift grid points is found. As a consequence, whereas the trends within the data set are qualitatively consistent with the expectation and with the TALOS-N approach, the exact values found in the various heterogeneity scores of the overall peak are generally lower. In contrast to the prediction by TALOS, where statistics are generally smoother, many zero values are obtained, and added care has to be taken to not overinterpret the resulting quantitative scores.

Conformational analysis driven by database search

A rather different approach to reconstructing the ensemble of conformations from a heterogeneously broadened peak shape is the utilization of a database that directly associates backbone dihedral angles from PDB structures with chemical shifts. We used PACSY (Lee et al. 2012), a relational database that contains over 6000 protein chains to allow this correlation directly from individual entries. For this purpose, we constructed the following workflow (see Fig. 2, left). All data are taken from the PACSY table X_CS_DB2 (where X stands for one-letter residue code, here X = A), which relates chemical shift, dihedral angles, secondary structure classification (according to STRIDE algorithm (Frishman and Argos 1995)), and other information. Residues that belong to the proteins marked as not passed PIQC were excluded from further analysis. All remaining entries within the populated 4D area of chemical shifts (the peak envelope) were used to reconstruct the underlying conformational ensemble. In order

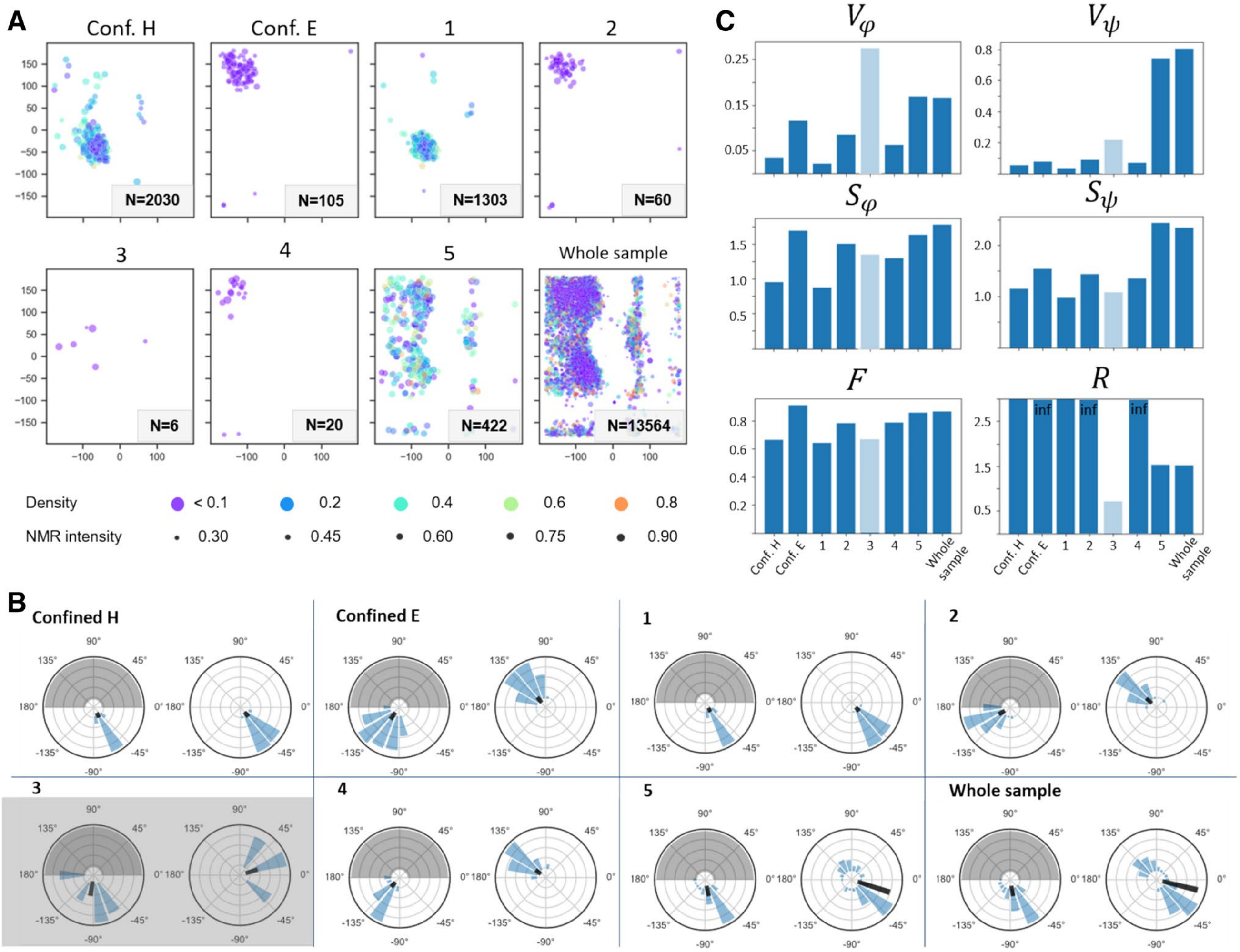

Fig. 6 Distributions of dihedral angles for the entries of PACSY database selected by their chemical shifts to represent the scenarios considered in TALOS-based approach and their parameters. **A** Sets of entries in φ/ψ space. Width of each interval is 0.8 ppm for ^{1}H, 2 ppm for ^{13}C$^\alpha$, and 3 ppm for ^{13}C$^\beta$ and ^{15} N dimensions. Selection for the entire heterogeneous peak as described in the main text. **B** Same distributions of PACSY entries, folded and represented in polar coordinates. In each pair, the left and right plot represent the individual φ and ψ distribution, respectively. Black bars represent circular variance and point to the mean direction. Entries around Point 3 are grayed out because they comprise too few points (N = 6). **C** Heterogeneity parameters for each set: circular variance *V*, entropy *S*, flatness *F*, and secondary-structure ratio *R* (see main text for the formulae). Results for the set around Point 3 are again bleached out due to too few database entries

