technische universität
dortmund

fakultät für chemie
und chemische biologie

Dissertation zur Erlangung des Doktorgrades

Qualitative and quantitative characterization of protein backbone heterogeneity by solid-state NMR spectroscopy

Ekaterina Burakova

aus

Moskau, Russland

Dortmund 2023

Kontakt

MSc. Ekaterina Burakova
Fakultät für Chemie und Chemische Biologie
Technische Universität Dortmund
Otto-Hahn Str. 4a D-44227 Dortmund
Email: ekaterina.burakova@tu-dortmund.de
ORCID 0009-0006-5122-3265

Prüfungskommision

1. Gutachter: Prof. Dr. Rasmus Linser
2. Gutachter: Prof. Dr. Henrike Heise

Vorsitz: Prof. Dr. Carsten Strohmann

Tag der Einreichung: 12.07.2023

Tag der mündlichen Prüfung: 18.10.2023

Contents

Acknowledgements

I would like to express my sincere gratitude to my supervisor Prof. Dr. Rasmus Linser who gave me the opportunity to grow as a scientist and NMR spectroscopist. This work would not have been possible without his persistent optimism, support, light-hearted attitude, and empathy. He consistently took his time to delve into the finest details of the projects and always encouraged scientific exploration and creativity. I would also like to acknowledge the liberty he gave to the group in attending NMR schools and conferences. It greatly helped me to validate my ideas and become a part of the global NMR family.

AK Linser remained a healthy and supportive environment throughout my entire time in the group. I can't thank enough all my fellow colleagues for creating a wonderful atmosphere. I would like to extend special thanks to Dr. Suresh K. Vasa, the spectrometer wizard and NMR theorist of the group, for his down-to-earth attitude, immense help with the spectrometers and for teaching me to troubleshoot and never give up. I would like to express my gratitude to the former postdoc in the group, Dr. Petra Rovó for her invaluable commitment to teaching the basics of solid-state NMR, help with the spectrometer, her encouragement and inspiration. I also thank Dr. Himanshu Sign for his help in the lab and Dr. Laura Kukuk for the countless discussions, scientific inspiration and for always cheering me up. Last but not least, I thank all the past and present fellow PhD students – Alex, Sara, Kristof, Romeo, Julia, Benedikt, Anja, and Suchandra - for being wonderful people and for all the professional and life experience I gained from each of you.

I am grateful to Prof. Dr. Henrike Heise (Heinrich-Heine-Universität Düsseldorf) for taking the time to review my thesis and agreeing to serve as my opponent during the defense.

I also thank the people outside the formal collaborations who nevertheless participated in discussions regarding the present work and inspired the directions in which this project developed. These people include Dr. Johannes Dietschreit (formerly LMU Munich) and Prof. Dr. Christian Ochsenfeld (LMU Munich), Dr. Anton Ivanov (Keldysh Institute of Applied Mathematics), Prof. Dr. Jonas M. Peters (University of Copenhagen) and Dr. Łukasz Czajka (formerly TU Dortmund) as well as Nicole Hufnagel (TU Dortmund) and the Department of Statistical Consulting and Analysis of TU Dortmund.

Part of this work was done using NMRbox, a Biomedical Technology Research Resource maintained by National Center for Biomolecular NMR Data Processing and Analysis (USA), which is supported by NIH grant P41GM111135 (NIGMS).

I would like to acknowledge the generous financial and infrastructural support from the SFB1309 consortium.

I would like to say a big “Thank you!” to all my friends (including flat-mates) for all the joyful moments in the life outside work as well as during writing the thesis. The thanks go out to all my friends back in Moscow, to all the Dortmund Board Game group and to Martin and his cozy Wattenscheid community.

Finally, I thank my mother Elena who has encouraged open-mindness and critical thinking since my early ages, and my husband Anton for being there and keeping my spirit up.

Abstract

Proteins were long believed to be rigid, well-defined structures and specificity of their interactions to be determined only by unique geometric match. This paradigm began to shift with increased evidence of protein dynamics on the broad range of time scales and discovery of intrinsic protein disorder. The recent breakthrough in protein structure predictions allows to make a very cheap initial guess by simply submitting the amino acid sequence to neural networks such as AlphaFold2. However, flexibility of the polypeptide chains plays the key role in protein functions, their un-, re- and misfolding pathways and is far from being predicted by purely *in silico* methods. Understanding the conformational landscape which protein chain occupies statically and dynamically is essential for understanding of cellular processes and ultimately, designing efficient drugs, safe pesticides, and industrial biotechnological processes. NMR spectroscopy is an indispensable technique studying disordered molecules site-specifically, both in solution and in the solid state.

Protein disorder covers the continuum between the static set of defined states and dynamic ensembles. The position within this spectrum is also defined by the conditions of the chosen experiment and the timescale it accesses. Dynamic disorder can be converted into static disorder by freeze-trapping and studied in the solid phase. In solid-state NMR, static disorder manifests itself as the presence of additional peaks or, in case of a continuous distribution of a geometrical parameter, line broadening. Converting the distribution of the resonance frequencies into conformational ensembles is not a trivial task due to the multitude of factors that contribute to the nuclear resonance frequencies. Only few attempts of analysis have been made, however, the only information routinely extracted from the signals is the ensemble-average parameters such as chemical shifts at the peak maxima, average relaxation rates, distance restraints, etc.

This works proposes approaches to analyze residue-specific static disorder by interpretation and quantification of heterogeneously broadened lines in solid-state NMR spectra. The analysis is based on the dominant working hypothesis that the broadening of the backbone signals results from the backbone conformational distribution. The engineered analytical routines reconstruct the distributions of the backbone dihedral angles φ and ψ in two ways: on the basis of database analyses and by help of dihedral-angle predictors. The workflows are tested on a model sample as well as on the naturally heterogeneous sample of a functional amyloid (EAS$_{\Delta15}$ rodlets), where they are compared to estimating heterogeneity from peak widths. The analysis of the EAS$_{\Delta15}$ sample demonstrates the power of the proposed analysis for rather challenging systems where the only available high-resolution physico-chemical data are the peak shapes in the solid-state NMR spectra.

Zusammenfassung

Lange Zeit ging man davon aus, dass Proteine starre, wohldefinierte Strukturen sind und die Spezifität ihrer Wechselwirkungen nur durch ihre jeweilige geometrische Passung bestimmt wird. Dieses Paradigma begann sich mit zunehmendem Bewusstsein der Proteindynamik auf einer Vielzahl von Zeitskalen sowie der intrinsischen Proteinunordnung zu ändern. Der jüngste Durchbruch bei der Vorhersage von Proteinstrukturen ermöglicht eine sehr günstige, erste Abschätzung, indem einfach die Aminosäuresequenz an neuronale Netze wie AlphaFold2 übermittelt wird. Die Flexibilität der Polypeptidketten spielt jedoch eine Schlüsselrolle bei den Proteinfunktionen und ihren Ent-, Rück- und Fehlfaltungswegen und kann bislang durch reine in-silico-Methoden nicht zuverlässig vorhergesagt zu werden. Das Verständnis der Konformationslandschaft, die mit einer Proteinkette statisch und dynamisch assoziiert ist, ist für das Verständnis zellulärer Prozesse und letztlich für die Entwicklung effizienter Arzneimittel, sicherer Pestizide und industrieller biotechnologischer Prozesse von wesentlicher Bedeutung. Die NMR-Spektroskopie ist eine unverzichtbare Technik zur Untersuchung ungeordneter Moleküle sowohl in Lösung als auch im festen Zustand.

Die Unordnung von Proteinen umfasst das gesamte Kontinuum zwischen Verteilungen definierter statischer Zustände und vollends dynamischer Ensembles. Die Position in diesem Spektrum wird auch durch die Bedingungen des gewählten Experiments und dessen zugänglichen Zeitskalen bestimmt. Dynamische Unordnung kann prinzipiell durch Einfrieren in statische umgewandelt und in der festen Phase untersucht werden. In der Festkörper-NMR äußert sich statische Unordnung durch das Auftreten zusätzlicher Peaks oder, im Falle einer kontinuierlichen Verteilung geometrischer Parameter, durch Linienverbreiterung. Die Rekonstruktion einer Verteilung von Resonanzfrequenzen in die zugrunde liegenden Konformationsensembles ist aufgrund der Vielzahl von Faktoren, die zu den Kernresonanzfrequenzen beitragen, nicht trivial. Bisher sind nur wenige Versuche einer solchen Analyse unternommen worden. Die einzigen Informationen, die routinemäßig aus den Signalen extrahiert werden, sind durchschnittliche Ensemble-Parameter wie die chemische Verschiebung an den Peakmaxima, durchschnittliche Relaxationsraten, Abstandsrestriktionen usw.

In dieser Arbeit werden Ansätze zur Analyse von aminosäurespezifischer Heterogenität über Interpretation und Quantifizierung von heterogen verbreiterten Linien in Festkörper-NMR-Spektren vorgeschlagen. Die Analyse basiert auf der zugrundeliegenden Arbeitshypothese, dass die Verbreiterung der Rückgratsignale aus der Konformationsverteilung des Rückgrats resultiert. Die analytischen Routinen rekonstruktuieren die Verteilung der Rückgratdihedralwinkel φ und ψ auf zwei Arten: auf der

Grundlage von Datenbankanalyse und unter Verwendung von Diederwinkelprädiktoren. Die Arbeitsabläufe werden an einem Modellsample und am heterogenen Sample eines funktionellen Amyloids (EAS$_{\Delta15}$ “Stäbchen”) getestet, wo sie mit dem alternativen Ansatz der Heterogenitätsschätzung aus Linienbreiten verglichen werden. Die Analyse der EAS$_{\Delta15}$-Probe zeigt die Leistungsfähigkeit der vorgeschlagenen Analyse für anspruchsvolle Systeme, bei denen die einzigen verfügbaren restaufgelösten Daten die Verteilungen der chemischen Verschiebungen aus den Festkörper-NMR-Spektren sind.

Abbreviations

BMRB	Biological Magnetic Resonance data Bank
CP	Cross-Polarization
CSA	Chemical Shift Anisotropy
DANGLE	Dihedral ANgles from Global Likelihood Estimates, a protein backbone dihedral angle prediction program
DFT	Density Functional Theory
DSS	4,4-dimethyl-4-silapentane-1-sulfonic acid
DSSP	Define Secondary Structure of Proteins, an algorithm of secondary structure assignment
FID	Free Induction Decay
fMLF	formyl-Met-Leu-Phe-OH
FWHH	Full line Width at Half peak Height
H-bond	Hydrogen bond
hmsIST	Harvard Medical School implementation of Iterative Soft Thresholding algorithm
IDP	Intrinsically Disordered Protein
IDR	Intrinsically Disordered protein Regions
MAS	Magic Angle Spinning
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
NUS	Non-Uniform Sampling
PACSY	Protein structure And Chemical Shift NMR Spectroscopy, a database relating refined 3D protein models and experimentally obtained chemical shifts.
PDF	Probability Density Function
PDB	Protein Data Bank
RMSD	Root Mean Square Difference
ROI	Residue Of Interest
SNR	Signal-to-Noise Ratio
SSA	Signal Separation Algorithm, a method of reconstruction of NUS data
SW	Spectral Width or Spectral Window
SMILE	Sparce Multidimensional Iterative Lineshape-Enchanced Reconstruction

STRIDE	STRructural IDEntification, an algorithm of secondary structure assignment
TALOS	Torsion Angle Likelihood Obtained from Shifts and sequence similarity, a protein backbone dihedral angle prediction program
TALOS-N	TALOS’s successor, augmented by Neural networks

1 | INTRODUCTION

This chapter introduces the basic physical, mathematical, and chemical concepts underlying the methods used in the work. The text focuses on methods of solid-state NMR for protein structural studies.

The first part (*Section 1.1*) provides a brief overview of physical principles of nuclear magnetic resonance, interactions and parameters which can be observed or calculated, and, if necessary, suppressed.

The second part (*Section 1.2*) describes methods of recording and processing of NMR data. In particular, it focuses on the techniques of non-uniform sampling, which come essential when dealing with high-dimensional experiments. Experiments of high dimensionality are extensively in this work.

The third part (*Section 1.3*) gives an overview about protein structure and methods of studying thereof, as well as the relationships of protein structure and chemical shifts. First, it provides a basic overview of the spatial organization of polypeptide chains and discusses the occurrence and nature of protein disorder. Then, it briefly summarizes methodology of NMR spectroscopy for structural studies, focusing on chemical shift-based approaches and chemical shift-structure relationships. Finally, it provides an overview of methods to study protein disorder and the role of NMR-based approaches among them; it reviews the published strategies of elucidating protein disorder in various systems.

1.1. Physical basis of NMR spectroscopy

Nuclear magnetic resonance (NMR) was first observed on molecular beams in 1930s in experiments of measuring nuclear magnetic moment by Rabi *et al.* in 1938. In the bulk matter, the first successful observations were conducted in late 1945 – early 1946 independently by two groups of physicists: in Stanford by Felix Bloch *et al.* in water (1946) and at the MIT by the group of Edward M. Purcell a on paraffin sample (Purcell *et al.*, 1946).

NMR spectroscopy exploits the ability of nuclei with non-zero spin angular momentum to interact with external magnetic fields and with each other. Nuclear magnetism stems from the innate properties of atomic nuclei such as spin and spin angular and magnetic momenta. The origins of nuclear spin are complex and, according to the current understanding, it arises from the spins of nucleons and subnucleon particles as well as the strong interactions between them (extensively discussed in, for example, Engelke, 2022). A spin is described by a spin quantum number *I* and takes integer or half-integer values, *I* = 0, $^1/_2$, 1, $^3/_2$, etc. The present text focuses only on the spin-½ nuclei since these include all the most important isotopes for biomolecular NMR.

The spin quantum number defines the number of eigenstates of a nucleus as $2I + 1$. The two eigenstates, for spin-½ nuclei commonly referred to as α and β, are associated with the particular orientations of the vector of spin magnetic momentum $\boldsymbol{\mu}$ in the presence of an externally applied magnetic field $\boldsymbol{B}_0$, which in the case of I = +½ corresponds to the parallel and anti-parallel alignment. Thereby, the external $\boldsymbol{B}_0$ field vector is defined to be parallel to the z-axis. Hence its z-component is the scalar $\boldsymbol{B_0}$. The magnetic momentum depends on the nucleus' sensitivity to the magnetic field, characterized by the gyromagnetic ratio γ. The energies of the two spin states are given by:

$$E_\alpha = -\frac{1}{2}\boldsymbol{\mu B}_0 = -\frac{1}{2}\gamma\hbar\boldsymbol{B_0} \qquad (1.1) \qquad\qquad E_\beta = -\frac{1}{2}\boldsymbol{\mu B}_0 = +\frac{1}{2}\gamma\hbar\boldsymbol{B_0} \qquad (1.2)$$

and the corresponding energy gap is

$$\Delta E = E_\alpha - E_\beta = -\frac{1}{2}\gamma\hbar\boldsymbol{B_0} - \frac{1}{2}\gamma\hbar B_0 = \gamma\hbar\boldsymbol{B_0} \qquad (2)$$

According to the relation (2), the energy gap between the eigenstates is directly proportional to the strength of the external magnetic field $\boldsymbol{B_0}$. The two states are populated according to Boltzmann distribution:

$$^{n_\alpha}/_{n_\beta} = e^{\frac{-\Delta E}{k_B T}} \qquad (3)$$

where n_α and n_β are the populations of the two energy levels and k_B is the Boltzmann constant. It is the population difference that drives NMR spectroscopy, as it creates the macroscopic net polarization of the sample. According to the formula (3), at 298 K and a field strength of 16.5 T, typical for spectrometers used for structural biology, only one in 10 000 spins contributes to the bulk magnetic moment. This explains the low sensitivity of NMR spectroscopy and the incentive to build ever stronger magnets.

The energy required to perform the transition between the states can be expressed in frequency units:

$$\Delta E/_{h} = {}^{\gamma\hbar\boldsymbol{B_0}}/_{h} = \upsilon \tag{4}$$

This frequency υ is referred to as Larmor frequency. The energy transitions can be probed with the magnetic field $\boldsymbol{B}_1$, oscillating at the Larmor frequency and applied perpendicularly to the flux of the constant field $\boldsymbol{B}_0$. In the early years of NMR spectroscopy, the $\boldsymbol{B}_1$ field was varied continuously (in so-called “continuous wave” experiments). Invention of pulsed NMR techniques with Fourier transform data processing enabled the access to a large variety of experiments. All modern NMR experiments are combinations of pulses, characterized by duration (pulse *length*), power and frequency, and delays that are needed to evolve a particular spin state or interaction. Unlike any other type of spectroscopy, for instance, IR or UV, NMR manipulates the mixed states of the nuclei, which have a much longer lifetime than those of electrons (as in UV or optical spectroscopy) or vibrational states of the molecule (in IR spectroscopy) (Keeler, 2010).

1.1.1. Interactions in NMR spectroscopy

The exact resonance frequency of an isolated spin depends on its local magnetic field defined by the net contributions of the shielding electrons and the neighboring nuclei. Besides $\boldsymbol{B_0}$, the magnetic field at the specific nucleus is defined by the collective effect of other minor magnetic fields created by electrons or the neighboring NMR active nuclei. The observed frequency also depends on the timescale of the motion of the moiety in question or its interaction partners (sometimes referred to as the interaction of spins with phonons (McDermott and Polenova, 2012). Even the strongest local effects are orders of magnitude smaller than the Zeeman interaction and they shift the absolute resonance frequency of the spin by 10^{-3} - 10^{-8} of its Larmor frequency.

Chemical shifts

The nuclei in atoms are surrounded by electron clouds of different geometry. The electrons modulate the magnetic field at the nucleus due to the diamagnetic properties of molecular (or atomic) orbitals, thus providing the *shift* of the nuclear resonance frequency. In diamagnetic materials, such as most samples of biomolecules, this phenomenon is called *chemical shift*. The shielding effect scales linearly with the strength of the external magnetic field $\boldsymbol{B_0}$**,** which can be illustrated by an example of a closed coil in the magnetic field that creates an opposing magnetic field according to Lenz's law. The differences in electron shielding provide frequency dispersion of the range of the order of 10^{-5} to 10^{-3} of the absolute resonance frequency, the range depends on the isotope. Therefore, the values of chemical shifts are conventionally expressed in parts per millions (ppm) of the difference between the absolute frequency of the nucleus in question from the absolute frequency of a reference nucleus. This allows to avoid both, the use of large numbers and dependence of the external magnetic field $\boldsymbol{B_0}$. As the reference for biomolecules, IUPAC recommends to use the methyl signal of 2,2-dimethylsilapentane-5-sulphonic acid (DSS), dissolved in low (typically, 1%) concentration, for direct chemical-shift referencing of protons as well as indirect referencing of other nuclei (Markley *et al.*, 1998). Apart from being sensitive probes of the local environment, chemical shifts are easily measured and highly reproducible, which renders them a valuable parameter in identification of particular moieties and site-specific analysis of molecular structure and dynamics.

In the general case, the electronic environment of the nucleus is not uniform, and the observed resonance frequency will depend on the *orientation* of the moiety in the magnetic field – this phenomenon is called Chemical Shift Anisotropy (CSA). The shielding can be represented with the tensor σ:

$$\sigma = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{pmatrix} \tag{5}$$

This tensor can be decomposed into three terms: the isotropic, symmetric, and asymmetric contributions,

$$\sigma = \sigma_{\text{iso}} + \sigma_{\text{sym}} + \sigma_{\text{anti}} \tag{6}$$

but the antisymmetric part does not contribute to the observable signal to any great extent. If the coordinate system is oriented along one of the three principal axes of the interaction tensor, the matrix is diagonal:

$$\sigma = \begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{pmatrix} \tag{7}$$

Where σ_{11}, σ_{22} and σ_{33} are the principal components of the interaction tensor. By convention ("Mehring notation", Mehring, 1983), $\sigma_{11} \leq \sigma_{22} \leq \sigma_{33}$, where the σ_{11} shielding component corresponds to the highest resonance frequency of the given nucleus.

When the movement of the moiety is unrestricted, which is the case for solutions with low viscosity, the three spatial components are averaged and the observed frequency for all molecules converges to the single *isotropic* chemical shift σ_{iso}, which equals to the trace of the matrix:

$$\sigma_{iso} = \frac{\sigma_{11}+\sigma_{22}+\sigma_{33}}{3} \tag{8}$$

According to a popular convention (Duer, 2002), the shielding tensor is described by two parameters: anisotropy $\mathit{\Delta}$

$$\mathit{\Delta} = \sigma_{11} - \sigma_{iso} \tag{9}$$

and asymmetry η:

$$\eta = \frac{\sigma_{33}-\sigma_{22}}{\sigma_{11}} \tag{10}$$

These parameters are defined analogously for the chemical shift tensor; one should keep in mind that the strongest shielding corresponds to the lowest chemical shift value, so $\delta_{11} \geq \delta_{22} \geq \delta_{33}$. In solids and colloidal media, the motion is restricted and the resonance frequencies of the differently oriented components combine into the *powder pattern* (Fig. 1.1.1A). The observed chemical shift of the nucleus in this case will depend on the orientation of the interaction frame in the magnetic field (Duer, 2002):

$$\begin{aligned} \delta &= \delta_{iso} + \delta_{aniso} \\ &= \delta_{iso} + \frac{1}{2}\mathit{\Delta}_{CS}(3\cos^2\theta - 1 + \eta_{CS}\sin^2\theta\cos 2\phi) \end{aligned} \tag{11}$$

where θ and ϕ are the polar angles (defined on Fig. 1.1.1B), $\mathit{\Delta}_{CS}$ and η_{CS} are the anisotropy and asymmetry of the chemical shift tensor.

Whereas in liquid phase only the isotropic chemical shift can be observed, chemical shift anisotropy contributes to relaxation processes by providing change in the local magnetic field for nuclei upon vibration of their moieties.

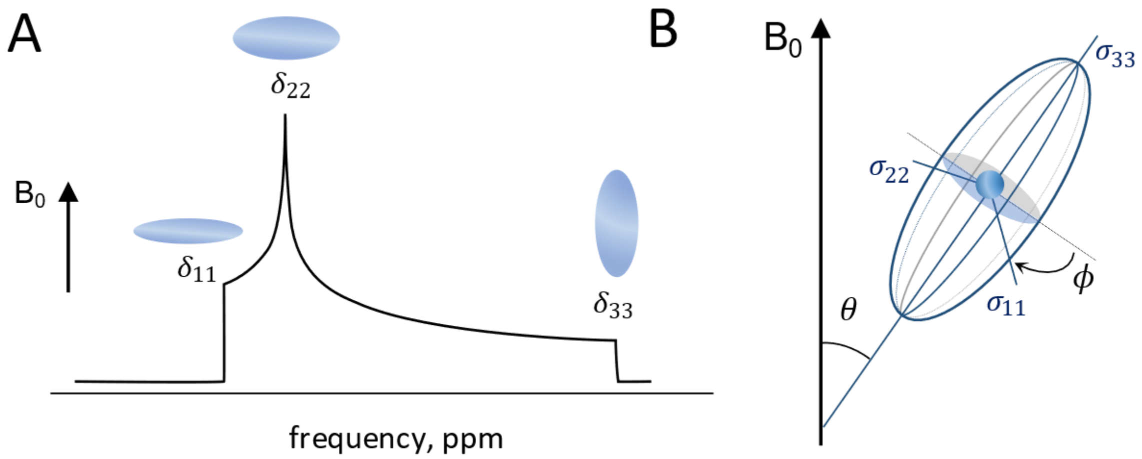


Figure 1.1.1. Illustration of chemical shift anisotropy. **A**: NMR Lineshape of the randomly distributed chemical shift tensors – the powder pattern. **B**: Definition of the polar angles θ and ϕ, which relate the principal components of the electron shielding tensor (σ_{11}, σ_{22}, σ_{33}) in the magnetic field. In the symmetric case, $\sigma_{11} = \sigma_{22}$.

Scalar couplings

The hyperfine interactions between the nuclei and the electrons in the covalent bonds between them induce splitting between energy levels of the states with different alignment of the electronic and nuclear magnetic momenta. The energy of the system is lower when the magnetic momenta of the electron and the nucleus are aligned antiparallelly. Given that the pair of bonding electrons can only have opposite spins (according to Pauli's principle), the slightly lower energy state corresponds to the antiparallel alignment of nuclear spins (Levitt, 2008). This interaction is referred as *indirect spin-spin coupling*, *J-coupling* or *scalar coupling,* reflecting its (overwhelmingly) isotropic character. Observed values of *J*-couplings are called coupling *constants* since the interaction is independent of the external magnetic fields. The effect of indirect spin interactions is relatively weak compared to the effect of electron shielding. The strength of the one-bond ^1J-couplings depends on the gyromagnetic ratios of the coupled nuclei and can reach up to a few hundred kHz.

Multiple-bond couplings depend on the angle between the bonds, which renders them valuable observables for molecular structure determination. Both homo- and heteronuclear three-bond ^3J-coupling constants can be quite accurately translated into structural information with Karplus relations (Karplus, 1959) that have been empirically parametrized for the most important dihedral angles in proteins (Li *et al.*, 2015) and nucleic acids (Marino *et al.*, 1999). Proton-proton couplings mediated by hydrogen bonds can be used to determine the H-bond length in solution (Cornilescu *et al.*, 1999b) as well as in the solid state (although in solid state the technique is far from being trivial) (Schanda *et al.*, 2009).

Indirect spin couplings enable one of the two essential mechanisms for magnetization transfer. The fundamental building block for solution-state multidimensional experiments – INEPT (Insensitive Nuclei Enhancement by Polarization Transfer) (Morris and Freeman, 1979) is used to increase polarization among the low-gamma nuclei (like ^{13}C and ^{15}N) and to conduct the magnetization along the desired pathway, for example, in a protein backbone.

Dipole-dipole interactions

The *direct* magnetic interaction between two nuclei is a purely anisotropic interaction, and is usually averaged out in liquid media as long as the molecular tumbling is not hampered. The mechanism of the spin-spin interactions is very similar to the classical interaction of two collinear bar magnets (Fig. 1.1.2A). The interaction strength between the magnetic dipoles depends on the distance between them and their mutual alignment. The lowest-energy configuration is achieved when both dipoles are oriented head-to-tail (opposite spin states), and positioned along the orientation of the two dipoles (and, correspondingly, the external magnetic field). If the dipoles are oriented side to side, the energy of the system is at maximum. If the spin states of the interaction partners are different, this effect is inverse – which is reflected in the splitting NMR signals of the interacting nuclei. In general, interaction between dipoles *i* and *j* can be expressed as

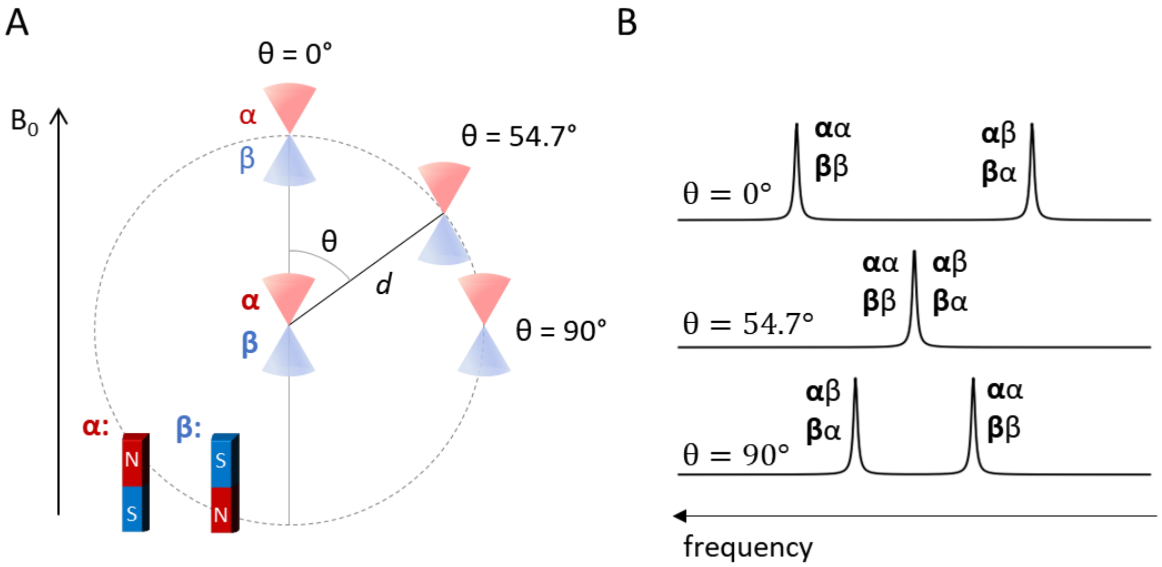


Figure 1.1.2. Illustration of the dipole-dipole interaction at different orientations of the interaction axis (**A**) and its manifestation in static, single-crystal NMR spectra (**B**). The state of the central spin is highlighted by bold font. The figure demonstrates the three key scenarios. The minimum interaction energy is achieved between the couple of nuclei in opposite states oriented collinearly to the magnetic field ($\theta = 0°$), which corresponds to the lowest resonance frequency of the spin (the leftmost doublet component). Degeneracy is achieved at the point of energy equivalence at $\theta = 54.7°$ (the *magic angle*). Interaction energy at $\theta = 90°$ is half of the maximum (corresponding to half the distance between the doublet components). The spins in α and β states can be compared to bar magnets.

$$d_{ij} = -\frac{\mu_0}{4\pi}\frac{\gamma_i\gamma_j\hbar}{r^3}\cdot\frac{1}{2}(3\cos^2\theta_{ij}-1) \tag{12}$$

where μ_0 is the magnetic permeability of vacuum ($\mu_0 = 1.26\ \mathrm{N}\cdot\mathrm{A}^{-2}$), γ is the gyromagnetic ratio of nuclei i and j, $\hbar$ is the reduced Plank constant ($\hbar = 1.05\,J\cdot s$), r the distance between the dipoles and θ is the angle between the internuclear vector and the orientation of the dipoles. The point of energy equivalence, which manifests in a singlet in the NMR spectrum (Fig. 1.1.2B), is achieved at $\theta \approx 54.74°$, which is known as the *magic angle.*

Like CSA, dipole-dipole interactions are averaged out in isotropic media (non-viscous liquid solutions) but cause a very important relaxation mechanism. The Nuclear Overhauser Effect (NOE) is the direct consequence of the dipole-dipole interaction and is the basis of measurements of interatomic distances (with a NOESY building block) and atomic mobility (for example, by measuring ^{15}N hetNOE). In the solid state, it is the dipole-dipole interaction that is exploited for magnetization transfer (for example, by the technique of cross-polarization, see below *Section 1.2.1*) as opposed to *J*-couplings for solution-state experiments.

1.1.2. Relaxation

Once the ***B***$_1$ pulse is switched off, the system starts to return to equilibrium. Relaxation of the bulk magnetization is split into two phenomena: recovery of the *longitudinal* component M_z along the direction of the magnetic field ***B***$_0$ and the loss of the *transverse* magnetization M_{xy}. Both of them can be described with the exponential decay from the initial states M_z^0 and M_{xy}^0 over time *t*:

$$M_z = M_z^0\left(1-e^{-\frac{t}{T_1}}\right) \qquad (13.1) \qquad\qquad M_{xy} = M_{xy}^0\,e^{-\frac{t}{T_2}} \qquad (13.2)$$

where *T*$_1$ and *T*$_2$ are the characteristic relaxation times. The first phenomenon is simply a recovery of the Boltzmann distribution of the spin α and β states (Eq. 3). The second one arises due to the loss of coherence between the spins. Both relaxation types are driven by fluctuations of the local effective magnetic field at a given spin that are induced by local motions.

The complex relationships between relaxation rates and the local dynamics of the spin – and hence the moiety – are the subject of thorough theoretical studies and are described elsewhere (Levitt, 2008; Kleckner and Foster, 2011; McDermott and Polenova, 2012).

The major contributing mechanisms are the dipole-dipole interactions and CSA, and the microscopic inhomogeneities of the ***B***$_0$ field additionally enhance the coherence loss in the transverse plane. Both relaxation mechanisms depend on the Larmor frequencies of the spin in question as well as its interaction partners, which means the higher the magnetic field of the spectrometer, the faster the relaxation. In the solid-state, relaxation rates are also a function of the sample spinning rate that averages anisotropic contributions (see below, *Section 1.1.3*). The longitudinal relaxation is more sensitive to the ns-µs motions, since it requires oscillations of the moieties (emissions of the radiofrequency photons) at the Larmor frequency to accomplish the spin flip. Loss of spin coherence however, occurs due to the random fluctuations, and therefore is sensitive to the broader range of timescales. As a consequence, in the systems with slow or absent molecular tumbling (the case of solid-state samples), T_2 is much smaller than T_1 and therefore determines the signal decay rate, or, converted to the frequency domain, the *homogeneous* line width.

1.1.3. Solid-state NMR with magic angle spinning

In the solid phase or in liquids with high viscosity, Brownian motion as well as molecular tumbling is hindered or absent, which leads to the anisotropic interactions to manifest themselves in the spectra as severely broadened signals. Whereas the angular-dependent parameters encoded in the observed patterns contain valuable information about molecular geometry, complex spectra of biological macromolecules become unintelligible. Since both CSA and DD-interaction depend on the term $3cos^2\theta - 1$ (Eq. 11, if the asymmetry η_{CS} is neglected, and 12), spinning the solid sample at the magic angle $\theta = \arccos\sqrt{\frac{1}{3}} = 54.74°$ to the magnetic field of the spectrometer would average out both effects (Fig. 1.1.3). This idea was proposed independently by Andrew *et al.*, (1958) and Lowe (1959). The necessary interactions can be further reintroduced by *recoupling* techniques.

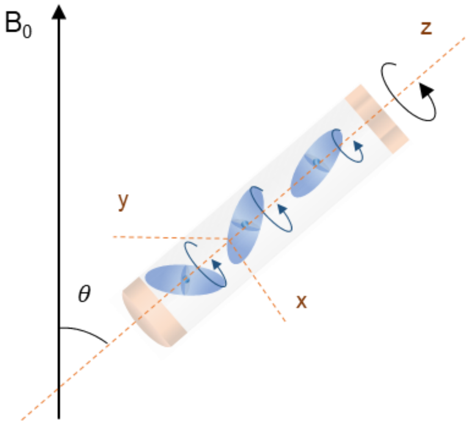


Figure 1.1.3 Upon spinning, anisotropic interactions average about the rotor coordinate system. If the rotor is positioned at the magic angle ($\theta \approx 54.7°$) to the $\boldsymbol{B_0}$ field, CSA and dipole-dipole couplings average around zero.

Table 1.1 presents all types of interactions of spin-½ nuclei in solid state. Dipole-dipolar interaction is overwhelmingly dominant, especially between protons due to their anomalously high gyromagnetic ratio. The effects of the anisotropic interactions of the spectral line widths are eliminated once the spinning rate achieves the interaction strength expressed in Hertz. Sufficient averaging of ^{1}H-^{1}H dipolar interactions the fastest probe heads available by now with spinning rates of 110 kHz or above, yet the complete eradication of the anisotropic effects would require MAS rates above 300 kHz (Xue *et al.*, 2018).

Table 1.1 Typical upper-border strength of anisotropic interactions in proteins given in frequency units (Bertini *et al.*, 2012). Strengths of CSA depend on the ***B*₀** field and therefore are given in ppm.

Interaction	**Typical strength**
^{1}H-^{1}H dipolar coupling, close contacts*	up to 140 kHz
^{1}H-^{1}H dipolar coupling (CH_3 group)	60 kHz
^{13}C-^{1}H dipolar coupling	23 kHz
^{15}N-^{1}H dipolar coupling	11 kHz
^{13}C-^{13}C dipolar coupling, directly bonded	3 kHz
^{13}C-^{15}N dipolar coupling	1 kHz
^{13}C CSA (carbonyls)	150 ppm
^{13}C CSA (aliphatic)	20 ppm
^{15}N CSA (amide)	150 ppm

* calculated with Eq. 12

1.2 NMR data acquisition and processing

1.2.1. Building blocks of NMR experiments

The advent of pulsed NMR allowed for design of a huge variety of complex experiments. Each NMR experiment is the application of a pulse sequence on a sample in the $\boldsymbol{B_0}$ field of the spectrometer. A pulse sequence is a combination of RF pulses and delays which would lead the magnetization through a specific pathway, evolve the target interactions and suppress the others.

A pulse sequence of a basic 2D hNH experiment, which combines all the principle building blocks of solid-state NMR pulse sequences, in shown in Fig. 1.2.1. A version of this pulse sequence is used in the present work.

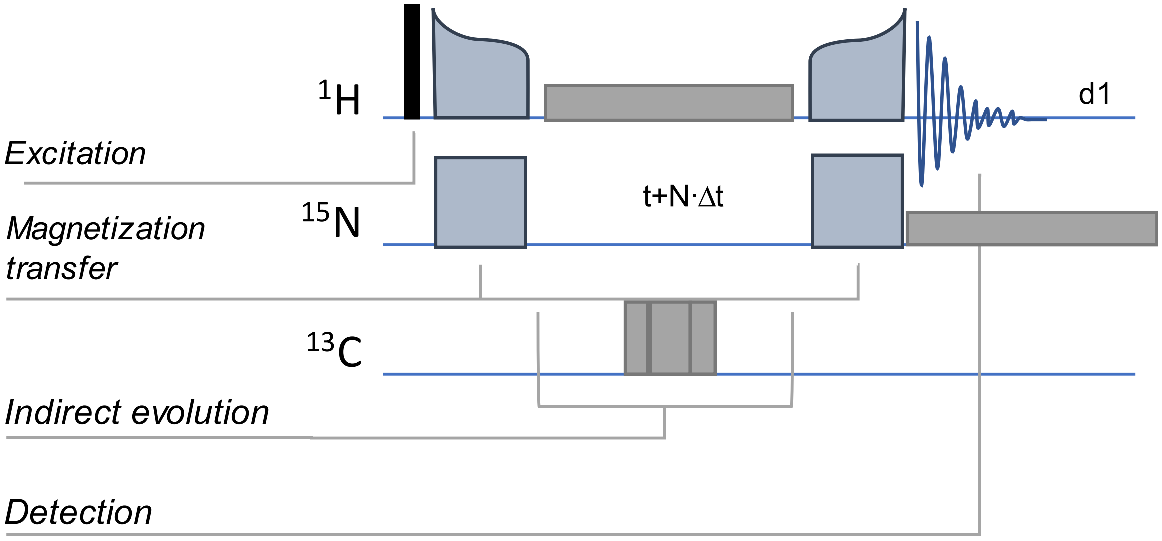


Figure 1.2.1. Building blocks of the solid-state NMR pulse sequence on example of a basic CP-hNH experiment. Solvent suppression block is omitted since in this work only solvent-free samples were studied.

Excitation

The absolute majority of experiments start with a 90° pulse that creates the spin coherence in the transverse plane (see *Section 1.1*) aside from specific applications like Inversion Recovery sequences. The amount of magnetization created – and hence the sensitivity of the experiment – directly depends on the difference between the populations of the two spin states. Therefore, it is always beneficial to excite nuclei with the highest gyromagnetic ratio (Eq. 3), which are ^{1}H in most of the measurements on proteins.

Magnetization transfer

As discussed in *Section 1.1.3*, solid-state experiments primarily utilize the dipolar interactions to transfer of magnetization, as opposed to *J*-couplings. (Solution state techniques will generally not be covered in this work.) A very popular technique referred to in the modern literature as cross-polarization (CP) was proposed by Pines *et al.* (1972). The transfer happens when the nuclei on the both, source (magnetization flipped into the transverse plane) and the target channels (nuclei in the thermal equilibrium) are hit by a long (order of ms) continuous wave. The CP pulse prevents any evolution of the magnetization, effectively refocusing all dephasing, and is called *spin lock*. The spin lock can be seen as the new external magnetic field $\boldsymbol{B}_1$, and, just like in the constant field of the spectrometer $\boldsymbol{B}_0$, it causes Zeeman splitting. It is possible to match the spin locks on the both channels *I* and *S* (in the example above, ^{1}H and ^{15}N) in such a way that the energy gaps for both groups of nuclei become equal. The matching condition was found "by the Wizard of Resonance Erwin Hahn and demonstrated by the Wizard and his Sorcerer's Apprentice Sven Hartmann" (quoting Slichter, 1990):

$$\gamma_I B_{1I} = \gamma_S B_{1S} \tag{14}$$

Upon sample spinning, however, the dipolar couplings become time-dependent. Thermodynamical (Slichter, 1990) or quantum mechanical derivations laid out in detail in other works (Wu and Zilm, 1993; Michel and Engelke, 1994) show that the polarization transfer can be achieved at the set of matching conditions:

$$\omega_I \pm \omega_S = n\omega_{MAS} \tag{15}$$

where ω_I and ω_S are the spin lock strength, n is an integer parameter; $n = -2, -1, 0, 1, 2,$ and ω_{MAS} is the angular frequency of MAS spinning. Condition n = 0 is known as Second Order CP (SOCP, Lange *et al.*, 2009) and comes into play at longer contact times (duration of the CP pulses, blue in Fig. 1.2.1). Sum or difference of the spin lock fields on both channels correspond to double-quantum (DQ, flip-flip) or zero-quantum (ZQ, flip-flop) interacting mechanisms.

For targeting of specific pathways, a variety of selective CP-based schemes has been developed, including SPECIFIC-CP (Baldus *et al.*, 1998), DREAM (Verel *et al.*, 2001), as well as non-CP recoupling techniques, PDSD (Szeverenyi *et al.*, 1982), DARR (Takegoshi *et al.*, 2001), REDOR (Gullion, 2006) or symmetry-based sequences (Levitt, 2012). The choice of the specific recoupling technique for the desired pathway highly depends on the set of the experimental conditions: MAS rate, field strength and probe specifications (Nielsen *et al.*, 2011). The most recent advancements in development of recoupling techniques for fast MAS are nicely reviewed by Ji *et al.* (2021).

The DREAM scheme that is used in the present work for the transfer between ^{13}Cα-^{13}Cβ atoms relies on the HORROR recoupling condition (Nielsen *et al.*, 1994):

$$2 \cdot \omega_{\mathrm{I}} = \omega_{\mathrm{MAS}} \tag{16}$$

Eq. 14 is the special case of Eq. 13 for a homonuclear ($\omega_S = \omega_I$) DQ transfer with $n = 1$. The key feature of the DREAM scheme is the adiabatic magnetization transfer which can be achieved by modulation of the spin lock amplitude according to the *tanh* function (the pulse shape is schematically depicted in Fig. 1.2.1 on the ^{1}H channel).

Indirect evolution

Multidimensional NMR spectroscopy created a vast range of opportunities for tackling complex problems of structure determination and the importance of this concept for biomolecular NMR is hard to overestimate. The main information encoded in the added dimensions is the isotropic chemical shift, which allows for correlations between different nuclei belonging to the particular sites. Parameters of the mixed spin states, like the coupling constants or double-quantum chemical shifts can also be recorded in the indirect dimensions. A remarkable feature of indirect evolution is that it makes it possible to observe and characterize states that are not detectable directly (as double-quantum coherences), which can be targeted by the specific research question or just serve as a convenient bypass for a more optimal pulse sequence (see HMQC experiment (Bax *et al.*, 1983). The unnecessary interactions can be suppressed by *decoupling* schemes (gray blocks in Fig. 1.2.1).

The conventional, uniform sampling of the chemical shift evolution curve is achieved by regular incrementation of the time delay (t is incremented with Δt in Fig. 1.2.1).

Detection

The directly detected NMR signal appears as oscillating current induced in the detection coil. The current that decays over time due to relaxation *(Section 1.1.2)* is conventionally referred to as Free Induction Decay (FID). The amount of the induced current is directly proportional to the gyromagnetic ratio of the detected nucleus, therefore it is always preferable to detect on the high-gamma nuclei (as shown in Fig. 1.2.1). In the solid state, for a long time the detection was done on ^{13}C channel, since proton detection suffered from severe broadening caused by anisotropic interactions (see *Section 1.1.1*). Owing to the recent advances in design of rotors and probeheads, which allowed to spin the MAS rotor at 40 kHz and higher, proton detection has become beneficial even though the complete averaging of proton-proton dipolar couplings is not even possible at spinning rates as high as 130 kHz (see *Section 1.1.3*, Table 1.1 and research of Xue *et al.* (2018).

The raw analog signal has a frequency on the order of hundreds of MHz. The sampling rate allowed by the modern hardware is on the order of hundreds kHz, and, as stated by the Nyquist-Shannon theorem, the highest detectable frequencies are twice as low. Therefore, before digitalization, the signal is passed through a radiofrequency mixer (Fig. 1.2.2) which creates a signal oscillating at the difference between the observed and the reference frequencies (on the order of 1-10^4 Hz). The receiver frequency is typically set to be in the middle of the spectrum, so each component in the NMR signal is *offset* from it by a positive or negative frequency. Combining the both *cos* (real) and *sin* (imaginary) components into complex numbers allows determining the sign of the offset; recording both of these components is referred to as *quadrature detection*. In the indirect dimensions, quadrature detection is achieved by *selecting* the orthogonal components of the magnetization dimension with a phase shift of 90° for pulses involved in magnetization transfer. Thus, the pulse sequence has to be run twice for each point of the indirect FID; acquisition of a point of an (N+1)-dimensional requires 2^N FIDs. This factor contributes to the large time cost of the high-dimensionality experiments, which is discussed in the following *Section 1.2.2*.

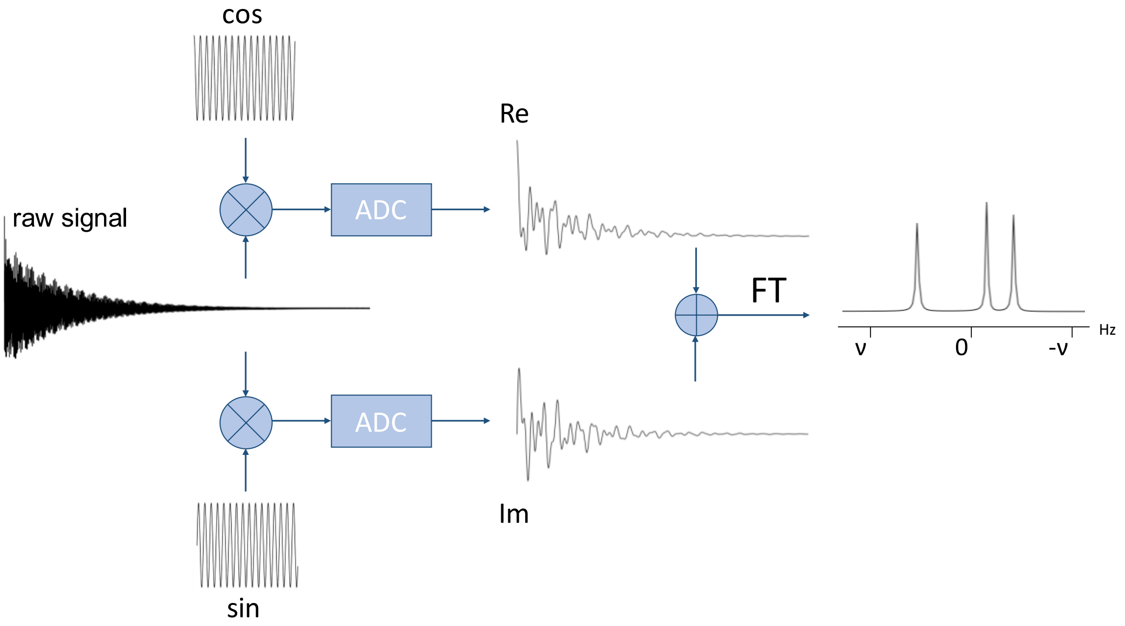


Figure 1.2.2. Conversion of the raw NMR signal into a spectrum. ADC stands for Analog-to-Digital Converter and FT for digital fast Fourier Transform.

Processing

Conversion from the time to frequency domain is most commonly done by Fast Fourier Transformation algorithms (FFT). Fourier transformation for discrete data is given by the formula:

$$s(\omega) = \sum s(t) * e^{-i\omega t}\Delta t \qquad (17.1) \qquad s(\nu) = \sum s(t) * e^{-i2\pi\nu t}\Delta t \qquad (17.2)$$

From the mathematical properties of the transformation, important characteristics of the dataset can be derived. The resolution of the frequency-domain data depends on the number of acquired points N and the increment Δt upon acquisition: $\Delta\nu = \frac{SW}{N} = \frac{1}{N\Delta t}$ with SW being the *spectral width*. It has been demonstrated by spectra modelling (Rovnyak *et al.*, 2004), that the optimal balance between resolution and signal-to-noise ratio can be achieved when the FID is sampled up to an acquisition time $T_{aq} = N\Delta t = 1.3\ T_2$. The problem of choosing the sampling length is acute in the indirect dimensions, where acquisition of each new point comes at a cost; however, processing of the direct dimension also benefits from omitting the points from the tail end of the FID.

Often prior to conversion from the time to the frequency domain, the signal is processed with various techniques to optimize the appearance of the resulting spectrum (Keeler, 2010). It is possible to extrapolate the time-domain data with *linear prediction* to increase the effective T_{aq} and thereby boost the resolution. *Weighting (*or *window) functions* are convoluted with the signal to improve signal-to-noise ratio (exponential or gaussian weighting) or resolution (most often squared cosine weighting).

1.2.2. Acquisition and processing with non-uniform sampling

Recording the indirect FID always requires a compromise between the experimental time and resolution. Given a linewidth of 0.5 ppm in ^{15}N (Zhou *et al.*, 2007, at 40 kHz MAS rate) and a range of 35 ppm for chemical shifts for the amide nitrogens in protein backbone, a total of 140 real and imaginary points of the indirect FID are needed to keep the digital resolution beyond the natural limit. Given a short 0.5 s `d1` relaxation delay and 8 scans per each FID, a 2D hNH correlation would take less than an hour to acquire (9 minutes in the assumed favorable conditions). Adding evolution in the ^{13}C dimension to the pulse sequence, a 3D hCBcaNH experiment with spectral window of 60 ppm in ^{13}C dimension would need about 14000x2^2 = 56000 points to resolve the homogeneous ^{13}C linewidth of 0.3 ppm (Zhou *et al.*, 2007), which corresponds to more than two days of measurement time. Full - *uniform* - sampling of spectra of even higher dimensionality becomes unfeasible: a similar 4D spectrum would take weeks,

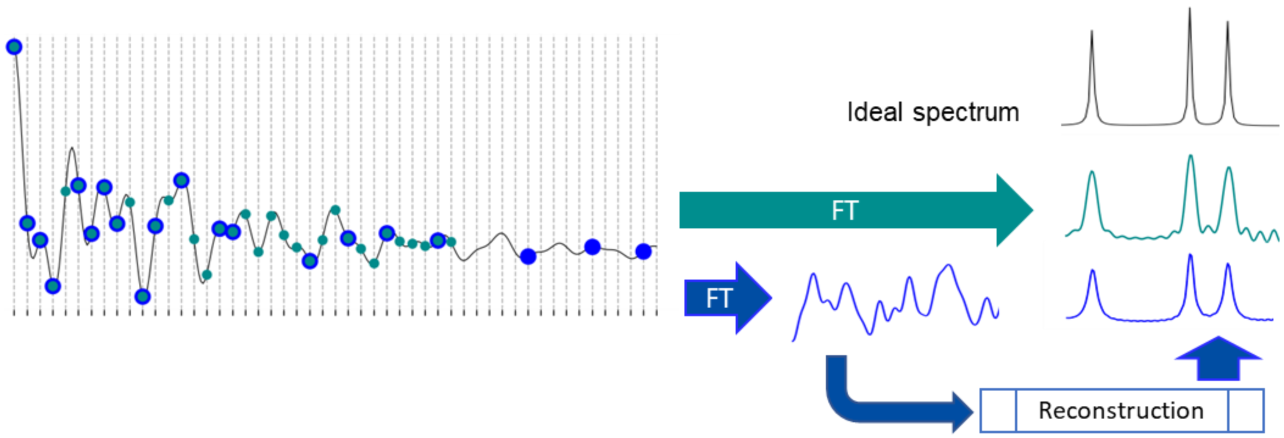


Figure 1.2.3. The concept and application of non-uniform sampling (NUS). Conventional processing of NMR data requires Fourier transformation of the FID, sampled on the Nyquist grid (teal). However, obtaining the fully sampling data in the high-dimensional experiments is heavily time demanding. A spectrum can be obtained from an FID even when not every point on the Nyquist grid is sampled (blue). Processed in a conventional way, the NUS data is heavily crowded with artifacts, arising from convolution of the signals and the sampling schedule. Artefacts are removed or suppressed by various reconstruction algorithms. The saved measurement time can be invested into more scans for better SNR and/or into acquisition of farther time points, thereby increasing resolution of the resulting spectrum. FT denotes conversion into the frequency domain regardless of the particular implementation of the Fourier transform (Fast, Multidimensional, etc.). Note that not every reconstruction procedure processes the frequency domain data as shown in this scheme and obtains it with a version of FT algorithm; however, this is the case for the majority of methods including the ones considered in this work.

and the time required for a fully sampled 5D rises to the order of years. The use of linear prediction methods to achieve better resolution as short acquisition times is limited (Stern *et al.*, 2002). This posed the incentive to develop techniques of acquisition and, more importantly, processing of *non-uniformly*, under-sampled FIDs.