to compensate for the random, shift-combination-specific sparsity of data bank entries and avoid the resulting bias that varying entry densities would have for reconstructing the ensemble, the elements were weighted in height by the inverted local 4D entry density. Then, to also reflect the actual peak shape (the experimental distribution of shift contributions), final weighting of the *i*th entry amounts to $w_i = I_i \bullet P_i^{-1}$, where I_i is the peak intensity and P_i the density of data bank entries at the respective chemical-shift positions. The selected points are shown in Fig. 5 with colors of the points corresponding to the original entry density and sizes corresponding to the relative intensity of the 4D peak; Fig. S10 shows the density map without normalization by point density for comparison. The reconstructed conformational ensemble is represented by the weighted entries depicted in the φ/ψ map (Fig. 5C). As an experiment, we calculated all mathematical scores for the extent of heterogeneity presented in the section *Quantification of heterogeneity* also for the PACSY approach. In order to represent test Points 1–5 (compare respective panels in Fig. 4A) we selected entries belonging to an interval centered at each of these chemical shift combinations (Fig. 6A). The interval was chosen equal to twice the TALOS grid resolution to represent a larger number of data bank entries (i. e., 0.8 ppm in ^{1}H, 2 ppm in ^{13}C$^{\alpha}$, and 3 ppm in both, ^{13}C$^{\beta}$ and ^{15} N). Test Point 3 (Fig. 3A) appeared so far away from the main clusters of PACSY entries that it included only 6 points and was excluded from further consideration; all the results are, however, still presented here for consistency (Fig. 6). "Clean" cases of confined helix and extended structure were represented here with a selection belonging to the box centered around the mode for chemical shifts of, correspondingly, H- and E-classified alanines. Finally, the selection for the entire heterogeneous peak ("whole sample") was made as described above.

Heterogeneity scores were calculated for the weighted and folded 10 × 18 maps converted to the format of the TALOS-based approach (Fig. S11). As in Fig. 4C, the bar charts in Fig. 6C (center row) only display absolute entropies, whereas in Table 2 we also list excess entropy for the entire Ramachandran map, which is the difference of individual scenarios from the case of a clean helical shift taken again as a reference.

Compared to the TALOS-based procedure, the obtained results for the heterogenized GGAGG sample are in fact similar. Each of the individual grid points yields a prediction qualitatively consistent between the methods as well as in line with what is expected from the known chemical-shift trends depicted in Fig. 3A. For the entire peak, the expected broad conformational distribution with the highest contribution for the angles corresponding to extended and helical conformation as well as a slight excess of helical conformations applies again. Even the quantitative assessment of heterogeneity across the full peak is highly consistent, which is gratifying to see given the various shortcomings for purely shift-based reconstruction mentioned throughout the manuscript. E.g., the parameter R, the ratio of populations associated with helical conformations and those associated with strand-like conformations, for the entire peak is 1.52, which is identical to the value obtained for TALOS-based maps (Table 2). The distribution of ϕ and ψ values adopted as measured by circular variance compares as 0.17 versus 0.22 and 0.80 versus 0.78, respectively. The excess entropy compares as 2.07 versus 1.78 and 2.56 versus 2.36 for ϕ and ψ, respectively. (Theoretically possible values for the 10 × 18 histogram lie between 0 and 2.3 for ϕ and between 0 and 2.9 for ψ.) Applying the scores for determining the "degree of heterogeneity" to the individual chemical-shift grid points—which is physically insensible—the trends are reasonably consistent and confirm the better predictability of more common vs. uncommon shifts. (Such test should in ideal cases yield low values and is expected to give more inhomogeneous predictions only for shift combinations outside the clean E or H conformations.) However, per pixel, only the predicted secondary structure is of interest in future applications, and the heterogeneity scores would be applied only to a whole heterogeneous peak. Eventually, future applications would always compare different residues of the same sample within the same methodology, and indeed these trends are fully in agreement with the expectation for all approaches tested.

Discussion

Even though different approaches have been proposed in the past, a detailed description of a conformational ensemble within a single sample has remained difficult to obtain. In previous studies, many of which are listed in the introduction, experimental means could often be used to disentangle disorder into individual samples for a reconstruction of their conformational properties. Alternatively, only few discrete conformations co-existed. In both cases, tailored spectroscopic methodology could be used to shed light on their site-specific properties. Limitations to such approaches occur in the case when the biological sample cannot be physically disentangled. The same applies when broad distributions within the data are expected that are not resolved via the chemical-shift space available. Ensemble properties from such heavily overlapping patterns have been reconstructed in other approaches from accordingly, *a-posteriori* reweighted conformational ensembles, obtained from MD in conjunction with shift prediction. Although this represents a very elegant approach, even with perfect performance of the available chemical-shift prediction tools, such data patterns can be underdetermined, leading to multiple (different) ensembles being in reasonable agreement. This is even more likely in case of peak overlap, which causes high-complexity distributional properties of multiple sites to be entangled in a single, lower-complexity pattern.

Using chemical shifts as a direct reporter of dihedral-angle properties is a rather straightforward and hence sensitive approach that dispenses encoding of angular features by dedicated pulse sequence elements. This, conversely, facilitates the addition of multiple chemical-shift dimensions for peak dispersion and voxel-specific interrogation. Higher-dimensionality chemical-shift correlations in the framework of proton detection in particular bear the prospect of creating sufficient space for peak dispersion, along with a high signal-to-noise ratio from small sample volumes. This largely facilitates a residue-specific assessment of disorder in solid preparations as a function of sequence that would be overlapped using lower dimensionality. The above data show that the correlation between shifts and dihedral angles is reasonably trustworthy for the extreme cases of secondary structure for carbon nuclei, which facilitates a faithful reconstruction of the secondary-structural distributions within a heterogeneously broadened but non-overlapped peak. In congruency with the longstanding shortcomings of shift-based dihedral-angle prediction for homogeneous samples, however, the data also show that reliable reconstruction of heterogeneity around "intermediate" angle combinations (turn structures) is compromised by the general uncertainty of chemical shifts adopted in these cases—even when other factors modulating the shift are ignored. A good prediction will hence be obtained when the relative amount of helix and sheet is to be determined, whereas a poor prediction may arise when a residue comprises a narrow angular distribution around an intermediate angle combination that is difficult to interpret.