A detailed overview of the immense literature body on non-uniform sampling (NUS) has been given by Mobli and Hoch (2014), Kazimierczuk and Orekhov (2015), Delaglio *et al.* (2017) and Robson *et al*. (2019).

General concept

As opposed to conventional acquisition of the indirect FID by uniform sampling of the Nyquist grid (Fig. 1.2.3, teal, dashed lines), non-uniform sampling (NUS) is a way to reduce the measurement time by acquiring only a fraction of points (Fig. 1.2.3, blue). The Fourier transform of the obtained NUS data will be a convolution of the NMR signal and the point spread function (PSF) of the sampling scheme (see examples in (Kazimierczuk and Orekhov, 2015) and in Fig. S1). Therefore, such datasets require special *reconstruction* techniques that extract the signals from the *artefact noise* produced by the sampling schedule. Analogously to the signal-to-noise ratio, the signal-to-artifact ratio S/A scales with the number of sampled points N,

$S/A = \sqrt{N}$. There is no definitive general solution to the question how low N may be in a particular experiment; yet there are a few considerations to be made. According to the general theory of non-uniform sampling proposed by Landau in 1960s, the average sampling rate must be at least twice the occupied bandwidth of the signal (Landau, 1967). This means that the number of points to be sampled depends rather on the number of signals and the linewidths rather than on the spectral window or desired resolution. In the ideal case of a signal with no noise ($SNR \to \infty$) and pure Lorentzian peak shape, the number of points needed to be sampled in the time domain is exactly equal to the number of expected signals. Without any assumptions of the peak shape, a very sparse NMR data of size N with K expected peaks can possibly be reconstructed from

$$m \geq \; cK \log\left(\frac{N}{K}\right) \tag{18}$$

points, where c is a small empirically determined constant (Sun *et al.*, 2015). This formula is given by the theory of compressed sensing (Donoho, 2006) and a successful reconstruction of a dramatically undersampled (0.8% of the total number of points) data of high sparsity has been first demonstrated in the proof-of-concept work of Hyberts *et al* (2012). In practice, sufficient data sparsity for protein spectra is achieved only in high-dimensional, especially ≥ 4D experiments. A general rule of thumb is that good results could be achieved while sampling 20 % of points per dimension (Hyberts *et al.*, 2014). The optimal sampling density and schedule, however, have to be carefully considered for each individual case.

Optimization of sampling patterns and development of faithful reconstruction algorithms are the two intertwined aspects of the development of NUS techniques.

Sampling schedules

The attempts to reduce the number of points sampled from the indirect FIDs can be traced back to the 1980s. Already in that decade, two classes of approaches emerged. The earliest ACCORDEON experiments (Bodenhausen and Ernst, 1981) were based on sampling projections of the multidimensional spectrum at different angles (radial sampling schemes); this technique inspired future development of Automated Projection SpectroscopY (APSY) (Hiller *et al.*, 2005). Several other pattern-based sampling schemes like radial or spiral sampling (Kazimierczuk *et al.*, 2007) were proposed in an attempt to improve performance of Fourier transform of NUS data. The other take on non-uniform sampling is recording randomly selected points both on and off-Nyquist grid (Robson *et al.*, 2019). Performance of the uniformly random schedule ranges greatly from one set of sampled points to another (Robson *et al.*, 2019), which makes it rather improbable to find the optimal combination and provides an incentive to find optimal constrains. The obvious idea is that the contribution of noise can be reduced if sampling density is skewed towards the beginning of the FID where the signal intensity is at maximum. The first weighting scheme used an exponential distribution of the point density that followed the general exponential decay of the signal (Barna *et al.*, 1987). It was shown later that some reconstruction schemes benefit from a more uniform distribution of gap lengths between the sampled points. This problem was solved by restricting the gap lengths to Poisson distribution (Hyberts *et al.*, 2010).

The choice of the sampling schedule depends on the reconstruction algorithm of choice as well as on the properties of the resulting data, such as peak density or dynamic range.

Reconstruction algorithms

Reconstruction of the full dataset on the basis of undersampled data is a task similar to finding the optimal solution to an underdetermined system of equations. Identification of the solution that is the most likely correct requires making some assumptions about the underlying data, for example, about the data sparsity or the line shapes. All reconstruction algorithms are typically iterative and terminate either upon reaching a specified stopping criterion, such as a maximum number of iterations or convergence based on accuracy or change in the objective function.

The reconstruction algorithms used in this work are described in the following.

Iterative Soft Thresholding implemented by Harvard Medical School (hmsIST)

Iterative thresholding methods originate from the field of image processing (reviewed in (Mobli and Hoch, 2014)). The general approach is simple: the NUS FID, with the missing data points being set to zero, is converted to the frequency domain with discrete Fourier transform; all points that exceed the given threshold are recorded into memory and scaled down (Iterative Soft Thresholding, IST) or set to zero (hard thresholding). The remainder is then converted back to the time domain, the non-recorded data points are set to zero again, and the procedure is repeated.

Application of IST to NMR data was first suggested by Drori (Drori, 2007). It used wavelet transform for obtaining frequency domain data. In 2012, Harvard Medical School reimplemented the algorithm (Hyberts *et al.*, 2012) replacing the wavelet transform with the more time-efficient Fast Fourier Transform and its inverse. Thus, the time needed per iteration of hmsIST is essentially determined by the time complexity of FFT and scales as $\mathrm{N_n} \log N$ for N number of points in n dimensions.

There are three parameters of the algorithm that may be potentially adjusted: the height of the threshold, termination condition (value of the l2 norm) or the number of iterations. However, the first value is defined by the developers (threshold of 98 %) and the necessary and sufficient number of iterations for reconstruction of weakest signals was shown to be 250 (Hyberts *et al.*, 2012). Termination of the program is then achieved by reaching the maximum number of iterations, and setting the value of l2 norm is not required.

Signal Separating Algorithm - SSA

The Signal Separating Algorithm was proposed by J. Stanek and W. Koźmiński, (2010, 2012) and is based on the CLEAN scheme initially used in radioastronomy (Högbom, 1974) and later reimplemented for NMR data (Kazimierczuk *et al.*, 2007; Coggins and Zhou, 2008). SSA operates with the data being transformed into the frequency domain in all dimensions. Identification and separating the signals from the residual is performed iteratively by identifying the potential signals in the crude spectrum and subtracting them one by one to yield the residual time-domain data. Since initial signal identification relies on statistical methods, it is crucial that the artifacts from the sampling would be distributed randomly, which can be provided only by a random schedule (with exponential or cosine weighting of sampling density, which does not create sampling artifacts and is simply equivalent to applying a window function). The identified peaks are considered one by one in the individual frames. The points in each frame that exceed the user-set threshold are then fitted with a Lorentzian, each point of which is then corrected to reflect the true peak shape (or shapes of overlapped peaks); the obtained function is then converted to the time-domain and subtracted from the residual. The use of the analytical model and its subsequent correction is the cornerstone of the SSA for the efficient extraction as well as artifacts removal. The algorithm terminates once no more peaks can be found. The performance of the algorithm depends on the number of peaks in the spectrum and on the parameters of peak identification.

Sparse Multidimensional Iterative Lineshape Enhanced (SMILE) reconstruction

Of the three algorithms considered here, SMILE (Ying *et al.*, 2017) is the most recent one. The idea behind SMILE is very similar to SSA, however, the algorithms differ substantially in implementation. Like SSA, SMILE performs peak identification on the fully frequency-domain dataset and fits peaks individually in their frames. However, it assumes the signals to be of Lorentzian shape in the frequency domain and works with the spectrum, fully converted into the frequency domain. SMILE assumes all peaks to be purely Lorentzian and have purely absorptive phase (i.e. all signals are cosine oscillation decaying exponentially). The purely random schedule was found to outperform exponentially-weighted random sampling as well as all types of Poisson-weighted schedules.

Outline

Non-uniform sampling can be used to save measurement time and invest it into a longer indirect acquisition for better spectral resolution or to increase number of scans thereby increasing signal-to-noise (Fig. 1.2.3). It enables access to high-dimensional experiments, which would be either extremely time consuming (4D) or not possible at all (5Ds and higher).

It is however, not trivial to choose the best sampling and reconstruction method without trial and errors, since their performance strongly depends on the characteristics of NMR data and correspondingly, the nature of the sample.

To the author's best knowledge, there is yet no formal theory describing and quantifying performance of various NUS experiments nor is there a comprehensive overview. A large step into this direction was taken by the recent community effort and release of the NUScon (Nonuniform Sampling Contest) platform (Pustovalova *et al.*, 2021). This project provides tools for simulating solution-state-like NMR data sampled in a desired way and evaluates performance of reconstruction algorithms by several quantitative metrics.

There is no overview concerning performance of applicability of NUS to the complex cases of solid-state NMR data such as spectra of samples with high heterogeneous broadening. Test of the three schemes outlined above on a model heterogeneous solid-state sample is a part of this work. This project is presented in *Results, Section 2.1*.

1.3. Protein structural biology

Among all biological macromolecules, proteins perform by far the largest range of biological functions. Proteins constitute the intricate molecular machinery behind all processes in cells, allowing them to grow, reproduce and differentiate, forming tissues, organs, and organisms. Proteins perform the skeletal and motoric functions, maintain the pH balance, regulate energy metabolism, cell division, and enable synthesis of other proteins by executing the genetic and epigenetic code. The interplay of cellular processes is both amazing and overwhelming, and with all progress in medicine and biology, we are but scratching the surface of understanding the organization of life. The detailed knowledge of underlying molecular mechanisms is crucial for very practical applications, such as design of efficient and safe bioactive compounds, such as drugs or pesticides, and biotechnological processes.

Proteins are biological polymers whose monomers (in the naturally occurring systems) are amino acids of twenty types (*Supplement*, Fig. S2). The function of a protein is the direct result of its spatial fold, the distribution of electrostatic charge, and site-specific dynamics. This makes protein structure and dynamics of particular interest for fundamental and applied science.

1.3.1. Protein order and disorder

Levels of organization of polypeptide chains

Proteins adopt up to 4 levels of organization.

Primary structure, or amino acid sequence is the order in which amino acid residues build the chain. Amino acid residues are connected with each other by a peptide bond. By convention, residue count starts from the residue with the free -NH_2 group called the N-terminus and continues until a free carboxylic group – the C-terminus. The chain of the amide group, aliphatic carbon Cα and the carbonyl group constitute the *protein backbone* and the substituent at the α-position is referred to as the *side chain* (Fig 1.3.1).

Secondary structure is defined by the spatial arrangement of the protein backbone. The two major parameters of secondary structure are two backbone dihedral angles:

φ: the twist around the N-CA bond, defined as CO_{i-1}-N-Cα-CO,

and

ψ: twist around the Cα -CO bond, defined N-CA-CO-N_{i+1}

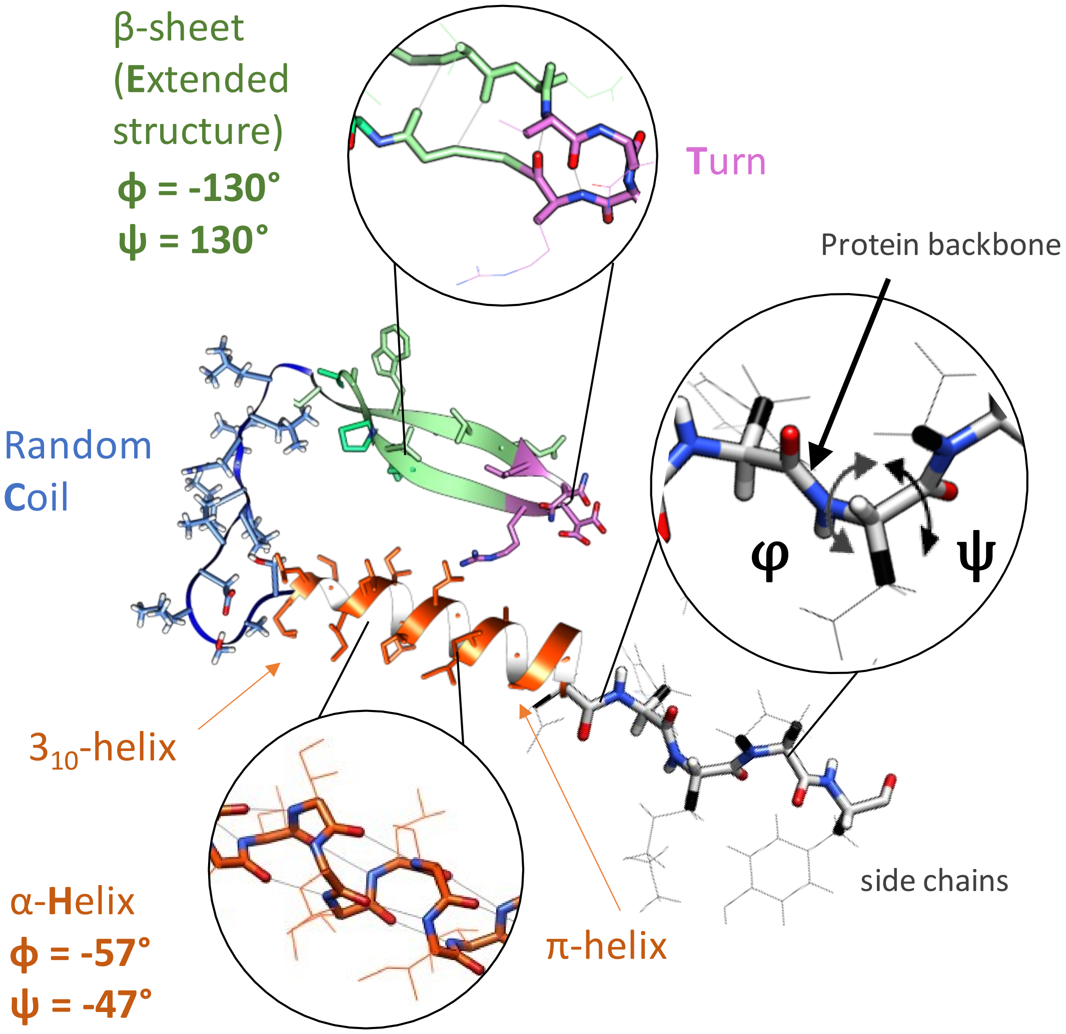


Figure 1.3.1. Organization of the amino acid chain into various secondary structure motifs. Hydrogen bonds shown on the inserts as gray lines. Highlights in bold font are the commonly used one-letter codes of the secondary structure classes (for example, in DSSP (Kabsch and Sander, 1983) and STRIDE (Frishman and Argos, 1995) classification). The figure was prepared using UCSF Chimera (Pettersen et al., 2004) as a compilation of hand-modeled fragments (helices) as well as fragments of real proteins (PDB IDs 2NUZ, 2FMC).

illustrated in Fig. 1.3.1. Rotation around the peptide N-CO bond is restricted because of the resonance interaction of sp^2-hybridized electrons of the carbon in the C=O group and the free electron pair on the nitrogen atom, therefore the third dihedral angle, ω is always close to 180°.

Classification of the secondary structure elements is based on the combination of hydrogen bonding pattern and the combination of the φ and ψ angles. In 1963, G. N. Ramachandran suggested visualization of the energetically allowed (φ, ψ) combinations on a two-dimensional diagram (Ramachandran *et al.*, 1963), and ever since, (φ, ψ) distributions of any kind (for residues of a particular protein or a statistical summary) bear his name. Due to the L-chirality of the nineteen natural proteinogenic amino acids, right-hand winding (negative φ values) is by far more preferential (Fig. 1.3.2). Glycines are unrestricted in rotations due to the absence of a side chain. Local conformational preference can be influenced by the type of the neighboring residue, most prominently, prolines (Fig. S3A), which occupy a particularly restricted conformational space (Fig. S3B).

There are two major categories of organized secondary structures: helices and extended strands (Figure 1.3.1, red and green). Among all studied proteins, helical structures are the most prevalent. The most abundant helical configuration is termed *α-helix*, first described by Linus Pauling *et al.* in (1951). Its key characteristic is hydrogen bonding between backbone N-H groups and a C=O group four residues along the chain ($i \rightarrow i+4$). One turn of α- helix takes on average 3.6 residues, i.e. each residue provides a 100° turn. The backbone typically adopts conformations around $\varphi = -60°$ and $\psi = -45°$ (Fig. 1.3.2). Stretching and contraction of the helix varies the (φ, ψ) combination along the line with a slope=1 on the (φ, ψ) plot, leading to the formation of other, rare helical forms. The stretched form is referred to as *3_{10} helix* denoting 3 residues per turn with 10 atoms being involved; this configuration creates $i \rightarrow i+3$ hydrogen bonding patterns. The contracted form, *π-helix*, is formed by 4.1 residues and is characterized by $i \rightarrow i+5$ H-bonding. Those rare structures are usually not longer than 7 residues and therefore cannot be described by a single combination of dihedral angles. Extended conformations can form individual flat β-strands or combine into β-sheets distributed in the range of $\varphi \in (-150°, -100°)$, $\psi \in (100°, 150°)$. Strands in β-sheets can orient in parallel or anti-parallel, which influences the preference of the (*φ, ψ*) combination: residues in the more common anti-parallel orientation are typically "flatter" with higher absolute angles. The individual strands in the sheets can sometimes be connected with very short fragments of the chain forming a *turn* and are often secluded into a separate class, when the two ends of the turn are bridged with an H-bond. Turns can be further subclassified with respect to the conformations of the involved residues. As turns are composite structures, no specific dihedral angle region can be attributed. Any structures that lack regularity and do not form regular patterns are referred to as *loops* or *coil*. A *very flexible chain* which randomly samples the entire thermodynamically allowed conformational space is called *random coil* (although a variety of definitions have been used in the literature depending on the context (Mielke and Krishnan, 2009). The Ramachandran map of random coils (Fig. S4) demonstrates the intrinsic tendency of polypeptide chains to adopt extended conformations. It has been shown, that in those unrestricted structures, residues of different types have different intrinsic preference for φ angle (Serrano, 1995). Conformations in the regions around φ = -75°, ψ = 150°

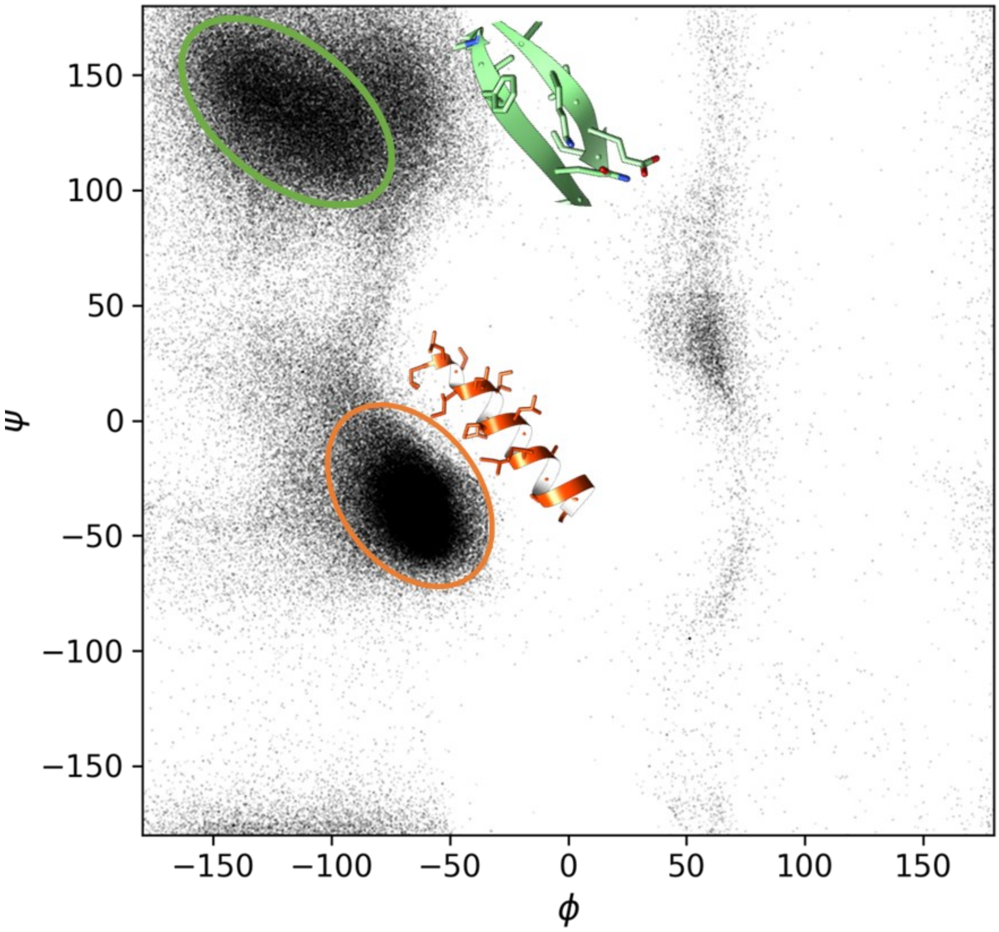


Figure 1.3.2. Ramachandran map of the amino acid residues of all proteinogenic types (except Gly). Data taken from the PACSY database (Lee *et al.*, 2012) (accessed on Dec., 2022). The major secondary structure motifs – helices and sheets – are depicted next to the regions of typically occupied backbone dihedral angles.

(the region of the major point density in Fig. S4) are often referred to as poly-proline II (PPII), named after the left-handed helices formed by poly-proline chains.

Formal assignment of secondary structure elements is important in some bioinformatical applications and scientific communication. Classification algorithms, such as DSSP (Kabsch and Sander, 1983) or STRIDE (Frishman and Argos, 1995), were designed to formalize the intuition of X-ray crystallographers. The classification algorithm is based on the combination of the backbone dihedral angles, arrangement of hydrogen bonds as well as the context between the amide and carbonyl groups N-H···O=C.

Tertiary structure is defined by the spatial arrangement of the secondary structure elements. It is usually stabilized by hydrophobic, staking and π–π interactions between the sidechains, as well as covalent disulphide bonds, ionic salt bridges, and hydrogen bonds (H-bonds).

The term *quaternary structure* refers to the assembly of the individual folded protein units into a molecular complex, the full functioning units of the molecular machinery. Proteins and protein complexes with catalytic function (*enzymes*) often require prosthetic groups: nonprotein elements like small organic molecule ligands (like flavin mononucleotide in oxidoreductases), metal ions (e. g., Zn^{2+} in carbonic anhydrases) or clusters (e. g., iron-sulfur clusters in oxidoreductases) or metalorganic ligands (e. g., chlorophyll in photosystems or heme in hemoglobins).

One sequence – one structure?

In the still-early stage of structural biology, it was widely assumed that the most, if not all, proteins possess well-defined structures (Epstein *et al.*, 1963). This assumption was based primarily on the early successes of protein X-ray crystallography and the realization that all at that time discovered proteins had unique amino acid sequences. Successful refolding of proteins from denaturating conditions by Anfinsen and colleagues lead to the *thermodynamic hypothesis* stating that the fold, natively adopted by a protein in its native physiological environment (including the presence of prosthetic groups, correct temperature and pH, etc.) is the one at which the free Gibbs energy of the whole system is lowest. This hypothesis further converged to the statement expressed in Anfinsen's Noble acceptance speech (Anfinsen, 1972): "The native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment", which sparked the endeavors of purely sequence-based predictions of protein fold. Despite a long track of largely underwhelming performance of such methods, advancements in machine learning enabled the very recent breakthrough in protein folding: the neural networks-based systems like RoseTTAFold and AlphaFold succeeded in predicting the lowest-energy structures of globular folds of single-domain proteins or even homo-multimers at an unprecedented accuracy (Baek *et al.*, 2021; Jumper *et al.*, 2021; Subramaniam and Kleywegt, 2022). The interactions between the domains, protein subunits and their ligands are only to be addressed in collaboration with experimental methods in the foreseeable future.

An environment often defines not only *if* the given sequence adopts a defined special fold but also *which* fold is adopted. Deviations from the major folding pathway lead to formation of protein aggregates, where the individual molecules can be either totally disordered or adopt a β-strand or -sheet form in amyloids – stable, insoluble filaments or fibrils. The ability to form amyloids seems to be natural for polypeptide chains since any protein, peptide or even single amino acids can apparently be converted into the amyloid form *in vitro*. This is also supported by the Ramachandran plot of random coils: the disordered flexible chains tend to adopt extended conformations (see above). At the same time, helical structures are prevalent in the known proteome, which can be attributed to the natural selection of proteins with higher helical content due to their lower amyloidogenic potential. *Polymorphism* (presence of fibrils of different architecture) and *heterogeneity* (disorder of the protein chain within a single fibril) of amyloids is the natural consequence of the physico-chemical properties of the protein chain and high energetic similarity between multiple arrangements. Amyloidogenesis *in vivo* is almost always associated with pathologies, but in some cases, it has been recruited by organisms, including bacteria, fungi and even humans, to produce functionally significant structures (Chiti and Dobson, 2006, 2017).

Whereas the thermodynamic hypothesis probably generally holds true for a large number of proteins, it has been challenged by discovery of intrinsically disordered proteins (IDPs) – the chains whose native conformational space has several free energy minima and low energy barriers between them. IDPs together with intrinsically disordered protein regions (IDRs) are found to constitute as much as half of the entire eucaryotic and viral proteome (Perdigão *et al.*, 2015). Intrinsic disorder gives proteins the ability to form weak yet highly specific complexes, which is important for regulatory pathways, where turning a signal off is as important as turning it on (Habchi *et al.*, 2014). Thus, IDPs and IDRs are essential elements of transcription factors (Sammak and Zinzalla, 2015), voltage-gated channels (Kjaergaard and Kragelund, 2017) or protein phase separation (Turoverov *et al.*, 2019).

It is yet unclear if the purely computational methods would be able to reliably predict the dynamic conformational ensembles and, furthermore, structures and order of aggregates. Further development of protein structural biology will rely profoundly on the synergy between experiment and prediction (Subramaniam and Kleywegt, 2022).

Disorder of ordered chains

Even within the well-defined highly ordered systems, like crystallized, well-folded globular proteins, the peptide chain is not static and exhibits constant switches between local energy minima on the ps-ns timescale. In the side chains, three rotameric forms exist for every single C-C bond. The most known type of backbone motions is the peptide flips in the tight turns (where residues *i*-1 and *i+2* are typically H-bonded), forcing interconversion between the βI/βII turn types. In more flexible regions not constricted by H-bonds, many other types of flips can occur, involving those perturbing and non-perturbing the neighboring peptide planes (Hayward, 2001). Collectively, peptide flips may synchronize into so-called crankshaft motions (Fadel *et al.*, 1995) propagating along the polypeptide chain. The backbone flips are clustering at four conformations, alternative to the ensemble-average structures solved from the electron density data (Fig. 1.3.3). According to Keedy *et al.*, (2015), the states are formed by two variants of rotations: two types of 180° flips (Fig. 1.3.3A), typically occurring in turns, and rotations by ±120° about the $C\alpha_i$-$C\alpha_{i+1}$ axis (Fig. 1.3.3B) occurring in ordered but irregular secondary structures. This analysis was on the electron densities from the Top8000 dataset, the collection of highest quality X-ray structures, performed with qFit 2.0. All four classes of flips often (>60 % of the cases) involve glycine as the second residue of the flipping peptide fragment. Other internal backbone motions include 'backrub' flips, typically around $C\alpha_{i-1}$-$C\alpha_{i+1}$ axis (Davis *et al.*, 2006), and shear motions in helices (Hallen *et al.*, 2013).

The fast internal peptide motions or peptide backbone flips are presumed to be the driving force of the motions on the larger (µs-ms) timescales, such as cross-correlated "flaps" in HIV-1 protease (suggested from NMR relaxation data by Nicholson *et al.* (1995) and newly found by Keedy *et al.* (2015) in the electron density map), "breathing" or domain motions (Mariño Pérez *et al.*, 2022) (note the profound example in (Mariño Pérez *et al.*, 2022) of the breathing motion of a "rigid" SH3 domain, caused by a tyrosine ring flip).

As will be discussed in the following *Section 1.3.2*, the conformational heterogeneity on either fast or slow timescales contributes to the NMR observables, such as relaxation rates or, most notably, chemical shifts. In the NMR experiments at cryogenic temperatures, slow motions are immobilized, chemical shifts of all conformers are observed at once, broadening the signals. Chemical shifts of the fast-exchanging conformers, however, still average out. These fast motions are likely one of the factors contributing to inaccuracy of structure-based chemical-shift predictions and vice versa, structure predictions from isotropic shifts. On the contrary,

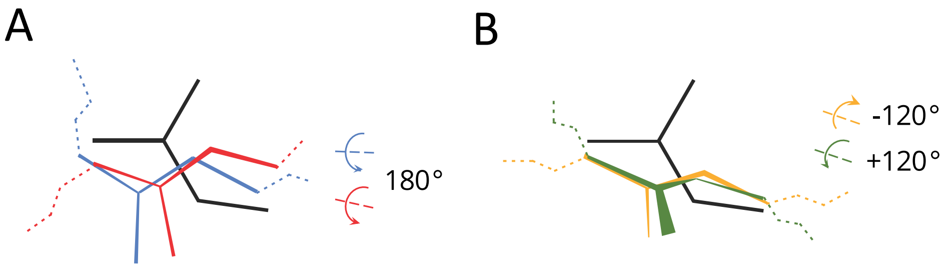


Figure 1.3.3. Schematic representation of the four classes of backbone peptide flips identified by analysis of electron densities of high-quality, high-resolution (<2 Å) X-ray structures. The flips occur in ordered polypeptide regions with irregular secondary structure. The reference peptide conformation is shown in black. **A**: Two ~180° rotated clusters with different translations in the peptide plane (blue vs. red). **B**: Two other clusters rotated by +120° (green) and -120° (yellow). Thickness of the line schematically represents the proximity of the region. Dotted lines represent the average direction of the backbone outside the flipping fragment. Adapted from Fig. 2 from (Keedy *et al.*, 2015) under Creative Commons Attribution 4.0 International License.

hybrid quantum mechanics/molecular mechanics/molecular dynamics (MD-QM/MM) methods are shown to calculate chemical-shift tensors with high accuracy (Kraus *et al.*, 2020).

1.3.2. NMR in protein structure determination

NMR enjoys wide recognition among other high-resolution techniques of structural biology. Although solving the protein structure solely by NMR takes much more human effort than calculating it from the X-ray diffraction or electron microscopy, NMR is indispensable in studying small proteins, protein-protein and protein-ligand interactions, site-specific dynamic information, and to approach significantly heterogeneous systems.

NMR does not observe the structures directly – instead, experimental geometrical parameters of the molecules are obtained from NMR observables and then further submitted to the structure calculation protocols. Distance information is routinely obtained from the variety of experiments exploiting dipole-dipole interaction: NOESY in liquid samples and REDOR, RFDR or other recoupling experiments in solids (Nielsen *et al.*, 2011), measurements of residual dipolar couplings (RDCs) in colloidal, partially aligned samples, among others. In addition, positions of specific atoms can be determined relative to a paramagnetic label introduced to the sample by the effect of paramagnetic relaxation enhancement (PRE). Dihedral-angle information can therefore be derived from the distances or obtained directly from the scalar couplings through Karplus equations (Karplus, 1959). The measurements of angles and distances are often limited to relatively small, homogeneous proteins since spectral crowding and peak broadening in more complex systems do not allow for sufficient spectral resolution and reduce signal-to-noise.

Structural and fast-timescale dynamic information is reflected in the observed shielding tensors and this gets encoded into chemical shifts. Accurate deciphering of the chemical shifts is thus an immensely complex problem due to the multiple intertwined factors affecting the nuclear resonance frequency. However, many patterns and dependencies have been identified, which allowed to translate chemical shifts into structural restraints (Nerli *et al.*, 2018) and even for obtaining high-resolution structures of small proteins based purely on chemical shift information (Robustelli *et al.*, 2008; Wishart, 2011; Berjanskii and Wishart, 2017). Interpretation of chemical shifts has been approached using *ab initio* computational methodology, statistical analysis, and hybrid techniques.

Factors affecting backbone chemical shifts

When molecules in liquid media are allowed to freely tumble in random directions, or when this tumbling is emulated for solid samples by fast spinning of the sample at the magic angle (see *Sections 1.1.1* and *1.1.3*), the anisotropic components of the chemical-shift tensor average into the isotropic value. Although anisotropic interactions bear useful information about local geometry, extracting isotropic chemical shifts is much more straightforward in solutions since it does not require additional sample preparations in anisotropic media. In solids, averaging of anisotropic observables is also the preferred condition in all residue-specific studies of the long polypeptide chains, since anisotropic peak broadening tremendously complicates the spectrum.

The isotropic chemical shift of a nucleus is influenced by several factors, which can be classified into four categories (Williamson and Asakura, 1997): (i) the close-range modulation of shielding by the electrons within a few bonds away, which depends on the covalent structure and local dihedral angles; (ii) the long-distance effects arising from van-der-Waals interactions and bond polarization effects; (iii) ring current effects and anisotropies of the neighboring bonds, and (iv) paramagnetic effects in case any unpaired electrons are present. In terms of structural effects, it can be broadly summarized as (Han *et al.*, 2011):

$$\delta = \delta_{rc} + \delta_{SS} + \delta_{\chi} + \delta_{el} + \delta_{HB} + \delta_{solv} + \delta_{ring} \tag{19}$$

Here, δ_{rc} denotes the random coil chemical shift, the “baseline” determined by the types of the residue of interest i and its direct neighbors $i+1$ and $i-1$. The term δ_{SS}, denoting the contributions of secondary structure, should be further broken down to dissern the influence of the conformation of the current, preceding, and succeeding residues as

$$\delta_{SS} = \delta_{SS}^{i} + \delta_{SS}^{-} + \delta_{SS}^{+} \tag{20}$$

The term δ_{χ} denotes contribution of the rotameric state of residue i. Terms δ_{el}, δ_{HB} and δ_{solv} describe the electrostatic effects, where the hydrogen bonding and solvent effects are special cases which are commonly secluded into the separate terms δ_{HB} and δ_{solv}. Finally, δ_{ring} describes ring current effect in case any aromatic rings are present. All the terms of Eq. 16 are rather conceptual than physical and often overlap.

Each of these effects contributes to chemical shifts of different nuclei with various weights, summarized in Table 1.2 according to the analysis presented in (Han *et al.*, 2011), the core publication of the hybrid chemical shift prediction method SHIFTX2 (see below). As such, excluding the random-coil term, the proton shifts $\delta^{1}H_{N}$ and $\delta^{1}H\alpha$ are influenced by a variety of factors, making them the most difficult to interpret and predict. On the contrary, $\delta^{13}C\alpha$ and

$\delta^{13}C\beta$ are predominantly affected by the conformation of the residue *i,* which makes them robust reporters on secondary-structure propensity. As such, Fig. S5 demonstrates this effect on statistical data for each residue type.

Table 1.2. Relative structural contributions to the protein chemical shifts of the given nuclei. '+++' denotes the key contribution (≥60%), '++' - medium effects (≥10%), '+' the minor effects (≥1%) and '-' denotes contributions that are mainly negligible for this class of atoms (<1%). Data taken from (Han *et al.*, 2011); contributions of δ_{rc} are excluded from the summary since they are relevant for peak assignments and not the structural analysis.

Effect	^1H_N	^{15}N	$^{13}C\alpha$	$^{13}C\beta$	$^{13}C'$	$^1H\alpha$
δ_{SS}^{i}	++	++	+++	+++	++	++
δ_{SS}^{-}	++	++	+	+	+	+
δ_{SS}^{+}	+	+	+	+	++	+
δ_{χ}	-	++	++	++	++	+
δ_{el}	+	-	-	-	-	++
δ_{HB}	++	+	+	+	+	+
δ_{solv}	-	+	-	+	+	-
δ_{ring}	++	+	+	+	-	++

The influence of hydrogen bonding parameters and patterns on chemical shifts has been extensively investigated by statistical and computational studies, yielding several empirical models (Fig. 1.3.4A). Generally, the shifts are negatively correlated with hydrogen bond lengths. The data collected on the crystalline amino acids (McDermott and Ridenour, 2002) formed a curve parametrized by Harris and Mildvan (1999) as a mixed logarithmic and linear dependance of proton chemical shift and the O--H-O distance (Fig. 1.3.4, gray line). Density functional theory (DFT) calculations of backbone amide proton shifts for spider silk (P. Holland *et al.*, 2013) formed a clear inverse cubic dependence on the H--O distance where coefficients differ for extended and 3_{10} helical structures (Fig. 1.3.4A, black solid and dashed lines). A thorough study conducted by Parker *et al.* (2006) used *ab initio* computations to explore behavior of proton chemical shifts for protons in different hydrogen bonding configuration, such as structures with and without secondary and tertiary hydrogen bond partners of various nature (water, amide or a charged side-chain group. The model was developed on the experimental chemical shift and geometrical data on two proteins (protein G, PDB ID: 1IGD, and human ubiquitin, PDB ID: 1UBQ) and demonstrated accuracy of 0.3 ppm, outperforming the tested prediction frameworks (SHIFTS (Xu and Case, 2001), SHIFTX (Neal *et al.*, 2003) and PROSHIFT (Meiler, 2003)). The model (Fig. 1.3.4A, the blue curve) includes increments depending on the secondary and tertiary bonding partners $\Delta\delta^1H(2°HB)$ and $\Delta\delta^1H(3°HB)$, backbone dihedral angles (combined into distance r_{ω} and angle ω as defined in Fig. 1.3.4; ω is not to be confused with the backbone dihedral angle denoted with the same letter, *Section 1.3.1*) well as rotation ρ of the acceptor and H-bond length r_{OH} and angle θ (definitions from (Parker *et al.*, 2006) are reproduced in Fig. 1.3.4B):

$$\delta^1 H = \delta^1 H(r_\omega, \omega) + \Delta\delta^1 H(r_{OH}, \theta, \rho) + \Delta\delta^1 H(2°HB) + \Delta\delta^1 H(3°HB) + \Delta\delta^1 H_{rc} \qquad (21)$$

As shown in Fig. 1.3.4A, C and D, the H-bond length r_{OH} and angle θ provide the greatest influence on the baseline shift defined by the first increment $\delta^1 H(r_\omega, \omega)$. The authors emphasize the importance of energy minimization of the X-ray structures after protonation for the reliability on the chemical shift calculations. The model formed the basis of the program ProCS (Christensen *et al.*, 2013).

The available models of the effect of hydrogen bonding on the amide ^{15}N chemical shifts primarily include only one parameter, the N--O interatomic distance, which estimates the length of the H-bond. The study of Kuroki *et al.* (1991) based on *ab initio* calculations of the backbone amide nitrogen shielding tensors in BocGly-containing dipeptides revealed the negative, non-linear relationship between the N--O distance and the absolute value of shielding with the strongest effect on σ_{11} component (defined as the least shielded direction). In addition, Kuroki *et al* obtained curves representing the effect of the H-bond angle on the nitrogen shift. Empirical formulas for neither relationship were suggested. The recent measurement of the principal shielding components by Paramasivam *et al.* (2018) did not confirm the non-linearity of the relationship and resulted in linear empirical models for α-helical and β-sheet structures for each principal component (the isotropic average is presented in Fig. 1.3.4E, top panel). Factors affecting ^{15}N shifts, including the N--O distance, were investigated by Xu and Case with DFT calculations. The contribution of the hydrogen bonding to the observed chemical shift was fitted using mixed exponential and hyperbolic curves (Fig. 1.3.4E, bottom panel), where coefficients depend on the role of the target residue (donor: 'direct' bond, black lines; acceptor 'indirect' bond, grey lines) and secondary structure. A combined model including fitting of other factors could predict ^{15}N shifts with root mean square difference of 2.27 ppm.

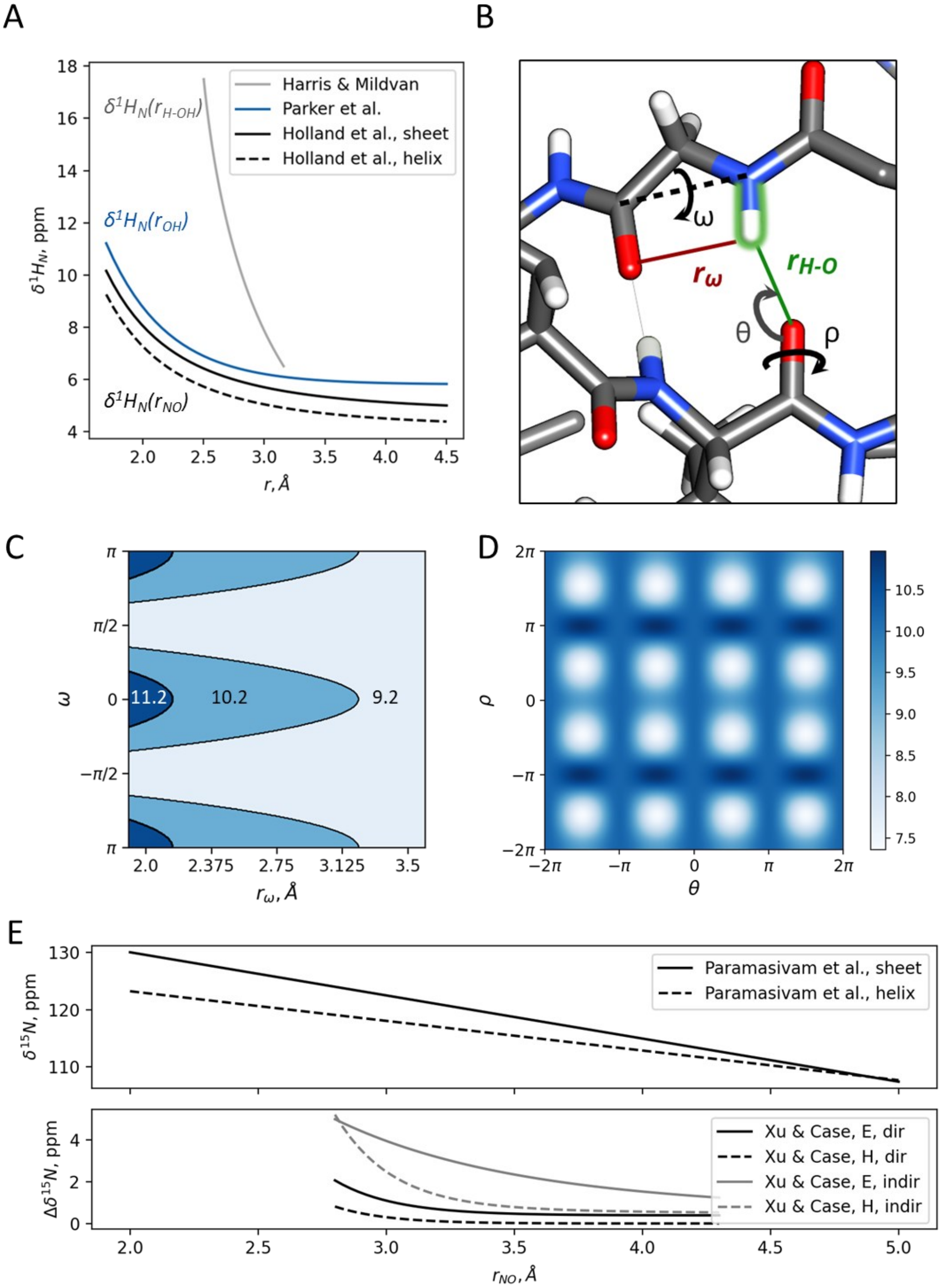


Figure 1.3.4. Empirical models of backbone amide chemical shifts δ^1H_N and $\delta^{15}N$. **A**: Comparison of the models for the amide proton shift as a function of H-bond length presented in the literature (Harris and Mildvan, 1999; Parker *et al.*, 2006; P. Holland *et al.*, 2013). In each source, the H-bond length *r* was defined based on the underlying data (see main text). **B**: Definition of the H-bond parameters of the model for δ^1H_N presented in Parker *et al.* (2006). The H and N in focus are highlighted in green. The angle H-N--C-O denoted with *ω* here and in the source is *not to be confused* with the third backbone dihedral angle Cα-C-N-Cα$_{i+1}$. **C** and **D**: Influence of the H – O distance r_ω angles *θ*, *ρ*, *ω* on δH_N according to the model of Parker *et al.* (2006). Note, that r_ω and *ω* are co-dependent and a large part of the (r_ω , *ω)*-space is energetically restricted. Color intensity corresponds to the chemical shift (in ppm). **E**: Influence of H-bonds on isotropic chemical shift $\delta^{15}N$ in helices and β-sheets. The models of Paramasivam *et al.* (2018) are calculated as an average of the models for the principle chemical shift components. Models of Xu & Case (2001) describe the difference between the shifts of the bonded and non-bonded nitrogen; the curves are calculated for the direct (O=C-N-H--O) and indirect (H--O=C-N-H) bonds in sheets ('E') and helices ('H').

Interpretations of chemical shifts

Discovery of correlation between chemical shifts and protein structure inspired emergence of a multitude of methods for predicting secondary structure and its derivatives (like protein flexibility or accessible surface area) from chemical shifts and vice versa, comprehensively reviewed in (Mielke and Krishnan, 2009; Wishart, 2011; Nerli *et al.*, 2018).

One strategy to develop understanding of the chemical shift-structure relationships is to try to predict the shifts from structural models and compare them with the experiment. The accuracy of quantum mechanical (QM) methods is greatly dependent upon the method and the level of theory used; their advantage is a particular sensitivity to the relative variations in the nucleus' environment, such as solvation and conformational changes (Sumowski *et al.*, 2014). However, the direct QM computation of protein chemical shifts is heavily demanding to the hardware and computational time. It motivated development of numerous hybrid methods, which rely on pre-computed libraries of chemical shifts that can be subsequently used to train the prediction models. Such approach is used by programs SHIFTS (Xu and Case, 2001), CheShift (Vila *et al.*, 2009) and ProCS (Christensen *et al.*, 2013). Despite a good precision of QM methods, i.e. consistency of the trends for small structural variations, their accuracy, i.e. prediction of the absolute shifts, has long been lower than accuracy of empirical methods, especially for the nuclei affected by electrostatic and H-bonding effects (Table 1.2). The inclusion of the dynamic factor in the simulations has been demonstrated to greatly improve accuracy (Dračínský *et al.*, 2013; Kraus *et al.*, 2020). With the advances in computational methodology and hardware, it can be foreseen that the QM methods will become more routinely used in many applications.

The empirical approaches combining sequential and structural information historically have outperformed the QM methods by both accuracy and speed. A strategy of combining classical equations and structure homology search was found to be efficient and was implemented in the most resent chemical shift predictors, SPARTA+ (Shen and Bax, 2010), SHIFTX2 (Han *et al.*, 2011) and UCBShift (Li *et al.*, 2020). SPARTA+ and UCBShift are powered by neural networks. Accuracy of the three methods was compared in Li *et al.* (2020) and found to be comparable with slight superiority of the newest UCBShift: RMSDs of the predicted shifts by UCBShift, i.e., H, N, C' Cα, Cβ, and Hα shifts, are on the level of 0.31, 1.81, 0.84, 0.81, 1.00 and 0.19 ppm, respectively.

The first methods tackling the inverse task of chemical shift-based structure predictions emerged in the early 1990s, as soon as the community accumulated sufficient amounts of data (Berjanskii and Wishart, 2017). Empirical approaches have generally been more successful in predicting the backbone structure than *ab initio* computations (Wishart, 2011). The first methods focused on qualitative assignment of secondary structure from the set of backbone chemical shifts such as chemical shift indices CSI (Wishart *et al.*, 1992; Wishart and Sykes, 1994) or Probability-based Secondary Structure Identification PSSI (Wang and Jardetzky, 2002) and others (reviewed in Mielke and Krishnan, 2009). Methods of prediction of protein flexibility, i.e. absence of defined secondary structure, consequently emerged on the basis of the aforementioned routines: as such, the chain flexibility can be estimated by the Random Coil Index RCI, (Berjanskii and Wishart, 2006). Machine learning opens the ways to find obscured patterns in the shift / structure relationships. A score calculated by the algorithm ShiftCrypt (Orlando *et al.*, 2020) is not as easily interpretable as, for example, CSI, but allows aligning structures with seemingly different chemical shifts impressively well.

The program TALOS (Torsion Angle Likelihood Obtained from Shift (Cornilescu *et al.*, 1999a) pioneered in "quantitative" predictions by providing the values of dihedral angles and has been followed by a variety of other software. TALOS is based on comparison of tripeptide fragments with the database by sequence and all available chemical shifts. This statistical analysis in TALOS' successors, TALOS+ (Shen *et al.*, 2009) and TALOS-N (Shen and Bax, 2013) is performed by two-layer neural networks; in addition to this and other differences, TALOS-N uses larger segments of the sequence and a large database of predicted chemical shifts. While TALOS-N, released in 2013, is to-date the latest and arguably most popular dihedral angle predictor, other implementations offer unique features and advantages. As such, PREDITOR (Berjanskii *et al.*, 2006) is able to identify and correct misreferencing of the submitted set of shifts (this function was later included in TALOS-N). DANGLE (Dihedral ANgles from Global Likelihood Estimates), unlike other algorithms, uses Bayesian inference for each (φ, ψ) combination on a fine (φ, ψ) grid with 10° resolution; the procedure resolves clusters of (φ, ψ) combinations ('islands') and allows to identify residues with multiple possible conformations, in particular reflected in its ability to accurately predict the (φ, ψ) angles for glycine and pre-proline residues (Cheung *et al.*, 2010; Wishart, 2011). TALOS-N and PREDITOR are able to predict side-chain angles χ_1. Performance of DANGLE, TALOS, TALOS+, PREDITOR (and two other algorithms not relevant for this review) were compared in (Wishart, 2011) on a set of 33 proteins by scoring parameters $A_{30}(\varphi/\psi)$ and $\Delta(\varphi/\psi)$. Both scores reflect the accuracy of the predictions: $A_{30}(\varphi/\psi)$

is defined as the percentage of residues where both predicted φ and ψ values fall within 30° of the observed values (in X-ray structures); $\Delta(\varphi/\psi)$ is the same for the sum of absolute differences between the predicted and observed values. PREDITOR showed the best results with $A_{30}(\varphi/\psi)$ = 94% and $\Delta(\varphi/\psi)$ = 85%.