We expect that, often, the level of accuracy obtained for individual ensemble members within this framework is sufficient to answer biological questions related to heterogeneous conformational distributions. Examples would be to residue-specifically quantify the ratio of extended to helical conformers within amyloid preparations, flash-frozen folding intermediates, or rigidified disordered parts within membrane proteins. In such cases, the approach enables analysis with relatively low effort and costs. In addition, analysis of all residues within one preparation without selective labeling also bears the advantage of avoiding differences between samples (Xiang et al. 2017), which can compromise a consistent analysis. The maximal length of the primary sequence that can be subjected to the approach with reasonable dispersion and measurement time depends on the degree of heterogeneity. This derives from the fact that wider peak shapes both, increase the probability of overlap and the signal to noise of the resultant data. Luckily, in most of the current studies on samples of biological interest by NMR, only part of the residues tend to be variable, which reduces the probably of overlap even for longer primary sequences.

The voxel-specific, higher-dimensionality methodology, in particular sample preparation/requirements and hardware involved, is substantially different from any of the carbon-detected approaches developed previously. As such, the results obtained here from the comparative implementation of prediction and direct data base approaches were only mutually validated by comparison with each other. Importantly, the introduced scores are designed as a measure for the *sequence-specific* degree of torsional variability over the ensemble, comparing different residues or samples within the same setting. Hence, self-consistency (i.e., the relative degree of disorder) within a single method is the most crucial property. A truly orthogonal (experimental) method would be desirable to benchmark or even just validate the findings from a non-NMR perspective. Unfortunately, however, methods that yield faithful distributions of conformations in mixtures are rare. In fact, various other biophysical techniques can principally be utilized to verify or support MD and NMR results. However, it is very difficult, if not impossible, to experimentally assess *site-specific* properties of *individual ensemble members* outside of NMR. Even when different conformations imprint themselves in the overall data, as for circular dichroism, powder diffraction, or FTIR, the obtained patterns are usually not sufficiently specific to reconstruct the underlying ensemble with residue resolution. The major alternatives to NMR as high-resolution structural-biology methods, cryo electron microscopy (cryoEM) and single-crystal X-ray diffraction, by contrast, may only deal with/quantify a limited degree of disorder (Nwanochie and Uversky 2019).

Despite the encouraging methodological results described above, in both approaches, however, the well-known systematic and general shortcoming of using chemical shifts for assessing torsional properties is the inherent sensitivity of the chemical shift to various physical effects other than backbone dihedral angles. A first factor is the influence of direct spin–spin interactions (homogeneous contributions) to the proton line. However, the role of protons will be largely restricted to further dispersing the peaks, given the rather loose association between their shifts and dihedral angles. The potential impact of differential contacts with the lattice represents a source of additional peak broadening potentially involving all nuclei. However, this drawback is expected again more strongly for those nuclei involved in H-bonds (H/N), whereas carbon shifts are more faithful reporters on angular properties. In particular, the *pair* of C^{α}/C^{β} shifts, probed in chemical-shift-based approaches usually as a shift *combination*, bears opposite secondary chemical-shift trends and is mostly influenced by dihedral angles. In fact, the populated shift space in the carbon/carbon plane, which reflects the anti-correlated trends of C^{α} and C^{β} for secondary structure both for the heterogeneous peak of this study as well as the data base entries (Figs. 2 and 6A), speaks against large additional contributions to the carbon pattern, both for our test sample as well as in general. This renders the chemical-shift-correlation approaches here (be it via prediction or data base matches) more resilient to other influences compared to the inhomogeneous chemical shift of a single dimension. Differential sidechain torsional angles, however, can be an added source of shift modulation. This effect is expected for fully rigidified, longer side chains (opposed to only the protein backbone or Ala residues) and is difficult to disentangle from secondary-structural modulation of the shift (Siemons et al. 2019). Lastly, on purpose, aromatic moieties, chemical or magnetic perturbations (e. g. differential oxidation states in cysteines/pseudo contact shifts etc.) were avoided for this low-molecular-weight peptide, as these can have longer-range effects on the chemical shifts of close-by nuclei (and in this case molecules). The C^{α} /C^{β} shift pair as the most informative/convergent source of information both for the prediction as well as for the data base approach will mostly be influenced in a similar way by nearby ring currents, such that the shift difference remains largely untouched. Nevertheless, residues in close vicinity of aromatics probably need to be treated with care. In this special sample, additionally, only a low level of residual solvent is likely present, which renders the H-bonding properties of the amides rather variable. For future samples with a high content of solvent, on the contrary, more consistent hydration properties of polar groups and larger spatial separation of individual molecules are expected. As such, the approach could turn out particularly useful for site-specific insights in solidified systems with lots of frozen water (e.g., from flash-freezing without lyophilization), looked at by DNP. At least as a qualitative measure for the degree of how defined conformational properties are within a given primary sequence, the described methodology should turn out helpful and—given the availability of all Python-based workflows as a download—easy to set up.

Conclusion

We have proposed a higher-dimensionality NMR approach to assess the site-specific conformational content in heterogeneous samples. Employing a single 4D hCBCANH spectrum, we enable heterogeneously broadened peak shapes in which the chemical shifts from individual conformers are correlated in the sense of a shift quadruple, from which the distribution of ϕ/ψ dihedral angles present for a given residue can be reconstructed. We demonstrate this reconstruction by two approaches, in a dihedral angle prediction-based and a data-base-derived manner, which—within the general limitations of shift-angle correlations—allow for reconstruction of the conformational distribution, in particular the ratio between sheet-like and helical conformers, of each residue that can be separated from other residues with four chemical-shift dimensions. As carbon shift combinations (i. e., the C^{α}/C^{β} shift difference) are comparably weakly affected by contributions other than secondary structure, the approaches should represent both, a reasonable measure for qualitative but self-consistent assessment of conformational heterogeneity, as well as enable sufficient dispersion for assessing inhomogeneity of proteins as a function of sequence without specific labeling. This may facilitate probing site-specific conformational heterogeneity in whole amyloids and freeze-trapped samples from protein folding. The mathematical approaches to analyze ϕ/ψ distributions shown here may be useful for quantification of variability in conformational ensembles of future research in general.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10858-022-00405-0.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The NMR spectrum analyzed during the current study is available from the corresponding author on reasonable request. All scripts along with sample data will be made available from the journal web site for facilitated future application.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Aeschbacher T, Schubert M, Allain FH-T (2012) A procedure to validate and correct the ^{13}C chemical shift calibration of RNA datasets. J Biomol NMR 52(2):179–190. https://doi.org/10.1007/s10858-011-9600-7

Asakura T, Ashida J, Zamane T, Kameda T, Nakazawa Y, Ohgo K, Komatsu K (2001) A repeated β-turn structure in poly(Ala-Gly) as a model for silk I of Bombyx mori silk fibroin studied with two-dimensional spin-diffusion NMR under off magic angle spinning and rotational echo double resonance. J Mol Biol 306(2):291–305. https://doi.org/10.1006/jmbi.2000.4394

Asakura T, Suzuki Y, Nakazawa Y, Holland GP, Yarger JL (2013a) Elucidating silk structure using solid-state NMR. Soft Matter 9(48):11440–11450. https://doi.org/10.1039/c3sm52187g

Asakura T, Suzuki Y, Nakazawa Y, Yazawa K, Holland GP, Yarger JL (2013b) Silk structure studied with nuclear magnetic resonance. Prog Nucl Magn Reson Spectrosc 69:23–68. https://doi.org/10.1016/j.pnmrs.2012.08.001

Burakova E, Vasa SK, Klein A, Linser R (2020) Non-uniform sampling in quantitative assessment of heterogeneous solid-state NMR line shapes. J Biomol NMR 74(1):71–82. https://doi.org/10.1007/s10858-019-00291-z