It is important to note that due to inherent degeneracy of the shift-structure relationships, it is fundamentally impossible to develop a chemical shift-based method with absolute accuracy. According to D. Wishart, “any method claiming to achieve a prediction accuracy of >90% is essentially over-trained and under-tested” (Wishart, 2011).

1.3.3. Methods of studying protein disorder

Systems with large degrees of disorder are inherently challenging for any method of structural analysis. The “dimension” of heterogeneity, static or dynamic, encodes itself into the measured parameters, thereby requiring development and adoption of new experimental techniques and novel methods of data analysis.

Experimental toolbox of protein structural biology

NMR is one of the three methods in the toolset of structural biology that provide high-resolution models. Unlike NMR, the two other methods X-ray crystallography and cryo-Electron Microscopy (cryo-EM) determine the atomic-resolution structures by fitting the atoms into the 3D maps (electron density maps or the maps of electrostatic potential). The resulting models reflect the spatial and temporal average.

X-ray crystallography was the first method that enabled a close look at the atomic-level features of molecules. It determines the electron density by analyzing the diffraction pattern of X-rays, scattered by the crystal. X-ray crystallography still remains the dominant method of protein structure elucidation. To date (April 2023), 85 % of all protein structures deposited into Protein Data Bank (RCSB.org, Berman, 2000) were determined by X-ray crystallography. Over the years, it has been proven to be a robust and fast structure elucidation technique for all molecules and molecular complexes that are able to form large, highly-ordered crystals. Finding crystallization conditions is not a trivial task which often requires approaches of trial and error. X-ray powder diffraction (XRPD, reviewed in Spiliopoulou *et al.* (2020)) can circumvent the requirement of large crystals to provide medium-resolution (3-10 Å) structures, but it demands large amounts of sample for the single measurement and raises the demand for sample homogeneity. XRPD is usually limited to simple, single-molecule systems.

Cryo-Electron Microscopy uses high-energy electrons as a source to illuminate specimens in a very thin layer of vitrified ice using a transmission electron microscope. Cryo-EM exploits the strong interactions of electrons with each atom's Coulomb potential to image the sample as well as to produce its diffraction pattern. Structure elucidation from the experimental cryo-EM data relies heavily on advanced methods of image processing. The 3D cryo-EM density is reconstructed from 2D images of randomly orientated single molecules. Thus, while it does not require the sample to be crystallized, faithful demands high consistency between the individual particles (i.e. sample homogeneity) (Wang and Wang, 2017).

Several low-resolution techniques can provide a good initial estimate of the protein fold or the lack thereof. Scattering techniques such as dynamic or static light scattering (DLS and SLS), small angle X-ray scattering (SAXS, reviewed in Kikhney and Svergun (2015) as well as analytic chromatography techniques are useful to determine the particle mass, size, and shape, which can be further used to infer the protein's aggregation state in native-like conditions. Circular dichroism (CD) reports on the secondary structural content and is widely used to characterize residual secondary structure in IDPs.

Förster resonance energy transfer (FRET) and double electron-electron resonance (DEER) are particularly useful techniques to study the dynamic interactions between molecules in the complex or the different flexible sites within a single chain. Both techniques can not only provide the average property over the dynamic ensemble but deliver a distribution of the observable parameter, which facilitates reconstruction of conformational ensembles. Both methods require introduction of a pair of labels, attached typically to lysine or cysteine side chains. Additional medium-resolution insights regarding the protein fold and dynamics can be obtained with methods of hydrogen-deuterium exchange (HDX) mass spectrometry (James *et al.*, 2022).

Manifestation of protein disorder in NMR spectra

Among the high-resolution methods of structural biology, NMR is by far the most tolerant to the static and dynamic sample heterogeneity. Existing NMR techniques cover almost the entire range of timescales (Fig. 1.3.5), allowing to approach any system, from the highly dynamic to highly rigid ones on multiple levels of organization.

Manifestation of protein dynamics in the NMR spectroscopic data depends on the timescale of motion and experimental conditions, such as spectrometer base frequency and, in some solid-state applications, MAS rate *(Sections 1.1.2* and *1.1.3).* From the chemical-shift perspective, the timescale of the exchange between the two electronic environments of the moiety is defined by the difference in the electron shielding at the two states. In other words, the definition of the timescale of the chemical exchange depends on the exchange rate and the differences of the chemical shifts of the two states (expressed in Hz).

The 'fast' motions, where the difference in resonance frequencies is significantly lower than the exchange rate, lead to chemical-shift averaging. Changes in the buffer conditions or presence of an interaction partner affects the relative populations of the two states, thereby affecting the observed chemical shift. Titration methods can help identifying the residues involved in the exchange processes. Alternatively, purely NMR-based approaches can provide the details on the relative populations (DEST, dark state exchange saturation transfer (Fawzi *et al.*, 2011), or CEST, chemical exchange saturation transfer, (Vallurupalli *et al.*, 2012)) and the exchange rates (EXSY, Exchange Spectroscopy, reviewed in (Nikitin and O'Gara, 2019).

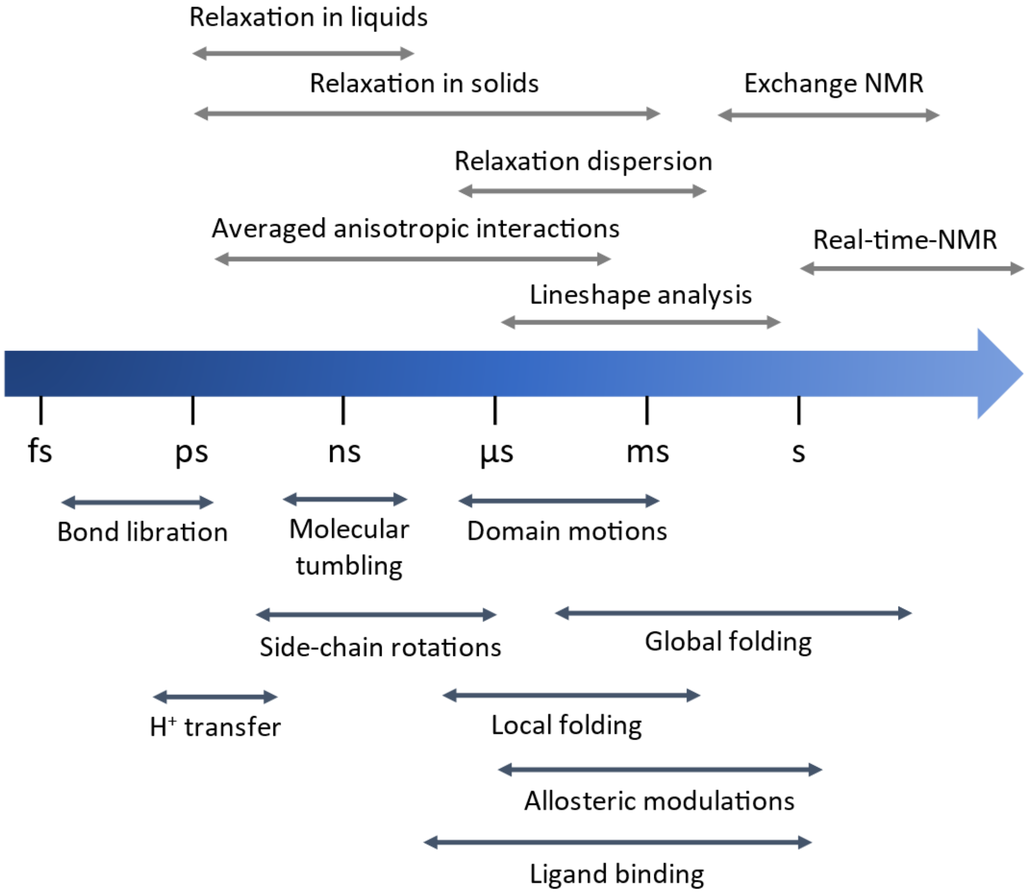


Figure 1.3.5. Timescales of protein motions and NMR approaches to study them.

Other NMR methods that provide structural restraints (*Section 1.3.2*) can be used to extract ensemble-average distance or angular information of proteins in solutions and solids. Dynamics on timescale faster than chemical exchange can be probed by examining moiety-specific relaxation rates R_1 and R_2 and can additionally guide MD simulations (see below).

Static disorder is the other extreme on the timescale of chemical exchange. Sometimes several distinct forms are present (Fig. 1.3.6). As such, peak doubling was observed for microcrystalline formyl-Met-Leu-Phe-OH (fMLF) tripeptide upon glass transition at 175 K (Bajaj *et al.*, 2009). Two forms were observed for microcrystalline arginine hydrochloride already upon slight cooling to 286 K (Su and Hong, 2011). A broadened but still distinct second peak was observed for the intermediate state of villin HP35 on the folding and denaturation pathways investigated by the group of Robert Tycko (Havlin and Tycko, 2005; Hu *et al.*, 2009). In their studies, conformational exchange between the folded, intermediate, and denaturated states was quenched by rapid freezing of the sample. Another example is the study of freeze-trapped photointermediates of proteorhodopsin (Becker-Baldus and Glaubitz, 2018). It should be noted that multiple resonances of the same site can appear not only due to quenched chemical exchange but also due to amyloid (Madine *et al.*, 2008) or crystal polymorphism (Harris, 2007).

The more common situation of trapped chemical exchange is characterized by simultaneous presence of an ensemble of structures, occupying a continuous conformational space (Fig. 1.3.6B), leading to severe line broadening. The inhomogeneous nature of line broadening can be tested by comparing the expected line width expected from the R_2 relaxation rate or by selectively saturating the frequency of the peak maximum, which would burn a hole in the spectrum (as done, for example, for frozen crystals of SH3 domain in (Linden *et al.*, 2011). The inhomogeneously broadened peaks are observed in the majority of experiments with dynamic nuclear polarization (DNP), which require presence of a paramagnetic agent as a source of polarization and a rigid homogeneous medium to achieve uniform polarization transfer. That

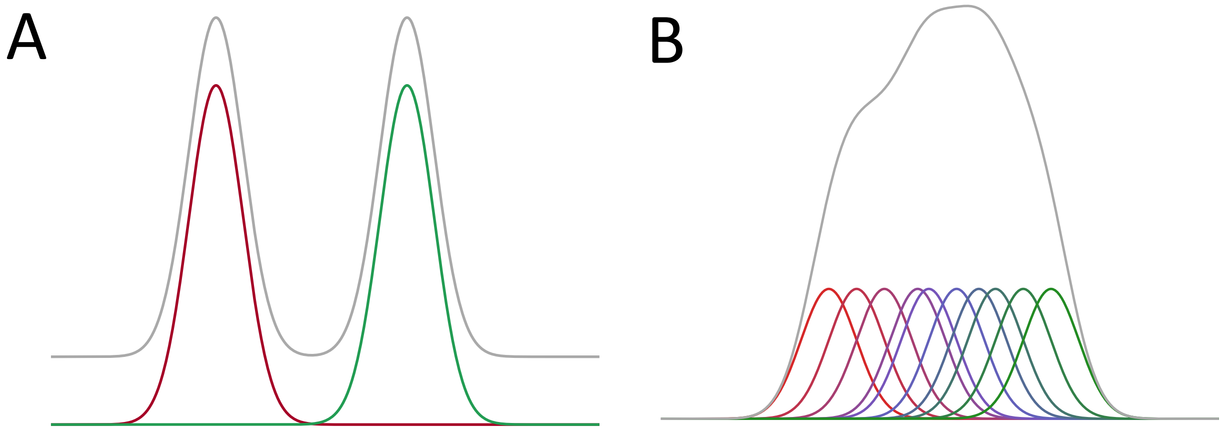


Figure 1.3.6. Sketch of NMR peaks of disordered residues. **A**: distinct forms are present; **B**: a continuum of similar members within a conformational ensemble forms a complex pattern which cannot be resolved. Situation **A** is common for amyloid and crystal polymorphs, trapped folding intermediates (Hu and Tycko, 2010) or different rotameric states of immobilized side chains. Situation **B** is common for trapped dynamic disorder, most notably seen in DNP applications. The gray lines represent the observed peaks; the colored lines represent the peaks of individual states of the residue.

defines the typical sample preparation protocol for DNP experiments, where the protein is dissolved in a water-glycerol mixture with the agent and the obtained solution is then rapidly frozen, such that formation of water crystals is avoided. Outside frozen solutions, the inhomogeneous peak broadening was observed in a frozen microcrystalline sample of SH3 domain, particularly below 253 K (Linden *et al.*, 2011) and membrane proteins in vesicles cooled to 238 K (Su and Hong, 2011). A special case of inhomogeneous peaks are the peaks formed by multiple (sometimes also presumably heterogeneously broadened) overlapped peaks of the same residue type due to residue-type-specific labelling scheme. This is the case for studies on spider silks, rich in poly-Ala and poly(Ala-Gly) motives (Asakura *et al.*, 2013a, 2013b). Another example can be found in the recent study of α-synuclein disorder in lipid bilayers by Uluca *et al.* (2018), where the collective inhomogeneous peak originated from all valines in the chain.

NMR-based approaches to studying protein disorder

The inhomogeneous broadening poses a large obstacle for site-specific studies due to severe peak overlaps. On the other hand, the inhomogeneous peak broadening contains valuable information about the conformational distribution of the flexible sites in the static ensemble, and extracting this information is a very attractive goal.

In the studies of HP35 folding pathways, Havlin *et al.* (2009) analyzed 2D ^{13}C-^{13}C spectra of partially folded HP35 by fitting them to linear combinations of 2D spectra of the folded state, the fully unfolded state, and partially denatured mixtures of HP35 fragments. Using this method, they were able to show that the unfolding of HP35 did not follow a simple two state model. Y. Su and M. Hong (2011) performed a qualitative analysis of relative secondary structure content in immobilized membrane proteins. They drew correspondence between single-quantum and double-quantum chemical shifts and chemical-shift regions typical for the different secondary-structure classes. A more sophisticated approach was taken by H. Heise and colleagues. In an early publication, Heise *et al.* (2005) used Monte-Carlo (MC) as well as molecular dynamics (MD) simulations to create a structural ensemble from which they predicted chemical shifts using SHIFTX (Neal *et al.*, 2003). The ensemble of structures was then reduced by discarding structures whose predicted shifts did not fit into the experimental data. The ensembles obtained with MC and MD were evaluated using principal-component analysis. A similar strategy was used in a later work by Uluca et al (2018) to reconstruct the conformational distribution of valines in the DNP spectra of α-synuclein with the help of MD and chemical-shift predictions by SPARTA+.

Several strategies exploiting anisotropic interactions have been followed to determine conformational and orientational distributions. In the studies of dragline spider silk, the groups of T. Asakura and B. Meier used static (no spinning at the magic angle) experiments, originally developed for synthetic polymers. They used double-quantum/single quantum (DOQSY) experiments (Schmidt-Rohr, 1996) to calculate the relative orientation of ^{13}CO carbon CSAs and obtain a probability density of dihedral angles P(φ, ψ) by comparison of the experimental data with the simulated DOQSY spectra for a range of backbone dihedral angle combinations (φ, ψ) (van Beek *et al.*, 2000, 2002). In addition, they applied the static DECODER experiment, also adapted from polymer science (Schmidt-Rohr *et al.*, 1992), to selectively ^{13}C-labeled silk, to determine the orientational distribution of the peptide chain within the fibril macrostructure (van Beek *et al.*, 2002). Using a similar strategy for proton-driven spin diffusion experiments under MAS, Kümmerlen *et al.* (1996) determined secondary structures for residue-type-labelled silk by matching simulated and experimental 2D correlated CSA patterns.

CSA-based approaches were applied to investigation of the conformational space of folding intermediates in the flash-frozen solutions. In the experiments on site-specifically isotopically labelled HP35, Hu *et al.* (2009) also determined (φ, ψ) distributions by measuring the correlation of CSA tensors (2DEXMAS), CSA-dependent DQ dephasing (DQCSA) of adjacent carbonyls, and ^{13}C-^{13}C dipolar couplings (CTDQFD). Simultaneous fitting of these data to the models of the different folding pathways of (φ, ψ) distributions allowed them to determine the conformations of the intermediates with high precision.

Integrative methods of ensemble reconstruction

The importance of representing proteins as structural ensembles instead of the average structure gradually becomes acknowledged by communities of all methods of structural biology. New techniques for reconstruction of dynamic disorder have been arising from NMR data in solution (eNORA, (Strotz *et al.*, 2017) and solid-state NMR (eRFDR, (Grohe *et al.*, 2019) , from electron density (Keedy *et al.*, 2015), and cryoEM maps (Kinman *et al.*, 2023). Extensive research based on fusion of the data obtained with different methods has yielded a variety of methods of ensemble reconstruction (Bonomi *et al.*, 2017). The methods typically take the ensemble-average restraints from liquid-state NMR methods (chemical shifts and, if available, RDCs), probability distributions from FRET and SAXS data, and electron densities from cryoEM maps. Integrating the orthogonal restraints turns the otherwise often underdetermined ("ill-posed") problem of ensemble determination into a well-determined one. Along with solution NMR, solid-state NMR could contribute restraints on trapped dynamic disorder. However, such practices have not received wide adoption yet.

1.4. Aims of the work

The present work aims to develop methodology to analyze conformational distributions in solid-state samples based on site-specific distributions of chemical shifts. Similar undertakings were pursued by Heise *et al*. (2005) and Uluca *et al*. (2018) with chemical-shift predictions for a set of structures obtained with MD simulations (see above). This work tries to establish a more direct translation between the chemical shift to the Ramachandran space using chemical-shift-based dihedral-angle prediction systems. The resulting (φ, ψ) distributions are compared with statistics on chemical shift-structure relations taken from PACSY database.

Other NMR-based methodology, such as CSA- and spin-diffusion based experiments, is severely hampered by low sensitivity and peak overlap. The very successful approach of the group of R. Tycko to investigate folding intermediates cannot be easily adapted for longer and more complex systems. Peptide synthesis, used to produce site-specifically labelled HP35, is not feasible for longer polypeptide chains and entails high costs of sample production. Also, site-specific labelling requires preparation of multiple samples with robust, highly reproducible sample preparation protocols, which are available for some of the existing research targets but would take long to develop for many newfound systems. In this work, individual sample sites are resolved in the chemical shift space by increasing spectrum dimensionality to 4D in an hCBCANH experiment. The dimensions of Cα and Cβ shifts are chosen as the ones most correlated with backbone geometry. Amide nitrogen and protons provide additional dimensions facilitating peak dispersion. Moreover, proton detection at high MAS rates, which is still a relatively rare technique in solid-state NMR, may boost experiment sensitivity, thus mediating the shortcoming of low signal-to-noise caused by heterogeneous broadening.

The full uniform sampling is not feasible for the high-dimensional data due to the high costs in measurement time. Non-uniform sampling has been successfully applied to reconstruction of NOESY and relaxation series. As a basis for the above subject, suitable conditions for acquisition and reconstruction of complex peak shapes that resemble no analytical function (Lorentzian, Gaussian or Voigt) are hence identified at first using test data obtained for a dehomogenized tripeptide.

The analysis presented here does not aim to provide a unique ensemble reconstruction. Any chemical shift-based approaches are intrinsically underdetermined due to complex structure-frequency relationships. Instead, the obtained dihedral angle probability maps distributions can be used as restraints on the way toward more comprehensive ensemble determination.

Additionally, the work seeks to provide quantitative metrics for degrees of disorder before the ensemble RMSD is available. Several heterogeneity scores describing the φ, ψ maps of each individual residue are suggested and tested for different scenarios.

The developed approaches to assess conformational distributions are applied to the functional amyloid of hydrophobin EAS$_{\Delta15}$. The heterogeneity metrics obtained are then compared with the trends of the linewidths of heterogeneous peaks.

2 | RESULTS AND DISCUSSION

This chapter presents the author's contribution to understanding and semi-quantitative evaluation of residue-specific static structural disorder in heterogeneous protein samples.

The preparatory part (*Section 2.1*) evaluates feasibility of non-uniform sampling for spectra of heterogeneous samples. The investigation of possibilities of acquiring NMR spectra of heterogeneous samples with non-uniform sampling techniques and selection of reconstruction procedure are the research contribution of the author, which was published in

> E. Burakova, A. Klein, S. K. Vasa, and R. Linser. *Non-uniform sampling in quantitative assessment of heterogeneous solid-state NMR line shapes.* Journal of Biomolecular NMR *(2020) 74: 71–82*

The main part (*Section 2.2*) investigates ways to analyze conformational distributions in the model solid-state sample of a pure u-(^{15}N, ^{13}C)-GGAGG pentapeptide comprising artificially introduced heterogeneity. Two approaches, one based on chemical shift predictions and one based on database search, are introduced and discussed. In addition, several numerical heterogeneity metrics derived from Ramachandran maps are suggested and discussed. This research was published in

> E. Burakova, S. K. Vasa, and R. Linser. *Characterization of conformational heterogeneity via higher-dimensionality, proton-detected solid-state NMR.* Journal of Biomolecular NMR *(2022)*

and is presented in this thesis in revised, restructured and, the author hopes, improved form.

The final part (*Section 2.3*) applies the developed methodology to residue-specific assessment of disorder of a functional amyloid formed by fungal hydrophobin EAS$_{\Delta15}$. The proposed φ, ψ -based metrics of disorder are compared with the trends formed by 1D linewidths obtained in dimensions representing different nuclei. The sample was obtained from Dr. Ann Kwan (The University of Sydney), and the NMR data that the approach was subjected to (spectra and assignments) were obtained in internal collaboration with Dr. Suresh K. Vasa and Prof. Dr. Rasmus Linser.

The manuscript including this work is in preparation.

2.1 Fidelity of reconstruction of non-uniformly sampled spectra of statically disordered samples

Acquisition of high-dimensional data, namely 3D spectra and higher, becomes unfeasible for a general case of a protein target: crowded spectral regions on one hand and large spectral windows in heteronuclear dimensions on the other hand lead to the necessity to acquire a large number of points of the indirect FIDs. Static disorder further complicates the spectra by severe peak broadening. In order to achieve sufficient resolution within a reasonable measurement time, NMR on homogeneous samples tends to resort to non-uniform sampling (NUS). Extensive research of the optimal ways to acquire the data and reconstruct them into the uniform grid has been done in the last decades (see *Introduction, Section 1.2.2*). Whereas it is shown to reliably reconstruct complex quadrupolar lineshapes (Rovnyak *et al.*, 2003) and sufficiently well preserve intensity ratios in NOESY spectra (Wieske and Erdélyi, 2021), it has not been clear until publication of this work whether the existing algorithms are capable of faithfully reconstructing the features of peaks that have been heavily distorted due to sample heterogeneity. The following section validates the performance of three algorithms – hmsIST, SSA and SMILE – on specifically generated model datasets (see the details on the implementation in *Section 1.2.2*). The data were obtained on a model sample of u-(^{13}C, ^{15}N)-fMLF powder, artificially heterogenized by freeze-drying from the dissolved state (*Materials and Methods, Section M.1.1*).

2.1.1. Overall approach

The reconstruction schemes were tested primarily on 3D datasets. Whereas the main strategy of the work (as outlined in *Introduction, Section 1.3.4*) involves acquisition of 4D spectra, 3D datasets are more practical in the sense of file size, while still providing good signal dispersion for the tripeptide sample; unlike 2Ds, 3D data are sufficiently robust to the variations in sampling density. Moreover, 3D is the most commonly used dimensionality in protein NMR, and the evaluation conducted here may best address the general interest for the community. Thus, the optimisation of sampling density and reconstruction parameters were done on 3D datasets, and 4D data were reconstructed by the three programs only once.

Unlike in many previous works, where the experiments with different subsampling and reconstruction schemes were carried out using simulated datasets, this study is performed on the experimentally obtained data to avoid any assumptions about the sample and the peak shapes required to produce simulated datasets. The powder of heterogenized fMLF gave rise to a reasonably representative yet not overly complicated set of distorted peaks in an hCONH spectrum (Fig. 2.1.1), which is used here as the reference dataset. The spectrum contains three groups of signals with a high dynamic range: the severely broadened and smeared out formyl/methionine crosspeak of the lowest overall height along with a small round signal at $\delta^{15}N$ = 128 ppm; the leucine/methionine group of two or three broad, overlapped ellipsoidal signals with different skew in the three dimensions of high to medium intensity; lastly, the high-intensity phenylalanine/leucine crosspeak of a standard ellipsoidal shape (Fig. 2.1.1A-D). The groups are labelled according to the assignments obtained for microcrystalline fMLF. From the study of Bajaj et al (2009), one can infer that the extra signal of the L/M group (centered at (8.0, 170, 120) ppm) likely belongs to an additional conformation of methionine. Assignment of each group component (overlapping NMR peak) requires collecting additional data; however, the nature of the peak broadening and the exact composition of the sample are not relevant for this particular study. The range of peak inhomogeneity, in combination with overall sufficient resolution and signal-to-noise ratio makes this spectrum a well-suited test case.

Subsampling from uniformly sampled spectra

The non-uniformly sampled datasets were created by subsampling the time-domain points from the uniformly sampled (US) dataset (*Materials and Methods, Section M.3.1*). This is a practical choice when working with the experimentally acquired data as opposed to the prevalent strategy of comparison between *time-equivalent* data, generated by reduced numbers of scans or addition of random noise to the more densely sampled datasets (as in (Hyberts *et al.*, 2013). Comparison of uniformly sampled and reconstructed datasets requires raw data to be highly consistent. Recording an array of experiments with varying sampling density could potentially introduce inconsistencies due to drift in the magnetic field, changes in probe tuning, and varying thermal noise patterns for different datasets. Subsampling from the US datasets makes it possible to attribute any differences between the reconstructed spectra and the reference data directly to the reconstruction process. The artifacts introduced by the various reconstruction methods depend on the sampling schedule (namely, number and distribution of points), so emulation of time-equivalent datasets by addition of noise to the reference data would not be straightforward.

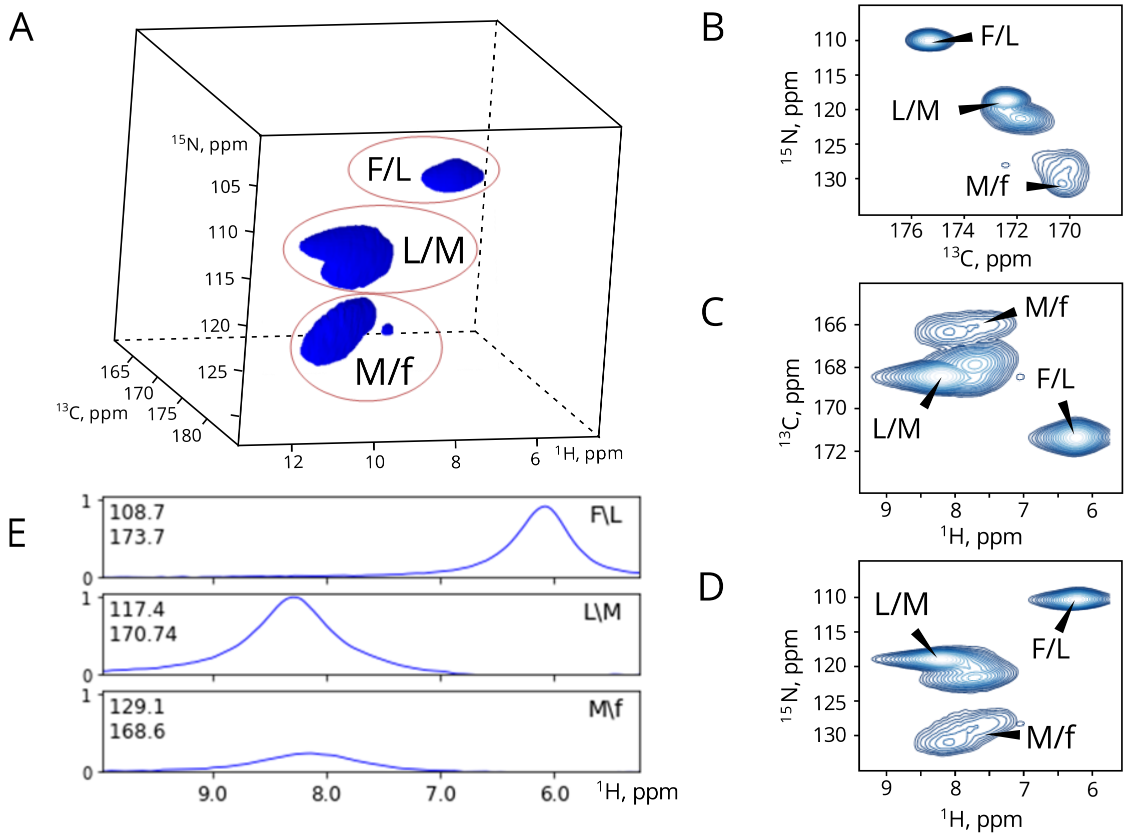


Figure 2.1.1. The hCONH spectrum of the dehomogenized fMLF sample. **A**: 3D view of the spectrum and assigned 2D projections onto **B**: the N/C, **C**: H/C, and **D**: H/N plane. The data were obtained at 55 kHz MAS in a 1.3 mm rotor at 700 MHz proton Larmor frequency. The two-letter labels reflect the assignments (first letter: residue *i*, H and N dimensions; second letter: residue *i-1*, CO dimension) of the main fMLF conformation as taken from the microcrystalline data and may or may not reflect the assignments of the newly emerged peaks. **E**: Cross sections through the peak maxima of the peaks

The subsampling was done with the home-made Python script (*Materials and Methods, Section M.3.1*), which extracts FIDs from the raw measurement file (`ser` for Bruker data) according to the defined schedule. Both hmsIST and SSA were run on the data subsampled by the Poisson-Gap sampling scheme (with sinusoidal weighting parameter `SSW=2`, explicitly recommended for hmsIST by its developers (Hyberts *et al.*, 2012). Since for SMILE fully random schedules have been proven the best (Ying *et al.*, 2017), it was run on randomly sampled FIDs). Reconstruction of alternatively sampled datasets (Poisson-gap for SMILE and random schedules with and without Gaussian weighting for SSA) yielded substantially worse outcomes (see examples for SMILE in Fig. S6).

Quantification of similarity between the spectra

In order to be able to compare the obtained spectra objectively, the simple measure of root mean square difference (RMSD) was applied to the pairs of spectra. The RMSD is a widely used score of similarity of multidimensional objects, applied, e.g., to series of structure obtained in MD simulations or NMR structure calculation. As a metrics of spectral similarity, the RMSD has been used by Wei Qiang *et al.* (2017) for quantification of variations between 2D spectra of Aβ fibrils. For the pair of spectra (frequency-domain data) the RMSD was calculated as:

$$RMSD = \sqrt{\sum_{i=1}^{N} (I_{i1} - I_{i2})^2 / N}\,, \qquad (22)$$

where I_{i1} and I_{i2} are the intensities of the i-th point (pixel) in the datasets 1 and 2, normalized by maximum intensity; N is the total number of points in the spectrum. The potential danger of using RMSD is that it could be strongly perturbed by artifacts at the border of the spectrum (upon States-TPPI acquisition mode) which would easily be recognizable by a spectroscopist and hence would not constitute a real problem. No such distinct artifacts far away from the obvious peak areas were observed (neither for US nor for NUS data), however. Conversely, if such artifacts were present, the spectra would need to be trimmed accordingly. The artifacts within and in proximity to the peaks, in particular within the reconstructed spectra, explicitly need to be taken into account: In applications focusing on sample conformational distributions, the exact features of the heterogeneously distorted peaks are of importance, and contributions like the chosen point-spread function will be inseparable from the conformationally defined peak features. Again, ridding the peaks from the artifacts by deconvolution is barely an option because features of the underlying chemical shift are unknown *a priori*.

Since the intensity of the points (pixels) in the spectra range from 0 to 100 %, the RMSDs can be understood as the average of relative deviations for the individual spectral points. In the case of the uniformly sampled 3D HNCO, the highest peak height has a signal-to-noise ratio (SNR) of 204 (see processing details in *Materials and Methods*, Table M1.1). Accordingly, the lowest contour level depicted in the figures (drawn at 10% of the highest intensity in the spectrum) represents a SNR of 20. The noise level was estimated as the standard deviation of signal-free data points in 1D slices of the direct dimension (Fig. 2.1.1E).

Important details of reconstruction procedures

Whereas some algorithms (within the selection tested here, this refers to hmsIST) have only few adjustable parameters that do not require optimization, some algorithms (here, SSA and SMILE) include parameters that can significantly influence the reconstruction quality. As such, in SSA there are two essential parameters: the Threshold T, the minimal SNR for a data point to be acknowledged as a peak maximum, and the Joint Threshold J, the SNR at the border of the peak frame. Analogously to J, SMILE parametrizes the data with nSigma, which defines the threshold above which all the points considered belong to a peak (SNR of the point must be ≥ nSigma). Optimization of those parameters (*Materials and Methods, Section M.3.1*) was done to minimize the RMSD between the reconstruction result and the US spectrum. As demonstrated in Figs. M2 and M3, the choice of parameters greatly impacts the quality of the reconstruction outcome.

In hmsIST, the only parameter that affects the depth of the reconstruction is the number of iterations; it has been shown by Hyberts *et al.*, (2012) that the necessary and sufficient number of iterations to recover a signal of SNR = 2 is 250 (where SNR is estimated as the standard deviation of noise).

2.1.2 Results

Reconstruction of the 3D data

The datasets were subsampled with 2, 5, 10, 30, 50, and 90% density, corresponding to 21, 105, 315, 525 and 945 points. According to the rule $m > K\ \log(N/K)$, where N is the total number of points in the US spectrum and K is the number of "significant points" (see *Section 1.2.2*), the theoretical minimal number of points m required for faithful reconstruction for the hCONH of a homogeneous fMLF would be as little as 16, which corresponds to about 1.4% sampling density. This very rough estimation is done assuming $K = 7$ (one 'significant' point per F/L peak, three points per M/L and three per f/M groups); however, such estimation is more complex to make for the "real" targets – biologically relevant, statically disordered samples. The resulting 3D data are represented by 2D H/C ‘skyline’ projections in Fig. 2.1.2. Apart from (expectedly) correct reconstruction of peak positions, we also observe a remarkable similarity for the *inhomogeneous* features of the peaks between the uniformly sampled spectrum and the spectra reconstructed from the dense datasets. Generally, the artifact level increases for lower sampling density, which becomes obvious in particular for lower-intensity regions of the peaks. A numerical comparison of these spectra with the uniformly sampled source data set via RMSD is shown as heatmaps in Fig. 2.1.3.

As has been noted (Hyberts et al. 2012), hmsIST can handle data sets recorded with relatively low density; this seems to holds true for the given 2 % dataset, which could be reconstructed to a spectrum of decent quality (RMSD of ca. 0.007, Figs. 2.1.2 and 2.1.3). By contrast, very poor reconstruction is obtained for the lowest density of 2% using the other two algorithms: at such a low density / number of points, SSA seems to misestimate the peak widths, and SMILE overestimates the number of signals while failing to recover the f/M peak. This results in particularly high RMSDs for the 2 and 5 % datasets. Data of good quality is generally obtained for 30 % and higher sampling density. Notably, unlike the cases of hmsIST and SMILE, in case of SSA increasing sampling density does not monotonously increase the quality of the spectrum. The 50 and 90 % density spectra, which are very close to the source data, seem to deviate more strongly from the reference than the 30 % sampled data set. Visually, this also becomes obvious due to the weak f/M peak disappearing (Fig. 2.1.2, panel SSA).

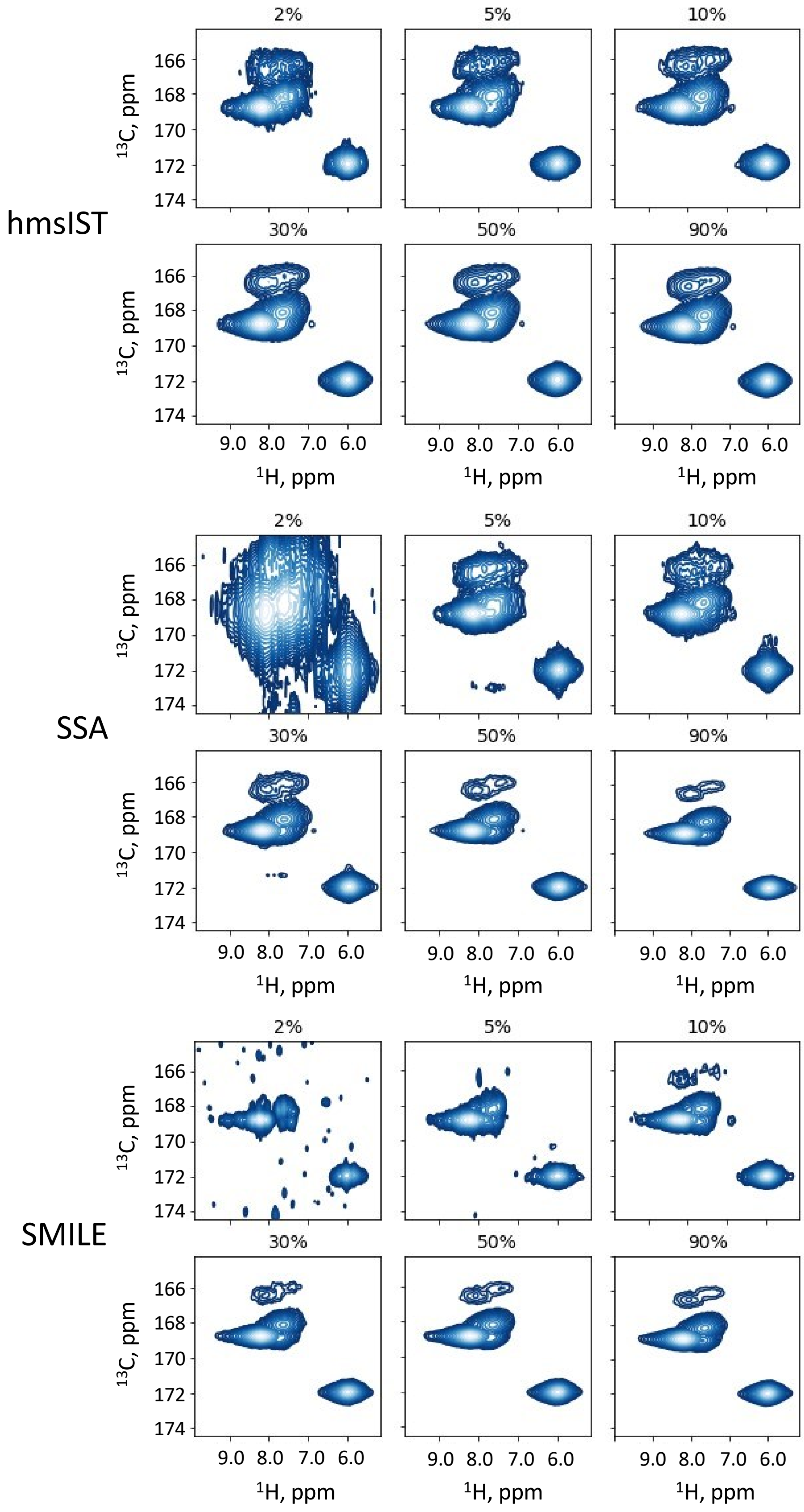


Figure 2.1.2. The effect of sampling density on the quality of reconstruction of the NUS datasets performed with the three algorithms: hmsIST, SSA and SMILE (see main text and *Materials and Methods* for details). The datasets were created by subsampling from the uniformly sampled 3D hCONH spectrum of dehomogenized fMLF powder (Fig. 1.2.1), the 3D spectra are visualized by H/CO projections. The lowest contour levels represent 10% of the highest signal intensity and succeed with a factor of 1.1.

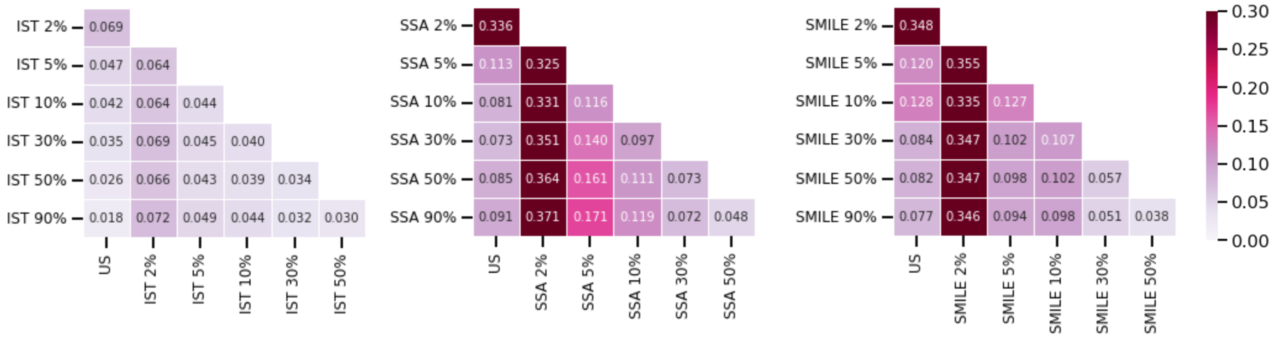


Figure 2.1.3. RMSD between US and NUS (subsampled) 3D HNCO spectra of dehomogenized fMLF, reconstructed with hmsIST, SSA and SMILE (see labels); the annotated values on the map have been multiplied by a factor of 10

Application to the 4D data

For proof of principle, the performance of all the considered techniques is demonstrated on a 4D hCOCANH spectrum (Fig. 2.1.3). Note, that in this particular case, owing to sensitivity considerations for this relatively insensitive experiment, US and NUS data sets were recorded as separate experiments, using generous 5 % sampling density for the NUS spectrum. Due to the time requirements of 4D US spectra, the US spectrum was recorded with half the number of scans and the T_1 relaxation delay being reduced compared to the NUS data, leading to a slightly lower detection sensitivity for the individual scans. Thus, opposed to the arrays of the reconstructed 3D data, an absolute identity of the NUS and the reference US data set cannot be expected here. However, the absolute intensity of the whole spectrum is not in focus. The total experimental time for US and NUS data compares as 4.5 d and 18 h, respectively. The RMSD of the spectrum reconstructed with hmsIST and SSA is equal to 0.017 and 0.035 respectively; note, that the values cannot be directly compared to the numbers obtained for the 3D datasets due to lower sensitivity of the experiment and thus relatively high noise in both US and NUS data (compare 1D traces in Fig. S6A). The SSA and SMILE parameters for 4D spectra were not optimized in such a thorough and systematic fashion as it was done for the 3D data sets; instead, realistically representing the application of the methods, a parameter estimation was based on the previous experience with the 3D data. Since the choice of a Poisson-Gap schedule has been shown not be ideal for SMILE reconstruction previously (Ying et al. 2017), these data are not representative and thus only shown in the *Supplement* (Fig. S6B).

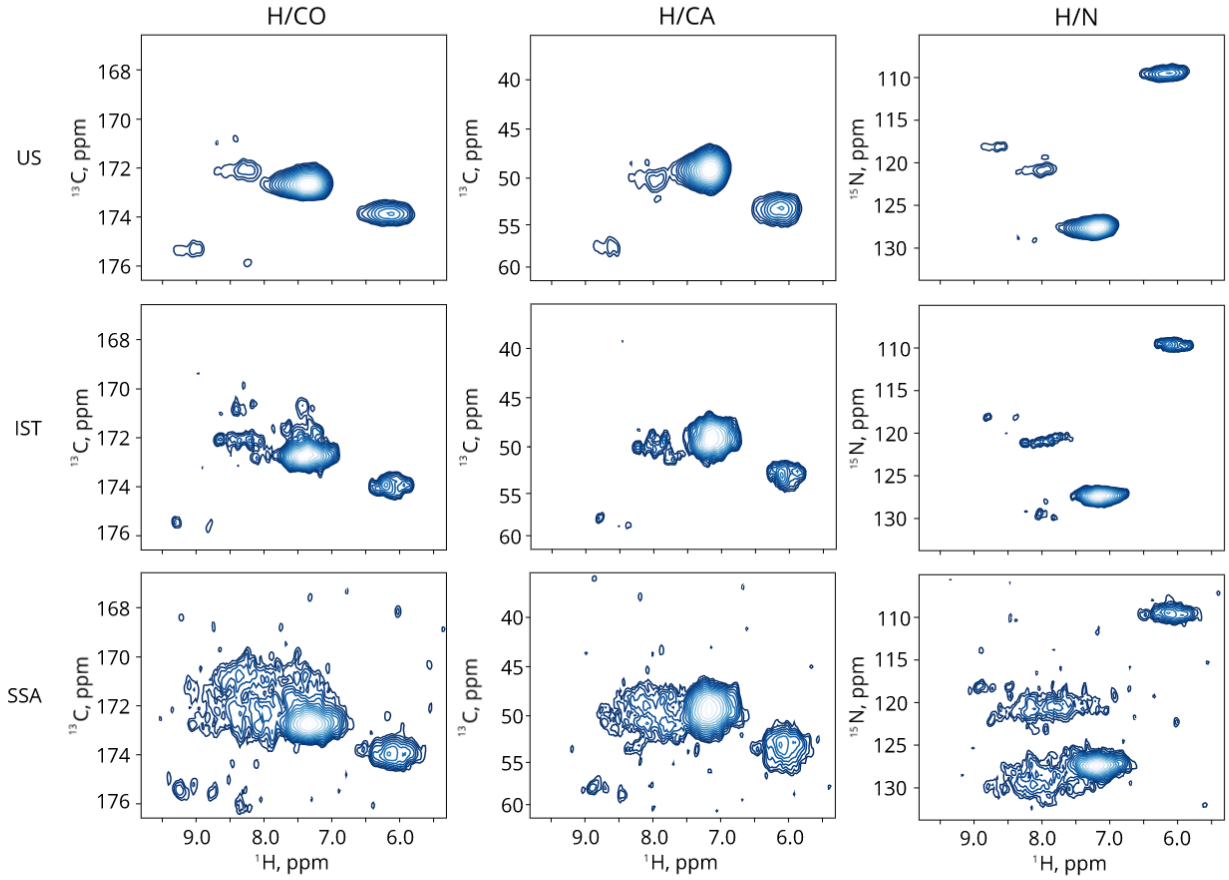


Figure 2.1.4. Projections of uniformly sampled and 5% sampled 4D hCOCANH of the dehomogenized fMLF powder reconstructed by hmsIST and SSA. The US spectrum was recorded with half as many scans per FID and a reduced recycle delay in 1/6th of the experimental time. See *Materials and Methods,* Table M1.2, for acquisition and processing details. In the plots, the first contour levels (10% of the maximum intensity within the spectrum) represent a SNR of 1.4 for the US spectrum, 3.19 for SSA, and 15.4 for hmsIST reconstructions.

Unlike in the inter-residual hCONH correlation, in the hCOCANH correlation, the 'M' group has the highest intensity, while of the 'L' group (split into the two lowest-intensity peaks, with a δ^{1}H of 9.0 and 8.7 ppm, respectively) barely reaches a signal-to-noise of 2 in the US spectrum. This low-intensity region is poorly reconstructed by both hmsIST and SSA. However, it is obvious that SSA has difficulties in artefact removal both, in the proximity of the peaks and in the "empty" spectral space, resulting in a heavily distorted dynamic range (the contours in Fig. 2.1.4 depict 10% of the projection's maximum intensity). These data altogether clearly motivate the choice of hmsIST as the reconstruction algorithm for inhomogeneous peak shapes.

2.1.3. Discussion

Since the series of the 3D NUS datasets were subsampled from the uniformly sampled one, we consider the latter as an ideal case. Correspondingly, the quality of reconstruction is to be taken as better whenever the RMSD between US and NUS spectra decreases. For comparison of individual peak shapes of nD objects sophisticated procedure have been described, e.g., in Chang and Kimia (2011). However, as a quantitative metrics of reconstruction quality, RMSD values are rather intuitively comprehensible, and the procedure allows a quick and straightforward numerical comparison. Alternatively, RMSD could be calculated separately for the individual peak boxes and the "empty space", but ultimately, similar results are anticipated. Numerical comparison by RMSD helps grasping the overall level of remaining artifacts that can be otherwise missed in the visual assessment of the multidimensional space.

For the series of the 3D data, an RMSD of 0.7% (Fig. 2.1.3) and lower is considered as corresponding to good quality of reconstruction; RMSD of 0.7–1% would reflect a decent reconstruction, RMSDs that exceed 1% correspond to spectra with a significant level of artifacts. The highest RMSDs (over 3%) reflect the strong deviations for the case of 2% sampled data processed by SSA and SMILE. Compared to the visual impression, the latter numbers still appear relatively low owing to the large "empty space", in which both reference and reconstructed data have close-to-zero intensities, thus a low level of absolute deviation.

The 3D spectrum of the test sample of the low molecular weight tripeptide provided much larger signal-to-noise of the individual FIDs then it would be expected in a regular experiment. The estimated measurement time needed to record the 2% subsampled dataset of ca. 11 min is equivalent to a few days of acquisition in case of a 60-residue protein, allowing the 3D US dataset to be subsampled. The poor site-specific sensitivity of the 4D experiment can be attributed to low CP transfer efficiency for the N-terminal residue, owing to remaining dynamics in the unfavorable intermediate timescale or other factors. In real-case studies, the general sensitivity of the experiment can be boosted by optimization of sample temperature, magnetization transfers (with, for example, optimal control strategies (Tošner *et al.*, 2018), or simply recording more scans.

Generally, the quality assessment conducted in the present work can neither be fully comprehensive nor representative for all possible NMR scenarios of possible future applications. Specifically, the reconstruction parameters used could potentially be optimized further. The study, however, conveys a positive message that the shapes of heterogeneously broadened peaks can be reconstructed given enough signal-to-noise (linked to the intrinsic sensitivity of the pulse sequence) and signal-to-artifact ratio (linked to the number of sampled points).

2.1.4. Conclusions

Feasibility of non-uniform sampling of NMR data on samples with a high degree of static disorder has been tested on 3D datasets, subsampled randomly or using Poisson-gap sampling schemes at 2-90 % density. The full data were reconstructed using three conceptually different algorithms: hmsIST, SSA and SMILE. All three in fact prove to be capable of handling complex peak shapes in multidimensional NMR spectra when sufficient number of points are sampled. hmsIST, representing the methods which do not make assumptions about the Lorentzian shape of peaks, expectedly performs most robustly for reconstructing the complex peaks both in 3D and 4D data; however, also SMILE and SSA provide good results on the 3D datasets of 30 % density and higher. For very low sampling density/data quantity, all methods are pushed to their limits, where different advantages and disadvantages of any method can be observed. Overall, however, non-uniform acquisition for NMR experiments on heterogeneous samples can reliably reconstruct the inhomogeneous patterns in an undistorted fashion. Any information content lying in the inhomogeneous contribution can thus faithfully be obtained in future higher-dimensionality studies based on NUS, given that parameters are chosen appropriately.

2.2 Development of analytical approaches on a model sample

2.2.1 The overall strategy

Heterogeneously broadened signals in NMR spectra require as much dispersion as possible. Without selective isotope labeling, which is often expensive and requires multiple consistent sample preparations, the only way to achieve this is introducing additional spectral dimensions. As discussed in *Introduction, Section 1.3.2*, the chemical shifts of ^{13}Cα and ^{13}Cβ carbons are the most sensitive to backbone dihedral angles. Although the relation between amide nitrogen chemical shifts and backbone conformations is more complex, they may serve as additional reporters and provide large chemical-shift dispersion. Proton detection is beneficial for experiment sensitivity, therefore the magnetization should ultimately land on protons *(Introduction, Section 1.2.1).* These considerations lead to the choices made here, in particular to employ a version of a 4D hCBCANH experiment, the magnetization transfer elements used here, as well as their particular arrangement (out-and-back vs. straight-through variants). This composition of the pulse sequence would need to be optimized for the individual sample.