Cheung MS, Maguire ML, Stevens TJ, Broadhurst RW (2010) DANGLE: a bayesian inferential method for predicting protein backbone dihedral angles and secondary structure. J Magn Reson. https://doi.org/10.1016/j.jmr.2009.11.008

Chimon S, Ishii Y (2005) Capturing intermediate structures OF Alzheimer's β-Amyloid, Aβ(1–40), by solid-state NMR spectroscopy. J Am Chem Soc 127(39):13472–13473. https://doi.org/10.1021/ja054039l

De Dios AC, Sears DN, Tycko R (2004) NMR studies of peptide T, an inhibitor of HIV infectivity, in an aqueous environment. J Pept Sci 10(10):622–630. https://doi.org/10.1002/psc.571

del Amo JML, Schmidt M, Fink U, Dasari M, Fändrich M, Reif B (2012) An asymmetric dimer as the basic subunit in Alzheimer's disease amyloid β fibrils. Angew Chem Int Ed 51(25):6136–6139. https://doi.org/10.1002/anie.201200965

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipe. J Biomol NMR 6(3):277–293

Elkins MR, Wang T, Nick M, Jo H, Lemmin T, Prusiner SB, DeGrado WF, Stör J, Hong M (2016) Structural polymorphism of Alzheimer's β-amyloid fibrils as controlled by an E22 switch:

a solid-state NMR study. J Am Chem Soc 138(31):9840–9852. https://doi.org/10.1021/jacs.6b03715

Endapally S, Frias D, Grzemska M, Gay A, Tomchick DR, Radhakrishnan A (2019) Molecular discrimination between two conformations of sphingomyelin in plasma membranes. Cell 176(5):1040-1053.e17. https://doi.org/10.1016/j.cell.2018.12.042

Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins Struct Funct Bioinforma 23(4):566–579. https://doi.org/10.1002/prot.340230412

Fritzsching KJ, Hong M, Schmidt-Rohr K (2016) Conformationally selective multidimensional chemical shift ranges in proteins from a PACSY database purged using intrinsic quality criteria. J Biomol NMR 64(2):115–130. https://doi.org/10.1007/s10858-016-0013-5

Harris CR et al (2020) Array programming with NumPy. Nature 585:357–362. https://doi.org/10.1038/s41586-020-2649-2

Havlin RH, Tycko R (2005a) Probing site-specific conformational distributions in protein folding with solid-state NMR. Proc Natl Acad Sci U S A 102(9):3284–3289. https://doi.org/10.1073/pnas.0406130102

Havlin RH, Tycko R (2005b) Probing site-specific conformational distributions in protein folding with solid-state NMR. Proc Natl Acad Sci 102(9):3284–3289. https://doi.org/10.1073/pnas.0406130102

Heise H, Luca S, De Groot BL, Grubmüller H, Baldus M (2005) Probing conformational disorder in neurotensin by two-dimensional solid-state NMR and comparison to molecular dynamics simulations. Biophys J 89(3):2113–2120. https://doi.org/10.1529/biophysj.105.059964

Helmus JJ, Jaroniec CP (2013) Nmrglue: an open source Python package for the analysis of multidimensional NMR data. J Biomol NMR 55:355–367. https://doi.org/10.1007/s10858-013-9718-x

Hu KN, Tycko R (2010) What can solid state NMR contribute to our understanding of protein folding? Biophys Chem 151(1–2):10–21. https://doi.org/10.1016/j.bpc.2010.05.009

Hu KN, Havlin RH, Yau WM, Tycko R (2009) Quantitative determination of site-specific conformational distributions in an unfolded protein by solid-state nuclear magnetic resonance. J Mol Biol 392(4):1055–1073. https://doi.org/10.1016/j.jmb.2009.07.073

Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9(3):90–95

Hutchinson EG, Thornton JM (1994) A revised set of potentials for beta-turn formation in proteins. Protein Sci 3:2207–2216. https://doi.org/10.1002/pro.5560031206

Jaroniec CP (2019) Two decades of progress in structural and dynamic studies of amyloids by solid-state NMR. J Magn Reson 306:42–47. https://doi.org/10.1016/j.jmr.2019.07.015