The reasonable strategy of acquisition of the 4D NMR data suggests non-uniform sampling. The choice of hmsIST as the reconstruction algorithm and Poisson-gap schedule is suggested on conclusions of the previous *Section 2.1.*

Based on the statistics of chemical shifts and the examples from the literature (Uluca *et al.*, 2018), the general shape of heterogeneous peak is expected to be uneven and distorted. Hence, the problem of deconvolution is expected to be generally underdetermined. Introducing additional constraints requires assumptions about the number of components and their relative intensity (i.e., concentration of individual conformers in the sample), but the physical basis for them is unclear. Thus, a peak should be treated as the whole rather than a set of components.

2.2.2 The model sample

The test NMR spectrum for methods development was obtained from a sample of a u-(^{13}C, ^{15}N)-GGAGG pentapeptide. The primary sequence was chosen such that no aromatic or hydrogen-bonding moieties would complicate the analysis of the chemical-shift patterns by long-range modulations; absence of all but one sidechain group would ensure the highest possible variability of backbone conformations. The snapshot of the conformational ensemble was made by flash-freezing and subsequent freeze-drying. High vacuum (10 mbar) upon freeze-drying allowed for the complete removal of water (no water peak is present in 1D proton MAS spectra, data not shown), thereby excluding all contributions to the NMR peaks from the potential differences in hydration shell. Intermolecular peptide-peptide contacts however, arise, but their manifestation is expected to be largely limited to the ^{15}N and ^{1}H dimensions rather than ^{13}Cα and ^{13}Cβ shifts. For the purpose of the present study, it is reasonable albeit not entirely vigurous to ignore all contributions to the 4D hCBCANH peak other than conformational differences.

Acquisition and processing parameters of the NMR spectra are listed in *Materials and Methods,* Tables M.2.1 and M2.2.

The ^{13}Cα/^{13}Cβ crosspeak in a carbon-carbon correlation (Fig 2.2.1A) appears severely broadened, as expected: The linewidth in both dimensions is larger than 280 Hz, whereas the expected homogeneous linewidths at the given $\boldsymbol{B}_0$-field, MAS rate and digital resolution amount to about 80 Hz. Statistical data superimposed on the carefully referenced spectrum confirms the presence of multiple conformers: the peak shape resembles the bean-like shape of the chemical shift distribution from the database (Fig. 2.2.1B).

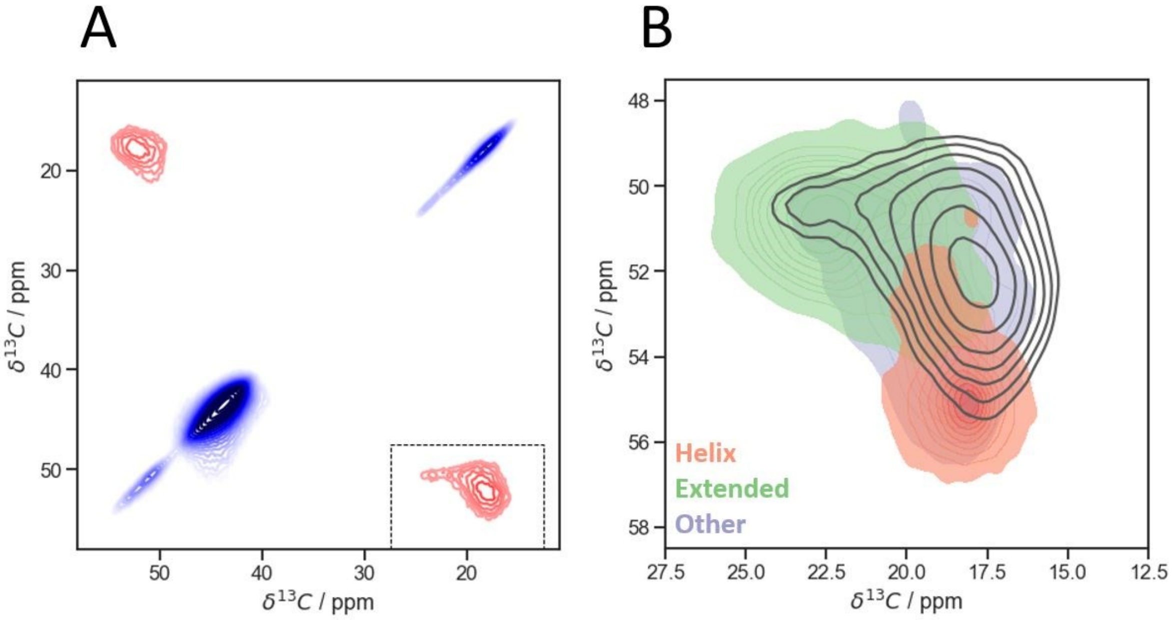


Figure 2.2.1. ^{13}C–^{13}C 2D DREAM correlation of an inhomogeneous sample of a GGAGG pentapeptide after freeze drying. A: Full spectrum (line broadening coefficient LB = 20 Hz). B: Overlay of Cα/Cβ Ala cross-peak (black contours, with exponential line broadening of 150 Hz) with expected chemical-shift 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 π-helices (H, G and I). The class "extended" 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. Random-coil chemical shifts result from fast averaging of different conformations in solution and have been omitted here

2.2.3 Predictions of collective (φ, ψ) distributions driven by prediction algorithms

Existing dihedral-angle prediction software, although based on different implementations, generally use similar overall approaches (see *Section 1.3.2*). Here we involved two of the most recent programs as such engines, TALOS-N (Shen and Bax, 2013) and DANGLE (Cheung *et al.*, 2010) utilizing different principles.

TALOS-N, a successor of the earlier programs TALOS and TALOS+, is the most popular program for the chemical-shift-based dihedral-angle predictions. The system predicts residue-wise Ramachandran maps at 20° resolution with an artificial neural network ((φ,ψ)-ANN), which was trained on a relatively small curated dataset of 580 X-ray structures and their experimentally obtained chemical shifts. The subsequent steps of TALOS-N evaluate the quality of these predictions and derive the most likely combination of φ and ψ angles. The presented approach to the evaluation of heterogeneity takes only the φ / ψ probability distribution maps from TALOS-N obtained by (φ,ψ)-ANN. All further program output is discarded.

DANGLE uses a different underlying mechanism of prediction. Like TALOS-N, it takes pentapeptide fragments of the chain and searches for similar fragments. For the likelihood estimate, instead of an ANN, it relies on Bayesian inference of the probable (φ, ψ) combinations from the data of pentapeptide fragments of 500 X-ray structures (Lovell *et al.*, 2003). The Ramachandran space is sampled with a matrix as fine as 10° per bin (360 × 360 points combined into bins of 10 × 10 points each).

The diagram of the workflow is presented in Fig. 2.2.2. The 4D hCBCANH peak assigned to the residue of interest (ROI) is subsampled into 4D grid coordinates with individual pixel intensities. Deconvolution of the hypervolume is specifically avoided, as in the general case this problem is underdetermined, so no unique solution would be found. All points of the resulting (4+1)D array that have an intensity below a given threshold (here: 20% of the peak maximum intensity, corresponding to the SNR = 15) are discarded. The grid must have sufficient resolution to retain the peak's features, at the same time increasing the number of points per dimension makes processing much longer. We consider this condition to be fulfilled when the distance between the neighboring grid points in each dimension is no larger than half the distance between the modes of the chemical shift clusters which correspond to particular areas on the *(φ, ψ)* map. Thus, in general, the grid resolution depends on the residue type. (See shapes of residue-wise chemical shift distributions on Fig S4.) In this work, the resolution in ^{1}H, ^{13}Cα, ^{13}Cβ and ^{15}N dimensions was set to 0.4, 1.0, 1.5 and 1.5 ppm, respectively, yielding a five-dimensional array of 1407 points. To determine the distributions, the curated PACSY database was used (state on Dec., 28th 2022). The choice of PACSY over the TALOS database was motivated by the larger number of 3D structures (7557 compared to 580), regular updates and the presence of not only crystal but also solution state NMR structures. The curation procedure and cluster analysis of PACSY was published in Fritzsching *et al.* (2016). Curation was performed by discarding proteins with presumably misreferenced NMR data; those entries are labeled as not passed by the Purging by Intrinsic Quality Criterion (`PIQC = 0` in the table `SEC_CS_DB.txt`). The pivotal "ideal" chemical shifts are defined as modes of each secondary structure cluster and were taken from the table `CS_STATS_DB.txt`.

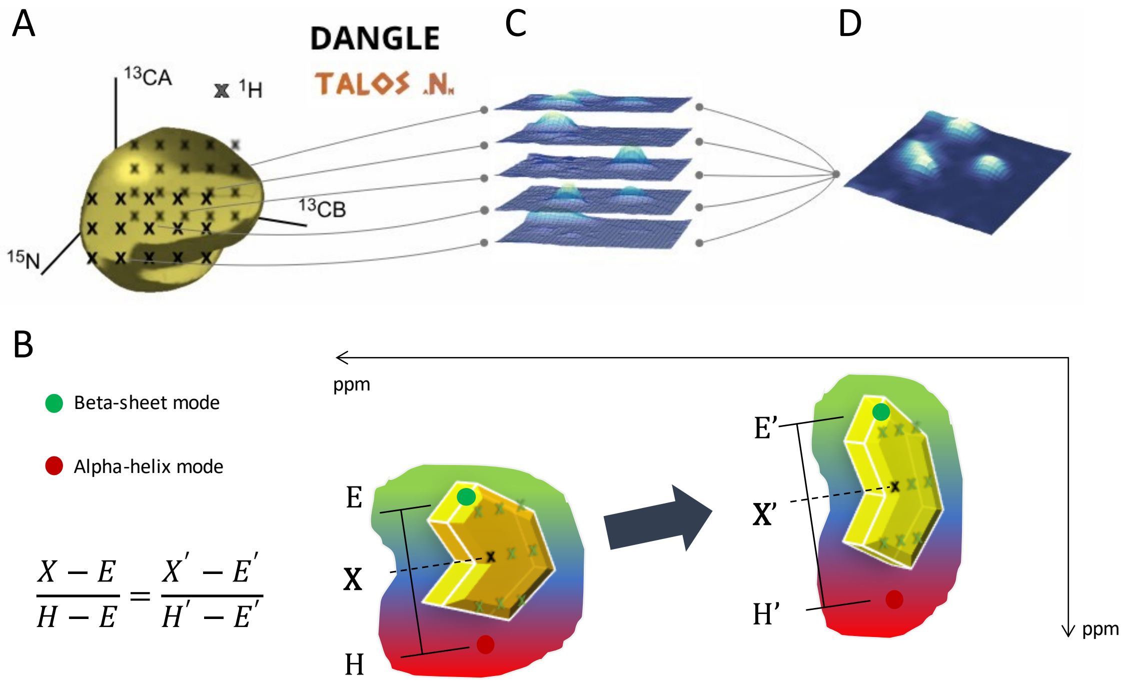


Figure 2.2.2 The workflow of heterogeneity analysis based on dihedral angle prediction software: here, TALOS-N and DANGLE. A: Sampling of the 4D hCBCANH peak of the Residue-Of-Interest with the regularly spaced grid. B: Translating the grid over ROI peak onto chemical shift distribution of the ROI's neighbors and preparing outputs; C: executing the dihedral angle prediction algorithm for every point; D: summarizing the predictions.

The structural context of the ROI, i.e., the secondary-structure propensity of its direct neighbors, manifested in chemical shifts, influences the prediction of the φ / ψ distribution of the ROI itself. In both, TALOS-N and DANGLE, the pentapeptide fragments are compared based on sequence and similarity of the chemical shifts of all five residues. This is highly reasonable in case of analysis of solution state peaks of a single conformation or conformational average, where solely the position of the peak maximum is considered. In the case of heterogeneous fragments of a protein chain, the sequentially connected peaks represent distributions of chemical shifts rather than a defined chemical-shift combination. Ideally, the features of the sequential peaks could be linked by interresidual experiments in the same manner as the peak maxima during a regular backbone walk. In practice, however, such experiments suffer from low sensitivity, and retrieving the interresidual correlation is almost always possible for the peak maximum only and not for the entire peak. Therefore, it seems the most reasonable to artificially generate those shifts for the neighboring residues that retain the secondary-structure propensities of each ROI pixel. The latter is ensured by scaling and translating the grid of ROI chemical shift according to residue type-specific parameters, i.e. width and position of the chemical shift distribution applicable for the residue type of the neighbors (Fig. 2.2.2B). The distribution widths are represented by the distance between the modes (or "ideal" values, as named in Fritzsching *et al.*, 2016) of the α -helical and β-sheet chemical-shift distributions, which are atom- and residue-specific.

For the test of TALOS-N performance, we artificially extended the sequence of the model peptide by two virtual glycine residues such that the pentapeptide becomes a heptapeptide – a fragment of the same size as the fragments of the database. Although from the description of the algorithm, no differences in the resulting 18 × 18 (φ, ψ) maps are anticipated, because the predictions are supposed to be based on the pentapeptide fragments only, some minute differences were found (Fig. S8A). The workflow was carried out for the extended sequence. Obviously, no artificial adjustments of the sequence would be needed for the real sequences in future applications. DANGLE uses for analysis only the pentapeptide window, so no sequence adjustment is necessary for GGAGG.

The grid that samples the 4D peak volume is shown in Fig. 2.2.3A (as two orthogonal 2D projections). To obtain a better feeling for the ensemble of points and predictions, five samples were taken from the array: one sample from the α-helical region (Point 1), two points from the sheet-like area (Points 2 and 4), one sample from the third corner of the triangular peak (Point 3), and the peak maximum (Point 5). Their TALOS-N predictions are displayed in Fig. 2.2.3B. As expected from the shape and size of the alanine hCBCANH cross-peak, the individual predictions cover the entire allowed Ramachandran space. Both Point 2 and Point 4 represent the extended conformation, where Point 2 yields a larger width of the probability density function (PDF) along the φ dimension. Point 1 yields predictions of helical propensity of both types, right- and left-wound. Given the primary sequence and absence of neighboring side-chains, the presence of such structures seems highly sensible; moreover, chemical shifts of the left- and right-handed helices are poorly distinguishable (Fig. S5).

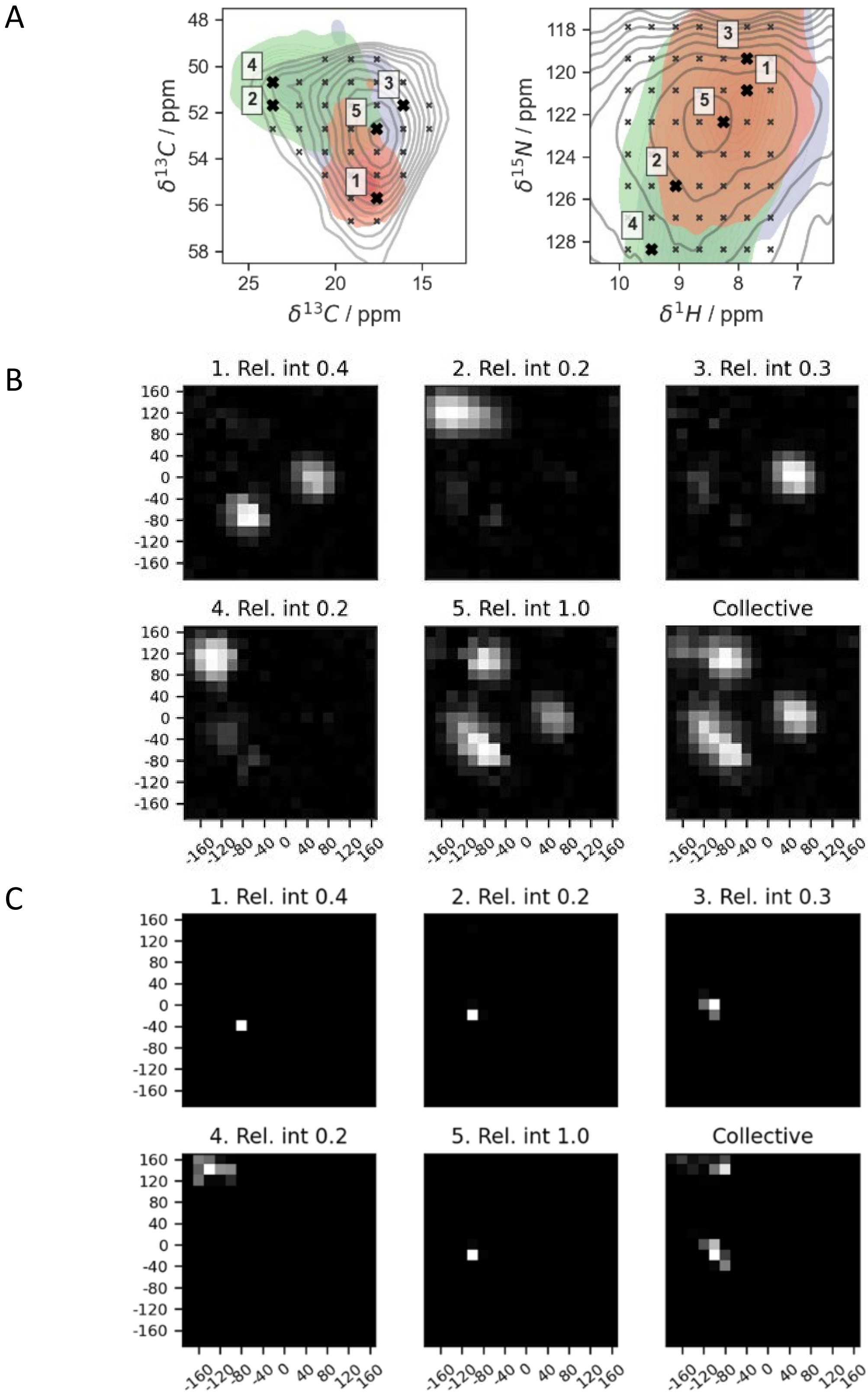


Figure 2.2.3. Individual and collective predictions of (φ, ψ)-PDFs made by dihedral-angle prediction software. **A**: Grid points within the peak (black crosses) and a selection of five test cases ("Points 1-5", bold black crosses) overlaid onto orthogonal projections of the 4D peak (Cα/Cβ and H/N projection, left and right respectively); **B, C:** Predicted PDF for the aforementioned grid points as well as the collective, weighted-average PDF obtained with TALOS-N (**B**) and DANGLE (**C**)**.** The PDFs are normalized by their maximum, brightness reflects the relative bin hight: white for 1 and black for 0.

Indeed, the chemical shifts observed for the left-handed structures occur in the same regions as their "conventional" counterparts, which can be illustrated by the statistics for the helical regions from PACSY (Fig 2.2.3). It is important to note that left-handed conformations formed by some other residue types (D, N, K, R) cluster in the chemical shift space as seen in Fig. S5. This should be investigated more closely in future statistical studies: Such segregation can be caused by the influence of the context of the primary sequence (for example, the context of proline-rich chains). The influence of the primary sequence on TALOS-N predictions with regard to the direction of winding is illustrated by an experiment (Fig. S8) where the glycine chemical shifts are translated to those of leucines, preserving the secondary structure propensity (as described above and sketched in Fig. 2.2.2B). The predictions for the LLALL and LLLALLL chains indeed closely resemble those for the GGAGG and (G)GGAGG(G), with the difference that the probability density in the right-handed helical region is now totally absent. Point 5 – the peak maximum – is located in the central region of overlapping tails of distributions of the two major classes (H and E), and in the center of the "Other" structures, dominated by residues involved in turns T (Fig. 2.2.1B). This element is labeled as "dynamic" due to very high proximity of all the four chemical shifts to the tabular random coil values initially determined for highly flexible structures under denaturating conditions (table `randcoil.tab` from the TALOS database). This map, however, differs from the distribution of random coil structures (Fig. S4), which is likely influenced by neighboring glycines: the experimental evidence for π-helices formed by GGAGG in solution (Ding *et al.*, 2003) supports this bias towards helical structures.

The collective prediction for the entire chemical shift space occupied by the peak is obtained by summing up the 1407 individual TALOS-N outputs, weighted by the intensity of the NMR peak at the position of the grid point:

$$D_k = \sum_{i=1}^{N} D_{ki} I_i, \tag{23}$$

where D_k is the 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. The collective Ramachandran map covers almost the entire allowed Ramachandran space (the panel "Collective" in Fig. 2.2.3B), which matches well the expectation of a conformational snapshot of a flexible polypeptide chain. High probability density is observed for "turn-like" and helical regions each, whereas sheet-like grid points (like Points 2 and 4) eventually contribute very little to the final picture.

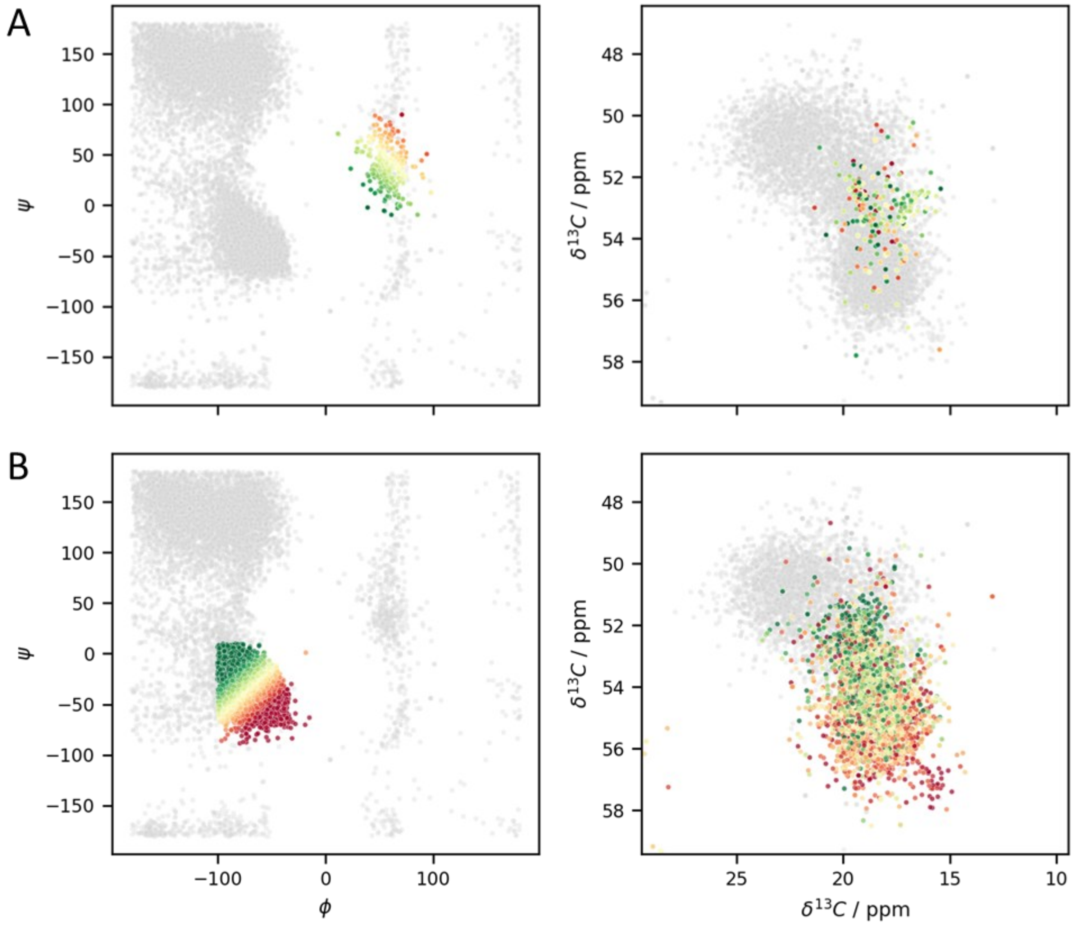


Figure 2.2.4. Overlap of chemical shift distributions of A: left- and B: right-handed helical conformations of amino acid residues illustrated by alanine entries of PACSY database. Residues belonging to the proteins not passed PIQC as well as classified by STRIDE as random coil ("C") are excluded.

The same evaluation of subsamples from the peak was performed with DANGLE as a dihedral-angle prediction engine. The 1407 assignment tables with glycine shifts being adjusted for equivalent propensity were rewritten in DANGLE input format and subjected to the algorithm; the results are shown in Fig 2.2.3. Noteworthily, the ROI neighbors' chemical shifts were found to be a decisive factor: without adjustment of the $i \pm (1,2)$ shifts for equivalent propensity (as discussed above and shown in Fig. 2.2.2), DANGLE, unlike TALOS-N, yields identical (φ, ψ)-maps for any chemical-shift combination of the central residue *(data not shown)*. Since DANGLE does not apply any Gaussian smoothing to the predicted maps, they appear significantly different from the ones generated by TALOS-N – for the same reason, the two outcomes shall not be quantitatively compared (see below, *Section 2.2.5*). To make the qualitative comparison easier, the resolution of the original DANGLE maps was lowered to 20° per pixel as in the TALOS-N maps, keeping the raw high-resolution data (Fig S9). Apart from Point 2, results are quite similar if the direction of winding is neglected the maxima of the predicted PDFs differ only by a systematic shift of +20° along the ψ axis of DANGLE predictions (see the high-resolution maps Fig S9). The DANGLE PDF for Point 2 is concentrated in the helical zone (φ=-90°, ψ=-30°), creating the only major discrepancy among the test points. This might result from the upfield-shifted amide H and N chemical shifts, although such strong influence from these nuclei is unexpected. The predicted (φ, ψ) combination for Point 5 falls into the most intense region in the TALOS-N map, yet the DANGLE prediction is strikingly unambiguous despite the overlap of chemical-shift distributions of all secondary structure classes at Point 5 and despite the general tendency of DANGLE to yield multi-island diagrams for dynamic residues (Cheung *et al.*, 2010). The collective maps obtained in both TALOS-N and DANGLE workflow are in good agreement.

2.2.4 Evaluation of the predicted maps by database analysis

The dihedral-angle prediction frameworks provide a thorough statistical analysis of each chemical-shift combination and sequential context and generally yield accurate and mutually consistent results. In the situation where the validation by a truly orthogonal physico-chemical method is close to impossible (see *Introduction*, *Section 1.3.3* as well as Discussion below), it is helpful to have an overview of the data accumulated hitherto to be able to verify the prediction results at least generally or resolve discrepancies, if any.

A plethora of relational tables can be obtained from the ReBoxitory at NMRbox (Maciejewski *et al.*, 2017). The data base on ReBoxitory is crude and requires more efforts to process but preserves many more structural properties, such as residue numbers, and experimental parameters.

PACSY (Lee *et al.*, 2012) relates the lowest-energy protein tertiary structures from the Protein Data Bank (PDB) along with the residue-wise secondary structure assignments (by STRIDE, Frishman and Argos, 1995) and their chemical shifts from the Biological Magnetic Resonance Data Bank (BMRB) (Hoch *et al.*, 2023). As of March 2023, PACSY contains 7557 proteins (alanine data available for 5088 proteins), among which at least 4131 structures are obtained from solution NMR data. Discarding of proteins with presumably misreferenced chemical shifts (PIQC, Fritzsching *et al.*, 2016) retains about 88 % of proteins as meeting both, the Intrinsic Quality Criterion and having the complete set of backbone assignments.

From the cleansed database, all residues (here, alanines for the analysis of the Ala in GGAGG) classified as random coil ('C') were further excluded from the pool. Presumably, this class includes only residues visible in solution NMR: disordered chains are usually barely observable in solids by any physico-chemical technique. The dynamic averaging cannot occur in fully rigidified solids, therefore the group of residues with chemical shifts close to random coil values will not necessarily occupy the same conformational space as the group unrestricted in motion. The conformational space of the central region of the chemical shift space and of the random-coil (‘C’) entries on one hand and entries belonging to more rigid structures (class ‘Other’ as Fig. 2.2.1) on the other are indeed different: residues unrestricted in movement tend to occupy regions of ψ > 50° (Fig. S10). Therefore, the chemical shifts of the middle region should be interpreted differently for solid than for solution samples. Worth noting, the solid nature of the sample is not taken into account explicitly by any of the dihedral-angle prediction algorithms due to scarcity of the solid-state NMR data.

The entire heterogeneous peak can be evaluated by looking at the collection of the same residue type entries that are weighted with coefficient w, which combines the intensity of the NMR peak I at the given entry i and a factor that accounts for the non-uniform distribution of the observed chemical shifts – the inverse of the density of the database entries P^{-1} at the given position in the 4D chemical shift space: $w_i = I_i \cdot P_i^{-1}$ (all points outside of the hyperspace occupied by the hCBCANH peak get a zero weight: $w_i = 0 \cdot P_i^{-1} = 0$, see *Materials and Methods, Section M.3,* for more details). This is essentially a very primitive dihedral-angle prediction method that does not (and, when the source database is PACSY, cannot) take into account the broad sequential and/or structural context, and it is not meant to compete with the sophisticated algorithms as TALOS-N and DANGLE in predicting *single* combinations for homogeneous residues. However, it is applicable in the context of this work, which is focused on the overall (φ, ψ) probability maps for the samples that exist in ensembles of conformations with presumably a number of higher-energy structures. Such unbiased overview gives a complementary perspective to the collective maps obtained from dihedral-angle prediction engines and can point out the presence of unusual conformations.

The selection of PACSY entries for the alanine crosspeak is shown in Fig. 2.2.5, the components of the resulting weight are denoted by color (point density P_i) and dot size (signal intensity I_i). The same selection plotted in the Ramachandran space constitutes a collective, weighted density map that may be seen as the representation of the conformational ensemble manifested in the broadened hCBCANH peak. For comparison, the non-weighted map is shown in Fig. S12. The major point density is concentrated in the α-helical and turn-like regions. Alanines with chemical shifts from the region typical for β-sheets (green area in Fig. 2.2.4A, upfield-Cα / downfield-Cβ, -N and -H regions.) provide a minor – yet noticeable – contribution to the density function resulting from the presence of some spectral intensity in this region. Thus, the obtained (φ, ψ) map is consistent with the collective PDFs predicted by TALOS-N or DANGLE.

It is interesting to look at the dihedral-angle distributions of the PACSY entries whose chemical shifts are close to the individual test points. Here, the Subsets were selected by the 4D boxes centered at the Points 1-5 in chemical-shift space. The width of the boxes in each dimension corresponds to twice the grid resolution – as such, the space was sampled with boxes of 0.8 ppm along the H-axis, 2 ppm in Cα, and 3 ppm in both, Cβ and N dimensions.

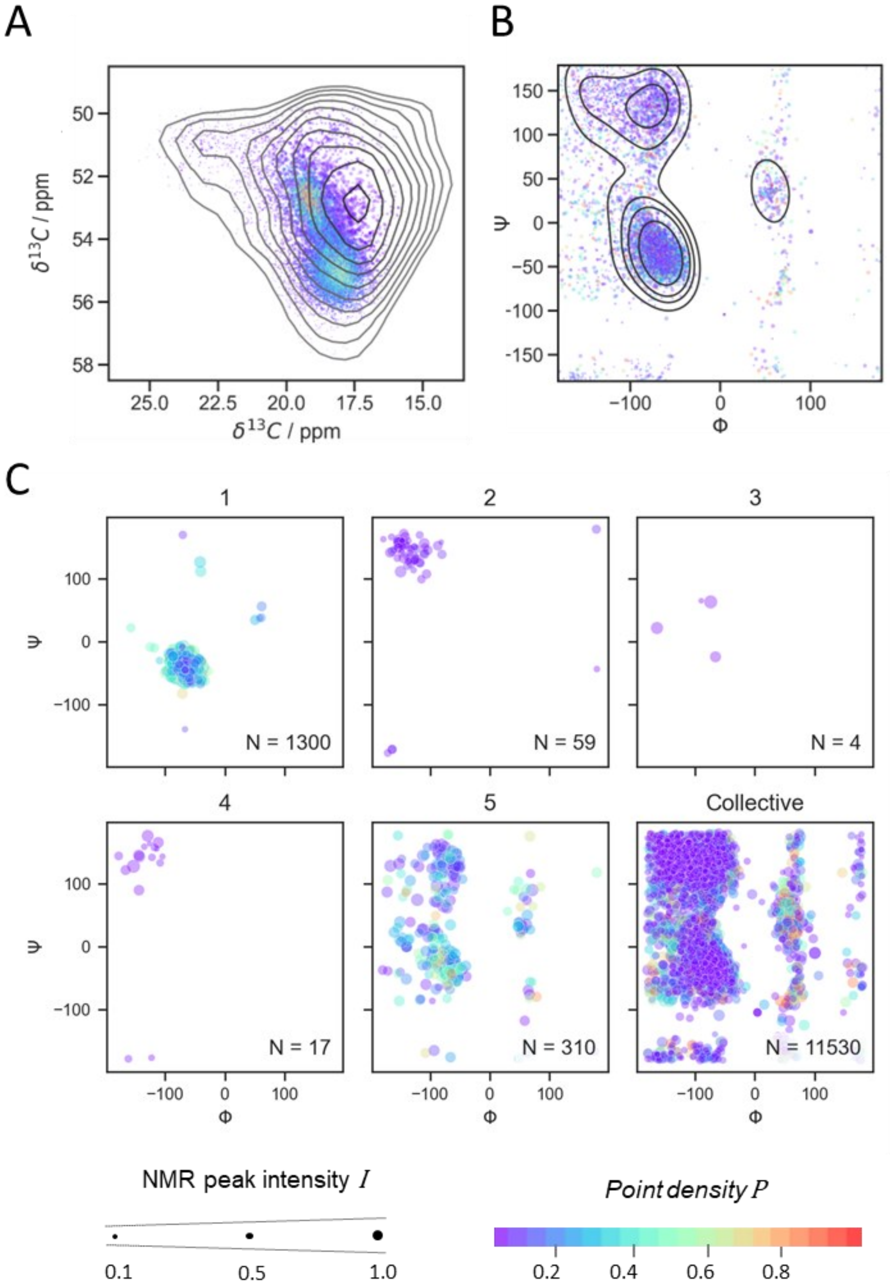


Figure 2.2.5. Statistical analysis of the conformational distribution of alanine in the static ensemble of GGAGG. **A**: PACSY entries belonging to the 4D hCBCANH alanine cross-peak. Color (see the bottom of the figure) denotes the density of the points in the 4D chemical-shift space, the point size representing the relative intensity of the peak. (The scale at the bottom of the figure serves illustrative purposes.) For the point density estimation see *Materials and Methods.* The contours show the two-dimensional projection of the cross-peak and start from 4% of the absolute intensity and succeed with the factor of 1.2. **B**: The selected points in the Ramachandran space and the bivariate weighted density estimate (gray contours). **C**: Ramachandran maps constructed from the above selections, along with the collective map for reference.

Ramachandran maps constructed from the selections are presented in Fig. 2.2.5C, along with the collective map for reference. TALOS-N and DANGLE predictions mismatch for two of the five points: Point 2 (a major conflict) and the peak maximum, Point 5. (For the latter, the DANGLE-predicted (φ, ψ) combination is strictly defined, in contrast to the mixed TALOS-N map). The Subset 2 (centered around Point 2) consists of 60 entries, all of which adopted extended conformations. This agrees well with the TALOS-N prediction and sharply conflicts with the result of DANGLE run for Point 2. Since DANGLE does not provide the details on the best-matching fragments used for the Global Likelihood Estimate, it is difficult to track the origin of this discrepancy.

The Subset 5 includes the points from both, high- and low-ψ values. A more detailed view on this region can be provided by the inverse view on structure-chemical shift relations as shown in Fig. 2.2.6 (and Fig. S2 for other residues). The central chemical shift region – hence, the same electronic environment – is populated by residues adopting all conformations, including the shoulders of the distributions of right- and left-handed helical (red and orange, Fig. 2.2.6B and E) and extended conformations (green, Fig. 2.2.6D). This illustrates the well-known challenge and shortcoming of purely chemical-shift-based secondary-structure prediction methods. The map 5 predicted by TALOS-N resembles the Subset 5 quite closely.

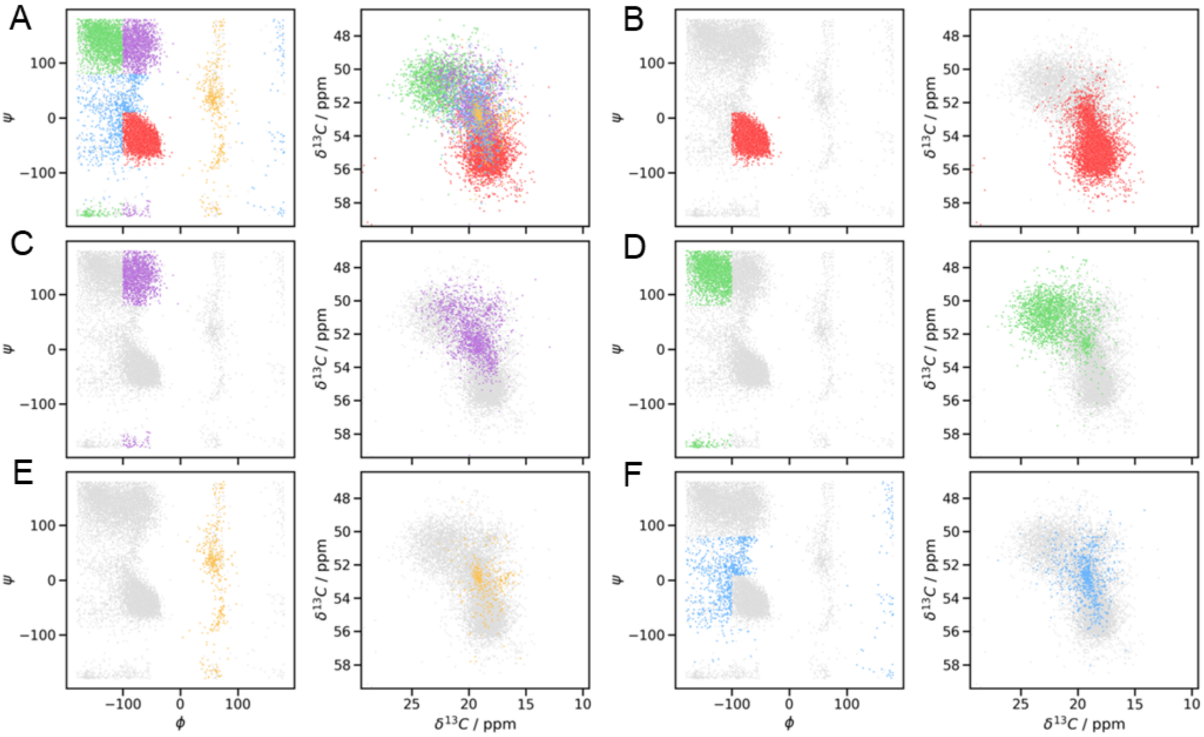


Figure 2.2.6. All alanine entries of PACSY, excluding those classified as random coil ("C"). The color highlights the dihedral-angle classes. Panel A shows all the entries; panels B-F show the five regions separately with all other points from the panel A being grayed out.

2.2.5 Quantification of heterogeneity

Apart from qualitative assessment, some application of heterogeneity analysis, for example the comparison of two preparations of a sample at different preparation conditions, would benefit from a simplified numeric comparison of degree of heterogeneity. A "heterogeneity score" should in some way reflect the *complexity* of conformational ensemble – a notion with no single formal definition. The following section aims to explore different metrics of residue-specific heterogeneity based on the parameters of the obtained collective (φ, ψ) distributions.

For such comparisons, the first question to answer is which maps to take into account for the quantification. The way the map is obtained matters, as it becomes obvious, for example, from comparison of TALOS-N and DANGLE outcomes (Fig. 2.2.3B and C). Such maps also cannot easily be simulated, because the allowed Ramachandran space is more complex than a linear combination of basic functions (Gaussian) at the local maxima of the PDF (Fig. 1.3.2). For these reasons, the approaches developed in the following are tested on the (φ, ψ) distributions obtained with TALOS-N, DANGLE, and PACSY for the GGAGG for the above Points 1-5 as well as the collective maps. A reference for purely homogeneous and well-defined cases of TALOS-N and DANGLE predictions for helical and extended conformation was obtained with a simulated sequence of (Leu)$_{10}$ for which the full set of chemical shifts matched the expected values of the corresponding structures (taken from Fritzsching *et al*. (2016)). For simplicity of the quantification, the direction of winding is omitted (see *Section 2.2.3*), because not only is it more challenging to discriminate between left- and right-handed structures, but the left-handed winding is in general extremely rare. For the residues with the most distinct clusters of left-handed entries in the chemical-shift space (D, N, Q, etc., Fig. S5) it may be a good direction of the future method development to include into consideration the "handness" of a residue, i.e. the sign of its φ angle. In this work, the reduction is performed by inverting the Ramachandran space of positive φ values through the central point (0, 0). The obtained eight test *scenarios* are shown in Fig. 2.2.7 ordered by intuition from the simplest to the most complex map, where the panels "Conf. H" and "Conf. E" correspond to the reference ("confined") helical and extended structure cases.

TALOS-N produces Gaussian-smoothened probability density functions that can be reasonably well described via the regular metrics for the spread of a circular distribution. Here, the spread is measured by the *circular variance* estimator for either of the two marginal distributions (projections onto φ or ψ axis):

$$V_\theta = 1 - |\vec{v}|_\theta = 1 - \sqrt{\overline{s\imath n^2}\theta_k + \overline{cos^2}\theta_k} =$$

$$1 - \sqrt{\left(\frac{\sum_{k=1}^{180} sin\theta_k \cdot D_k}{\sum_{k=1}^{180} D_k}\right)^2 + \left(\frac{\sum_{k=1}^{180} cos\theta_k \cdot D_k}{\sum_{k=1}^{180} D_k}\right)^2}, \tag{24}$$

where θ stands for either the φ or ψ coordinate and D_k is the predicted probability of the k-th (φ, ψ) combination. The second term of Eq. 24 reflects the vector sum of all angles and ranges from 0 to 1 (the radius of the unit circle). Thus, V ranges from 1 for confined distributions to 0 for uniform distributions. The marginal distributions (obtained by summing up all D_k values of the same φ or ψ coordinate) and their variance V are visualized in polar coordinates in Fig. 2.2.7B. In contrast to the reference cases and the clear shift combinations (Points 1 and 2), an increasingly large V is found for Points 3–5, i.e., when shifts differ drastically from 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 predicted (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) bin k of the Ramachandran map as an independent state of an amino acid residue, with its intensity D_k representing its likelihood to be adopted, the entropy of a prediction would be calculated as follows:

$$S_\phi = -\sum_{k=1}^{10} D_k \, ln(D_k) \tag{25.1}$$

$$S_\psi = -\sum_{k=1}^{18} D_k ln(D_k) \tag{25.2}$$

$$S_{total} = -\sum_{k=1}^{180} D_k ln(D_k) \tag{25.3}$$

The 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. Just as the variance, the entropy can be calculated for the marginal φ and ψ distributions (Eq. 25.1/25.3): In this case, $0 \leq k \leq 10$ for φ, $S_\varphi^{max} = \ln(10) \approx 2.3$; $0 \leq k \leq 18$ for ψ, $S_\psi^{max} = \ln(18) \approx 2.9$, where the reduced numbers of bins result from projecting all values of one row or column before application of Eq. 25. Worth noting, that this *discrete* entropy should be seen as a way to estimate *continuous* entropy of the underlying continuous probability distribution.

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, …, 180}, and for single-angle entropies (Eq. 3.1 and 3.2, k bearing 10 or 18 values for φ and ψ, respectively), D_k can also refer to probabilities for individual φ or ψ distributions in one-dimensional Ramachandran maps. For better contrast, it may be useful to subtract the baseline of, e.g., the well-defined case of *Confined Helix* and thereby consider only *excess entropy*: $\Delta S = S - S^{conf.\ H}$. For the collective map of the GGAGG sample, the overall ΔS amounts to 1.30, the highest-possible value would be 2.02.

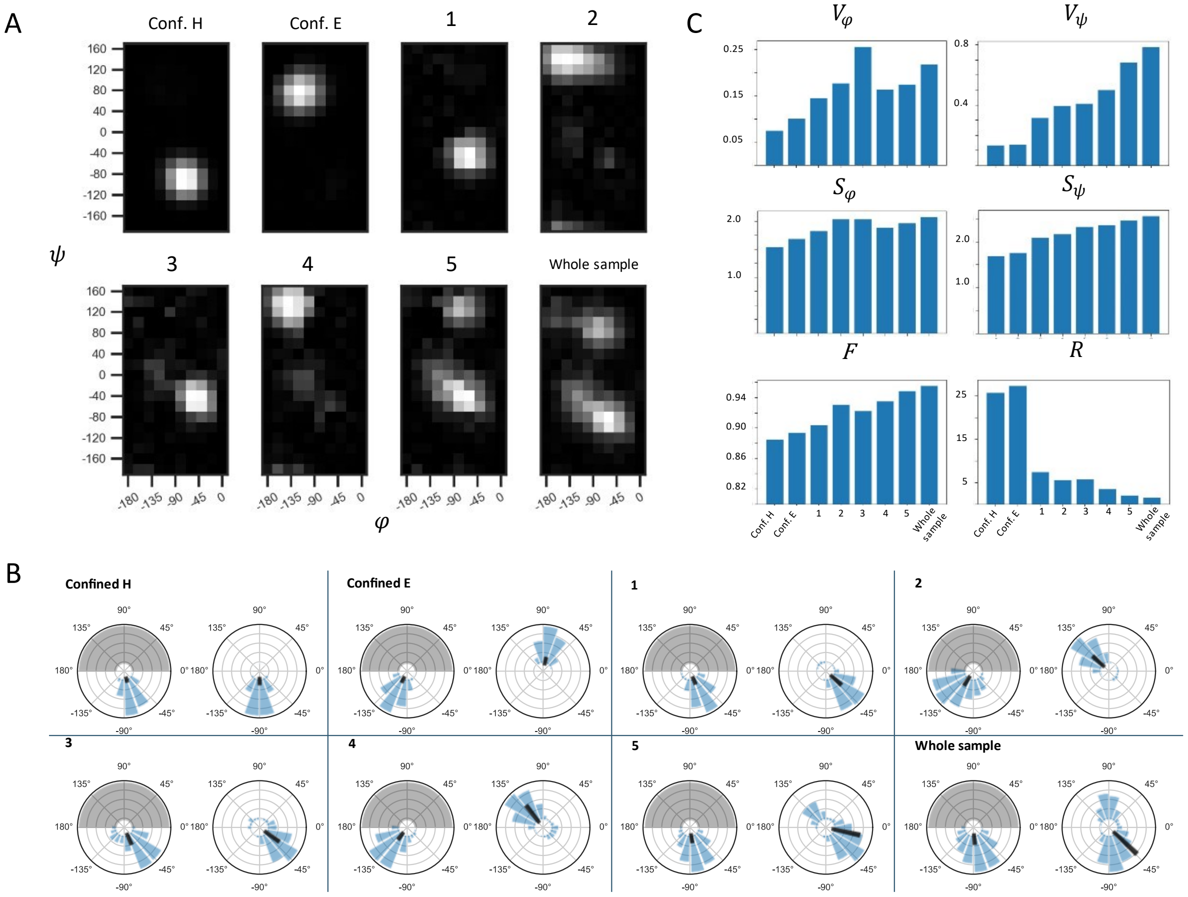


Figure 2.2.7. Exploration of various methods of quantification of heterogeneity in the solid-state NMR sample of this study, applied to the TALOS-N based reconstruction of conformational distributions. **A**: Folded Ramachandran maps of the test coordinates. Panel "Collective" corresponds to the weighted sum of predictions over the whole Ala peak of heterogeneous GGAGG. For generation of pure secondary structure (confined cases), predictions were made for 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 the φ and the right one to the ψ distribution. The gray area denotes the non-valid φ region for the calculations irradicated upon 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.

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 all different events of the prediction):

$$F = \max(D_k) \Big/ \textstyle\sum_{k=0}^{180}(D_k) \qquad (26)$$

Like the entropy, by definition, F is insensitive to the number of modes and rather characterizes how confined the distribution is overall (Fig. 2.2.7C). In addition, it may be interesting to consider the *ratio* between the population of helical and extended regions (R), which is *indifferent* to the prevalent propensity. Here, it is determined from the integral over relative densities in the typical areas of the Ramachandran plot (defined in Fig. M4).

$$R = \begin{cases} H/E & if\ H > E; \\ E/H & if\ H \leq E' \end{cases} \qquad (27)$$

where H and E are the integrals of the allowed regions in the folded (φ, ψ) maps. For the collective prediction, R amounts to ~1.53, with a slight excess of helical properties.

All measures described above and calculated for the eight test cases are summarized in Table 2.1 as well as in the plots in Fig. 2.2.7C.

Table 2.1. Heterogeneity parameters obtained for the folded Ramachandran maps predicted by TALOS-N for the two reference cases (H and E) and local test scenarios (‘Points 1-5’, from the Ala cross-peak in the hCBCANH experiment, defined in Fig. 2.2.3), as well as for the broad heterogeneous peak. 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	$\underline{\text{H}}$ + E	0.18	0.68	1.96	2.47	4.24	1.07	0.949	1.99
Collective	**$\underline{\text{H}}$+ E**	**0.22**	**0.78**	**2.07**	**2.56**	**4.46**	**1.30**	**0.955**	**1.53**

The DANGLE maps (Fig. 2.2.8) are more defined than the ones produced by TALOS-N, and the metrics of spread as well as the secondary structure ratio R make less sense. The entropy and flatness F, which do not require large clusters of pixels, respond well to the apparent visual complexity of the maps and become preferable metrics.

The circular variance of PACSY scatter patterns can be estimated with the formula in its more common notation:

$$V_\theta = 1 - \sqrt{\tfrac{1}{n^2}\left(\textstyle\sum_{i=1}^{n} w_i \cos^2 \theta_i + \sum_{i=1}^{n} w_i \sin^2 \theta_i\right)}, \qquad (28)$$

where n is the number of points and $w_i = I_i \cdot P_i^{-1}$ is the weight of each point that includes point density in the chemical shift space P^{-1} and the NMR peak intensity I. The (φ, ψ) histograms and the parameter charts are shown in Fig. 2.2.9, and listed in Table 2.1. The Confined H and E Subsets were created with the boxes centered at the modes of chemical shifts for 'H' and 'E' STRIDE classes. Although the Subsets 3 and 4 do not comprise enough points for any statistical evaluations (N=4 and 17, Fig. 2.2.5C) and furthermore the Subset 3 is not representative for a typical heterogeneity map of "real" heterogeneous samples, they are still presented for reference – in those cases the heterogeneity metrics become inapplicable, just like in case of DANGLE maps. For the other six maps, all scores – apart from R, which suffers in situations of well-defined propensity – reflect the apparent heterogeneity well; the trends appear different because of a broader range of φ angles as compared to (arbitarily arranged) TALOS-N predictions.

Interestingly, the secondary-structure ratios R obtained for the collective maps are similar (1.53, 1.52, and 1.17 for TALOS-N, PACSY, and DANGLE maps, respectively, Tables 2.1-2.3). This may be merely a special case for the extremely heterogeneous test sample, which was purposely created to sample as much of the Ramachandran space as possible. The fact that all three approaches reconstructed the broad conformational space not only qualitatively similarly but also at quantitatively similar proportions is gratifying and can be interpreted as a sign of mutual validation. The same applies to the variance of ψ angles (V_ψ equals 0.8, 0.78, and 0.8 for TALOS-N, PACSY, and DANGLE maps, respectively). V_φ is comparable for TALOS-N and PACSY (0.22 and 0.17) but is substantially lower for DANGLE (V_φ = 0.08). It should be noted that, the similarity between the relative entropies ΔS_{total} should be rather attributed to a coincidence, since in each case the floor value is different. Generally, the quantitative comparison of the results between the different frameworks is not intended, however, and only different sites of a sample, assessed in each case using the same framework, are the intended subject of study.

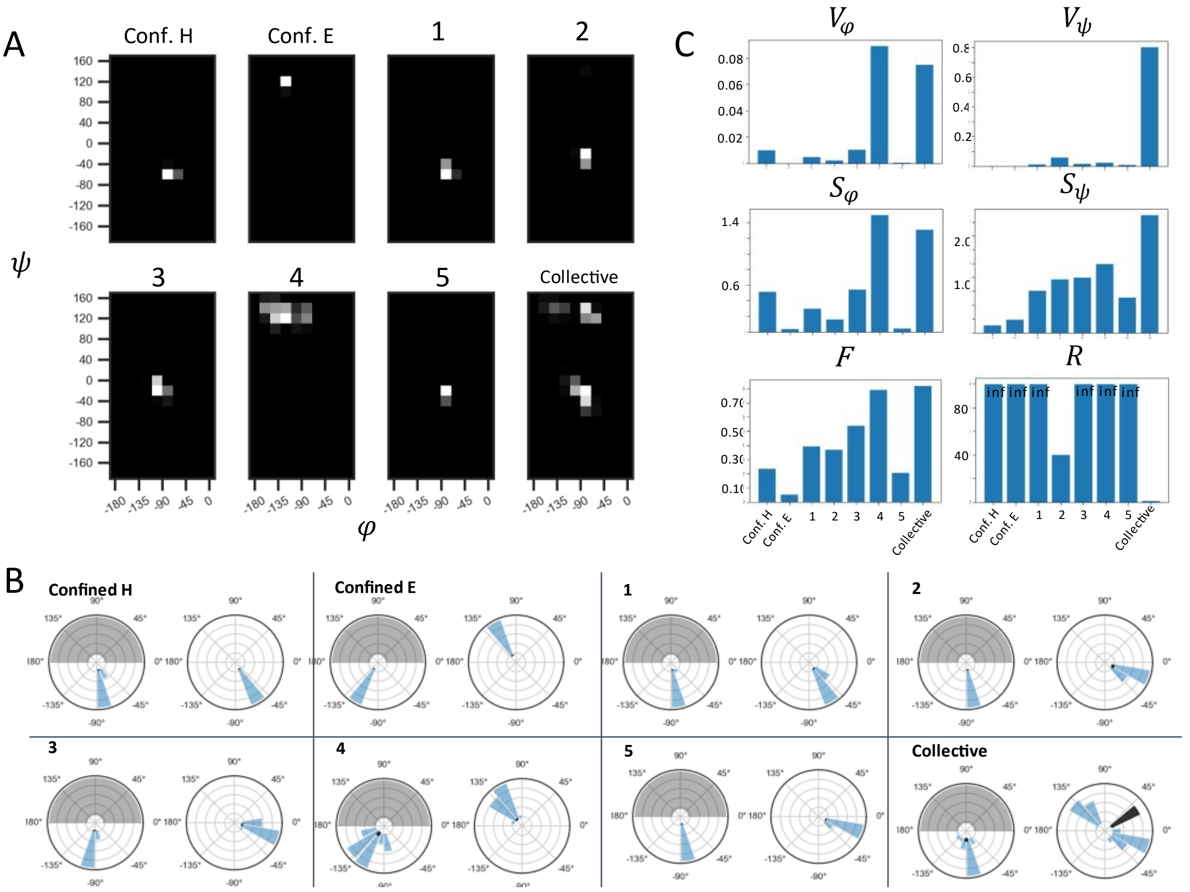


Figure 2.2.8. Tests and reconstruction of conformational distributions based on dihedral-angle predictions with DANGLE. The plots are analogous to those in Fig. 2.2.5. **A**: Folded Ramachandran maps of the eight test chemical-shift combinations. **B**: Ramachandran maps from A in polar coordinates. In each pair, the left plot corresponds to the φ and the right one to the ψ distribution. The 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. **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.

Table 2.2. Heterogeneity parameters obtained for the folded Ramachandran maps predicted by DANGLE for the two reference cases (H and E), local test scenarios ('Points 1-5', from the Ala cross-peak in hCBCANH experiment, defined in Fig. 2.2.3), as well as for the broad heterogeneous peak. 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	0.01	0.00	0.52	0.11	0.62	0.00	0.24	inf.
Conf. E	E	0.00	0.00	0.04	0.19	0.23	-0.39	0.53	inf.
Point 1	H	0.01	0.01	0.31	0.61	0.88	0.26	0.39	inf.
Point 2	H	0.00	0.06	0.17	0.77	0.91	0.29	0.37	40.2
Point 3	H	0.01	0.02	0.55	0.80	1.23	0.61	0.54	inf
Point 4	E	0.09	0.03	1.50	0.99	2.39	1.77	0.79	inf
Point 5	H	0.00	0.01	0.05	0.51	0.55	-0.07	0.21	364.5
Collective	**$\underline{\text{H}}$+ E**	**0.08**	**0.80**	**1.31**	**1.70**	**2.52**	**1.90**	**0.82**	**1.17**

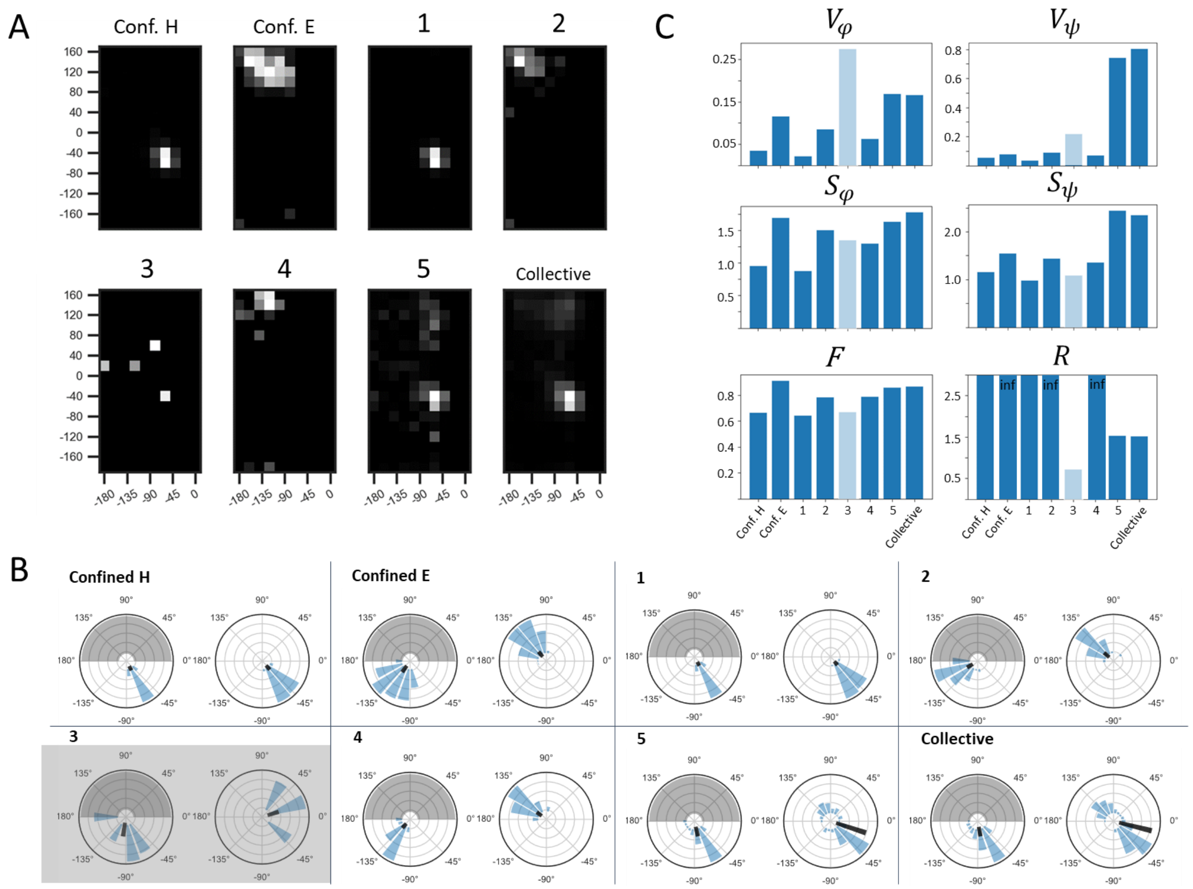


Figure 2.2.9. Exploration of various metrics of (φ, ψ) distributions applied to Ramachandran maps of test selections from the PACSY database (see main text and Fig. 2.2.4). **A**: Density histograms with dimensions of folded TALOS-N maps (18 × 10 bins). **B**: The same, represented by marginal distributions in polar coordinates. In each pair, the left plot represents the marginal φ and the right the marginal ψ distribution. Black bars represent circular variance and point to the mean direction. The plots of Subset 3 are grayed out because they comprise too few points (N=6). **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.

Table 2.3. Quantitative analysis of Ramachandran maps obtained using the PACSY approach, focusing on chemical-shift combinations of confined helix and sheet, positions 1-5, and the full heterogeneous GGAGG peak. 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), which hence represents a clean prediction. N stands for the number of PACSY entries in the respective Subset. 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..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
Collective	**H+ E**	**13565**	**0.17**	**0.80**	**1.78**	**2.36**	**3.94**	**1.92**	**0.869**	**1.52**

2.2.6 Discussion

Using chemical shifts as a direct measure of dihedral-angle properties is a straightforward and sensitive approach that avoids the need for specialized pulse sequences to encode angular features. It allows to include multiple chemical-shift dimensions to increase peak dispersion, thereby minimizing peak overlap and allowing capturing the features of multidimensional chemical-shift distributions precisely. Using proton as the direct dimension increases the sensitivity of the experiment and additionally boosts the SNR of the fine peak features of low intensity.

Translation of the chemical-shift space into the geometrical parameters is challenging due to the low correlation between the backbone chemical shifts and dihedral angles. However, it can be considered reliable enough to differentiate between the two extreme cases of secondary structure (extended and helical conformations). The "intermediate" chemical shifts translate into a mix of secondary structures, which comprises the "turn-like" conformations (Fig. 2.2.6C), naturally clustering in the center of chemical-shift distributions, as well as the traces of the extreme and "intermediate" structures (Fig. 2.2.6F), as demonstrated by the chemical-shift combination of the point of maximum intensity in the alanine crosspeak (Point 5, Figs. 2.2.7-9). Thus, purely chemical shift-based methods are not capable of yielding purely *thermodynamic ensembles* and rather provide *uncertainty ensembles* even for the solid conformational snapshots.

Further complications arise when the backbone chemical shifts are influenced by factors other than geometry. The ring current effects from neighboring residues can shift the entire heterogeneous peak far from the region of major concentration of chemical shifts, which may confuse the prediction algorithms. Furthermore, insufficiently averaged direct spin-spin interactions broaden the proton linewidths of the underlying individual components of the heterogeneous peak. 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), whose decreased correlation with the backbone conformation is already incorporated into the prediction models. The effect of differential dynamic properties on the measured chemical shifts can be neglected since backbone motion on the intermediate and fast time scales is extremely minimized in the statically disordered samples by definition.

The predicted conformational distributions for the model sample of a disordered GGAGG pentapeptide obtained involving the dihedral-angle prediction software (TALOS-N and

DANGLE) are reasonably consistent and mutually validated by the database analysis approach. A truly orthogonal validation by another method of NMR spectroscopy or even other physico-chemical technique, although desirable, is difficult to perform. Outside NMR, the existing high-resolution methods of structural biology – cryo-electron microscopy and diffraction methods – perform poorly on heterogeneous samples (see *Introduction, Section 1.3.1*), and fluorescent or paramagnetic labelling required for FTIR or electron spin resonance methods can be too large to fit into the natively tight structures. Perhaps the best validation of the predictions can be performed by selective isotope labelling of several presumably heterogeneous residues and further ensemble reconstruction with computational methodology involving molecular dynamics (MD), Monte-Carlo simulations and statistical analysis (Bonomi *et al.*, 2017; Bonomi and Vendruscolo, 2019). The test protein system for this experiment must be chosen carefully, such that the sample preparation procedure can be highly reproducible. Even in this case, however, NMR chemical-shift data likely remains the only experimental observable, which makes faithful ensemble reconstruction challenging.

The initial choice of dihedral-angle prediction "engines", TALOS-N and DANGLE, was arbitrary and motivated by their release dates (Among the backbone dihedral-angle prediction software, TALOS-N and DANGLE are the most recent) and convenience of usage, i.e. possibility of a local copy and a dihedral-angle distribution as the output. While within the TALOS family, TALOS-N has clearly been demonstrated to provide highest accuracy, other frameworks may show better performance in particular cases. As such, a comparison involving 33 test proteins performed in Wishart (2011) reported the absolute best accuracy of 94 % achieved by PREDITOR (Berjanskii *et al.*, 2006) (measured by A_{30} score, see Cheung *et al.* (2010) or Wishart (2011) for a definition), as compared to TALOS, TALOS+, DANGLE and SHIFTOR (see *Section 1.3.2*). However, PREDITOR is only available as a server (*which still could not be accessed by the author even by the time of writing the thesis*), which prevents embedding it into customized pipelines. If its source code were available, the workflow presented in this section could be built around it, in a manner that the resulting probability distribution is constructed from the single combinations of (φ, ψ) values provided by PREDITOR as the output. Ultimately, the accuracy of the predicted collective maps depends on the accuracy of the underlying engine, including its robustness in interpretating "exotic" chemical-shift combinations.

Nevertheless, the author believes that the current level of accuracy is sufficient to answer biological questions such as quantifying the ratio between the conformations adopted by disordered residues and obtaining ideas about the structural changes upon changes of physico-chemical conditions (temperature, pH). Unlike the series of selectively labeled samples, this approach does not require the preparations to be reproducible. The maximal length of the primary sequence that can be subjected to the approach depends on the degree of heterogeneity, as wider peak shapes increase the probability of overlap and decrease the signal-to-noise ratio. Fortunately, in most current studies on biological samples by NMR, only part of the residues tends to be variable, reducing the probability of overlap even for longer primary sequences. For the semi-quantitative analysis 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.

2.3 Asssessment of residue-specific heterogeneity in the functional amyloid of hydrophobin $EAS_{\Delta15}$

2.3.1. Introduction

Hydrophobins are a family of small fungal proteins with a high content of hydrophobic residues that often are able to polymerize into filamental structures. The filaments – or rodlets – self-assemble at water-air interfaces, forming a hydrophobic monolayer, thereby altering the surface activity of conidia (asexual spores) or airborne hyphae (filaments that constitute the mycellium and fungal fruit bodies) (Ball *et al.*, 2019). The rodlets are usually approx. 10 nm in diameter. The hydrophobin family includes about 1000 proteins, which are further subdivided into three classes based on their consensus primary sequences and chemical stability of the rodlets. Monomers of all classes are small, 70-150 residues long proteins (5-20 kDa) with long, disordered loops and small, ordered cores, typically rich in β-strands. The two best studied classes of hydrophobins include eight cysteins that form four disulphide bonds, which stabilise the monomeric structures. Class I hydrophobins form highly stable functional amyloids resistant to surfactants and boiling alchohols. Assemblies formed by class II hydrophobins are stabilized only by hydrophobic interactions, without formation of a stable amyloid (i.e. β-sheet rich) structre, and therefore these rodlets require milder conditions for depolymerisation. Class III hydrophobins are characterized by a consensus sequence containing nine cysteins and are not well studied yet. Fragments of the sequences of class I and class II hydrophobins between pairs of cysteins are referred to as loops. Thus, for class I, the segment between the third and forth cystein, Cys3-Cys4, is referred to as L_1, segment Cys4-Cys5 as L_2 and Cys7-Cys8 as L_3.

Hydrophobin EAS from *Neurospora crassa*, named after the "EASily wettable" phenotype of EAS-mutant spores, belongs to the class I with its ability to form chemically stable rodlets whose amyloidic nature was first revealed by FTIR spectroscopy, CD spectropolarimetry and X-ray fiber diffraction (Kwan *et al.*, 2006). Natural EAS rodlets involve four EAS isoforms with slightly different length of the C- or N-termini, which influences the rodlet length (Mackay *et al.*, 2001). Mutation studies revealed that deletion of up to 15 residues in the L_1 loop does not impact the formation and any physico-chemical properties of the resulting amyloids (Kwan *et al.*, 2008). The three-dimensional structures of both, full-length EAS and its truncated construct $EAS_{\Delta15}$, have been characterised by solution NMR (PDB IDs 2FMC and 2K6A) and were found to have a small β-barrel core; the truncation does not affect the overall 3D fold (Fig. 2.3.1B).

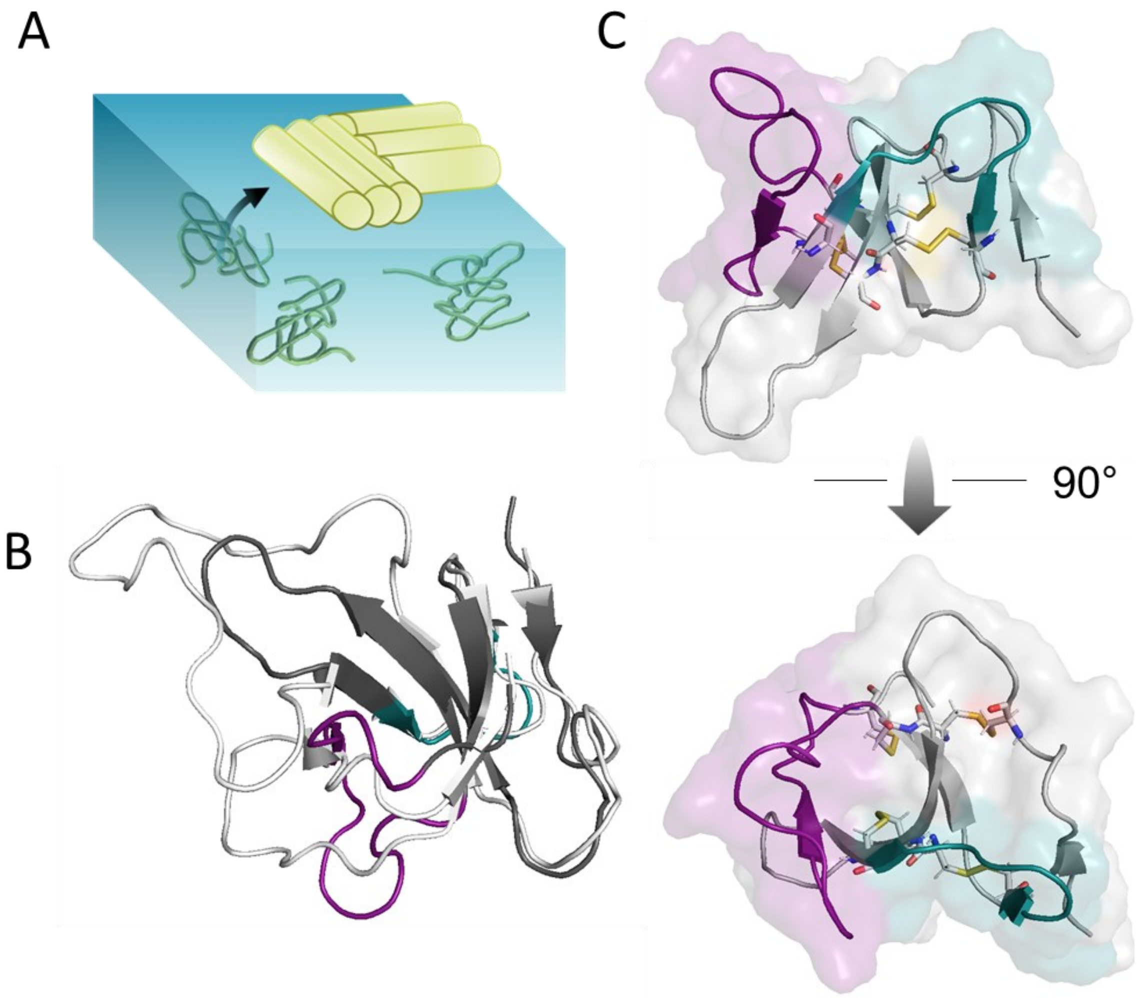


Figure 2.3.1. A: Sketch of the process of rodlet formation. **B**: Superimposed 3D models of monomeric hydrophobin EAS (PDB ID 2FMC, light gray) and EAS$_{\Delta15}$ (PDB ID 2K6A, dark gray) **C**: Orthogonal views on monomeric form of EAS$_{\Delta15}$ stabilized by the disulphide bonds: Cys1-Cys6, Cys2-Cys5, Cys3-Cys4, and Cys7-Cys8. On B and C, loops L$_2$ and L$_3$ are highlighted in teal and purple correspondingly.

Therefore all future structural investigations of the EAS amyloid were performed on the truncated EAS$_{\Delta15}$ construct to simplify the spectra.

The current hypothesis of rodlet formation by EAS or EAS$_{\Delta15}$ suggests the detachment of the L$_3$ loop from the core of the monomer and its exposure to the water-air interface (Kwan *et al.*, 2006; Macindoe *et al.*, 2012; Morris *et al.*, 2013). This proposed mechanism is based on the results of mutation studies, which revealed the critical role of L$_3$ (Fig. 2.3.1C, purple) in rodlet formation. Given that the self-assembley is promoted by a rather oxidative environment, it is most reasonable to assume that no disulphide bridge breaks upon structural rearrangements. The most prominent model of the rodlet core is antiparralel zippers. All attempts of determining the rodlet structure on the basis of this assumption, via NMR and microscopy data, however, have not yield a converged model at the moment of writing *(unpublished data, private* *discussions with S.K.V. and R.L.)*, which raises the demand for more insights based on experimental evidence.

Severe broadening and a diagonal tilt of the hNH peaks is typical to amyloids (Fig 2.3.2A, compare with spectra of nitrogen shifts of FOR005 (Pradhan *et al.*, 2020) and proton-detected hNH spectra of fibrils formed by amyloid β (Linser *et al.*, 2011) or Tau (Xiang *et al.*, 2017)) and is the case for EAS$_{\Delta15}$ rodlet sample, too (Morris *et al.*, 2012). The phenomenon could be attributed to a continuous distribution of the charachteristics of an H-bond network, e.g. H-bond lengths or angles. As opposed to other well-studied amyloids, for example formed by α-synuclein (Tuttle *et al.*, 2016), Tau (Xiang *et al.*, 2017) and Aβ (Niu *et al.*, 2020), EAS$_{\Delta15}$ rodlets, as well as those of another class I hydrophobin, DewA, show a significant degree of structural heterogeneity, manifested in severely broadened peaks in the carbon dimensions (Fig. 2.3.2B). The tunnel electron micrographs of EAS$_{\Delta15}$ (Kwan *et al.*, 2008), EAS as well as other class I fibrils (Pham *et al.*, 2016) do not reveal any non-uniformity of the fibrils, suggesting that disorder manifests itself only on the atomic scale.

The work presented in this chapter is a part of the collective effort of revealing the structure of EAS$_{\Delta15}$ rodlets. Analysis by all high-resolution techniques of structural biology is hampered by the strong structural disorder at the atomic level. Full structure elucidation by solid-state NMR is challenging for the exact same reasons. However, even very basic NMR data – chemical shifts and peak shapes – are rich in information and can provide substantial insights into secondary structure as well as the degree of order and disorder along the amino acid sequence, which would contribute to an understanding of the amyloid core packing and organisation of side chains. The analysis presented below aims to evaluate the residue-specific structural disorder of EAS$_{\Delta15}$ rodlets and generate hypotheses about its nature. The heterogeneity analysis routine presented in *Section 2.2* is applied to a lengthy amyloid sequence for the first time and compared with the alternative approaches based purely on peak shape information. The author expresses hope that heterogeneity assesment can be implemented (here and for other targets of future research) into routines of protein structure calculation (e.g., as weights for other restraints).

2.3.2. General initial assessment of the sample

A sample of 100% back-exchanged u-(^{2}H, ^{13}C, ^{15}N)-EAS$_{\Delta15}$ rodlets was prepared by the group of Dr. Ann Kwan *(The University of Sydney)* as described in Kwan *et al.*, (2006). The severely broadened hNH spectrum (Fig. 2.3.2A) is close-to-identical to the spectrum published earlier in Morris *et al.* (2012). A few peaks in the plane stand out from the subset of low-intensity overlapping peaks, suggesting that some fragments are more ordered than others. For some residues, ^{1}H and ^{15}N chemical shifts are correlated, resulting in a visible tilt of the hNH peaks. Some peaks are broadened and tilted in ^{13}C dimensions (Fig. 2.3.2B).

Altogether, 44 peaks were found in the intraresidual 3D correlations hCANH and hCOcaNH, of which 40 were assigned *(in collaboration with R.L. and S.K.V., TU Dortmund)* (see Fig. 2.3.2A and B). Fewer (distinct) signals are present in the corresponding ^{13}Cα-CO 4Ds (hCACONH, hCOCANH, Fig. 2.3.2C) and only 15 peaks were found in the 4D hCBCANH. These dramatic losses of magnetization are presumed to occur during the chemical-shift evolution periods, which is the only (major) difference between the 3D hcaCBcaNH and the 4D hCBCANH sequences. As a result, the backbone walk could not be done using the 4D data, and assignments were merely transferred from the 3D experiments, which reminds of the limitations of the high-dimensional experiments.

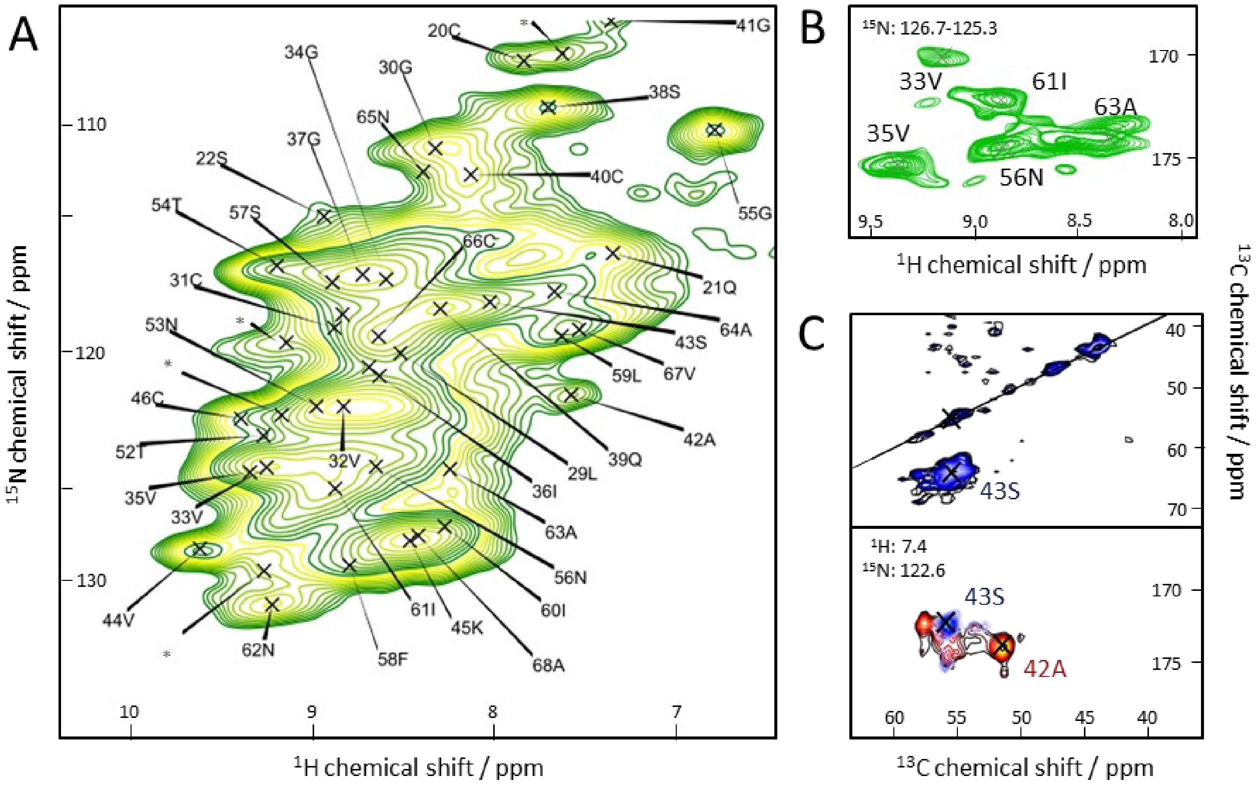


Figure 2.3.2. Solid-state NMR spectra of EAS$_{\Delta15}$ rodlets. **A**: Fingerprint hNH correlation, conservatively apodized (see *Materials and Methods, Table M3.1*). Unassigned peaks are marked with asterisks ('*'). **B**: A slice from a 3D hCONH spectrum, demonstrating notable peak broadening. **C**: Slices of the 4D hCBCANH, hCOCANH (blue) and hCACONH (black-red-yellow color scheme). For acquisition and processing details see *Matherials and Methods*.

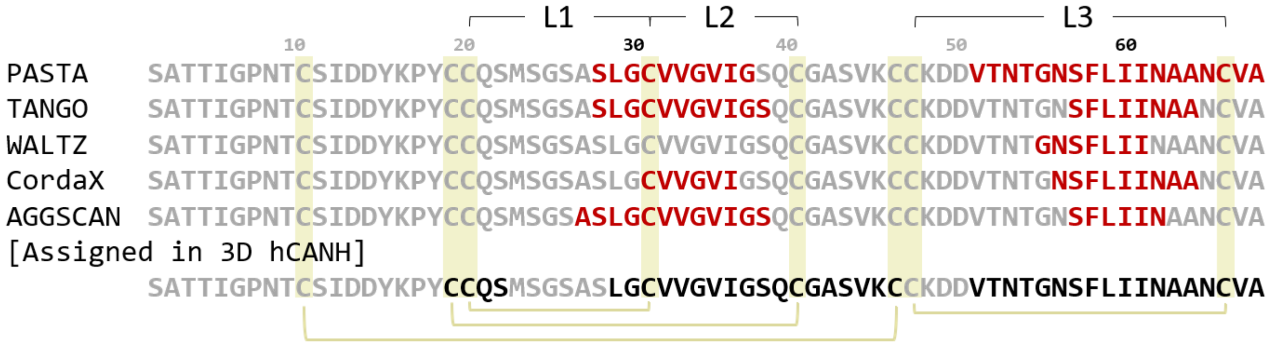

L1
L2
L3
10
20
30
40
50
60
PASTA SATTIGPNTCSIDDYKPYCCQSMSGSASLGCVVGVIGSQCGASVKCCKDDVTNTGNSFLIINAANCVA
TANGO SATTIGPNTCSIDDYKPYCCQSMSGSASLGCVVGVIGSQCGASVKCCKDDVTNTGNSFLIINAANCVA
WALTZ SATTIGPNTCSIDDYKPYCCQSMSGSASLGCVVGVIGSQCGASVKCCKDDVTNTGNSFLIINAANCVA
CordaX SATTIGPNTCSIDDYKPYCCQSMSGSASLGCVVGVIGSQCGASVKCCKDDVTNTGNSFLIINAANCVA
AGGSCAN SATTIGPNTCSIDDYKPYCCQSMSGSASLGCVVGVIGSQCGASVKCCKDDVTNTGNSFLIINAANCVA
[Assigned in 3D hCANH]
SATTIGPNTCSIDDYKPYCCQSMSGSASLGCVVGVIGSQCGASVKCCKDDVTNTGNSFLIINAANCVA

Figure 2.3.3. Amiloidogenic regions of EAS$_{\Delta 15}$ predicted by various algorithms (highlighted in red) on the sole basis of primary sequence. Positions of cysteins are highlighted by yellow shades. The bottom line provides comparison with those residues that could be identified in the sets of 3D assignment spectra. (hCANH contains the most assigned peaks.) Brackets at the connected cysteins form disulphide bridges.

Several methods of primary sequence analysis – TANGO (Fernandez-Escamilla *et al.*, 2004), AGGRESCAN (Conchillo-Solé *et al.*, 2007), WALTZ (Maurer-Stroh *et al.*, 2010), PASTA (Walsh *et al.*, 2014) and CordaX (Louros *et al.*, 2020) – consistently predict aggregation propensity for the segments in the L$_1$-L$_2$ region – most notably, C31-I36 – and for the pentapeptide S57-I61 in the L$_3$ region (Fig. 2.3.3). The predicted amyloidogenic potential of the L$_3$ region is consistent with the findings from mutation analysis, which revealed its key role for rodlet formation (Macindoe *et al.*, 2012). PASTA additionaly predicts disorder of the first ten residues and β-strand propensity for residues C19-Q21, S28-I36, S43-C47 and S57-A63. All of those residues were identified by the 3D assignment experiments.

2.3.3. Analysis of residue-specific static disorder based on TALOS-N routine

The regular TALOS-N analysis of EAS$_{\Delta 15}$ chemical shifts (i.e., prediction of each residue's most likely (φ, ψ) combination from its chemical shift combination via database analysis) expectedly shows the prevalence of β-sheet propensity ($\varphi \sim$ -150°, $\psi \sim$ 150°) in all of the identified sequence segments (Fig. 2.3.4A). Prediction of more extended conformations (i.e., towards φ = ±180°, ψ = ±180°) occur exclusively for glycines (see G34, G41, G55). "More helical" propensity (a decrease of absolute φ and ψ values) is predicted for some residues in the L$_1$-L$_2$ segment and for two residues in L$_3$, although in each such case, the result is marked as "Warn", meaning there were not enough examples of similar sequences and (φ, ψ) distributions in the TALOS database. The six central residues in L$_1$, the remnant of the long, disordered loop in EAS that has been shortened in the Δ15 mutant, escape assignment, which may point to their maintained flexibility in the rodlets.

The heterogeneity scores (variances, entropy, flatness and *R* for the collective maps (Fig. S13), closed circles in Fig. 2.3.4B-E) imply an increasing order towards the N-end of the L$_3$

segment. The fact that for residues C47-V51 the assignments are missing does not necessarily speak for flexibility of this segment since there are three unassigned peaks remaining in the hCANH that may belong there. However, given the flexibility of this loop in the monomer structure, a remnant degree of flexibility, deteriorating CP transfer efficiency is not unlikely. "Uncertainty scores" (open circles in Fig. 2.3.4B-E) – the same metrics as calculated for the collective TALOS-N maps but now applied to the predictions made for peak maximuma only (Fig. S14) – are provided in the figure for reference. Whereas the heterogeneity metrics implicitly include the linewidth information, the "uncertainty scores" for the position of the peak maximum are not influenced by the peak shape. The differences between the uncertainty and heterogeneity scores (bars in Fig. 2.3.4A-E) show the added effect from heterogeneous line broadening to the predicted probabilities / occupancies of backbone dihedral angles in the static ensemble. Interestingly, the trends formed by the uncertainty scores are overall in line with the heterogeneity trends.

The heterogeneity measures fluctuate in both L_1-L_2 and L_3 segments. Residues C40, G41, G55 and A64 show the highest variance V of both, φ and ψ, angles in the collective predictions. C40 and G41 lie in the middle of a short turn in the monomer structure (between strands 4 and 5), and it seems that they are converted into a poorly ordered fragment upon rodlet formation. The same applies for G55, which marks the center of the loop connecting strands 5 and 6. A64 lies in a turn just before strand 7, and the residues in between 55 and 64 – the residues with the highest amyloidogenic properties, which have been predicted to be responsible for rodlet formation (Fig. 2.3.3) – show remarkable low values of heterogeneity.

Collective (φ, ψ) predictions are more defined than the predicted distribution based on peak maxima for G30, S57, A63 and C66 (Fig. 2.3.4, also compare the source maps in Figs. S13 and S14). In particular, the central two flank the well-defined amyloidogenic region in between residues G55 and A64 with a high degree of static disorder. Larger variance of both angles is estimated for G34, as well as again for G55 and A64. Notable differences between the regular and collective maps occur in the L_2, when the variance of one backbone angle increases and that of the other decreases (G37, C40, G41). Interestingly, when compared to the maps predicted from peak maximum information only ('regular' analysis), it is revealed that the variance both rises for some residues and decreases for others when the entire peak is taken into consideration. The differences in the scores effectively report on the effect of the peak broadening and partially discard the uncertainty resulting from the peak position, as in cases when the chemical shift does not belong to the expected ranges for either helices or extended structures (compare *Sections 2.3-2.4*, test Point 5). All in all, the changes in variance are relatively large for the residues that are suspectedly close to the ends of the ordered strands. Another visible trend is the steady decline of disorder from residues 40/41 towards the unassigned (likely flexible) residues 48 – 51. The same applies for the residues flanking the flexible L1 loop. Complementing the first trend, no static disorder is possible for flexible residues, which sample the different conformations faster than manifested in differential chemical shifts.

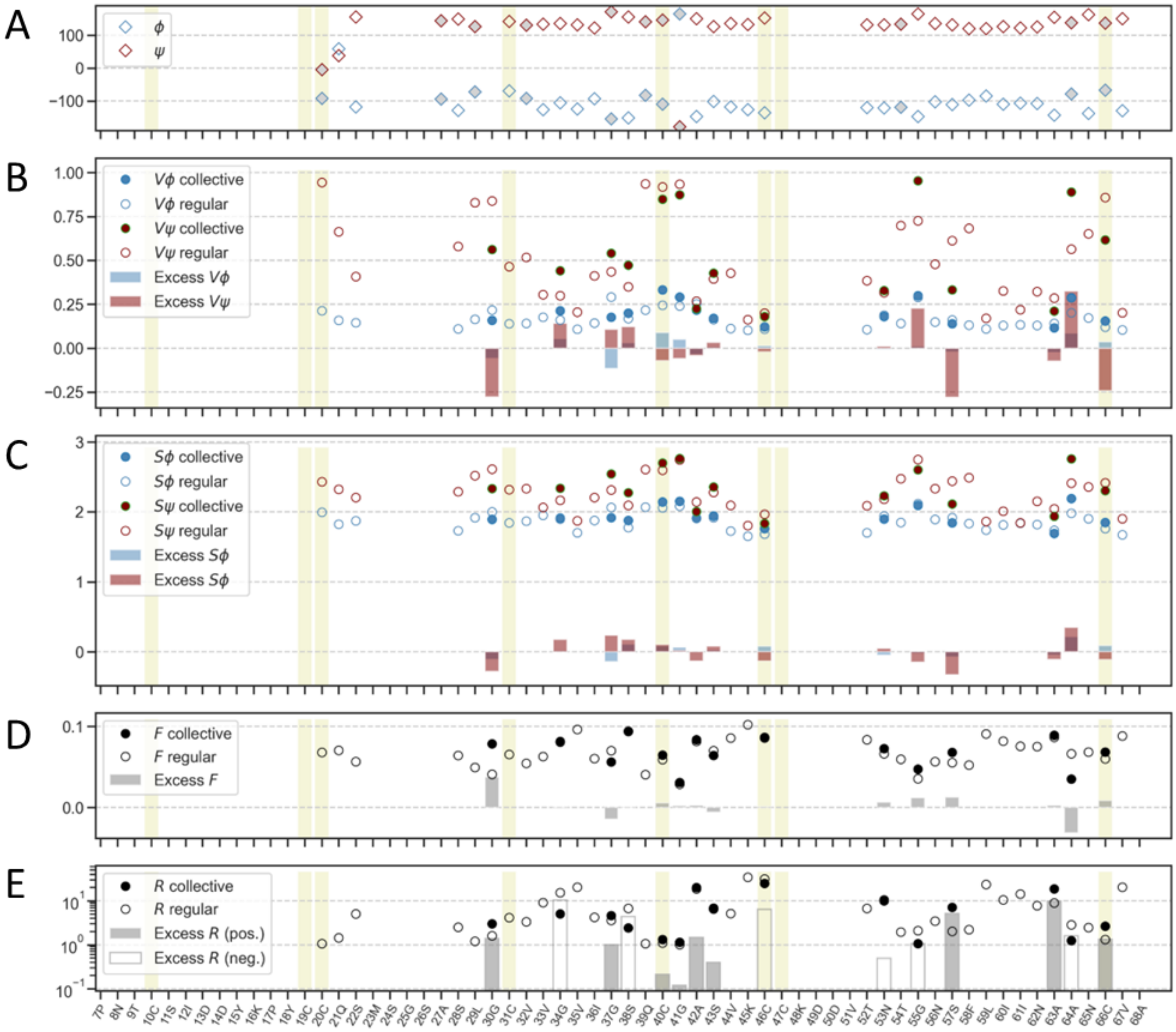


Figure 2.3.4. Residue-specific heterogeneity assessment on the basis of TALOS-N individual and collective predictions for hydrophobin EAS$_{\Delta15}$ rodlets. Positions of cysteins are highlighted with yellow shades. **A**: φ and ψ angles predicted by TALOS-N in the regular workflow. **B-E**: Heterogeneity metrics and “uncertanty scores” (see main text): **B**: circular variance, **C**: entropy, **D**: flatness, **E**: ratio between the helical an dextended structures.

2.3.4. Peak shape analysis

An alternative approach to estimate residue-specific heterogeneity may utilize the raw peak parameters such as line widths and intensity. The collective TALOS-N predictions implicitly include the linewidth information, and the differences between the regular and collective maps should theoretically yield the same patterns as the analysis of the raw linewidths but in the geometrical domain. Extracting the data from the more sensitive 3D experiments is beneficial for the present sample, whereas the 4D hCBCANH yielded only a small fraction of the expected signals. However, this approach has other fundamental advantages. Separate evaluation of the various peak dimensions for an atom allows to discern the *types of disorder* that drive the peak broadening. As such, carbon ^{13}Cα and ^{13}Cβ chemical shifts are more sensitive to the backbone conformational differences, whereas ^{13}CO, ^{15}N and ^{1}H shifts are largely influenced by H-bonding (see *Chapter 1, Section 1.3.2* for a detailed overview). The peak diagonal tilt in the ^{1}H/^{15}N plane – a peak shape parameter not appearing in the literature – encodes the distribution of the H-bond parameters, such as length or angles.

Although it is intuitive to expect the peak width and intensity to be anticorrelated (i.e., that the peak volume for all amide peaks is roughly constant), this is not always the case. For example, in the S38-C46 segment of L$_2$, the intensity varies largely (from around 0.4 up to the absolute maximum of 1.0 in both spectra), building an uptrend towards the C-terminal end of the fragment (at C46), whereas the peak widths in all dimensions remain nearly constant (Fig. 2.3.5A-C). The spikes in signal intensity at S38, Q39 and V44 with the relative spike of V32 imply high rigidity of those residues, as these notable differences can be explained with the differences in CP transfer intensities, which decrease with increasing dynamics. For S38, this coincides with the low heterogeneity scores derived from TALOS-N predictions (Fig. 2.3.4); for the other residues no data is available. Peaks of the three pre-glycine residues of L$_2$ (V33, I36, and C40 or Cys5) are particularly weak and broadened, especially in the ^{15}N dimension of the hcaCBcaNH spectrum. The combinations of narrow lines and low intensity can result from not a static but rather dynamic disorder. Most notably, this occurs for cysteins C31 (Cys4) and C40 (Cys5), which form disulphide bridges with the presumably mobile C20 (Cys3) and C19 (Cys2) (see Fig. 2.3.3) from the N-terminus, which is entirely absent in all NMR spectra.

In the L$_3$ segment, the hcaCBcaNH peaks of the last four residues in a row – N62-N65 – are broadened (N65 – beyond detectability) in all dimensions. Peaks of the same four residues in the hCANH spectrum, on the contrary, sharpen. This peculiar divergence might be explained by the ring-current effect of the side chain of F58, which could be arranged in such a way that it affects specifically the sidechains of the two asparagine and the two alanine residues as well as both Cα and Cβ lines of the preceding I61. However, ring current effects usually entail anomalous chemical shift combinations (i.e., positions of the peak maximum), which is not the case for the fragment I61-N65. Another argument against the attribution to the ring influence,

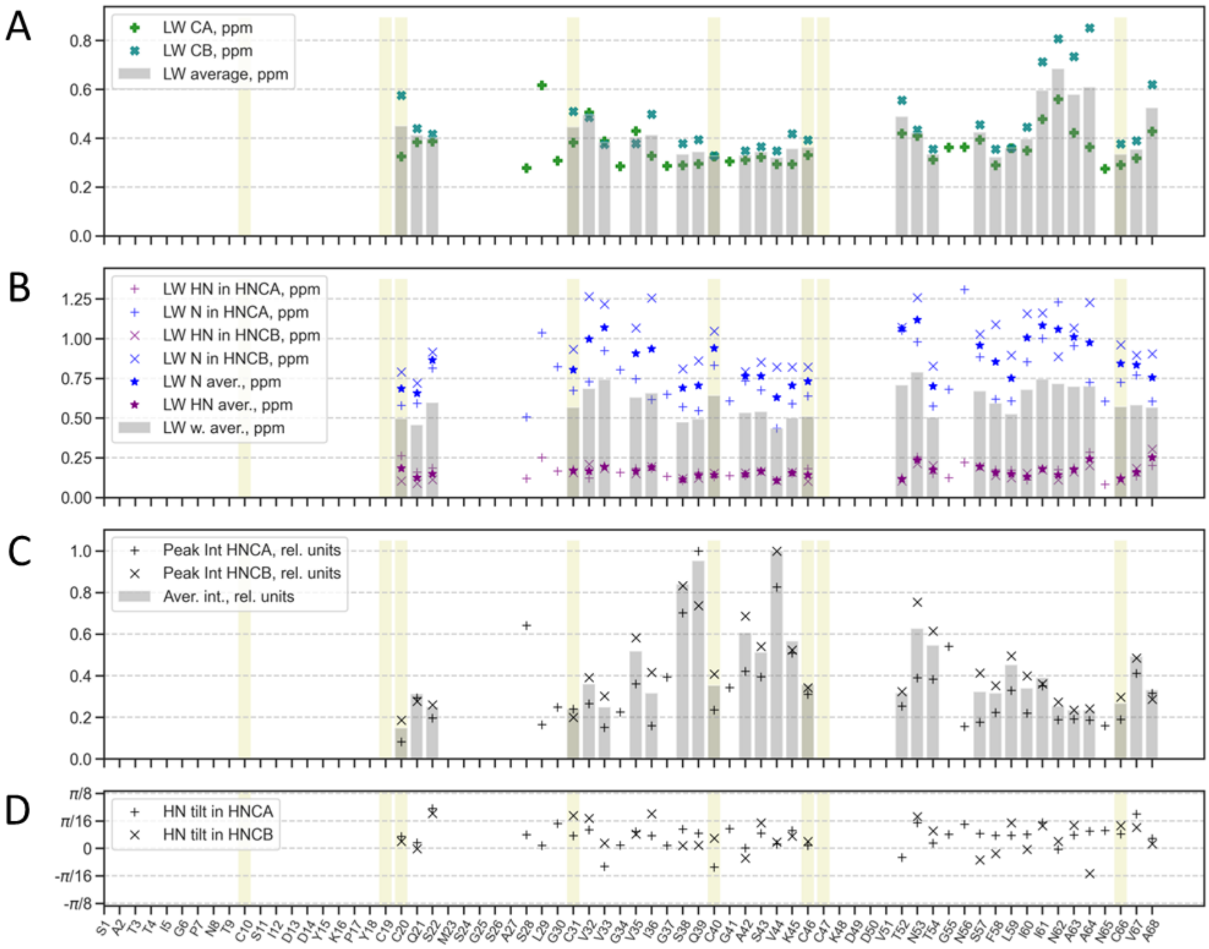


Figure 2.3.5. Charachteristics of 3D peaks in hCANH and hcaCBcaNH experiments (shortly aliased as HNCA and HNCB). **A**: ^{13}C line widths (green '+' for Cα, teal 'x' for Cβ); **B**: amide ^{1}H and ^{15}N linewidths; **C**: relative and average peak intensities; **D**: peak tilt in H/N plane ('+' for hCANH, teal 'x' for hcaCBcaNH spectra). Gray bars on the plots **A-C** represent the average values (in **B** – weighted average). For NMR data processing details and details on parameter calculations, see *Materials and Methods, Section M.3.5*; the raw peaks are shown in Figs S15 and S16; all fits are shown in Figs. S17-S22; the 2D fit residuals are shown in Fig. S23 and S24.

is that theoretically (*Section 1.3.2*), the ring current effect on carbon shielding – hence ^{13}Cβ and ^{13}Cα shifts – is extremely weak (Wishart and Case, 2002).

The TALOS-N analysis (Fig. 2.3.4, *Section 2.3.3*) indeed suggests increasing disorder towards the C-end of the L$_3$ fragment, but implies a high local order of A63, which contradicts the severe peak broadening. Disorder of the two subsequent residues, A64 and N65 is, however, suggested by all estimations. Consensus about aggregation propensity between the prediction methods is achieved for S57-N62 (Fig. 2.3.3). It is possible that N62 is indeed the edge residue of the amyloid core, which precedes the semi-flexible, disordered loop at the C-terminus. C66 may be partially conserved due to the C-C bridge with C47 (whose data is not available) as indicated by small peak widths and lower heterogeneity score.

The last peak shape parameter was extracted from the fits of the two-dimensional peak projections (Figs. S21 and S22; the residuals are shown in Figs. S23 and S24) and presented as a function of residue in Fig. 2.3.5D. Comparatively high positive tilt in the HN-plane is observed for residues I36 (hcaCBcaNH) and V67 (both spectra). An interesting negative tilt is observed for V33 and C40 from the L$_2$ region (both in HNCA spectrum only) and for S57, A64 and the most remarkably T52 from the L$_3$ fragment (in each case in hcaCBcaNH only). The tilt of the L$_3$ residues may be caused by the ring-current effect of the F58.

2.3.5. Discussion

Residue-specific heterogeneity in EAS$_{\Delta15}$ rodlets

The large degree of structural disorder in surface-active amyloids may be attributed to the evolutionary advantages of optimized fibril packing at surfaces even with largely variable curvatures (Morris *et al.*, 2012). Understanding of this packing on the molecular level can provide inspiration for designing novel surface-active materials or antimicotic agents.

Despite the apparent visual homogeneity of the fibrils at the assembly level (Kwan *et al.*, 2008), the monomers of EAS$_{\Delta15}$ rodlet assemblies show pronounced disorder at the atomic level. The N-terminus and the L$_1$ segment are largely abscent in the CP-based experiments, suggesting either high mobility of those regions on the intermediate timescale or highly random static arrangments. The heterogeneity analysis based on the TALOS-N probability maps as well as the linewidth analysis suggest a gradually rising order at the N-end of the L$_2$ segment. Among all considered indicators, only the peak intensities drop at the C-terminal end of the C40-C46 fragment (Fig. 2.3.5C), keeping the linewidths constant and the heterogeneity scores dropping, suggesting increasing conformational order. The dropping intensity may result from the decreasing CP efficiency due to the increasing mobility in the segment towards C46. This is also in line with the hypothesis of a dynamic N-terminal end: C46 forms a disulphide bridge with C10 (see Fig. 2.3.3). Also, no aggregation potential is predicted from the primary sequence for the C40-C46 segment whereas four out of five algorithms suggest some aggregation within L$_2$ (Fig. 2.3.3).