Jeon J, Thurber KR, Ghirlando R, Yau WM, Tycko R (2019) Application of millisecond time-resolved solid state NMR to the kinetics and mechanism of melittin self-assembly. Proc Natl Acad Sci USA 116(34):16717–16722. https://doi.org/10.1073/pnas.1908006116

Kjaergaard M, Kragelund BB (2017) Functions of intrinsic disorder in transmembrane proteins. Cell. Mol. Life Sci. 74(17):3205–3224. https://doi.org/10.1007/s00018-017-2562-5

Kümmerlen J, van Beek JD, Vollrath F, Meier BH (1996) Local structure in spider dragline silk investigated by two-dimensional spin-diffusion nuclear magnetic resonance. Macromolecules 29(8):2920–2928. https://doi.org/10.1021/ma951098i

Lee W, Yu W, Kim S, Chang I, Lee W, Markley JL (2012) PACSY, a relational database managemen system for protein structure and chemical shift analysis. J Biomol NMR 54:169–179. https://doi.org/10.1007/s10858-012-9660-3

Linden AH, Franks WT, Akbey Ü, Lange S, van Rossum B-J, Oschkinat H (2011) Cryogenic temperature effects and resolution upon slow cooling of protein preparations in solid state NMR. J Biomol NMR 51(3):283–292. https://doi.org/10.1007/s10858-011-9535-z

Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M (2004) Simultaneous determination of protein structure and dynamics. Nature 433(7022):128–132. https://doi.org/10.1038/nature03199

Linser R, Dasari M, Hiller M, Higman V, Fink U, Lopez del Amo J-M, Markovic S, Handel L, Kessler B, Schmieder P, Oesterhelt D, Oschkinat H, Reif B (2011) Proton-detected solid-state NMR spectroscopy of fibrillar and membrane proteins. Angew Chemie - Int Ed 50(19):4508–4512. https://doi.org/10.1002/anie.201008244

Luo X, Yu H (2008) Protein metamorphosis: the two-state behavior of Mad2. Structure 16(11):1616–1625. https://doi.org/10.1016/j.str.2008.10.002

McKinney W *et al*. (2010) Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference*. pp. 51–56

Morris VK, Linser R, Wilde KL, Duff AP, Sunde M, Kwan AH (2012) Solid-state NMR spectroscopy of functional amyloid from a fungal hydrophobin: a well-ordered β-sheet core amidst structural heterogeneity. Angew Chemie Int Ed 51(50):12621–12625. https://doi.org/10.1002/anie.201205625

Nielsen JT, Mulder FAA (2020) Quantitative protein disorder assessment using NMR chemical shifts. Methods Mol Biol 2141:303–317. https://doi.org/10.1007/978-1-0716-0524-0_15

Nwanochie E, Uversky VN (2019) Structure determination by single-particle cryo-electron microscopy: only the sky (and intrinsic disorder) is the limit. Int J Mol Sci. https://doi.org/10.3390/ijms20174186

Paravastu AK, Leapman RD, Yau WM, Tycko R (2008) Molecular structural basis for polymorphism in Alzheimer's-amyloid fibrils. Proc Natl Acad Sci USA 47:18349–18354

Potapov A, Yau WM, Ghirlando R, Thurber KR, Tycko R (2015) Successive stages of amyloid-β self-assembly characterized by solid-state nuclear magnetic resonance with dynamic nuclear polarization. J Am Chem Soc 137(25):8294–8307. https://doi.org/10.1021/jacs.5b04843

Sammak S, Zinzalla G (2015) Targeting protein-protein interactions (PPIs) of transcription factors: challenges of intrinsically disordered proteins (IDPs) and regions (IDRs). Prog Biophys Mol Biol 119(1):41–46

Shen Y, Bax A (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. J Biomol NMR 56(3):227–241. https://doi.org/10.1007/s10858-013-9741-y

Shen Y, Delaglio F, Cornilescu G et al (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223. https://doi.org/10.1007/s10858-009-9333-z

Siemer AB (2020) Advances in studying protein disorder with solid-state NMR. Sol St Nucl Magn Reson 106:101643. https://doi.org/10.1016/j.ssnmr.2020.101643

Siemer AB, Huang KY, McDermott AE (2012) Protein linewidth and solvent dynamics in frozen solution NMR. PLoS One 7:10. https://doi.org/10.1371/journal.pone.0047242

Siemons L, Uluca B, Pritchard RB, McCarthy S, Heise H, Hansen DF (2019) Determining isoleucine side-chain rotamer-sampling in proteins from 13C chemical shift. Chem Commun 55:14107–14110. https://doi.org/10.1039/c9cc06496f

Su Y, Hong M (2011) Conformational disorder of membrane peptides investigated from solid-state NMR line widths and line shapes. J Phys Chem B 115(36):10758–10767. https://doi.org/10.1021/jp205002n

Tamiola K, Acar B, Mulder FAA (2010) Sequence-specific random coil chemical shifts of intrinsically disordered proteins. J Am Chem Soc 132(51):18000–18003. https://doi.org/10.1021/ja105656t

Turoverov KK, Kuznetsova IM, Fonin AV, Darling AL, Zaslavsky BY, Uversky VN (2019) Stochasticity of biological soft matter: emerging concepts in intrinsically disordered proteins and biological phase separation. Trends Biochem. Sci. 44(8):716–728. https://doi.org/10.1016/j.tibs.2019.03.005

Tycko R (2011) Solid-state NMR studies of amyloid fibril structure. Annu Rev Phys Chem 62:279–299. https://doi.org/10.1146/annurev-physchem-032210-103539

Tycko R (2014) Physical and structural basis for polymorphism in amyloid fibrils. Protein Sci. 23(11):1528–1539. https://doi.org/10.1002/pro.2544

Uluca B et al (2018) DNP-enhanced MAS NMR: a tool to snapshot conformational ensembles of α-synuclein in different states. Biophys J 114(7):1614–1623. https://doi.org/10.1016/j.bpj.2018.02.011

Uversky VN (2013) A decade and a half of protein intrinsic disorder: biology still waits for physics. Protein Sci. https://doi.org/10.1002/pro.2261

Uversky VN (2018) Intrinsic disorder, protein-protein interactions, and disease. Adv Prot Chem Struct Biol 110:85–121

van Beek JD, Beaulieu L, Schäfer H, Demura M, Asakura T, Meier BH (2000) Solid-state NMR determination of the secondary structure of Samia cynthia ricini silk. Nature 405(6790):1077–1079. https://doi.org/10.1038/35016625

Verel R, Ernst M, Meier BH (2001) Adiabatic dipolar recoupling in solid-state NMR: The DREAM scheme. J Magn Reson 150(1):81–99. https://doi.org/10.1006/jmre.2001.2310

Virtanen P et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in python. Nature Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M. *et al*. (2017) mwaskom/seaborn: v0.8.1, *Zenodo*. 10.5281/zenodo.883859

Xiang SQ, Biernat J, Mandelkow E, Becker S, Linser R (2016) Backbone assignment for minimal protein amounts of low structural homogeneity in the absence of deuteration. Chem Commun 52(21):4002–4005. https://doi.org/10.1039/c5cc09160h

Xiang S et al (2017) A two-component adhesive: tau fibrils arise from a combination of a well-defined motif and conformationally flexible interactions. J Am Chem Soc 139(7):2639–2646. https://doi.org/10.1021/jacs.6b09619

Zhou M, Morais-Cabral JH, Mann S, MacKinnon R (2001) Potassium channel receptor site for the inactivation gate and quaternary amine inhibitors. Nature 411(6838):657–661. https://doi.org/10.1038/35079500

Zhou DH, Shah G, Cormos M, Mullen C, Sandoz D, Rienstra CM (2007) Proton-detected solid-state nmr spectroscopy of fully protonated proteins at 40 kHz magic-angle spinning. J Am Chem Soc 129(38):11791–11801. https://doi.org/10.1021/ja073462m