TALOS-N-based scores drop in the middle of L$_{3,}$ for the segment F58-N62, which agrees with the amyloidogenic potential predicted for these residues by all algorithms. The increasing linewidths imply some disorder of the segment N62-N65, manifesting in peak broadening in ^{13}Cβ and proton dimensions. Interestingly, in the hCANH spectrum ^{13}Cα and nitrogen linewidths (Fig. 2.3.5A and B, '+' markers) peak at N62 and steeply decrease towards N65 whereas those dimensions in hcaCBcaNH continue growing. A64 shows exceptionally high heterogeneity scores. This information combined and given the direct proximity of this segment to the C47-C66, one can infer structural distortions towards the two ends of the well-ordered amyloid core formed by the central residues of L$_3$ loop.

All information combined, the residue-specific analysis of the hydrophobin rodlets sheds further light on the previously proposed structural model of the fibrils (Kwan *et al.*, 2006). All residues bear strand-like secondary-structure predictions, denoting a maintained extended character of those residues that are in strands in the monomer form and some rigidification of the loops in between. Only residues 22 to 28 and 47 to 51 maintain their flexibility known from the monomer structure. The putative fibril core, formed by several central residues of the L$_3$ segment, is a well-defined sheet, and a totally disordered outer coat is formed by the N-terminus up to residue 19. Those segments that form strands in the monomer structure tend to be associated again with well-defined secondary-structural (sheet-like) predictions. For the residues that in the monomer structure belong to loops in between these strands (other than the above-mentioned two), in particular around 30, 40, 56, and 66, these scores are substantially ambiguous. Given the limitations of the TALOS-N-based engine discussed above, this may either denote disorder or reflect the unusual shift combinations of the distribution for these residues. This hints to the character of the individual residues from the monomer (loop character but now rigidified) to be largely maintained. The increased linewidths just after the putative amyloidic segment is an interesting observation, which hints to a genuine heterogeneity in the region flanking the amyloid core before the final, C-terminal β-strand.

Interpretation of the peak tilt

Peak tilts in the H/N plane are common – and specific – for amyloid samples (see above, *Section 2.3.1*). Generally, proton and nitrogen chemical shifts are positively correlated ($\delta^1H \approx 14 \cdot \delta^{15}N$, see below), which is also reflected in the overall shape of a typical HSQC spectrum. The peaks of amyloids, including the EAS$_{\Delta15}$ sample, are tilted with an individual slope each, including sporadic negative tilts (hCANH: V33, C40; hcaCBcaNH: A42, S57, A64, both: T52). Since chemical shifts of both, protons and nitrogens, are known to be affected by the H-bonds (see *Introduction, Section 1.3.2*), the slope of the heterogeneous H/N peaks may report on the distribution of the H-bond parameters. These parameters are barely getting attention in the literature due to multiple factors, including the scarcity of proton-detected experiments on amyloids hitherto. This subsection aimes to investigate the factors that may influence H/N peak slopes and evaluates their interpretability.

As follows from a review of the empirical models for the relationships of chemical shifts to the H-bond geometry presented in the *Section 1.3.2*, the major factors affecting the proton and nitrogen shifts are the H-bond angle and interatomic H--O and N--O distance. The heterogeneous peaks with positive and negative slopes can be modeled as an ensemble footprint of homogeneous components with a continuous variation of the H-bond parameters.

For investigation of the influence of each H-bond parameter on the backbone amide chemical shifts, chemical shifts were calculated based on a realistic parameter space using the model of Parker *et al* (2006) for the proton shifts (this is the only model that includes bond angles) and either of the two models (Xu and Case, 2001; Paramasivam *et al.*, 2018) for the nitrogen shifts. The realistic parameter ranges were estimated manually from a solid-state NMR structure of $A\beta_{42}$ fibrils (Xiao *et al.*, 2015, PDB ID: 2MXU); the inter-strand twist angle ρ was set within the range of (0°, ±20°) based on the data from Periole *et al.* (2018). Since the intra-strand twist parameters angle ω and r_ω vary in a narrow range and are related to the chemical shift via a step function (Fig. 1.3.4C), they were assigned to constants ($\omega = 0, r_\omega = 2.65$ Å). The results for the range of ρ are presented in Fig. S25: its influence of the strand tilt angle ρ was found to be only minor, while other parameters were kept constant (also see Fig. 1.3.4D). Instead, high ρ (ρ = 20°) amplifiies the influence of the bond angle θ by increasing the chemical-shift span (Fig. S26). The dependance of the amide chemical shifts on the bond angle and lengths was investivated at a high ρ for better visibility (Fig. 2.3.6A).

As demonstrated in Fig. 2.3.6B and C, neither positive nor negative peak slope contradicts the existing models. The linear dependance reported by Paramasivam *et al.* was used to calculate the nitrogen shifts; the mixed model of Xu and Case can be used to obtain similar trends. The slopes can be modeled assuming different scenarios of structural heterogeneity along the fibril length. As such, the heterogeneous peaks with positive slopes can be formed by the homogeneous components of the monomers with a distribution of intermonomeric distance and constant (or increasing) H-bond angle (Fig. 2.3.6B). A negative tilt was obtained for "flattening" H-bond angle (increasing θ) at fixed positions of the monomers, leading to the decrease of r_{OH} and increase of r_{NO}. In both cases, a normal distribution of the conformation was assumed (i.e. intensities of the individual components follow a gaussian curve).

It is worth emphasizing that the simulated peak shapes may not accurately resemble the observed peaks, their shape and scale and only serve as a first attempt to quantify amyloid peak slopes. Multiple other parameters of the fibril geometry influence proton as well as nitrogen chemical shifts: positions of other electron donors and acceptors, rotations of the charged or

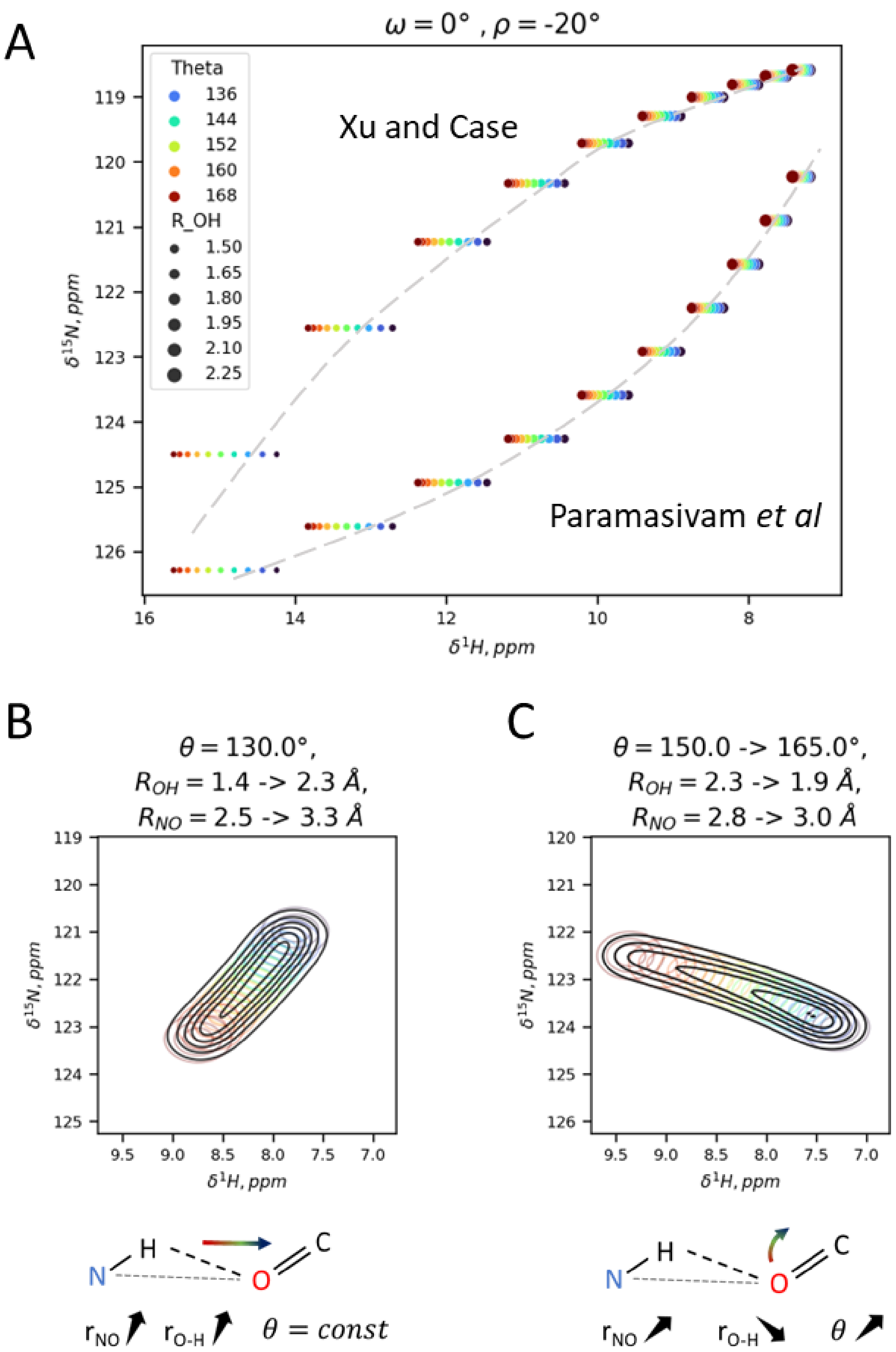


Figure 2.3.6. Influence of the H-bond parameters on the backbone amide chemical shifts: interatomic distances O--H and N--O (r_{OH} and r_{NO}), the H-bond angle (∠H-O-C) θ, inter-strand tilt ρ and the intra-strand tilt ω. The parameters are defined in Fig. 1.3.4B (*Section 1.3.2*) and in the corresponding sources for the models: (Parker *et al.*, 2006) for proton chemical shifts, (Xu and Case, 2001) and (Paramasivam *et al.*, 2018) for nitrogen chemical shifts (in β-sheets). **A**: Chemical shifts calculated with either pair of models for the range of bond lengths and angles. The interatomic N--O distance was approximated as the sum of H--N and H--O bond lengths (r_{OH} + 1 Å). **B** and **C**: Simulated heterogeneous peaks in the H/N plane and sketches of possible structural variations causing the shift differences among invidual ensemble members. **B**: The positive slope caused by the increased distance between the monomers in a fibril; **C:** the negative slope caused by rotation of the amide and carbonyl bonds. Intensities of the individual conformers are normally distributed; the linear model of Paramasivam *et al.* for nitrogen shifts was used to simulate the peaks in **B** and **C.**

aromatic sidechains (consider the proximity of F58 and its effect on the carbon linewidths discussed above), φ and ψ angles of the donor and acceptor residues.

The same reasons make it is difficult to understand the effect of the H-bond configuration on the broad range of experimental chemical-shift data coming from different proteins. An attempt was made to investigate statistical data for the residues from the extended structures (class 'E') with respect to the residue type. The chemical shifts and the N-O interatomic distances collected from 2837 proteins (see *Materials and Methods*, *Section M.3.5* for details on the data collection and distance estimation) show weak negative correlations (Figs. 2.3.7 and S26, the underlying data shown in Fig. S27). Stronger correlations are observed for the proton shifts, as expected from the empirical models. Chemical-shift correction for the sequential context (data for alanines shown in Fig. S28) did not lead to any significant changes. Althogh the sensitivity of both, δ^1H and δ^{15}N shifts of different residue types, may seem different at first, the data should be further denoised before making conclusions. As such, the patterns resulting from the variation of a small group of parameters (Fig. 2.3.6) can be obfuscated by the sequential context (Wishart *et al.*, 1995). An additional crucial factor, as emphasized by Parker *et al.* (2006), is a non-optimized geometry of structures, which have been solved without protons (as it is the case for all X-ray structures). This may lead to worng estimates of interatomic distances. Thus, the author highly encourages continuation of this research which, due to the time restrictions, she could not complete herself.

Linewidths and TALOS-N scores as heterogeneity metrics

It was shown that sequential peak characteristics, such as linewidths and peak intensities, can be used as simple metrics for residue disorder that provide complementary information to the heterogeneity metrics obtained with TALOS-N. Most valuably, the linewidth approach allows to focus on different nuclei and their combinations, thereby allowing evaluating the linebroadening effects separately and making hypotheses about the nature of structural disorder.

The trends formed by either set of indicators, the TALOS-N scores or peak characteristics, should be interpreted with caution due to the large variety of factors affecting the predictions as well as the chemical shifts and hence linewidths themselves. Rising TALOS-N heterogeneity scores for narrow linewidths can occur due to adoption of defined conformations that, however,

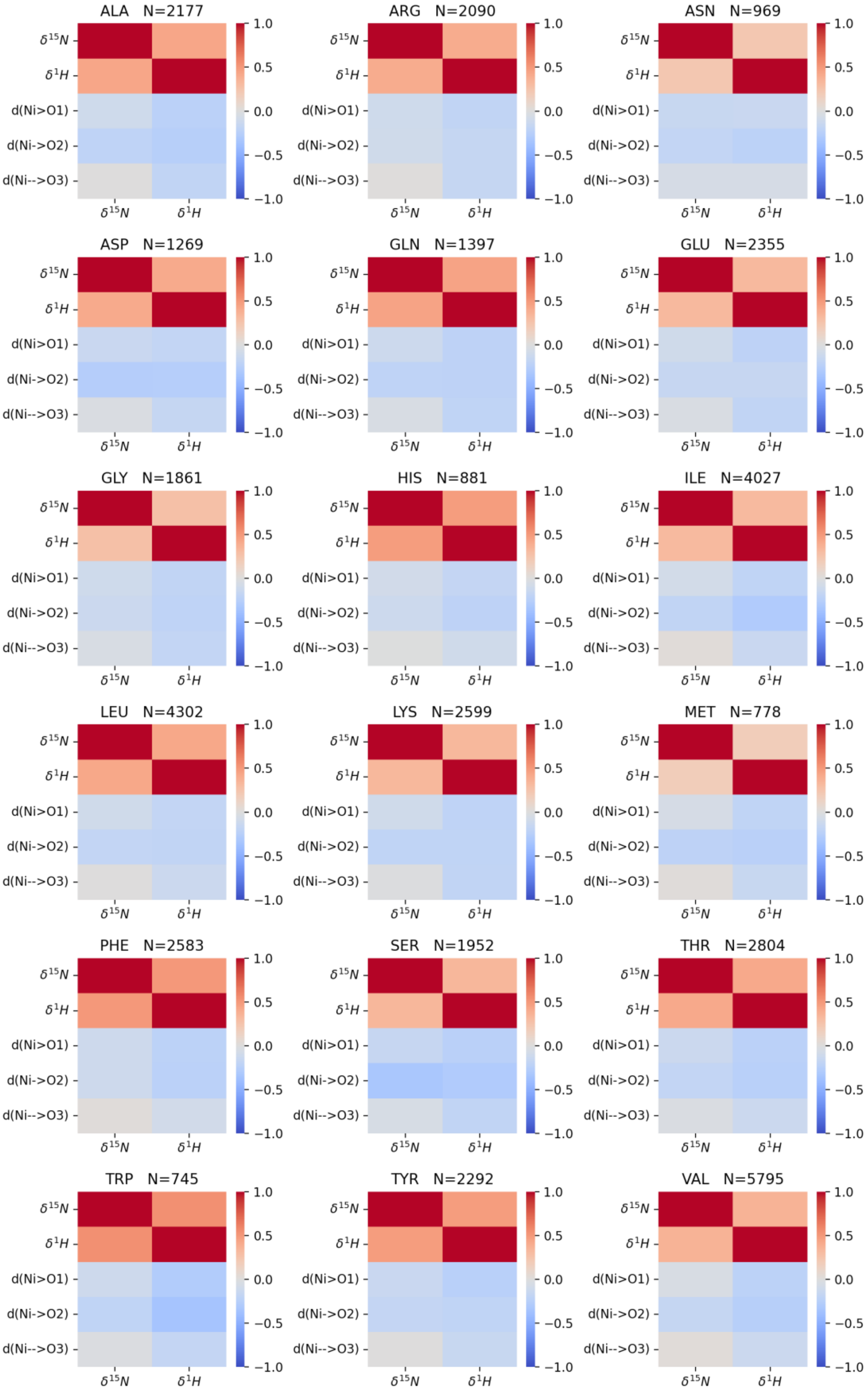


Figure 2.3.7. Pearson correlation coefficients for amide protons and nitrogens belonging to extended-structure ('E') chemical shifts and the three lowest N--O interatomic distances N_i>O_1, N_i->O_2. N_i-->O_3. The PDB and BMRB data of 2116 proteins are related by the tables from ReBoxitory (Maciejewski *et al.*, 2017). Note, that *N* reflects only the number of the shortest N--O contacts (N_i>O_1,), number of longer contacts (N_i->O_2 and N_i-->O_3) is substentially lower.

do not fall into the typical chemicals-hift regions (i. e. α-helical or β-sheet shifts). The peak tilts provide an additional perspective on the disorder within the assembley, differentiating between heterogeneous H-bond distances or rather heterogeneous angles. However, the slopes are not always reproducable from the transfer scheme to transfer scheme (hCANH vs. hcaCBcaNH), which points to the signal-to-noise shortcomings for heterogeneous samples. The behaviour of all proposed metrics should be further explored on a flash-frozen heterogeneous polipeptide with intrinsically disordered elements, whose conformational ensemble in solution has been reconstructed based on orthogonal high-resolution (NOESY) as well as medium resolution data (FRET or EPR). However, this would assume that the shock-freezing quantitatively retains all secondary-structural properties, which is unlikely to happen. All in all, even if not perfect, the proposed approaches to measure and evaluate the residue-specific disorder can provide valuable insights into the organisation of the polipeptide chain at the molecular level and come of use for the validation of propsed biological models when other physico-chemical data are limited.

3 | OUTLINE AND PERSPECTIVES

This thesis presents approaches for i) extracting conformational distributions from heterogeneously broadened peaks; ii) measuring static disorder based on the obtained (φ, ψ)-domain data and linewidths parameters. Despite the apparent simplicity of the presented procedures, as it is common in science, the devil is always in the details. In case of a workflow involving the many-to-one parameter mappings, a part of the challenge is to properly apply the underlying methods and to define how to not over- and underinterpret the outcome. The heterogeneity trends (formed by either score) should be considered in the light of the residue type as well as keeping in mind the possible special arrangement of the chain. The obtained collective (φ, ψ) distributions can be passed downstream as restraints for MD-driven ensemble reconstruction either as maps or as, perhaps, pairs of circular variance (V_φ, V_ψ).

The presented workflows of residue-specific heterogeneity assessment pursue the "top-down" strategy of going from the observables (here, chemical shifts) into the underlying structural features. As opposed to the other "top-down" studies where interatomic distances/angles obtained with dipolar-recoupling techniques were average, the grid sampling of the peak allows preserving the distribution parameters.

In the sight of improving QM chemical-shift prediction offered by hybrid MD-QM/MM techniques (Yi *et al.*, 2023b), the "bottom-up" approaches for peak shape analysis akin to those by Heise et al (2005) and Uluca et al (2018) (*Section 1.3.3*) are more involved but may gain advantage over the dihedral-angle prediction methods.

The combination of NUS and high dimensional experiments allows circumventing the expensive and time-consuming preparation of selectively labelled samples. While here the analysis was based on 4D spectra, the chemical-shift correlations might be further expanded into even higher dimensionality (5D+) by techniques of projection reconstruction developed for rendering spatial models from the images at different angles (Cremers *et al.*, 2011). Applied to NMR, reconstruction of spectra was implemented as GAPRO algorithm and APSY (Automated Projection SpectroscopY) (Hiller *et al.*, 2005) - however, as of now, these techniques recover only positions of peak maxima without the peak shape information. Projection spectroscopy of statically disordered samples might allow resolving features of heterogeneous peaks in the dimensions of all backbone nuclei (plus ^{13}Cβ) at once, providing better chemical-shift inputs. Sensitivity is expected to remain the major shortcoming – however, emerging methods of pulse optimization based on optimal control (Tošner *et al.*, 2021) can be tested for heterogeneous samples and, if successful, would significantly improve signal-to-noise. Even more pronounced advantages are expected for the emerging cryo-MAS probes (Hassan *et al.*, 2020). Application of projection spectroscopy or NUS may allow expanding dimensionality for the DNP approaches pursued hitherto, currently recorded at most as the 3Ds. This study shows (*Section 2.1*), that faithful reconstruction of the complex peak shapes is possible, so it is at least worth testing on the DNP data.

Another fundamental question for future research is how directly the immobilized conformational ensembles represent the underlying dynamic ensembles? In the very recent study of Yi *et al.*(2023a), the authors make the following observation for Ile60 in selectively labeled DHFR. They report that at room temperature in solution, this residue is structured by all indicators, however, at cryogenic DNP conditions, at least three conformations and a lower order parameter were observed. As suggested by the authors, this may imply that the cryogenic linewidths also capture internal motions that happen in solution on fast timescales. Therefore, future studies should study the chemical-shift distributions in tandem with relaxation measurements.

All in all, such holistic approach to protein structural biology, including investigation of structure, including the distribution of conformational parameters, in addition to dynamics within a large range of time scales, promises a greater understanding of the patterns in the organization of the polypeptide chain, opening additional ways to better target protein complexes using medical drugs, improve biotechnological processes and design protein machinery *de novo*.

M | MATERIALS AND METHODS

M.1. Sample preparation

M.1.1. fMLF

Micro-crystalline powder fMLF (Merck) was dehomogenized by dissolving it in 50% ethanol and freeze-drying at high vacuum. The powder was packed into a 1.3 mm ZrO_2 MAS rotor with rubber plugs.

M.1.2. GGAGG

The dehomogenized sample of u-(^{13}C, ^{15}N)-GGAGG pentapeptide was purchased as a lyophilized powder from Thermo Fisher Scientific Inc. The provided HPLC profile and the mass-spectrum confirmed >95% purity. To ensure a high degree of heterogeneity, the sample was redissolved in ddH_2O and freeze-dried over 18 hours under absolute pressure of 10 mbar. The obtained slightly yellow, glassy material was centrifuged into a 1.3 mm ZrO_2 MAS rotor overnight.

M.2. Acquisition and processing of NMR data

All NMR experiments were recorded on Bruker spectrometers with AVANCE NEO consoles. Spectra of GGAGG were recorded on a spectrometer with a proton Larmor frequency of 700 MHz (magnetic field of 16.54 T) at an MAS rate of 40 kHz. Spectra of fMLF were recorded on a 800 MHz spectrometer with an MAS rate of 55.5 kHz. Target cooling temperature in all cases was set to -25 °C (248.15 K), which corresponded to effective sample temperatures of approx. 10 °C at a MAS rate of 40 kHz (for the GGAGG sample) and 15 °C at 55.5 kHz (for the fMLF sample). Shimming and adjustment of the rotor angle inside the probe were pursued using a separate KBr sample. The spectrometers were operated with TopSpin (v. 4.0.8).

Poisson-gap sampling schedule for non-uniform sampling was generated via the web tool http://gwagner.med.harvard.edu/intranet/istHMS/gensched_new.html.

The acquisition and processing parameters of uniformly sampled spectra of fMLF are listed in Tables M1.1 and M1.2. The parameters for experiments on the GGAGG sample are listed in Tables M2.1 and M2.2. The acquisition and processing parameters of the $\mathrm{EAS}_{\Delta15}$ data are listed in Tables M3.1-M3.4.

M.2.1. External chemical-shift referencing

Chemical-shift referencing was done externally on a sample of 1 % 4,4-dimehyl-4-silapentane-1-sulfonic acid (DSS) solution in D_2O using an MAS rotor without spinning.

The direct dimension was calibrated by the command `cal` with option `manual calibration` and setting of the methyl signal to 0 ppm. As the result, TopSpin saves the absolute methyl resonance frequency into the parameter `SF`. In all the other experiment, the `SF` for heteronuclei was set as `SF` (^{1}H) divided by the ratio of the relative proton $\Xi_{1\mathrm{H}}$ and heteronucleus Ξ_{X} Larmor frequencies (Markley *et al.*, 1998). For the carbon dimension, $\frac{\Xi_{1\mathrm{H}}}{\Xi_{13\mathrm{C}}} = 3.97$, for nitrogen, $\frac{\Xi_{1\mathrm{H}}}{\Xi_{15\mathrm{N}}} = 9{,}87$. The nitrogen dimension can also be referenced with the default settings by simply dividing the automatically calculated `SR` (^{1}H) value by the correction factor of $\frac{\Xi_{1\mathrm{H}}}{\Xi_{15\mathrm{N}}}$. However, analogous calculation of the `SR` (^{13}C) parameter would result in a 2.66 ppm offset. The reason for this is the default referencing of the Larmor frequency of ^{13}C on Bruker spectrometers to TMS standard as opposed to DSS. Therefore, for the carbon dimension the author recommends the procedure described above.

M.2.2. Data processing

All lower-dimensional (1D and 2D) experiments were processed in TopSpin. Uniformly as well as non-uniformly sampled 3D and 4D spectra reconstructed by hmsIST and SMILE were processed in NMR Pipe (v. 10.9, rev. 2021.258.11.26 64-bit) (Delaglio *et al.*, 1995), apart from the 3D hCANH, 3D hcaCBcaNH and other spectra of the $\mathrm{EAS}_{\Delta15}$ rodlets that were used for the assignments: those were recorded and processed with TopSpin (*the data were obtained in the final form from R.L.).* The 4D and 3D data that were subjected to reconstruction by SSA were processed by the in-built functions of the SSA package. Baseline correction was performed with in-built functions.

Apodization of the data were done using the standard functions similarly implemented into every software:

Exponential: $\exp(-\mathrm{K}\pi t/w)$ denoted “Exp., K” in Tables M1.1-3.4;

Sine squared: $\sin\left(\frac{\pi t}{2K}\right)^2$ denoted "Sine sq., K" in Tables M1.2 and 3.1.

where t is the time-domain data point and w is the spectral window of the direct dimension; $w = 20833.334$ Hz; K is the weighting coefficient.

Table M1.1. Acquisition and processing parameters of the 3D hCONH spectrum of the u-(^{13}C, ^{15}N)-fMLF sample

	^{13}C (F1)	^{15}N (F2)	^{1}H (F4)
Base frequency, MHz	176.11	70.97	700.17
Number of points	35 x 2	30 x 2	2048
Spectral width, ppm	10.16	29.8	29.75
Offset, ppm	174.4	120.0	6.1
Number of scans	24		
CP H → Cα	^{1}H 15.1 kHz, tang 3 m; ^{13}C 7.93 kHz, rect.;		
CP Cα → N	^{13}C 14.0 kHz; ^{15}N 33.0; 8.2 m ramp 90100		
CP N → H	^{15}N 35.2 kHz; ^{1}H 2.83 kHz; 0.14 m ramp 70100		
Apodization	-	-	Exp., 50 Hz
Zero filling up to, points	512	512	2048

Table M1.2. Acquisition and processing parameters of the 4D hCOCANH spectrum of the u-(^{13}C, ^{15}N)-fMLF sample

	^{13}CO (F1)	^{13}Cα (F2)	^{15}N (F3)	^{1}H (F4)
Base frequency, MHz	176.11	176.11	70.97	700.17
Number of points	14 x 2	23 x 2	27 x 2	2048
Spectral width, ppm	11.3	21.8	54.2	29.75
Offset, ppm	172.0	52.55	123.0	6.1
Number of scans	US: 8; NUS: 16			
Points in NUS schedule	435			
CP H → CO	Pldb27 6.3 kHz spdb6 15.8 kHz tang 2 ms			
BSH-CP CO → Cα	19.9 kHz, ramp100->70, 5m			
CP Cα → N	Pldb8 15.8 kHz Spdb5 29.4 kHz 9 m ramp90100			
CP N → H	Pldb5 35.2 kHz spdb2 2.75 kHz ramp 70100			
Apodization	US, IST, SMILE: Sine sq., 0.3; SSA: Exp., 50 Hz			Exp., 50 Hz
Zero filling up to, points	128	128	128	2048

Table M2.1. Acquisition and processing parameters of the 2D $^{13}C/^{13}C$ correlation on the u-(^{13}C, ^{15}N)-GGAGG sample. CP power levels do not account for the shape effects.

	^{13}C (F1)	^{13}C (F2)
Base frequency, MHz	201.24	201.24
Number of points	59 x 2	2856
Spectral width, ppm	42	405
Offset, ppm	35	35
Number of scans	256	
CP H → Cα	^{1}H: 8.16 kHz, ramp100->80; ^{13}C: 14.9 kHz; 1.71 ms	
DREAM Cα → Cβ	35.9 kHz, 5 m, tangential 75%	
Apodization	Exp., 70 Hz	Exp., 70 Hz
Zero filling up to, points	1024	4096

Table M2.2. Acquisition and processing parameters of the 4D hcaCBCANH spectrum on the u-(^{13}C, ^{15}N)-GGAGG sample. CP power levels do not account for the shape effects.

	$^{13}C\beta$ (F1)	$^{13}C\alpha$ (F2)	^{15}N (F3)	^{1}H (F4)
Base frequency, MHz	201.24	201.24	81.10	800.3
Number of points	2 x 48	2 x 48	2 x 21	2048
Spectral width, ppm	59.9	59.9	33.1	26
Offset, ppm	14.1	29.0	112.5	4.40
Number of scans	8 x 7			
Points in NUS schedule	2401 (≈ 5%)			
CP H → Cα	^{1}H: 69.09 kHz, ramp100->80; ^{13}C: 21.02 kHz, rect; 0.48 ms			
DREAM Cα → Cβ	35.9 kHz, tangential, 75%			
CP Cα → N	^{13}C: 19.3 kHz, rect.; ^{15}N: 16.82 kHz, ramp90->100; 6 ms			
CP N → H	^{15}N: 30 kHz, rect.; ^{1}H: 7.9 kHz, tangential, 75%; 0.15 ms			
Apodization	Exp, 200 Hz	Exp, 200 Hz	Exp, 400 Hz	Exp, 500 Hz
Zero filling up to, points	128	128	128	2048

Table M3.1. Acquisition and processing parameters of the 2D hNH spectrum of the u-(^{2}H, ^{13}C, ^{15}N)-EAS$_{\Delta15}$ rodlet sample (100% back-exchanged)

	^{15}N (F1)	^{1}H (F2)
Base frequency, MHz	70.95	700.17
Number of points	90 x 2	2048
Spectral width, ppm	55.92	39.7
Offset, ppm	120.0	6.7
Number of scans	8	
CP H → N, N → H	^{15}N: 30 kHz, rect.; ^{1}H: 7.9 kHz, tangential, 75%; 0.15 ms	
Apodization	Sine squared	Exp. 30 Hz
Zero filling up to, points	2048	4096

Table M3.2. Acquisition and processing parameters of the 3D hCANH spectrum of the u-(^{2}H, ^{13}C, ^{15}N)-EAS$_{\Delta15}$ rodlet sample (100% back-exchanged)

	^{13}C (F1)	^{15}N (F2)	^{1}H (F3)
Base frequency, MHz	201.21	81.08	800.15
Number of points	32 x 2	20 x 2	2048
Spectral width, ppm	29.8	41.1	39.7
Offset, ppm	54.27	119.6	5.2
Number of scans	48		
Apodization	Sine squared	Sine squared	Gauss. -10
Zero filling up to, points	256	256	2048

Table M3.3. Acquisition and processing parameters of the 3D hcaCBcaNH spectrum of the u-(^{2}H, ^{13}C, ^{15}N)-EAS$_{\Delta15}$ rodlet sample (100% back-exchanged)

	^{13}C (F1)	^{15}N (F2)	^{1}H (F3)
Base frequency, MHz	201.21	81.08	800.15
Number of points	80 x 2	20 x 2	2048
Spectral width, ppm	79.99	41.1	39.7
Offset, ppm	54.27	119.6	5.2
Number of scans	32		
Apodization	Sine squared	Sine squared	Gauss., -10
Zero filling up to, points	512	256	2048

Table M3.4. Acquisition and processing parameters of the 4D hcaCBCANH spectrum on the sample of u-(^{2}H, ^{13}C, ^{15}N)-EAS$_{\Delta15}$ rodlets (100% back-exchanged)

	^{13}Cβ (F1)	^{13}Cα (F2)	^{15}N (F3)	^{1}H (F4)
Base frequency, MHz	176.1	176.1	70.96	700.175
Number of points	29 x 2	47 x 2	70 x 2	2048
Spectral width, ppm	1388.9	9259.3	2525.3	20833.3
Offset, ppm	45	45	120	6.70
Number of scans	8 x 7			
Points in NUS schedule	1991 (≈ 2%)			
CP H → Cα	^{1}H: 34.0 kHz, tangential, 75 %; ^{13}C: 16.4 kHz, rect.; 1.71 ms			
DREAM Cα → Cβ	32.5 kHz, tangential, 75 %; 4 ms			
CP Cα → N	^{13}C: 18.2 kHz, rect.; ^{15}N: 33.75 kHz, ramp90->100.; 11 ms			
CP N → H	^{15}N: 37.8 kHz, rect.; ^{1}H: 12.0 kHz, tangential 75 %; 0.45 ms			
Apodization	Exp, 70 Hz	Exp, 70 Hz	Exp, 70 Hz	Exp, 50 Hz
Zero filling up to, points	128	128	128	1024

M.3. Data handling and analysis

The NMR spectra were visualized and analyzed in NMRFAM Sparky (v. 1.414) (Lee *et al.*, 2015) CCPNMR (v.2.4.5) (Skinner *et al.*, 2016)

All numerical operations, calculations, fitting and plotting were done in Jupyter Notebooks (v. 6.4.11), IPython v.8.4.6, using the Python 3.8-11 kernel. The following packages were used (versions as last tested):

numpy	v.1.22.4	(Harris *et al.*, 2020)
pandas	v.1.4.2	(McKinney, 2010)
scipy	v.1.7.2	(Virtanen *et al.*, 2020)
nmrglue	v.0.8	(Helmus and Jaroniec, 2013)
matplotlib	v.3.5.1	(Hunter, 2007)
seaborn	v.0.11.2	(Waskom *et al.*, 2017)

All procedures described in this work are available at the GitHub repository github.com/eburakova/protein_heterogeneity_ssnmr. In the following, code excerpts are provided for reference.

M.3.1. Tests of the NUS reconstruction algorithms

Main text: Section 2.1

Subsampling from a uniformly sampled dataset

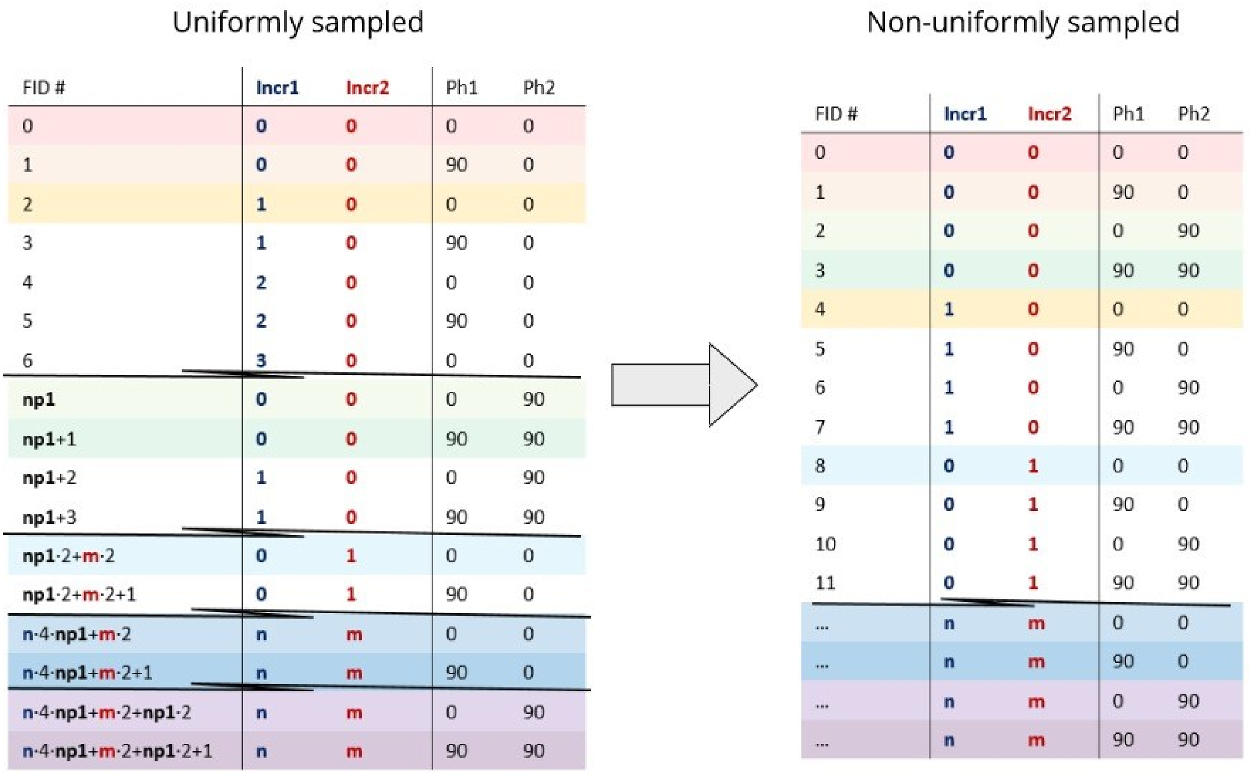
NUS schedules with Poisson-gap weighting were generated at http://gwagner.med.harvard.edu/intranet/hmsIST/gensched_new.html. Random schedules with Gaussian weighting were generated at random schedule with the exponential weighting was generated with a home-made script.

The procedure was done with the following script:

```
import sys, numpy, os
try:
       import pandas as pd
except ModuleNotFoundError:
       print("Error: Pandas library is required. Please, install it with \n pip install pandas \n")
       sys.exit()

inFilePath = "~/Spectra/fMLF_HNCO/6/ser"
outFilePath = "~/NUS_3D/fMLF_HNCO/ser_90_r"
SchedPath = "~/NUS_3D/fMLF_HNCO/fMLF_HNCO_random_90pc"

try:
    os.remove(outFilename)
```

```
except OSError:
    pass

NUSlist = pd.read_csv(SchedPath, sep = "\s+", header=None, names = ["1nd_Indir", "2nd_Indir"])

td1 = 60
td2 = 70

TDindirect = td1*td2
TDdirect=2048

dt_in = numpy.dtype(("i4", (TDindirect,TDdirect)))
inData = numpy.fromfile(inFilename, dtype=dt_in, count=1)

dt_out = numpy.dtype(("i4", (len(NUSlist)*4, TDdirect)))
outData = numpy.zeros(1, dtype=dt_out)

for i in range(len(NUSlist)):
    ln_US = NUSlist.iloc[i,:]["2nd_Indir"]*2*td1 + NUSlist.iloc[i,:]["1nd_Indir"]*2
    outData[0,i*4,:] = inData[0,ln_US,:]
    outData[0,i*4+1,:] = inData[0,ln_US+1,:]
    outData[0,i*4+2,:] = inData[0,ln_US+td1,:]
    outData[0,i*4+3,:] = inData[0,ln_US+td1+1,:]

outData.tofile(outFilename)
```

A schematic representation of the procedure is shown in Fig. M1.



Uniformly sampled

FID #	Incr1	Incr2	Ph1	Ph2
0	0	0	0	0
1	0	0	90	0
2	1	0	0	0
3	1	0	90	0
4	2	0	0	0
5	2	0	90	0
6	3	0	0	0
np1	0	0	0	90
np1+1	0	0	90	90
np1+2	1	0	0	90
np1+3	1	0	90	90
np1·2+m·2	0	1	0	0
np1·2+m·2+1	0	1	90	0
n·4·np1+m·2	n	m	0	0
n·4·np1+m·2+1	n	m	90	0
n·4·np1+m·2+np1·2	n	m	0	90
n·4·np1+m·2+np1·2+1	n	m	90	90

Non-uniformly sampled

FID #	Incr1	Incr2	Ph1	Ph2
0	0	0	0	0
1	0	0	90	0
2	0	0	0	90
3	0	0	90	90
4	1	0	0	0
5	1	0	90	0
6	1	0	0	90
7	1	0	90	90
8	0	1	0	0
9	0	1	90	0
10	0	1	0	90
11	0	1	90	90
...	n	m	0	0
...	n	m	90	0
...	n	m	0	90
...	n	m	90	90

Figure M1. Illustration of rearranging FIDs according to the standard Bruker order when artificially sampled from a uniformly acquired, multidimensional dataset. NUS methods generally require all combinations of each complex time increment in one block, i. e. real/real, imaginary/real, real/imaginary, and imaginary/imaginary.

Parameter screening

The key parameters of SMILE and SSA were optimized for each density by minimizing the RMSD to the US spectrum. The results for 2% and 30% are shown in Figs. M2 and M3. For SMILE, only the parameter `nSigma` was varied; for SSA, parameters `T` and `J` were optimized on a grid.

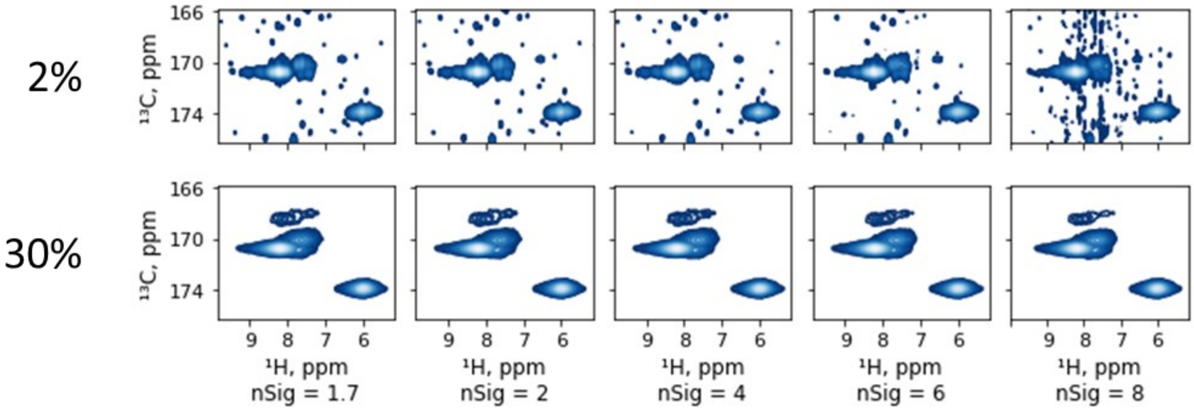


Figure M2: Results of SMILE parameter optimization for the NUS hCONH spectrum, acquired on dehomogenized fMLF powder, represented by H/CO projections. Parameter optimization was pursued via screening through values of `nSigma` (shortened as "`nSig`") for reconstruction of datasets. Depictions represent A) 2 % sampling density and B) 30 % sampling density.

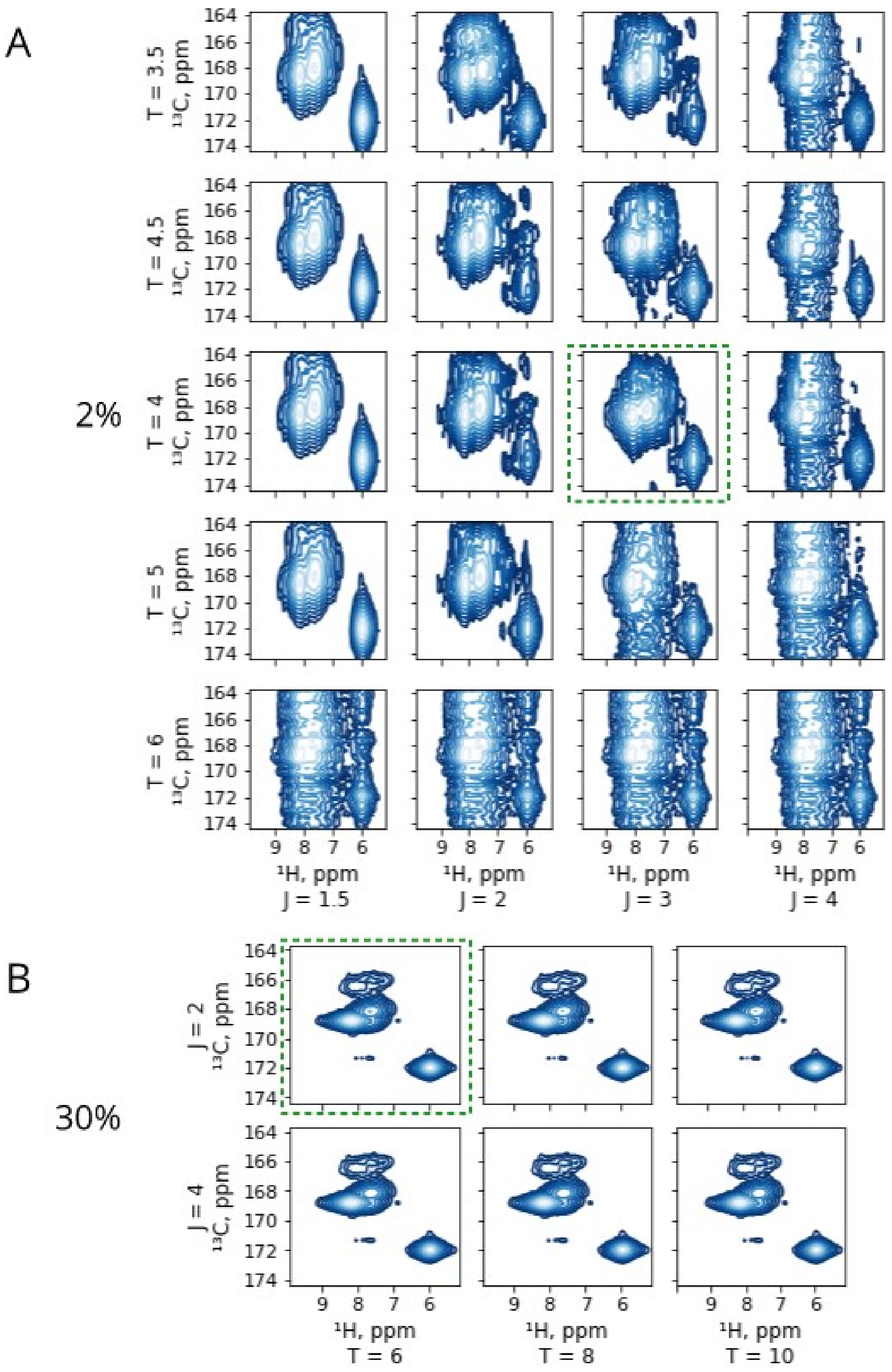


Figure M3. Results of SSA parameter optimization for the NUS hCONH spectrum for dehomogenized fMLF powder, represented by H/CO projections. Screening was done in a grid fashion through T and J values for cleaning and reconstruction of datasets, shown in **A**: for 2 % sampling density and in **B**: for 30 % sampling density. Default values are T=10 and J=4. The green frames highlight the reconstruction with the lowest RMSD with respect to the uniformly sampled 3D spectrum.

M.3.2. Analysis of heterogeneous peaks with TALOS-N and DANGLE

Main text: Section 2.2.3

Generation of individual inputs

Chemical shifts of the $i \pm (1,2)$ neighbors (where i is the Residue-Of-Interest) were replaced with combinations that matched the secondary structure propensity of the ROI. The procedure illustrated by Fig. 2.2.2B was implemented as:

```
   def transform(distrA, limitsA, limitsB):
'''Transforms distribution of unevenly distributed points in a space A to space B"
    Input:
    distrA - numpy 2D array [[arrdim1 ...], [arrdim2 ...], [arrdim3 ...], [arrdim4 ...]] - distribution
to be transformed.
    limitsA and limitsB (list of tuples) - limits of space A and B, correspondingly, in the form (left,
right) - mind the order!
    Output:
    distrB - transformed distribution'''
       shape=distrA.shape
       distrB = np.empty(shape=distrA.shape)
       for i in range(shape[0]):
           spanA = limitsA[i][1] - limitsA[i][0]
           spanB = limitsB[i][1] - limitsB[i][0]
           for j in range(shape[1]):
               distrB[i, j] = spanB * (distrA[i, j]-limitsA[i][0]) / spanA + limitsB[i][0]
       return distrB
```

Batch execution

TALOS-N was run with the following Bash script:

```
   #!/bin/sh

   for i in $(seq 0 1406)
          do
                    echo "$i"
                    mkdir $i
                    cd $i
                    talosn -in ./../../TALOS_inputs_A4/$i.tab -noclip - np 4
                    cd ..
          done
```

DANGLE was run on Windows with PowerShell with the command

```
   for ($i=0; $i -le 1406; $i++) {C:/Python27/python.exe dangle.py               \
   "C:/Users/Admin/Documents/DANGLE experimets GGAGG/DANGLE_Inputs/$i.tab"     \
   -dir "C:/Users/Admin/Documents/PhD/DANGLE experimets GGAGG/DANGLE_Outputs/$i/"}
```

M.3.3. Assignment of weights to PACSY entries

Main text: Section 2.2.4

The NMR intensity *I* at the each PACSY point in 4D chemical-shift space (Cα, Cβ, N and H dimensions of the hCBCANH spectrum) was determined by linear interpolation of the spectral intensity with `scipy.ndimage.map_coordinates` function. This function requires translation of the data coordinates (chemical shifts) into array index coordinates (i.e. *point positions in the spectrum).* This was done in the same manner as translation of chemical shifts from one residue type to the other as described directly above. Points with intensity lower than 15% (SNR = 15) of the peak maximum were excluded.

```
# Dimension order: HN, CA, N, CB
# pacsy_purged – the table with chemical shift – ϕ and ψ angles relation, A_CS_DB2.txt
# transform – function for coordinate transformation, see above

from scipy.ndimage import map_coordinates

limits_ppm = [(11.71, 5.995), (61.72, 1.79), (129.79, 96.79), (61.72, 1.79)]
limits_points = [(0, spectrum_shape[i]) for i in range(4)]

ala_ppm = pacsy_purged[['H', 'CA', 'N', 'CB']].values.T
ala_points = transform(ala_ppm, limits_ppm, limits_points)

purged['Weights'] = map_coordinates(the4D, ala_points, order=1) #intensity

threshold = the4D[238, 19, 29, 94]*0.15 # Maximum of the peak of interest; the peak is positive.

included_points=purged[['H','CA','N',CB']].where(purged['Weights']>threshold).dropna().values.T
included_points
```

```
   OUT: array([[   7.91 ,   8.06 ,   7.88 , ...,    7.9   ,   7.9   ,   7.9   ],
          [ 55.1   ,  54.1   ,  54.9   , ...,  54.671,  54.12 ,  54.12 ],
          [122.3   , 122.9   , 122.8   , ..., 124.61 , 124.61 , 119.281],
          [ 17.8   ,  18.1   ,  17.5   , ...,  17.562,  17.562,  17.562]])
```

Point density *P* in 4D space was estimated using `scipy.stats.gaussian_kde` function on a 70×70×70×70 grid with the limits at the minimum and maximum chemical shift of every selected PACSY point in each of the 4 dimensions. This took about 40 minutes on AMD Ryzen 3 1300X Quad-Core Processor (3500 Mhz, 4 Core(s) with 4 logical processors).

```
   # selected – the dataframe of ‘included points’ (see above)

   xmin, xmax = selected['CB'].min(), selected['CB'].max()
   ymin, ymax = selected['CA'].min(), selected['CA'].max()
   zmin, zmax = selected['N'].min(), selected['N'].max()
   amin, amax = selected['H'].min(), selected['H'].max()

   X, Y, Z, A = np.mgrid[xmin:xmax:70j, ymin:ymax:70j, zmin:zmax:70j, amin:amax:70j]
   positions = np.vstack([X.ravel(), Y.ravel(), Z.ravel(), A.ravel()])
   values     =    np.vstack([selected['CB'].values,    selected['CA'].values,    selected['N'].values,
selected['H'].values])

   go = input('Type "go" for performing KDE').lower() == 'go'
   if go:
       kernel = stats.gaussian_kde(values)
```

```
    selected['Density'] = kernel(values) # Add a new column with the result
```

M.3.4. Calculation of heterogeneity scores

Main text: Section 2.2.5

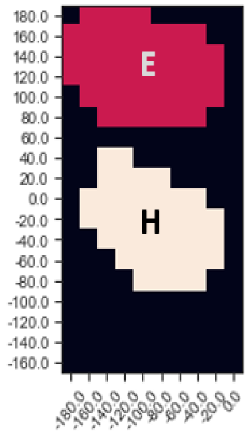


Figure M4. Regions used in calculation of the *R* score. The red region and tan regions, marked with E and H correspondingly, denote pixels of the 18×10 φ/ψ plot that were integrated into E and H parameters

M.3.5. Analysis of peak parameters

Main text: Section 2.3.4.

In order to cope with the un-evenness of the heterogeneous peak shapes, each individual peak was extracted from the 3D spectra in a “peak box” and projected one or two times on the target axis (axes) for the 1D or 2D fits (Fig. M5). An additional benefit of this approach is an increased signal-to-noise ratio as compared to fitting of a higher-dimensional shape or a peak slice. The 3D boxes were adjusted for every peak manually. Before fitting of the 2D projections, each box was stripped from the remainders of other peaks and noise: all data points below a threshold (set at SNR=15) were replaced with zeros; then the target peak was identified as the cluster of the highest intensity points, and all other clusters were discarded:

```
    from scipy import ndimage
    s = ndimage.generate_binary_structure(3,2)

    for res_num in peak_list:
        print(residue_number)

        fname_peak = os.path.join(os.path.join(dataDir, str(roi)), str(roi)+"_Peak_0.15_int.npy")
        peak = np.load(fname_peak)
        labels, num_features = ndimage.label(peak, structure=s)
        amount, _ = np.histogram(labels.ravel(), bins=num_features)
        print("Number of clusters \t", num_features)
        if num_features>1:
            main_cluster_id = np.argmax(amount[1:])+1 #starting from one because zero is not a cluster,
it is the floor
        else:
            main_cluster_id = 1
        print("Main cluster id \t", main_cluster_id)
        print('\n')
        peak_stripped = np.where(labels==main_cluster_id, peak, 0)
        np.save(os.path.join(os.path.join(dataDir,str(res_num),      str(res_num)+"_Peak_stripped.npy"),
peak_stripped) ## writing the result into the file
```

The Gaussian functions were implemented as:

```
import numpy as np

def gauss1d(x, I, x0, sigma):
    return I * np.exp(-1/2*(x-x0)**2/(sigma**2))

def gauss2d(xdata_tuple, amplitude, x0, y0, sigma_x, sigma_y, theta, offset):
    (x, y) = xdata_tuple
    a = (np.cos(theta)**2)/(2*sigma_x**2) + (np.sin(theta)**2)/(2*sigma_y**2)
    b = -(np.sin(2*theta))/(4*sigma_x**2) + (np.sin(2*theta))/(4*sigma_y**2)
    c = (np.sin(theta)**2)/(2*sigma_x**2) + (np.cos(theta)**2)/(2*sigma_y**2)

    g = offset + amplitude*np.exp( - (a*((x-x0)**2) + 2*b*(x-x0)*(y-y0)
                                        + c*((y-y0)**2)))
    return g.ravel()
```

To the author's experience, no available NMR software could reliably fit all heterogeneously broadened peaks in a conventional way (i.e. without extraction and projection). The 1D and 2D fits were reviewed and corrected manually.

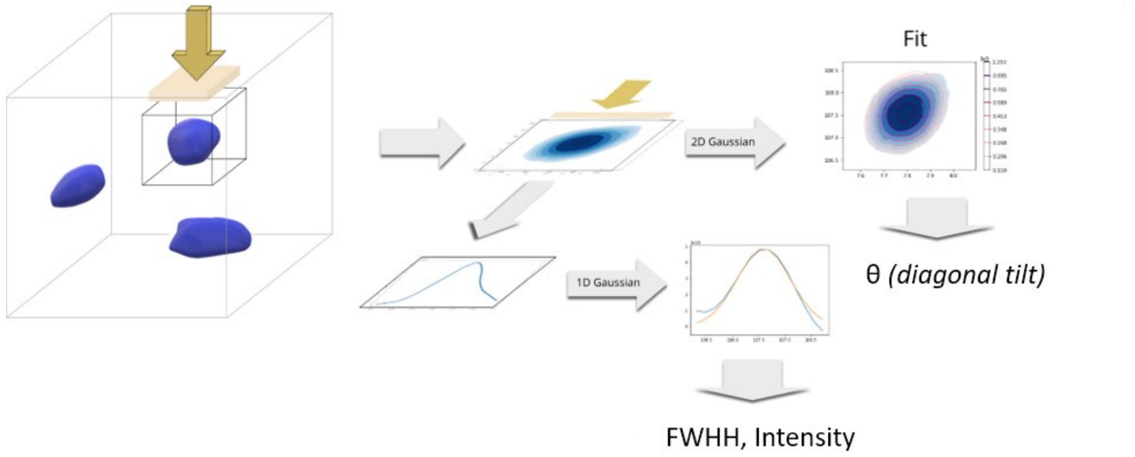


Figure M5. Workflow of fitting the heterogeneous peaks.

The linewidths were measured as full width at half height of the corresponing 1D or 2D gaussian fit, $FWHH_{DIM} = 2.355 * \sigma_{DIM}$ and then further normalized by the maximum. The peak slope ϑ on the HN plane was obtained from the correlation coefficient `theta`: $\vartheta = \theta \cdot {}^{2}/_{\pi}$. See Fig. M5 and the code snippet below:

```
res_num = 67
atom = "HN"
manual = True

dataDir = "~\\workdir\\IndPeaks\\3D\\HNCA\\"

path_proj = os.path.join(os.path.join(dataDir, str(res_num)), str(res_num) +atom+"_proj.npy")
path_xdata = os.path.join(os.path.join(dataDir, str(res_num)), str(res_num)+atom+"_X.npy")

xdata = np.load(path_xdata)
```

```
xdata_trunk = xdata[0 : xdata.shape[0] // 2+10]
ydata = np.load(path_proj)
ydata_trunk = ydata[0 : ydata.shape[0] // 2+10]

p0 = (ydata.max(), xdata[np.argmax(ydata)], 0.5) #initial guess

## p1 = (ydata.max(), 9.5, 0.5) #adjusted guess
bounds = ([2e9, 7.5, -0.5], [1e11, 7.7, 0.99]) #fit boundaries

if manual:
    # Adjusted manually
    popt, pcov = curve_fit(gauss1d, xdata_trunk, ydata_trunk,
                           bounds=bounds)
else:
    # Automatic fit
    popt, pcov = curve_fit(gauss1d, xdata, ydata,
                           p0=p0)

fit = gauss1d(xdata, *popt)

#print("Linewidth by fit: ", 2.355*popt[2], " ppm")

print("Linewidth by fit: ", 2.355*popt[2]*HNCA_dic["FDF1OBS"], " Hz")
print(f"Int\t{popt[0]}\ncenter\t{popt[1]}\nsigma\t{popt[2]}")
```

```
OUT    Linewidth by fit:  86.89353968896644 Hz
       Int        10059840686.369625
       center     7.605180115303766
       sigma      0.18338003066063832
```

The weighted average of the amide linewidths (gray bars in Fig. 2.3.5B) was calculated as $\overline{LW} = \frac{3}{8}\overline{LW_H} + \frac{5}{8}\overline{LW_N}$ [ppm], where $\overline{LW_H}$ and $\overline{LW_N}$ are the average ^{1}H and ^{15}N linewidths in the two spectra (hCANH and hCBNH). The weights were derived from the ratio of expected homogeneous linewidths of ^{1}H (0.5 ppm) and ^{15}N (0.3 ppm). The ppm units were preferred by the author over the traditional Hz scale to a) be able to derive the average score between the two isotopes and b) to make the measure independent of the external magnetic field: The *heterogeneous* line broadening in either dimension scales with the differences (in Hz) between *isotropic* shifts of individual conformations proportionally to the strength of the $\boldsymbol{B_0}$ field (compare to chemical shift *anisotropy*).

M.3.6. Amide chemical shifts and hydrogen bond lengths in β-sheets

Main text: Section 2.3.4.

The list of PDB structures was taken from a similar study (Baskaran *et al.*, 2021) and parced for elements of extended structures (β-sheets or strands, 'E' or 'B') as classified by DSSP (original version published by Kabsch and Sander, 1983; Maarten L. Hekkelman's version `mkdssp`, v. 3.0.1 used in this work). The relational tables of PDB IDs and BMRB data were taken from ReBoxitory on NMRBox (Maciejewski *et al.*, 2017). The PDB structures and DSSP output were analysed by a byopython.PDB module (Hamelryck and Manderick, 2003). Hydrogen bonds were identified by presence of an oxygen atom within a radius of ≤3.5 Å from the query backbone amide nitrogen. Correlations of the amide chemical shift and the interatomic N-O distances for 18 residue types are presented in Fig. S15.

Implementation of the search for the N-O contacts is performed by the following code snippet:

```
   pdb_id='aaaa'
   H_bond_max_length = 3.5 # Angstrom
   #amide_N_H_bonds=pd.DataFrame()

   pdb_N_df = pd.DataFrame()
   problemsB=[]
   counter=0
   for problem in problems: # searching all problematic PDBs

       pdb_id = problem
       print(pdb_id)

       # Retrieving DSSP data
       dssp_df = pd.read_csv('D://dssp/sheet_csvs//'+pdb_id+'.tab', sep='\t', index_col=[0,1])
       if dssp_df.empty:
           continue

       chains = dssp_df.index.get_level_values(level=0).unique()
       if chains[0] != 'B':
           continue
       else:
           ## Which residues of the given pdb (only chain A by definition) have E or B conformation?
           dssp_df = dssp_df.loc['B', :]
           pdb_E_idx = dssp_df.index
           bmrb_data                                                                        =
amide_N.where(amide_N['pdb_id']==pdb_id.upper()).dropna().set_index('Seq_ID').drop_duplicates()
           ## Which entries in this huge table belong to our pdb and which residues have assignments?
           pdb_res = bmrb_data.index

           ## Which of them are in E or B conformation?
           pdb_res_E_idx = pdb_res.intersection(pdb_E_idx)

           ## Creating a table with the residues of interest
           pdb_N_df = pd.DataFrame(index=pdb_res_E_idx)
           try:
               pdb_N_df['RES_type'] = bmrb_data.loc[pdb_res_E_idx, 'Comp_ID']
               pdb_N_df['R_type'] = dssp_df['0']
               pdb_N_df['Shift'] = bmrb_data.loc[pdb_res_E_idx, 'Val']
               pdb_N_df['Atom'] = 'N'
               pdb_N_df['Phi'] = dssp_df['3']
               pdb_N_df['Psi'] = dssp_df['4']
               pdb_N_df['DSSP'] = dssp_df['1']
               pdb_N_df['PDB_ID'] = pdb_id.upper()
               pdb_N_df['BMRB_ID']                                                          =
amide_N.where(amide_N['pdb_id']==pdb_id.upper()).dropna().index.unique()[0]
           except ValueError:
               continue

           #pdb_N_df.loc[]

           # Find potential H-bonds and measure their distance!
           ## Get the 3D structure (always chain A!)
           structure = parser.get_structure("_", pdbdir+pdb_id+".cif")
           try:
               chain = structure[0]["B"]
           except KeyError:
               problemsB.append(pdb_id)
               continue
```

```
        O_atom_list=[]
        for res in structure[0]["B"]:
            if res.has_id("O"):
                O_atom_list.append(res["O"]) # Gathering all oxygens in the structure
            else:
                continue

        try:
            for rnum in pdb_res_E_idx.values: # Check every E residue in our PDB!
                search_set = O_atom_list
                ## search for each Nitrogen individually

                search_set.append(chain[int(rnum)]['N'])
                H_bond_finder=NeighborSearch(search_set)
                raw_out = H_bond_finder.search_all(radius=4.1, level='R')
                i=0
                for pair in raw_out:
                    print(pair)
                    for res in pair:
                        if res.get_id()[1] == rnum:
                            # Potential partners found
                            ## Initialize atoms
                            if pair[0].get_id()[1] == rnum:
                                partner_id = pair[1].get_id()[1]
                                n = pair[0]['N'] # our N
                                o = pair[1]['O'] # other's O
                            else:
                                partner_id = pair[0].get_id()[1]
                                n = pair[1]['N'] # our N
                                o = pair[0]['O'] # other's O

                            # Exclude direct sequential neighbors!
                            print('Residue:', rnum, 'Partner', partner_id)
                            if abs(partner_id-rnum) < 2:
                                continue
                            else:
                                dist = n - o
                                print(dist)
                                if dist < H_bond_max_length:
                                    # Found an H-bond! Now write it down
                                    i+=1
                                    pdb_N_df.at[rnum, f'H_bond_partner_{i}'] = partner_id
                                    pdb_N_df.at[rnum, f'H_bond_{i}_len'] = dist
                print(pdb_N_df.head(5))

            pdb_N_df.reset_index(inplace=True)
            pdb_N_df.rename(columns={'index': 'Res_ID'}, inplace=True)
            pdb_N_df.to_csv(f'D://H_bonds//{pdb_id}.csv')
        except KeyError:
            print('Problem with PDB ', pdb_id)
            problemsB.append(pdb_id)
            continue
    counter+=1

# amide_N_H_bonds.to_csv('D://amide_N_H_bonds.csv') # Writing the table to the disk if desired
```

S | SUPPLEMENT

References

Andrew E.R., Bradbury A., Eades R.G., 1958. *Nuclear Magnetic Resonance Spectra from a Crystal rotated at High Speed*. Nature. 182(4650): 1659–1659. doi.org/10.1038/1821659a0

Anfinsen C., 1972. *Christian Anfinsen – Nobel Lecture.*

Asakura T., Suzuki Y., Nakazawa Y., Holland G.P., Yarger J.L., 2013a. *Elucidating silk structure using solid-state NMR*. Soft Matter. 9(48): 11440–50. doi.org/10.1039/c3sm52187g

Asakura T., Suzuki Y., Nakazawa Y., Yazawa K., Holland G.P., Yarger J.L., 2013b. *Silk structure studied with nuclear magnetic resonance*. Progress in Nuclear Magnetic Resonance Spectroscopy. 69: 23–68. doi.org/10.1016/j.pnmrs.2012.08.001

Baek M., DiMaio F., Anishchenko I., Dauparas J., Ovchinnikov S., Lee G.R., Wang J., Cong Q., Kinch L.N., Schaeffer R.D., Millán C., Park H., Adams C., Glassman C.R., DeGiovanni A., Pereira J.H., Rodrigues A.V., van Dijk A.A., Ebrecht A.C., Opperman D.J., Sagmeister T., Buhlheller C., Pavkov-Keller T., Rathinaswamy M.K., Dalwadi U., Yip C.K., Burke J.E., Garcia K.C., Grishin N.V., Adams P.D., Read R.J., Baker D., 2021. *Accurate prediction of protein structures and interactions using a three-track neural network*. Science. 373(6557): 871–6. doi.org/10.1126/science.abj8754

Bajaj V.S., van der Wel P.C.A., Griffin R.G., 2009. *Observation of a Low-Temperature, Dynamically Driven Structural Transition in a Polypeptide by Solid-State NMR Spectroscopy*. Journal of the American Chemical Society. 131(1): 118–28. doi.org/10.1021/ja8045926

Baldus M., Petkova A.T., Herzfeld J., Griffin R.G., 1998. *Cross polarization in the tilted frame: assignment and spectral simplification in heteronuclear spin systems*. Molecular Physics. 95(6): 1197–207. doi.org/10.1080/00268979809483251

Ball S.R., Kwan A.H., Sunde M., 2019. *Hydrophobin Rodlets on the Fungal Cell Wall.* . In: Latgé, J.-P. (ed.), The Fungal Cell Wall, Current Topics in Microbiology and Immunology. Springer International Publishing, Cham: 29–51. doi.org/10.1007/82_2019_186

Barna J.C.J., Laue E.D., Mayger M.R., Skilling J., Worrall S.J.P., 1987. *Exponential sampling, an alternative method for sampling in two-dimensional NMR experiments*. Journal of Magnetic Resonance (1969). 73(1): 69–77. doi.org/10.1016/0022-2364(87)90225-3

Baskaran K., Wilburn C.W., Wedell J.R., Koharudin L.M.I., Ulrich E.L., Schuyler A.D., Eghbalnia H.R., Gronenborn A.M., Hoch J.C., 2021. *Anomalous amide proton chemical shifts as signatures of hydrogen bonding to aromatic sidechains*. Magnetic Resonance. 2(2): 765–75. doi.org/10.5194/mr-2-765-2021

Bax A., Griffey R.H., Hawkins B.L., 1983. *Correlation of proton and nitrogen-15 chemical shifts by multiple quantum NMR*. Journal of Magnetic Resonance (1969). 55(2): 301–15. doi.org/10.1016/0022-2364(83)90241-X

Becker-Baldus J., Glaubitz C., 2018. *Cryo-trapped Intermediates of Retinal Proteins Studied by DNP-enhanced MAS NMR Spectroscopy*. 7: 79–92. doi.org/10.1002/9780470034590.emrstm1552

Berjanskii M., Wishart D.S., 2006. *NMR: prediction of protein flexibility*. Nature Protocols. 1(2): 683–8. doi.org/10.1038/nprot.2006.108

Berjanskii M.V., Neal S., Wishart D.S., 2006. *PREDITOR: a web server for predicting protein torsion angle restraints*. Nucleic Acids Research. 34(Web Server): W63–9. doi.org/10.1093/nar/gkl341

Berjanskii M.V., Wishart D.S., 2017. *Unraveling the meaning of chemical shifts in protein NMR*. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 1865(11): 1564–76. doi.org/10.1016/j.bbapap.2017.07.005

Berman H.M., 2000. *The Protein Data Bank*. Nucleic Acids Research. 28(1): 235–42. doi.org/10.1093/nar/28.1.235

Bertini I., McGreevy K.S., Parigi G. (eds.), 2012. *NMR of biomolecules: towards mechanistic systems biology*. Wiley-VCH, Weinheim, Germany.

Bloch F., 1946. *Nuclear Induction*. Physical Review. 70(7–8): 460–74. doi.org/10.1103/PhysRev.70.460

Bloch F., Hansen W.W., Packard M., 1946. *The Nuclear Induction Experiment*. Physical Review. 70(7–8): 474–85. doi.org/10.1103/PhysRev.70.474

Bodenhausen G., Ernst R.R., 1981. *The accordion experiment, a simple approach to three-dimensional NMR spectroscopy*. Journal of Magnetic Resonance (1969). 45(2): 367–73. doi.org/10.1016/0022-2364(81)90137-2

Bonomi M., Heller G.T., Camilloni C., Vendruscolo M., 2017. *Principles of protein structural ensemble determination*. Current Opinion in Structural Biology. 42: 106–16. doi.org/10.1016/j.sbi.2016.12.004

Bonomi M., Vendruscolo M., 2019. *Determination of protein structural ensembles using cryo-electron microscopy*. Current Opinion in Structural Biology. 56: 37–45. doi.org/10.1016/j.sbi.2018.10.006

Cheung M.-S., Maguire M.L., Stevens T.J., Broadhurst R.W., 2010. *DANGLE: A Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure*. Journal of Magnetic Resonance. 202(2): 223–33. doi.org/10.1016/j.jmr.2009.11.008

Chiti F., Dobson C.M., 2017. *Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade*. Annual Review of Biochemistry. 86(1): 27–68. doi.org/10.1146/annurev-biochem-061516-045115

Chiti F., Dobson C.M., 2006. *Protein Misfolding, Functional Amyloid, and Human Disease*. Annual Review of Biochemistry. 75(1): 333–66. doi.org/10.1146/annurev.biochem.75.101304.123901

Christensen A.S., Linnet T.E., Borg M., Boomsma W., Lindorff-Larsen K., Hamelryck T., Jensen J.H., 2013. *Protein Structure Validation and Refinement Using Amide Proton Chemical Shifts Derived from Quantum Mechanics*. PLOS ONE. 8(12): e84123. doi.org/10.1371/journal.pone.0084123

Coggins B.E., Zhou P., 2008. *High resolution 4-D spectroscopy with sparse concentric shell sampling and FFT-CLEAN*. Journal of Biomolecular NMR. 42(4): 225–39. doi.org/10.1007/s10858-008-9275-x

Conchillo-Solé O., de Groot N.S., Avilés F.X., Vendrell J., Daura X., Ventura S., 2007. *AGGRESCAN: a server for the prediction and evaluation of 'hot spots' of aggregation in polypeptides*. BMC Bioinformatics. 8(1): 65. doi.org/10.1186/1471-2105-8-65

Cornilescu G., Delaglio F., Bax A., 1999a. *Protein backbone angle restraints from searching a database for chemical shift and sequence homology*. Journal of Biomolecular NMR. 13(3): 289–302. doi.org/10.1023/A:1008392405740

Cornilescu G., Hu J.-S., Bax A., 1999b. *Identification of the Hydrogen Bonding Network in a Protein by Scalar Couplings*. Journal of the American Chemical Society. 121(12): 2949–50. doi.org/10.1021/ja9902221

Cremers D., Pock T., Kolev K., Chambolle A., 2011. *Convex Relaxation Techniques for Segmentation, Stereo and Multiview Reconstruction*. Markov Random Fields for Vision and Image Processing. MIT Press.

Davis I.W., Arendall W.B., Richardson D.C., Richardson J.S., 2006. *The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances*. Structure. 14(2): 265–74. doi.org/10.1016/j.str.2005.10.007

Delaglio F., Grzesiek S., Vuister GeertenW., Zhu G., Pfeifer J., Bax A., 1995. *NMRPipe: A multidimensional spectral processing system based on UNIX pipes*. Journal of Biomolecular NMR. 6(3). doi.org/10.1007/BF00197809

Delaglio F., Walker G.S., Farley K.A., Sharma R., Hoch J.C., Arbogast L.W., Brinson R.G., Marino J.P., 2017. *Non-Uniform Sampling for All: More NMR Spectral Quality, Less Measurement Time*. American pharmaceutical review. 20(4): 339681.

Ding L., Chen K., Santini P.A., Shi Z., Kallenbach N.R., 2003. *The pentapeptide GGAGG has PII conformation*. Journal of the American Chemical Society. 125(27): 8092–3. doi.org/10.1021/ja035551e

Donoho D.L., 2006. *Compressed sensing*. IEEE Transactions on Information Theory. 52(4): 1289–306. doi.org/10.1109/TIT.2006.871582

Dračínský M., Möller H.M., Exner T.E., 2013. *Conformational Sampling by Ab Initio Molecular Dynamics Simulations Improves NMR Chemical Shift Predictions*. Journal of Chemical Theory and Computation. 9(8): 3806–15. doi.org/10.1021/ct400282h

Drori I., 2007. *Fast Minimization by Iterative Thresholding for Multidimensional NMR Spectroscopy*. EURASIP Journal on Advances in Signal Processing. 2007(1): 020248. doi.org/10.1155/2007/20248

Duer M.J., 2002. *Solid-state NMR spectroscopy: Principles and applications*. Blackwell Science.

Engelke F., 2022. *On the origin of proton spin and its magnetic dipole moment*. doi.org/10.13140/RG.2.2.15078.96325

Epstein C.J., Goldberger R.F., Anfinsen C.B., 1963. *The Genetic Control of Tertiary Protein Structure: Studies With Model Systems*. Cold Spring Harbor Symposia on Quantitative Biology. 28(0): 439–49. doi.org/10.1101/SQB.1963.028.01.060

Fadel A.R., Jin D.Q., Montelione G.T., Levy R.M., 1995. *Crankshaft motions of the polypeptide backbone in molecular dynamics simulations of human type- z transforming growth factor*. 6: 221–6.

Fawzi N.L., Ying J., Ghirlando R., Torchia D.A., Clore G.M., 2011. *Atomic-resolution dynamics on the surface of amyloid-β protofibrils probed by solution NMR*. Nature. 480(7376): 268–72. doi.org/10.1038/nature10577

Fernandez-Escamilla A.-M., Rousseau F., Schymkowitz J., Serrano L., 2004. *Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins*. Nature Biotechnology. 22(10): 1302–6. doi.org/10.1038/nbt1012

Frishman D., Argos P., 1995. *Knowledge-based protein secondary structure assignment*. Proteins: Structure, Function, and Bioinformatics. 23(4): 566–79. doi.org/10.1002/prot.340230412

Fritzsching K.J., Hong M., Schmidt-Rohr K., 2016. *Conformationally selective multidimensional chemical shift ranges in proteins from a PACSY database purged using intrinsic quality criteria*. Journal of Biomolecular NMR. 64(2): 115–30. doi.org/10.1007/s10858-016-0013-5

Grohe K., Nimerovsky E., Singh H., K. Vasa S., Söldner B., Vögeli B., M. Rienstra C., Linser R., 2019. *Exact distance measurements for structure and dynamics in solid proteins by fast-magic-angle-spinning NMR*. Chemical Communications. 55(55): 7899–902. doi.org/10.1039/C9CC02317H

Gullion T., 2006. *Rotational-Echo, Double-Resonance NMR*. . In: Webb, G.A. (ed.), Modern Magnetic Resonance. Springer Netherlands, Dordrecht: 713–8. doi.org/10.1007/1-4020-3910-7_89

Habchi J., Tompa P., Longhi S., Uversky V.N., 2014. *Introducing Protein Intrinsic Disorder*. Chemical Reviews. 114(13): 6561–88. doi.org/10.1021/cr400514h

Hallen M.A., Keedy D.A., Donald B.R., 2013. *Dead-end elimination with perturbations (DEEPer): A provable protein design algorithm with continuous sidechain and backbone flexibility*. Proteins: Structure, Function, and Bioinformatics. 81(1): 18–39. doi.org/10.1002/prot.24150

Hamelryck T., Manderick B., 2003. *PDB file parser and structure class implemented in Python*. Bioinformatics. 19(17): 2308–10. doi.org/10.1093/bioinformatics/btg299

Han B., Liu Y., Ginzinger S.W., Wishart D.S., 2011. *SHIFTX2: significantly improved protein chemical shift prediction*. Journal of Biomolecular NMR. 50(1): 43–57. doi.org/10.1007/s10858-011-9478-4

Harris C.R., Millman K.J., van der Walt S.J., Gommers R., Virtanen P., Cournapeau D., Wieser E., Taylor J., Berg S., Smith N.J., Kern R., Picus M., Hoyer S., van Kerkwijk M.H., Brett M., Haldane A., del Río J.F., Wiebe M., Peterson P., Gérard-Marchant P., Sheppard K., Reddy T., Weckesser W., Abbasi H., Gohlke C., Oliphant T.E., 2020. *Array programming with NumPy*. Nature. 585(7825): 357–62. doi.org/10.1038/s41586-020-2649-2

Harris R.K., 2007. *Applications of solid-state NMR to pharmaceutical polymorphism and related matters**. Journal of Pharmacy and Pharmacology. 59(2): 225–39. doi.org/10.1211/jpp.59.2.0009

Harris T.K., Mildvan A.S., 1999. *High-Precision Measurement of Hydrogen Bond Lengths in Proteins by Nuclear Magnetic Resonance Methods*. Proteins: Structure, Function, and Genetics. 35(3): 275–82. doi.org/10.1002/(SICI)1097-0134(19990515)35:3<275::AID-PROT1>3.0.CO;2-V

Hassan A., Quinn C.M., Struppe J., Sergeyev I.V., Zhang C., Guo C., Runge B., Theint T., Dao H.H., Jaroniec C.P., Berbon M., Lends A., Habenstein B., Loquet A., Kuemmerle R., Perrone B., Gronenborn A.M., Polenova T., 2020. *Sensitivity boosts by the CPMAS CryoProbe for challenging biological assemblies*. Journal of Magnetic Resonance. 311: 106680. doi.org/10.1016/j.jmr.2019.106680

Havlin R.H., Tycko R., 2005. *Probing site-specific conformational distributions in protein folding with solid-state NMR*. Proceedings of the National Academy of Sciences. 102(9): 3284–9. doi.org/10.1073/pnas.0406130102

Hayward S., 2001. *Peptide-plane flipping in proteins*. Protein Science : A Publication of the Protein Society. 10(11): 2219–27.

Heise H., Luca S., de Groot B.L., Grubmüller H., Baldus M., 2005. *Probing Conformational Disorder in Neurotensin by Two-Dimensional Solid-State NMR and Comparison to Molecular Dynamics Simulations*. Biophysical Journal. 89(3): 2113–20. doi.org/10.1529/biophysj.105.059964

Helmus J.J., Jaroniec C.P., 2013. *Nmrglue: an open source Python package for the analysis of multidimensional NMR data*. Journal of Biomolecular NMR. 55(4): 355–67. doi.org/10.1007/s10858-013-9718-x

Hiller S., Fiorito F., Wüthrich K., Wider G., 2005. *Automated projection spectroscopy (APSY)*. Proceedings of the National Academy of Sciences. 102(31): 10876–81. doi.org/10.1073/pnas.0504818102

Hoch J.C., Baskaran K., Burr H., Chin J., Eghbalnia H.R., Fujiwara T., Gryk M.R., Iwata T., Kojima C., Kurisu G., Maziuk D., Miyanoiri Y., Wedell J.R., Wilburn C., Yao H., Yokochi M., 2023. *Biological Magnetic Resonance Data Bank*. Nucleic Acids Research. 51(D1): D368–76. doi.org/10.1093/nar/gkac1050

Högbom J.A., 1974. *Aperture Synthesis with a Non-Regular Distribution of Interferometer Baselines*. Astronomy and Astrophysics Supplement Series. 15: 417.

Hu K.-N., Havlin R.H., Yau W.-M., Tycko R., 2009. *Quantitative Determination of Site-Specific Conformational Distributions in an Unfolded Protein by Solid-State Nuclear Magnetic Resonance*. Journal of Molecular Biology. 392(4): 1055–73. doi.org/10.1016/j.jmb.2009.07.073

Hu K.N., Tycko R., 2010. *What can solid state NMR contribute to our understanding of protein folding?* Biophysical Chemistry. 151(1–2): 10–21. doi.org/10.1016/j.bpc.2010.05.009

Hunter J.D., 2007. *Matplotlib: A 2D graphics environment*. Computing in science & engineering. 9(03): 90–5.

Hyberts S.G., Arthanari H., Robson S.A., Wagner G., 2014. *Perspectives in magnetic resonance: NMR in the post-FFT era*. Journal of Magnetic Resonance. 241(1): 60–73. doi.org/10.1016/j.jmr.2013.11.014

Hyberts S.G., Milbradt A.G., Wagner A.B., Arthanari H., Wagner G., 2012. *Application of iterative soft thresholding for fast reconstruction of NMR data non-uniformly sampled with multidimensional Poisson Gap scheduling*. Journal of Biomolecular NMR. 52(4): 315–27. doi.org/10.1007/s10858-012-9611-z

Hyberts S.G., Robson S.A., Wagner G., 2013. *Exploring signal-to-noise ratio and sensitivity in non-uniformly sampled multi-dimensional NMR spectra*. Journal of Biomolecular NMR. 55(2): 167–78. doi.org/10.1007/s10858-012-9698-2

Hyberts S.G., Takeuchi K., Wagner G., 2010. *Poisson-Gap Sampling and Forward Maximum Entropy Reconstruction for Enhancing the Resolution and Sensitivity of Protein NMR Data*. Journal of the American Chemical Society. 132(7): 2145–7. doi.org/10.1021/ja908004w

James E.I., Murphree T.A., Vorauer C., Engen J.R., Guttman M., 2022. *Advances in Hydrogen/Deuterium Exchange Mass Spectrometry and the Pursuit of Challenging Biological Systems*. Chemical Reviews. 122(8): 7562–623. doi.org/10.1021/acs.chemrev.1c00279

Ji Y., Liang L., Bao X., Hou G., 2021. *Recent progress in dipolar recoupling techniques under fast MAS in solid-state NMR spectroscopy*. Solid State Nuclear Magnetic Resonance. 112: 101711. doi.org/10.1016/j.ssnmr.2020.101711

Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., Bates R., Žídek A., Potapenko A., Bridgland A., Meyer C., Kohl S.A.A., Ballard A.J., Cowie A., Romera-Paredes B., Nikolov S., Jain R., Adler J., Back T., Petersen S., Reiman D., Clancy E., Zielinski M., Steinegger M., Pacholska M., Berghammer T., Bodenstein S., Silver D., Vinyals O., Senior A.W., Kavukcuoglu K., Kohli P., Hassabis D., 2021. *Highly accurate protein structure prediction with AlphaFold*. Nature. 596(7873): 583–9. doi.org/10.1038/s41586-021-03819-2

Kabsch W., Sander C., 1983. *Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features*. Biopolymers. 22(12): 2577–637. doi.org/10.1002/bip.360221211

Karplus M., 1959. *Contact Electron-Spin Coupling of Nuclear Magnetic Moments*. The Journal of Chemical Physics. 30(1): 11–5. doi.org/10.1063/1.1729860

Kazimierczuk K., Orekhov V., 2015. *Non-uniform sampling: Post-Fourier era of NMR data collection and processing*. Magnetic Resonance in Chemistry. 53(11): 921–6. doi.org/10.1002/mrc.4284

Kazimierczuk K., Zawadzka A., Koźmiński W., Zhukov I., 2007. *Lineshapes and artifacts in Multidimensional Fourier Transform of arbitrary sampled NMR data sets*. Journal of Magnetic Resonance. 188(2): 344–56. doi.org/10.1016/J.JMR.2007.08.005

Keedy D.A., Fraser J.S., Bedem H. van den, 2015. *Exposing Hidden Alternative Backbone Conformations in X-ray Crystallography Using qFit*. PLOS Computational Biology. 11(10): e1004507. doi.org/10.1371/journal.pcbi.1004507

Keeler J., 2010. *Understanding NMR spectroscopy*, 2nd ed. ed. John Wiley and Sons, Chichester, U.K.

Kikhney A.G., Svergun D.I., 2015. *A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins*. FEBS Letters, Dynamics, flexibility, and intrinsic disorder in protein assemblies. 589(19, Part A): 2570–7. doi.org/10.1016/j.febslet.2015.08.027

Kinman L.F., Powell B.M., Zhong E.D., Berger B., Davis J.H., 2023. *Uncovering structural ensembles from single-particle cryo-EM data using cryoDRGN*. Nature Protocols. 18(2): 319–39. doi.org/10.1038/s41596-022-00763-x

Kjaergaard M., Kragelund B.B., 2017. *Functions of intrinsic disorder in transmembrane proteins*. Cellular and Molecular Life Sciences. 74(17): 3205–24. doi.org/10.1007/s00018-017-2562-5

Kleckner I.R., Foster M.P., 2011. *Biochimica et Biophysica Acta An introduction to NMR-based approaches for measuring protein dynamics ☆*. BBA - Proteins and Proteomics. 1814(8): 942–68. doi.org/10.1016/j.bbapap.2010.10.012

Kraus J., Gupta R., Lu M., Gronenborn A.M., Akke M., Polenova T., 2020. *Accurate Backbone 13C and 15N Chemical Shift Tensors in Galectin-3 Determined by MAS NMR and QM/MM: Details of Structure and Environment Matter*. ChemPhysChem. 21(13): 1436–43. doi.org/10.1002/cphc.202000249

Kümmerlen J., D. van Beek J., Vollrath F., H. Meier B., 1996. *Local Structure in Spider Dragline Silk Investigated by Two-Dimensional Spin-Diffusion Nuclear Magnetic Resonance*. Macromolecules. 29(8): 2920–8. doi.org/10.1021/ma951098i

Kuroki S., Asakawa N., Ando S., Ando I., Shoji A., Ozaki T., 1991. *Hydrogen bond length and 15N NMR chemical shift of the glycine residue of some oligopeptides in the solid state*. Journal of Molecular Structure. 245(1–2): 69–80. doi.org/10.1016/0022-2860(91)87007-5

Kwan A.H., Macindoe I., Vukašin P.V., Morris V.K., Kass I., Gupte R., Mark A.E., Templeton M.D., Mackay J.P., Sunde M., 2008. *The Cys3–Cys4 Loop of the Hydrophobin EAS Is Not Required for Rodlet Formation and Surface Activity*. Journal of Molecular Biology. 382(3): 708–20. doi.org/10.1016/j.jmb.2008.07.034

Kwan A.H.Y., Winefield R.D., Sunde M., Matthews J.M., Haverkamp R.G., Templeton M.D., Mackay J.P., 2006. *Structural basis for rodlet assembly in fungal hydrophobins*. Proceedings of the National Academy of Sciences. 103(10): 3621–6. doi.org/10.1073/pnas.0505704103

Landau H.J., 1967. *Necessary density conditions for sampling and interpolation of certain entire functions*. Acta Mathematica. 117(0): 37–52. doi.org/10.1007/BF02395039

Lange A., Scholz I., Manolikas T., Ernst M., Meier B.H., 2009. *Low-power cross polarization in fast magic-angle spinning NMR experiments*. Chemical Physics Letters. 468(1–3): 100–5. doi.org/10.1016/j.cplett.2008.11.089

Lee W., Tonelli M., Markley J.L., 2015. *NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy*. Bioinformatics. 31(8): 1325–7.

Lee Woonghee, Yu W., Kim S., Chang I., Lee Weontae, Markley J.L., 2012. *PACSY, a relational database management system for protein structure and chemical shift analysis*. Journal of Biomolecular NMR. 54(2): 169–79. doi.org/10.1007/s10858-012-9660-3

Levitt M.H., 2012. *Symmetry-Based Pulse Sequences in Magic-Angle Spinning Solid-State NMR*. Solid-State NMR Studies of Biopolymers. John Wilwy & Sons, Ltd.: 229–71.

Levitt M.H., 2008. *Spin dynamics: basics of nuclear magnetic resonance*, 2nd ed. ed. John Wiley & Sons, Chichester, England ; Hoboken, NJ.

Li F., Lee J.H., Grishaev A., Ying J., Bax A., 2015. *High Accuracy of Karplus Equations for Relating Three-Bond J Couplings to Protein Backbone Torsion Angles*. ChemPhysChem. 16(3): 572–8. doi.org/10.1002/cphc.201402704

Li J., Bennett K.C., Liu Y., Martin M.V., Head-Gordon T., 2020. *Accurate prediction of chemical shifts for aqueous protein structure on "Real World" data*. Chemical Science. 11(12): 3180–91. doi.org/10.1039/C9SC06561J

Linden A.H., Franks W.T., 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*. Journal of Biomolecular NMR. 51(3): 283–92. doi.org/10.1007/s10858-011-9535-z

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*. Angewandte Chemie - International Edition. 50(19): 4508–12. doi.org/10.1002/anie.201008244

Louros N., Orlando G., De Vleeschouwer M., Rousseau F., Schymkowitz J., 2020. *Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities*. Nature Communications. 11(1): 3314. doi.org/10.1038/s41467-020-17207-3

Lovell S.C., Davis I.W., Arendall W.B., de Bakker P.I.W., Word J.M., Prisant M.G., Richardson J.S., Richardson D.C., 2003. *Structure validation by Cα geometry: ϕ,ψ and Cβ deviation*. Proteins: Structure, Function, and Bioinformatics. 50(3): 437–50. doi.org/10.1002/prot.10286

Lowe I.J., 1959. *Free Induction Decays of Rotating Solids*. Physical Review Letters. 2(7): 285–7. doi.org/10.1103/PhysRevLett.2.285

Maciejewski M.W., Schuyler A.D., Gryk M.R., Moraru I.I., Romero P.R., Ulrich E.L., Eghbalnia H.R., Livny M., Delaglio F., Hoch J.C., 2017. *NMRbox: A Resource for Biomolecular NMR Computation*. Biophysical Journal. 112(8): 1529–34. doi.org/10.1016/j.bpj.2017.03.011

Macindoe I., Kwan A.H., Ren Q., Morris V.K., Yang W., Mackay J.P., Sunde M., 2012. *Self-assembly of functional, amphipathic amyloid monolayers by the fungal hydrophobin EAS*. Proceedings of the National Academy of Sciences. 109(14). doi.org/10.1073/pnas.1114052109

Mackay J.P., Matthews J.M., Winefield R.D., Mackay L.G., Haverkamp R.G., Templeton M.D., 2001. *The Hydrophobin EAS Is Largely Unstructured in Solution and Functions by Forming Amyloid-Like Structures*. Structure. 9(2): 83–91. doi.org/10.1016/S0969-2126(00)00559-1

Madine J., Jack E., Stockley P.G., Radford S.E., Serpell L.C., Middleton D.A., 2008. *Structural Insights into the Polymorphism of Amyloid-Like Fibrils Formed by Region 20−29 of Amylin Revealed by Solid-State NMR and X-ray Fiber Diffraction*. Journal of the American Chemical Society. 130(45): 14990–5001. doi.org/10.1021/ja802483d

Marino J.P., Schwalbe H., Griesinger C., 1999. J *-Coupling Restraints in RNA Structure Determination*. Accounts of Chemical Research. 32(7): 614–23. doi.org/10.1021/ar9600392

Mariño Pérez L., Ielasi F.S., Bessa L.M., Maurin D., Kragelj J., Blackledge M., Salvi N., Bouvignies G., Palencia A., Jensen M.R., 2022. *Visualizing protein breathing motions associated with aromatic ring flipping*. Nature. 602(7898): 695–700. doi.org/10.1038/s41586-022-04417-6

Markley J.L., Bax A., Arata Y., Hilbers C.W., Kaptein R., Sykes B.D., Wright P.E., Wüthrich K., 1998. *Recommendations for the presentation of NMR structures of proteins and nucleic acids (IUPAC Recommendations 1998)*. Pure and Applied Chemistry. 70(1): 117–42. doi.org/10.1351/pac199870010117

Maurer-Stroh S., Debulpaep M., Kuemmerer N., de la Paz M.L., Martins I.C., Reumers J., Morris K.L., Copland A., Serpell L., Serrano L., Schymkowitz J.W.H., Rousseau F., 2010. *Exploring the sequence determinants of amyloid structure using position-specific scoring matrices*. Nature Methods. 7(3): 237–42. doi.org/10.1038/nmeth.1432

McDermott A., Ridenour C.F., 2002. *Proton Chemical Shift Measurements In Biological Solids*. Encyclopedia of Nuclear Magnetic Resonance. Willey.

McDermott A.E., Polenova T., 2012. *Solid State NMR Studies of Biopolymers*. John Wiley & Sons.

McKinney W., 2010. *Data structures for statistical computing in Python*. Proceedings of the 9th Python In Science Conference. Presented at the Scipy 2010: 56–61.

Mehring M., 1983. *Principles of high resolution NMR in solids*, Second, revised and enlarged edition, softcover reprint of the hardcover 2nd edition 1983. ed. Springer, Berlin Heidelberg New York.

Meiler J., 2003. *PROSHIFT: Protein chemical shift prediction using artificial neural networks*. 26: 25–37. doi.org/10.1023/A:1023060720156

Michel D., Engelke F., 1994. *Cross-Polarization, Relaxation Times and Spin-Diffusion in Rotating Solids*. . In: Blümich, B. (ed.), Solid-State NMR III Organic Matter. Springer Berlin Heidelberg, Berlin, Heidelberg: 69–125. doi.org/10.1007/978-3-642-61223-7_2

Mielke S.P., Krishnan V.V., 2009. *Characterization of protein secondary structure from NMR chemical shifts*. Progress in nuclear magnetic resonance spectroscopy. 54(3–4): 141–65. doi.org/10.1016/j.pnmrs.2008.06.002

Mobli M., Hoch J.C., 2014. *Nonuniform sampling and non-Fourier signal processing methods in multidimensional NMR*. Progress in Nuclear Magnetic Resonance Spectroscopy. 83: 21–41. doi.org/10.1016/j.pnmrs.2014.09.002

Morris G.A., Freeman R., 1979. *Insensitive nuclei enhanced by polarization transfer (INEPT)*. (101): 760.

Morris V.K., Kwan A.H., Sunde M., 2013. *Analysis of the Structure and Conformational States of DewA Gives Insight into the Assembly of the Fungal Hydrophobins*. Journal of Molecular Biology. 425(2): 244–56. doi.org/10.1016/j.jmb.2012.10.021

Morris V.K., Linser R., Wilde K.L., Duff A.P., Sunde M., Kwan A.H., 2012. *Solid-state NMR spectroscopy of functional amyloid from a fungal hydrophobin: A well-ordered β-sheet core amidst structural heterogeneity*. Angewandte Chemie - International Edition. 51(50): 12621–5. doi.org/10.1002/anie.201205625

Neal S., Nip A.M., Zhang H., Wishart D.S., 2003. *Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts*. 26(3): 215–40. doi.org/10.1023/A:1023812930288

Nerli S., McShan A.C., Sgourakis N.G., 2018. *Chemical shift-based methods in NMR structure determination*. Progress in Nuclear Magnetic Resonance Spectroscopy. 106–107: 1–25. doi.org/10.1016/j.pnmrs.2018.03.002

Nicholson L.K., Yamazaki T., Torchia D.A., Grzesiek S., Bax A., Stahl S.J., Kaufman J.D., Wingfield P.T., Lam P.Y.S., Jadhav P.K., Hodge C.N., Domaille P.J., Chang C.-H., 1995. *Flexibility and function in HIV-1 protease*. Nature Structural Biology. 2(4): 274–80. doi.org/10.1038/nsb0495-274

Nielsen N.C., Bildso/e H., Jakobsen H.J., Levitt M.H., 1994. *Double-quantum homonuclear rotary resonance: Efficient dipolar recovery in magic-angle spinning nuclear magnetic resonance*. The Journal of Chemical Physics. 101(3): 1805–12. doi.org/10.1063/1.467759

Nielsen N.Chr., Strassø L.A., Nielsen A.B., 2011. *Dipolar Recoupling*. . In: Chan, J.C.C. (ed.), Solid State NMR, Topics in Current Chemistry. Springer Berlin Heidelberg, Berlin, Heidelberg: 1–45. doi.org/10.1007/128_2011_129

Nikitin K., O'Gara R., 2019. *Mechanisms and Beyond: Elucidation of Fluxional Dynamics by Exchange NMR Spectroscopy*. 25: 4551–89.

Niu Z., Sarkar R., Aichler M., Wester H., Yousefi B.H., Reif B., 2020. *Mapping the Binding Interface of PET Tracer Molecules and Alzheimer Disease Aβ Fibrils by Using MAS Solid-State NMR Spectroscopy*. ChemBioChem. 21(17): 2495–502. doi.org/10.1002/cbic.202000143

Orlando G., Raimondi D., Kagami L.P., Vranken W.F., 2020. *ShiftCrypt: a web server to understand and biophysically align proteins through their NMR chemical shift values*. Nucleic Acids Research. 48(W1): W36–40. doi.org/10.1093/nar/gkaa391

Paramasivam S., Gronenborn A.M., Polenova T., 2018. *Backbone amide 15N chemical shift tensors report on hydrogen bonding interactions in proteins: A magic angle spinning NMR study*. Solid State Nuclear Magnetic Resonance. 92: 1–6. doi.org/10.1016/j.ssnmr.2018.03.002

Parker L.L., Houk A.R., Jensen J.H., 2006. *Cooperative Hydrogen Bonding Effects Are Key Determinants of Backbone Amide Proton Chemical Shifts in Proteins*. Journal of the American Chemical Society. 128(30): 9863–72. doi.org/10.1021/ja0617901

Pauling L., Corey R.B., Branson H.R., 1951. *The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain*. Proceedings of the National Academy of Sciences. 37(4): 205–11. doi.org/10.1073/pnas.37.4.205

Perdigão N., Heinrich J., Stolte C., Sabir K.S., Buckley M.J., Tabor B., Signal B., Gloss B.S., Hammang C.J., Rost B., Schafferhans A., O'Donoghue S.I., 2015. *Unexpected features of the dark proteome*. Proceedings of the National Academy of Sciences. 112(52): 15898–903. doi.org/10.1073/pnas.1508380112

Periole X., Huber T., Bonito-Oliva A., Aberg K.C., van der Wel P.C.A., Sakmar T.P., Marrink S.J., 2018. *Energetics Underlying Twist Polymorphisms in Amyloid Fibrils*. The journal of physical chemistry. B. 122(3): 1081–91. doi.org/10.1021/acs.jpcb.7b10233

Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., Ferrin T.E., 2004. *UCSF Chimera — A visualization system for exploratory research and analysis*. Journal of Computational Chemistry. 25(13): 1605–12. doi.org/10.1002/jcc.20084

Pham C.L.L., Rey A., Lo V., Soulès M., Ren Q., Meisl G., Knowles T.P.J., Kwan A.H., Sunde M., 2016. *Self-assembly of MPG1, a hydrophobin protein from the rice blast fungus that forms functional amyloid coatings, occurs by a surface-driven mechanism*. Scientific Reports. 6(1): 25288. doi.org/10.1038/srep25288

P. Holland G., Mou Q., L. Yarger J., 2013. *Determining hydrogen-bond interactions in spider silk with 1 H– 13 C HETCOR fast MAS solid-state NMR and DFT proton chemical shift calculations*. Chemical Communications. 49(59): 6680–2. doi.org/10.1039/C3CC43737J

Pines A., Gibby M.G., Waugh J.S., 1972. *Proton-Enhanced Nuclear Induction Spectroscopy. A Method for High Resolution NMR of Dilute Spins in Solids*. The Journal of Chemical Physics. 56(4): 1776–7. doi.org/10.1063/1.1677439

Pradhan T., Annamalai K., Sarkar R., Huhn S., Hegenbart U., Schönland S., Fändrich M., Reif B., 2020. *Seeded fibrils of the germline variant of human λ-III immunoglobulin light chain FOR005 have a similar core as patient fibrils with reduced stability*. Journal of Biological Chemistry. 295(52): 18474–84. doi.org/10.1074/jbc.RA120.016006

Purcell E.M., Torrey H.C., Pound R.V., 1946. *Resonance Absorption by Nuclear Magnetic Moments in a Solid*. Physical Review. 69(1–2): 37–8. doi.org/10.1103/PhysRev.69.37

Pustovalova Y., Delaglio F., Craft D.L., Arthanari H., Bax A., Billeter M., Bostock M.J., Dashti H., Hansen D.F., Hyberts S.G., Johnson B.A., Kazimierczuk K., Lu H., Maciejewski M., Miljenović T.M., Mobli M., Nietlispach D., Orekhov V., Powers R., Qu X., Robson S.A., Rovnyak D., Wagner G., Ying J., Zambrello M., Hoch J.C., Donoho D.L., Schuyler A.D., 2021. *NUScon: a community-driven platform for quantitative evaluation of nonuniform sampling in NMR*. Magnetic Resonance. 2(2): 843–61. doi.org/10.5194/mr-2-843-2021

Qiang W., Yau W.-M., Lu J.-X., Collinge J., Tycko R., 2017. *Structural variation in amyloid-\textbeta fibrils from Alzheimer's disease clinical subtypes*. Nature. 541: 217–21. doi.org/10.1038/nature20814

Rabi I.I., Zacharias J.R., Millman S., Kusch P., 1938. *A New Method of Measuring Nuclear Magnetic Moment*. Physical Review. 53(4): 318–318. doi.org/10.1103/PhysRev.53.318

Ramachandran G.N., Ramakrishnan C., Sasisekharan V., 1963. *Stereochemistry of polypeptide chain configurations*. Journal of Molecular Biology. 7(1): 95–9. doi.org/10.1016/S0022-2836(63)80023-6

Robson S., Arthanari H., Hyberts S.G., Wagner G., 2019. *Nonuniform Sampling for NMR Spectroscopy*. Methods in Enzymology. Elsevier: 263–91. doi.org/10.1016/bs.mie.2018.09.009

Robustelli P., Cavalli A., Vendruscolo M., 2008. *Determination of Protein Structures in the Solid State from NMR Chemical Shifts*. Structure. 16(12): 1764–9. doi.org/10.1016/j.str.2008.10.016

Rovnyak D., Filip C., Itin B., Stern A.S., Wagner G., Griffin R.G., Hoch J.C., 2003. *Multiple-quantum magic-angle spinning spectroscopy using nonlinear sampling*. Journal of Magnetic Resonance. 161(1): 43–55. doi.org/10.1016/S1090-7807(02)00189-1

Rovnyak D., Hoch J.C., Stern A.S., Wagner G., 2004. *Resolution and sensitivity of high field nuclear magnetic resonance spectroscopy*. Journal of Biomolecular NMR. 30(1): 1–10. doi.org/10.1023/B:JNMR.0000042946.04002.19

Sammak S., Zinzalla G., 2015. *Targeting protein–protein interactions (PPIs) of transcription factors: Challenges of intrinsically disordered proteins (IDPs) and regions (IDRs)*. Progress in Biophysics and Molecular Biology. 119(1): 41–6. doi.org/10.1016/j.pbiomolbio.2015.06.004

Schanda P., Huber M., Verel R., Ernst M., Meier B.H., 2009. *Direct Detection of ^{3h} J $_{NC'}$ Hydrogen-Bond Scalar Couplings in Proteins by Solid-State NMR Spectroscopy*. Angewandte Chemie. 121(49): 9486–9. doi.org/10.1002/ange.200904411

Schmidt-Rohr K., 1996. *A Double-Quantum Solid-State NMR Technique for Determining Torsion Angles in Polymers*. Macromolecules. 29(11): 3975–81. doi.org/10.1021/ma9517106

Schmidt-Rohr K., Hehn M., Schaefer D., Spiess H.W., 1992. *Two-dimensional nuclear magnetic resonance with sample flip for characterizing orientation distributions, and its analogy to x-ray scattering*. The Journal of Chemical Physics. 97(4): 2247–62. doi.org/10.1063/1.463116

Serrano L., 1995. *Comparison between the φ distribution of the amino acids in the protein database and NMR data indicates that amino acids have various φ propensities in the random coil conformation*. Journal of Molecular Biology. 254(2): 322–33. doi.org/10.1006/jmbi.1995.0619

Shen Y., Bax A., 2013. *Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks*. Journal of Biomolecular NMR. 56(3): 227–41. doi.org/10.1007/s10858-013-9741-y

Shen Y., Bax A., 2010. *SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network*. Journal of Biomolecular NMR. 48(1): 13–22. doi.org/10.1007/s10858-010-9433-9

Shen Y., Delaglio F., Cornilescu G., Bax A., 2009. *TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts*. Journal of Biomolecular NMR. 44(4): 213–23. doi.org/10.1007/s10858-009-9333-z

Skinner S.P., Fogh R.H., Boucher W., Ragan T.J., Mureddu L.G., Vuister G.W., 2016. *CcpNmr AnalysisAssign: a flexible platform for integrated NMR analysis*. Journal of Biomolecular NMR. 66(2): 111–24. doi.org/10.1007/s10858-016-0060-y

Slichter C.P., 1990. *Double Resonance*. Principles of Magnetic Resonance, Springer Series in Solid-State Sciences. Springer Berlin Heidelberg, Berlin, Heidelberg: 247–366. doi.org/10.1007/978-3-662-09441-9_7

Spiliopoulou M., Valmas A., Triandafillidis D.-P., Kosinas C., Fitch A., Karavassili F., Margiolaki I., 2020. *Applications of X-ray Powder Diffraction in Protein Crystallography and Drug Screening*. Crystals. 10(2): 54. doi.org/10.3390/cryst10020054

Stanek J., Koźmiński W., 2010. *Iterative algorithm of discrete Fourier transform for processing randomly sampled NMR data sets*. Journal of Biomolecular NMR. 47(1): 65–77. doi.org/10.1007/s10858-010-9411-2

Stern A.S., Li K.-B., Hoch J.C., 2002. *Modern spectrum analysis in multidimensional NMR spectroscopy: comparison of linear-prediction extrapolation and maximum-entropy reconstruction*. Journal of the American Chemical Society. 124(9): 1982–93. doi.org/10.1021/ja011669o

Strotz D., Orts J., Chi C.N., Riek R., Vögeli B., 2017. *eNORA2 Exact NOE Analysis Program*. Journal of Chemical Theory and Computation. 13(9): 4336–46. doi.org/10.1021/acs.jctc.7b00436

Su Y., Hong M., 2011. *Conformational disorder of membrane peptides investigated from solid-state NMR line widths and line shapes*. Journal of Physical Chemistry B. 115(36): 10758–67. doi.org/10.1021/jp205002n

Subramaniam S., Kleywegt G.J., 2022. *A paradigm shift in structural biology*. Nature Methods. 19(1): 20–3. doi.org/10.1038/s41592-021-01361-7

Sumowski C.V., Hanni M., Schweizer S., Ochsenfeld C., 2014. *Sensitivity of ab initio vs empirical methods in computing structural effects on NMR chemical shifts for the example of peptides*. Journal of Chemical Theory and Computation. 10(1): 122–33. doi.org/10.1021/ct400713t

Sun S., Gill M., Li Y., Huang M., Byrd R.A., 2015. *Efficient and generalized processing of multidimensional NUS NMR data: The NESTA algorithm and comparison of regularization terms*. Journal of Biomolecular NMR. 62(1): 105–17. doi.org/10.1007/s10858-015-9923-x

Szeverenyi N.M., Sullivan M.J., Maciel G.E., 1982. *Observation of spin exchange by two-dimensional fourier transform 13C cross polarization-magic-angle spinning*. Journal of Magnetic Resonance (1969). 47(3): 462–75. doi.org/10.1016/0022-2364(82)90213-X

Takegoshi K., Nakamura S., Terao T., 2001. *13C–1H dipolar-assisted rotational resonance in magic-angle spinning NMR*. Chemical Physics Letters. 344(5): 631–7. doi.org/10.1016/S0009-2614(01)00791-6

Tamiola K., Acar B., Mulder F.A.A., 2010. *Sequence-Specific Random Coil Chemical Shifts of Intrinsically Disordered Proteins*. Journal of the American Chemical Society. 132(51): 18000–3. doi.org/10.1021/ja105656t

Tošner Z., Brandl M.J., Blahut J., Glaser S.J., Reif B., 2021. *Maximizing efficiency of dipolar recoupling in solid-state NMR using optimal control sequences*. Science Advances. 7(42): eabj5913. doi.org/10.1126/sciadv.abj5913

Tošner Z., Sarkar R., Becker-Baldus J., Glaubitz C., Wegner S., Engelke F., Glaser S.J., Reif B., 2018. *Overcoming Volume Selectivity of Dipolar Recoupling in Biological Solid-State NMR Spectroscopy*. Angewandte Chemie International Edition. 57(44): 14514–8. doi.org/10.1002/anie.201805002

Turoverov K.K., Kuznetsova I.M., Fonin A.V., Darling A.L., Zaslavsky B.Y., Uversky V.N., 2019. *Stochasticity of Biological Soft Matter: Emerging Concepts in Intrinsically Disordered Proteins and Biological Phase Separation*. Trends in Biochemical Sciences. 44(8): 716–28. doi.org/10.1016/j.tibs.2019.03.005

Tuttle M.D., Comellas G., Nieuwkoop A.J., Covell D.J., Berthold D.A., Kloepper K.D., Courtney J.M., Kim J.K., Barclay A.M., Kendall A., Wan W., Stubbs G., Schwieters C.D., Lee V.M.Y., George J.M., Rienstra C.M., 2016. *Solid-state NMR structure of a pathogenic fibril of full-length human α-synuclein*. Nature Structural and Molecular Biology. 23(5): 409–15. doi.org/10.1038/nsmb.3194

Uluca B., Viennet T., Petrović D., Shaykhalishahi H., Weirich F., Gönülalan A., Strodel B., Etzkorn M., Hoyer W., Heise H., 2018. *DNP-Enhanced MAS NMR: A Tool to Snapshot Conformational Ensembles of α-Synuclein in Different States.* Biophysical journal. 114(7): 1614–23. doi.org/10.1016/j.bpj.2018.02.011

Vallurupalli P., Bouvignies G., Kay L.E., 2012. *Studying "Invisible" Excited Protein States in Slow Exchange with a Major State Conformation*. Journal of the American Chemical Society. 134(19): 8148–61. doi.org/10.1021/ja3001419

van Beek J.D., Beaulieu L., Schäfer H., Demura M., Asakura T., Meier B.H., 2000. *Solid-state NMR determination of the secondary structure of Samia cynthia ricini silk*. Nature. 405(6790): 1077–9. doi.org/10.1038/35016625

van Beek J.D., Hess S., Vollrath F., Meier B.H., 2002. *The molecular structure of spider dragline silk: Folding and orientation of the protein backbone*. Proceedings of the National Academy of Sciences. 99(16): 10266–71. doi.org/10.1073/pnas.152162299

Verel R., Ernst M., Meier B.H., 2001. *Adiabatic Dipolar Recoupling in Solid-State NMR: The DREAM Scheme*. Journal of Magnetic Resonance. 150(1): 81–99. doi.org/10.1006/jmre.2001.2310

Vila J.A., Arnautova Y.A., Martin O.A., Scheraga H.A., 2009. *Quantum-mechanics-derived 13Cα chemical shift server (CheShift) for protein structure validation*. Proceedings of the National Academy of Sciences. 106(40): 16972–7. doi.org/10.1073/pnas.0908833106

Virtanen P., Gommers R., Oliphant T.E., Haberland M., Reddy T., Cournapeau D., Burovski E., Peterson P., Weckesser W., Bright J., van der Walt S.J., Brett M., Wilson J., Millman K.J., Mayorov N., Nelson A.R.J., Jones E., Kern R., Larson E., Carey C.J., Polat İ., Feng Y., Moore E.W., VanderPlas J., Laxalde D., Perktold J., Cimrman R., Henriksen I., Quintero E.A., Harris C.R., Archibald A.M., Ribeiro A.H., Pedregosa F., van Mulbregt P., 2020. *SciPy 1.0: fundamental algorithms for scientific computing in Python*. Nature Methods. 17(3): 261–72. doi.org/10.1038/s41592-019-0686-2

Walsh I., Seno F., Tosatto S.C.E., Trovato A., 2014. *PASTA 2.0: an improved server for protein aggregation prediction*. Nucleic Acids Research. 42(W1): W301–7. doi.org/10.1093/nar/gku399

Wang H.-W., Wang J.-W., 2017. *How cryo-electron microscopy and X-ray crystallography complement each other: Cryo-EM and X-Ray Crystallography Complement Each Other*. Protein Science. 26(1): 32–9. doi.org/10.1002/pro.3022

Wang Y., Jardetzky O., 2002. *Probability-based protein secondary structure identification using combined NMR chemical-shift data*. Protein Science. 11(4): 852–61. doi.org/10.1110/ps.3180102

Waskom M., Botvinnik O., O'Kane D., Hobson P., Lukauskas S., Gemperline D.C., Augspurger T., Halchenko Y., Cole J.B., Warmenhoven J., 2017. *mwaskom/seaborn: v0. 8.1 (September 2017), Zenodo*. Available at: doi. 10.

Wieske L.H.E., Erdélyi M., 2021. *Non-uniform sampling for NOESY? A case study on spiramycin*. Magnetic Resonance in Chemistry. 59(7): 723–37. doi.org/10.1002/mrc.5133

Williamson M.P., Asakura T., 1997. *Protein Chemical Shifts*. Protein NMR Techniques. Humana Press, New Jersey: 53–70. doi.org/10.1385/0-89603-309-0:53

Wishart D.S., 2011. *Interpreting protein chemical shift data*. Progress in Nuclear Magnetic Resonance Spectroscopy. 58(1–2): 62–87. doi.org/10.1016/j.pnmrs.2010.07.004

Wishart D.S., Bigam C.G., Holm A., Hodges R.S., Sykes B.D., 1995. *1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects*. Journal of Biomolecular NMR. 5(1): 67–81. doi.org/10.1007/BF00227471

Wishart D.S., Case D.A., 2002. *Use of Chemical Shifts in Macromolecular Structure Determination*. Methods in Enzymology. Elsevier: 3–34. doi.org/10.1016/S0076-6879(02)38214-4

Wishart D.S., Sykes B.D., 1994. *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(2): 171–80. doi.org/10.1007/BF00175245

Wishart D.S., Sykes B.D., Richards F.M., 1992. *The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy*. Biochemistry. 31(6): 1647–51. doi.org/10.1021/bi00121a010

Wu X.L., Zilm K.W., 1993. *Cross Polarization with High-Speed Magic-Angle Spinning*. Journal of Magnetic Resonance, Series A. 104(2): 154–65. doi.org/10.1006/jmra.1993.1203

Xiang S., Kulminskaya N., Habenstein B., Biernat J., Tepper K., Paulat M., Griesinger C., Becker S., Lange A., Mandelkow E., Linser R., 2017. *A Two-Component Adhesive: Tau Fibrils Arise from a Combination of a Well-Defined Motif and Conformationally Flexible Interactions*. Journal of the American Chemical Society. 139(7): 2639–46. doi.org/10.1021/jacs.6b09619

Xiao Y., Ma B., McElheny D., Parthasarathy S., Long F., Hoshi M., Nussinov R., Ishii Y., 2015. *Aβ(1–42) fibril structure illuminates self-recognition and replication of amyloid in Alzheimer's disease*. Nature Structural & Molecular Biology. 22(6): 499–505. doi.org/10.1038/nsmb.2991

Xu X.P., Case D.A., 2001. *Automated prediction of 15N, 13Cα, 13Cβ and 13C′ chemical shifts in proteins using a density functional database*. Journal of Biomolecular NMR. 21(4): 321–33. doi.org/10.1023/A:1013324104681

Xue K., Sarkar R., Motz C., Asami S., Decker V., Wegner S., Tosner Z., Reif B., 2018. *Magic-Angle Spinning Frequencies beyond 300 kHz Are Necessary To Yield Maximum Sensitivity in Selectively Methyl Protonated Protein Samples in Solid-State NMR*. The Journal of Physical Chemistry C. 122(28): 16437–42. doi.org/10.1021/acs.jpcc.8b05600

Yi X., Fritzsching K.J., Rogawski R., Xu Y., McDermott A.E., 2023a. *Contribution of protein conformational heterogeneity to NMR lineshapes at cryogenic temperatures* (preprint). Biophysics. doi.org/10.1101/2023.01.24.525358

Yi X., Zhang L., Friesner R.A., McDermott A., 2023b. *Predicted and Experimental NMR Chemical Shifts at Variable Temperatures: The Effect of Protein Conformational Dynamics* (preprint). Biophysics. doi.org/10.1101/2023.01.25.525502

Ying J., Delaglio F., Torchia D.A., Bax A., 2017. *Sparse multidimensional iterative lineshape-enhanced (SMILE) reconstruction of both non-uniformly sampled and conventional NMR data*. Journal of Biomolecular NMR. 68(2): 101–18. doi.org/10.1007/s10858-016-0072-7

Zhou D.H., Shah G., Cormos M., Mullen C., Sandoz D., Rienstra C.M., 2007. *Proton-Detected Solid-State NMR Spectroscopy of Fully Protonated Proteins at 40 kHz Magic-Angle Spinning*. Journal of the American Chemical Society. 129(38): 11791–801. doi.org/10.1021/ja073462m

Figures

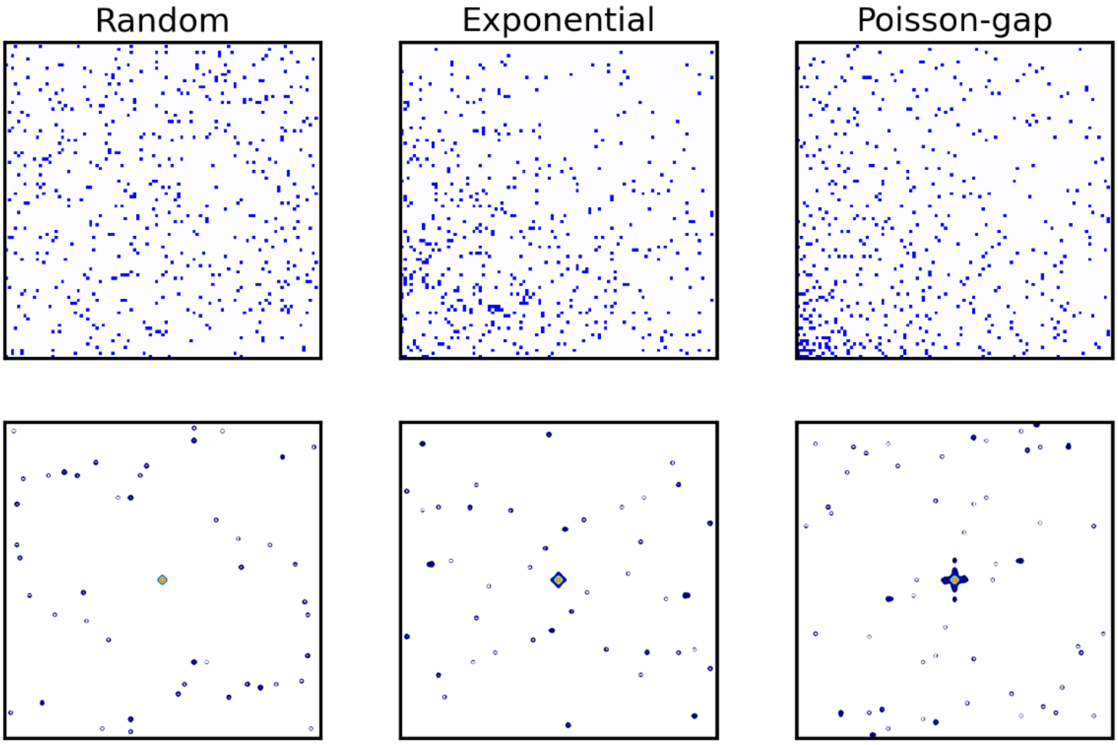


Figure S1. NUS schedules covering 5% points (top) and their Point Spread Functions (bottom). Schedules were generated using the `nus-tool` program at NMRbox (Maciejewski *et al.*, 2017). Contours on the PSF plots are set to 0.05% of maximum intensity and increase linearly with an increment of 0.1

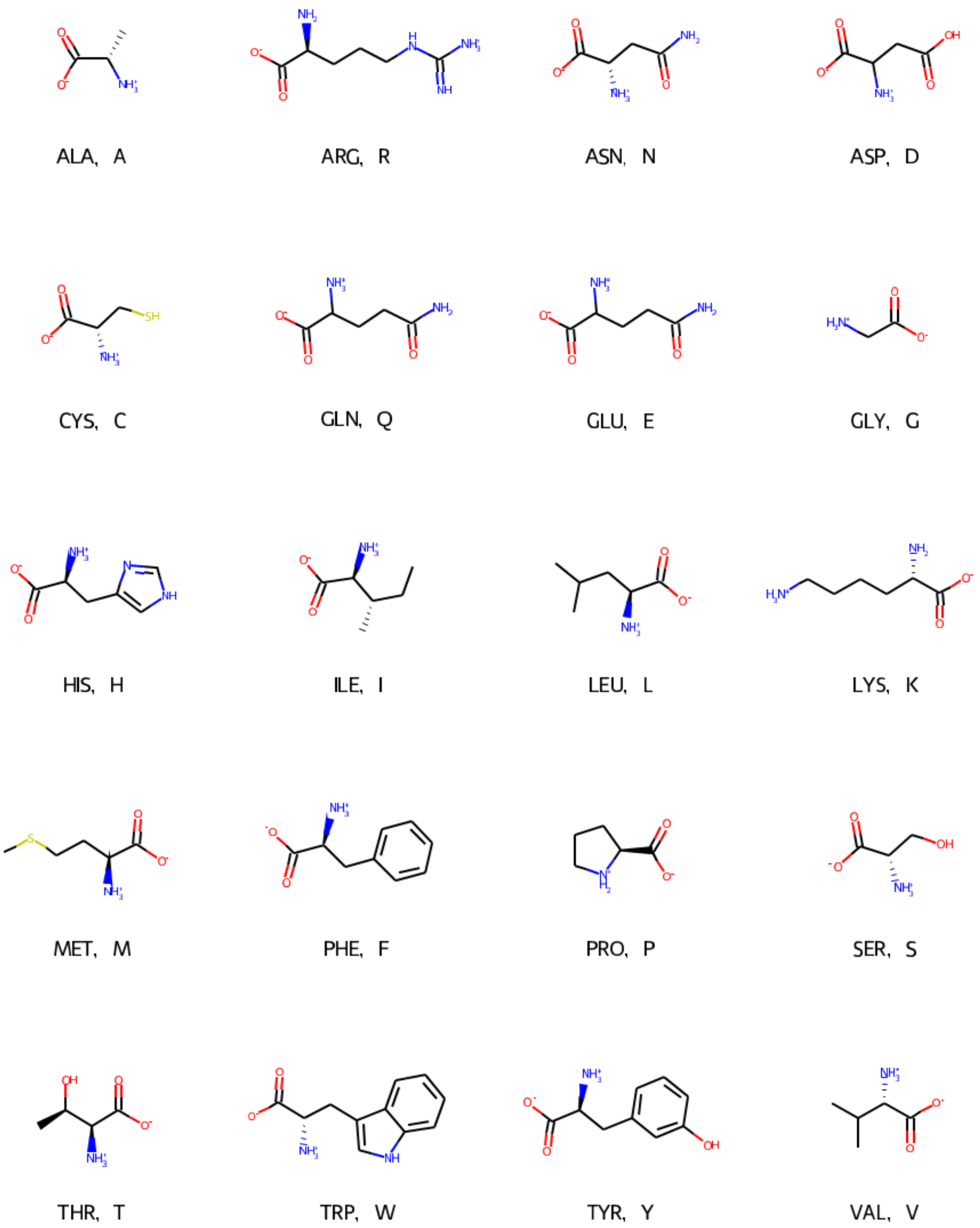


Figure S2. Chemical structures and letter codes of the 20 proteinogenic amino acids in zwitterionic form.

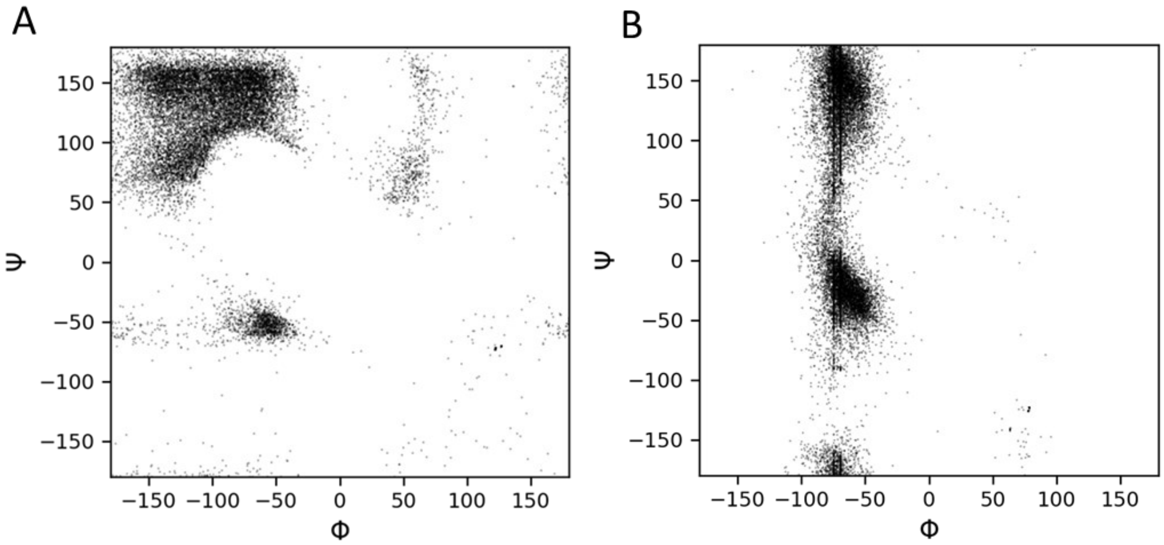


Figure S3. Backbone dihedral-angle distributions of **A**: pre-Pro residues (Gly excluded) and **B**: prolines. Data taken from PACSY (accessed in March 2023). The nature of the vertical artefacts (at *φ*≈75.5° and *φ*≈70°) in **B** is unclear.

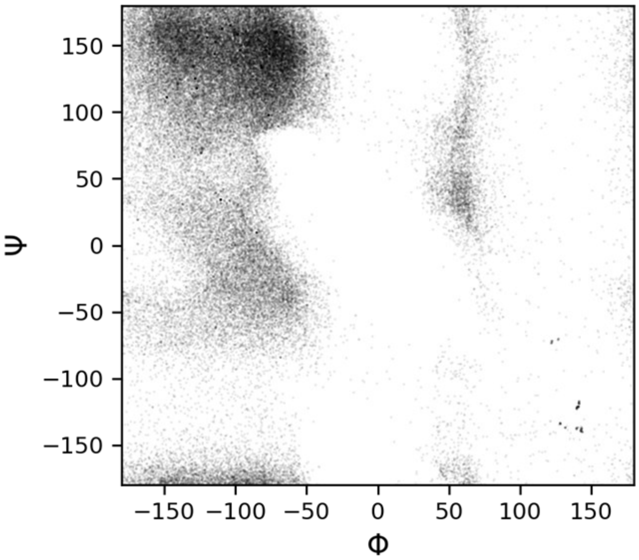


Figure S4. Backbone dihedral-angle distribution of residues identified as random coil (‘C’) by STRIDE. Data taken from PACSY (accessed in March 2023).

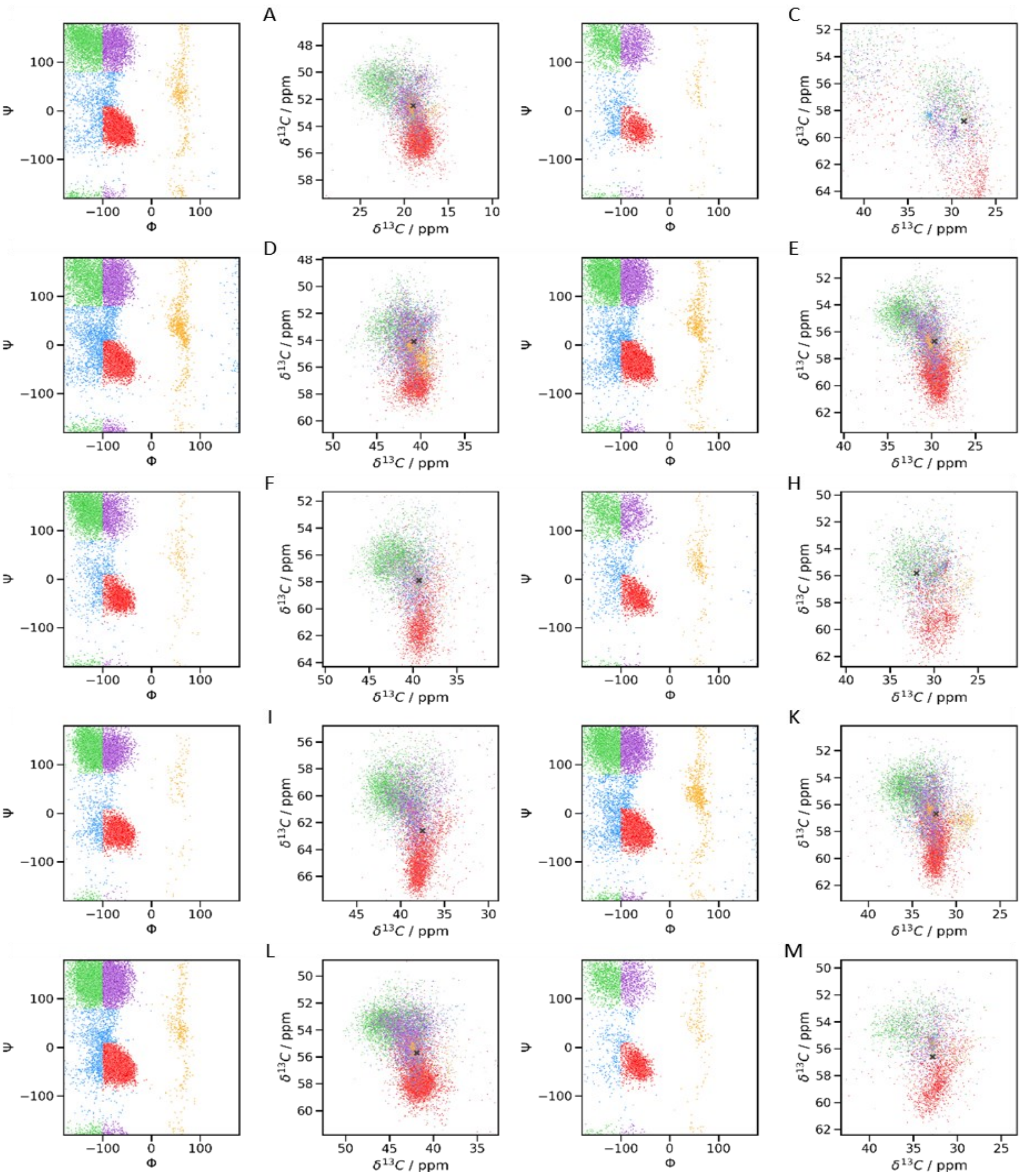


Figure S5. (Continued on the next page)

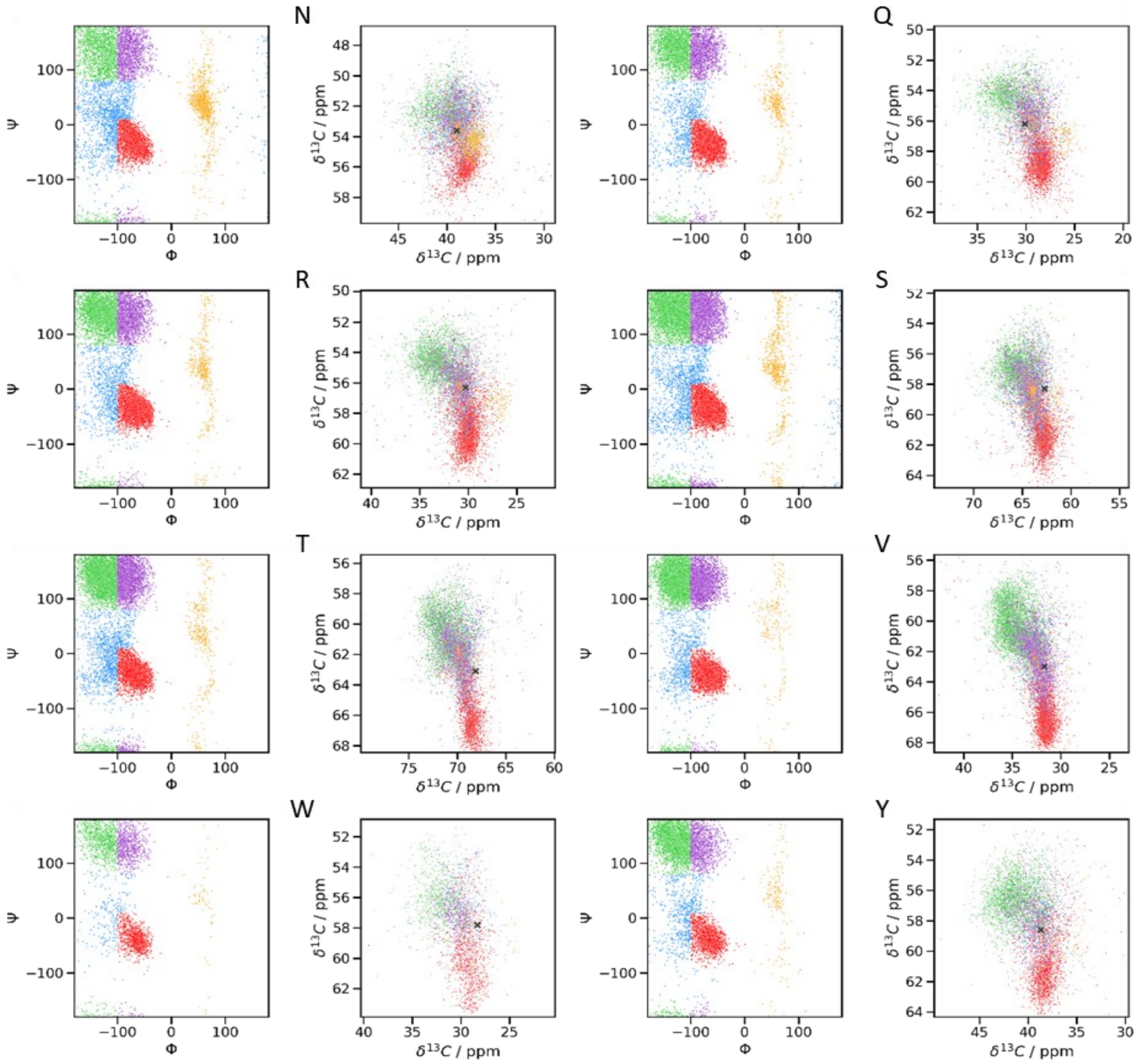


Figure S5. Backbone dihedral angles adopted by amino acid residues in proteins and their Cα and Cβ chemical shifts. Proline and glycine are omitted. Data is taken from the relational database PACSY (Lee *et al.*, 2012), which has been Purged through Intrinsic Quality Criterion (PICS) (Fritzsching *et al.*, 2016). Colors correspond to the ranges of φ and ψ angles. Random-coil chemical shifts are marked with black 'x' (data taken from TALOS database).

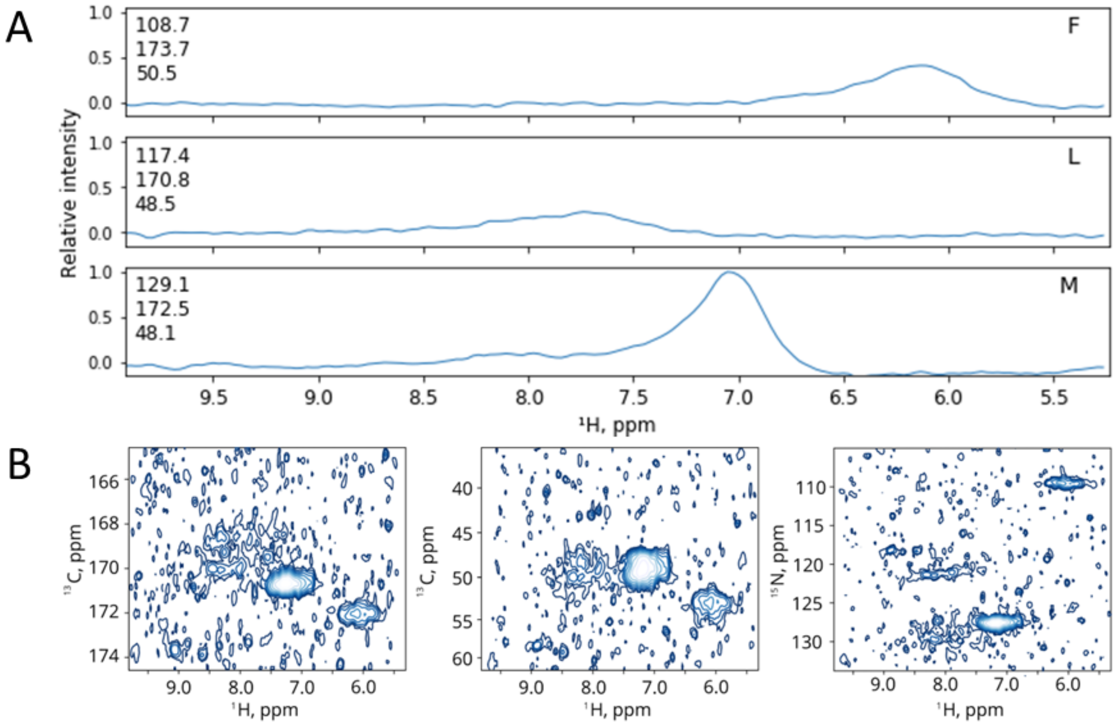


Figure S6. Additional data to the 4D hCOCANH spectrum of dehomogenized fMLF. **A**: 1D slices of the uniformly sampled spectrum; **B**: H/CO (left), H/CA (middle), and H/N (right) projections of the 4D spectrum acquired with 5% sampling density using the Poisson-Gap sampling scheme (default setting of sinusoidal weighting parameter, SSW=2) and reconstructed with SMILE.

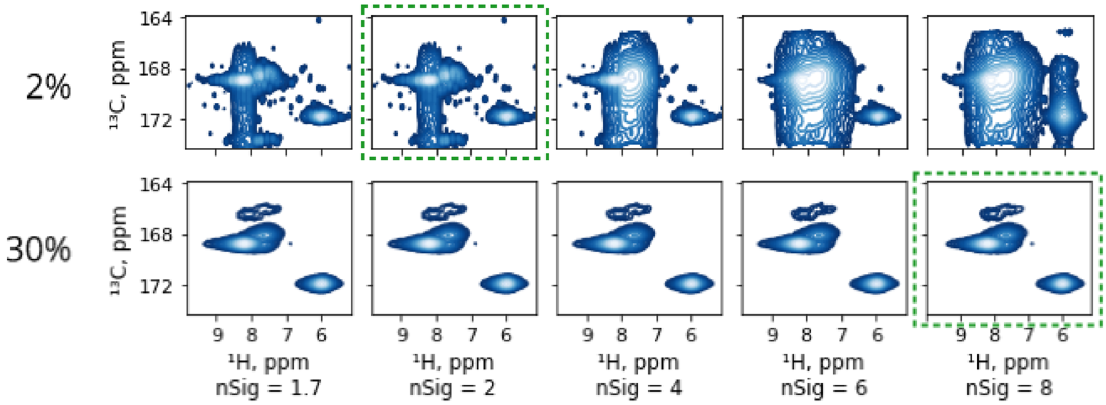


Figure S7. Optimization of SMILE reconstruction using data subsampled from a uniformly sampled hCONH spectrum of fMLF with the Poisson-gap sampling scheme. The figure demonstrates poorer quality of the reconstructed datasets as compared to the randomly sampled data.

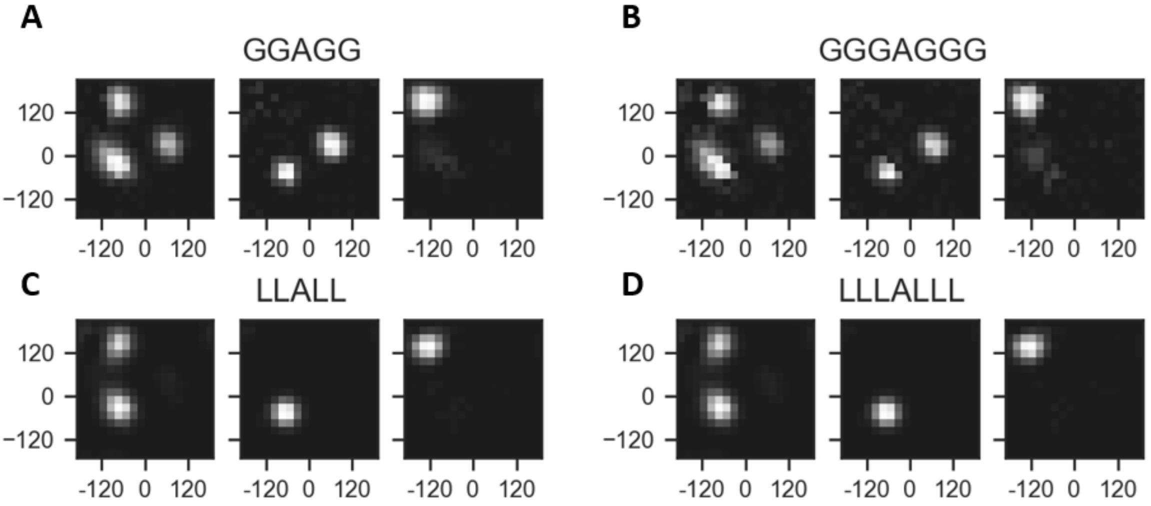


Figure S8: TALOS-N predictions depend on the amino acid context. Every panel (**A-D**) comprises predictions of (φ, ψ) maps for Ala of the peak maximum / mixed propensity point (left), a helix-like point (middle) and a sheet-like point (right). Chemical shifts for other residues were propensity- and residue-type corrected as described in the main text. In run **B** additional glycines were prepended and appended. In runs **C** and **D**, glycines were artificially replaced (for the TALOS input) with leucines. This experiment demonstrates that TALOS-N returns predictions without the right-handed helical component for a (normal) sequence with (bulkier) sidechains. In case of no sidechains (**A** and **C**), the right-handed helical component is lower in the case of a 7-mer, which is probably related to the overall confidence of TALOS-N prediction.

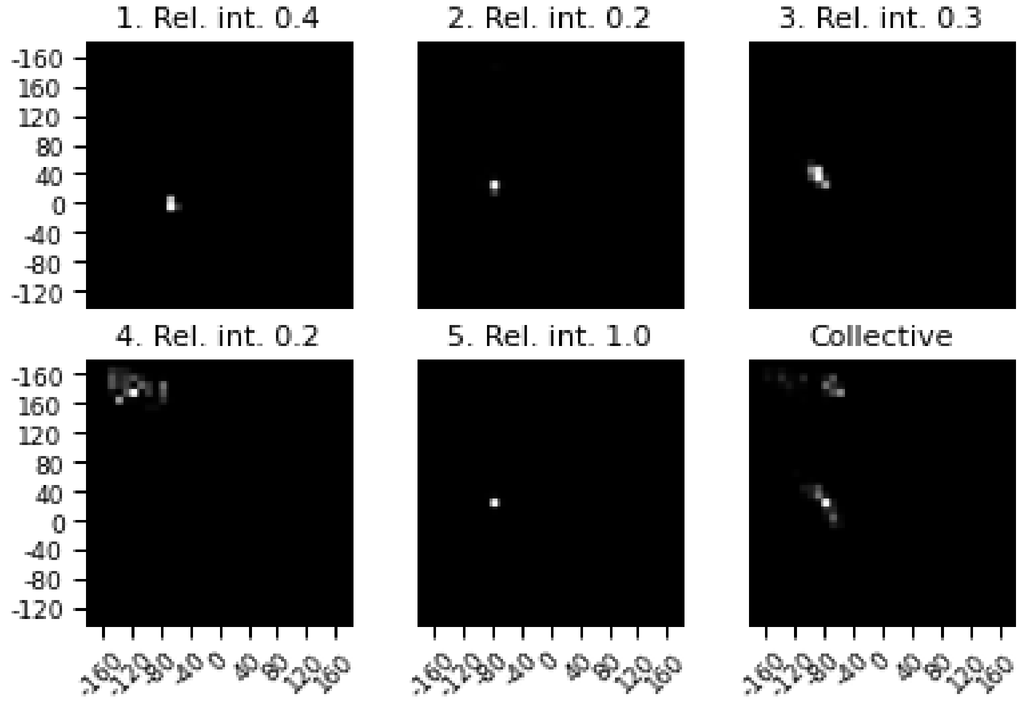


Figure S9. Individual and collective (φ, ψ)-maps obtained with DANGLE (Global Likelihood Diagrams in original terminology of Cheung *et al.*, (2010)) in original 10° resolution. Compare with Fig. 2.2.3.

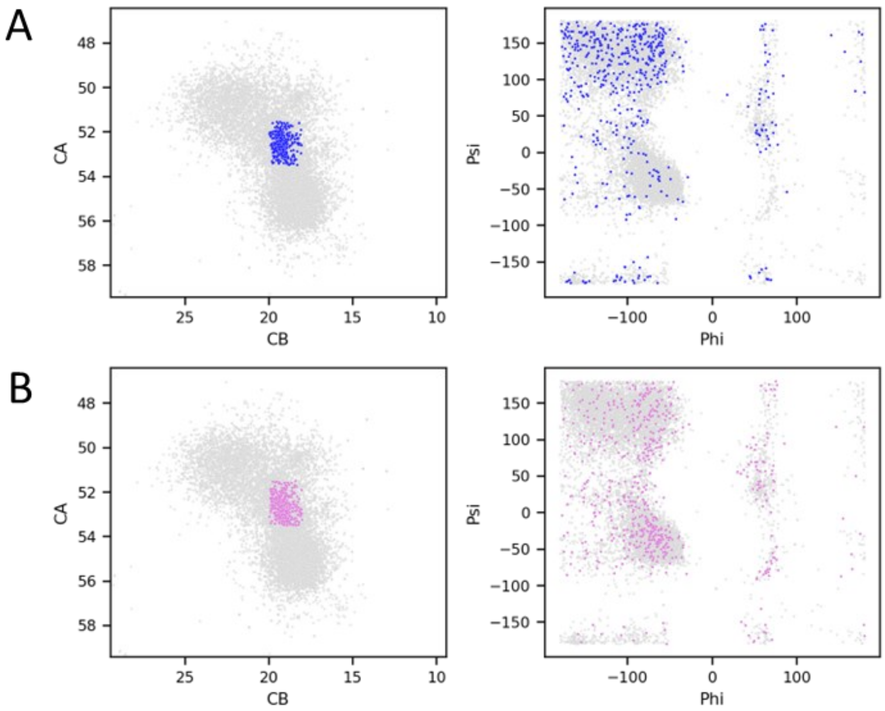


Figure S10. Chemical shifts (left) and adopted backbone conformations (right) of alanines with chemical shifts close to the random-coil values. **A**: Alanines from random coils (class ‘C’, presumably dynamic fragments) and **B**: alanines of all other secondary-structure classes (presumably rigid fragments). Gray dots are all alanine entries of PACSY (accessed March, 2023)

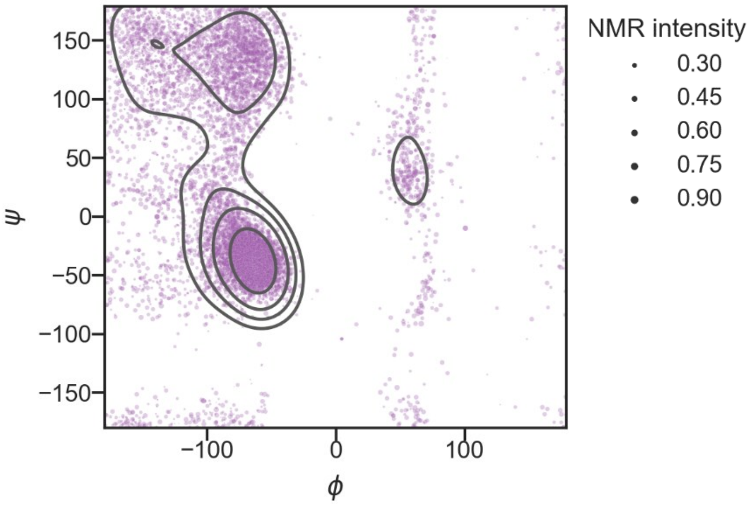


Fig. S11: PACSY entries, included in the volume of the 4D CACB crosspeak in GGAGG HNCACB spectrum. Sizes of the points represent the peak intensity (relative to peak maximum). Contours represent weighted density estimate (starting from 0.15 relative point density and succeeding with the factor of 1.1)

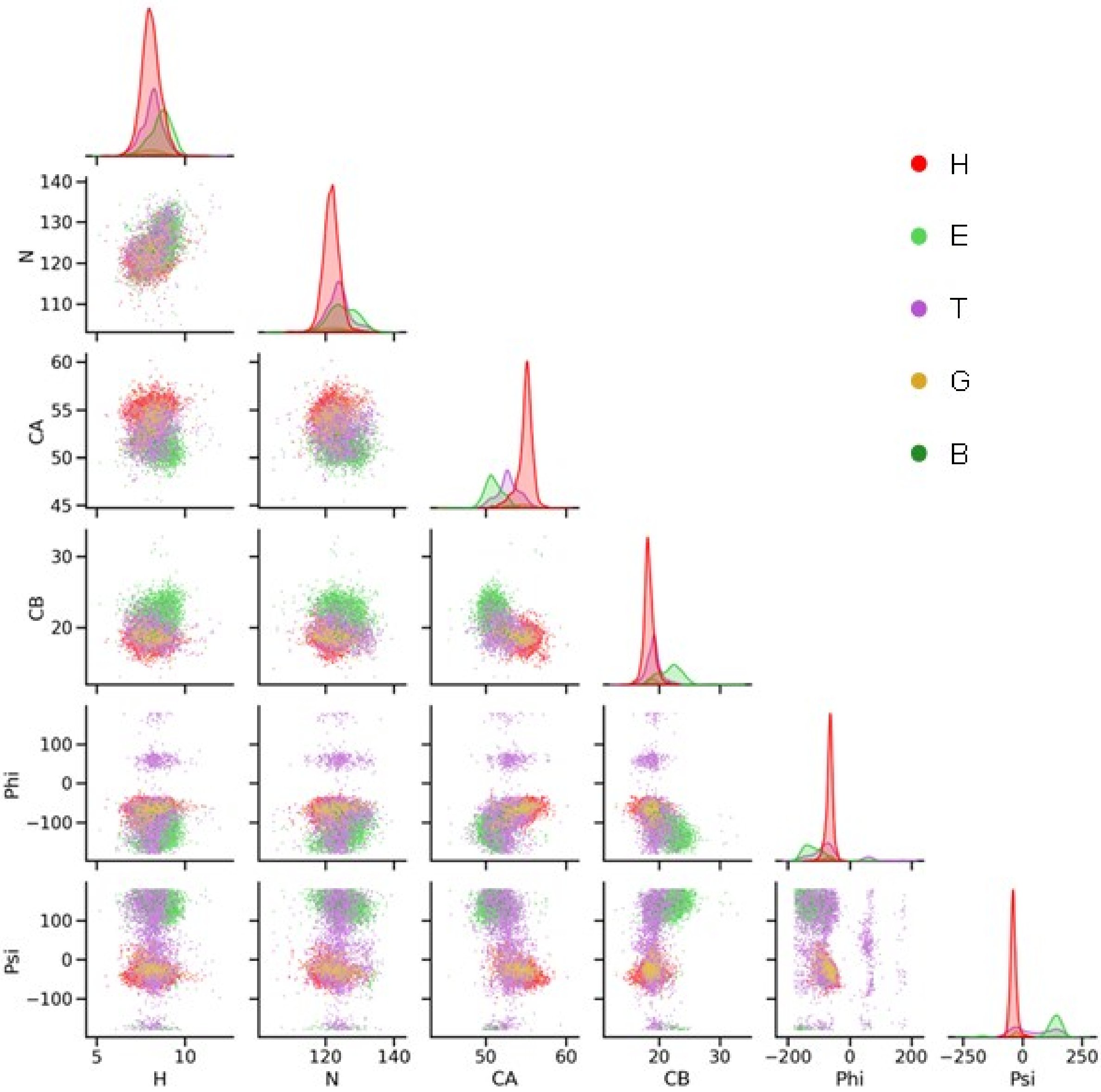


Figure S12. Pair-wise relations of the four chemical shifts and the two backbone dihedral angles (N=14137) for alanine. Points are colored according to the STRIDE classification of the residue: H – α-helix, E – β-sheet, T – turn, B – isolated β-strand, G – 3-10 helix. Rare classes ("I", "b") as well as random coil ("C") are omitted.

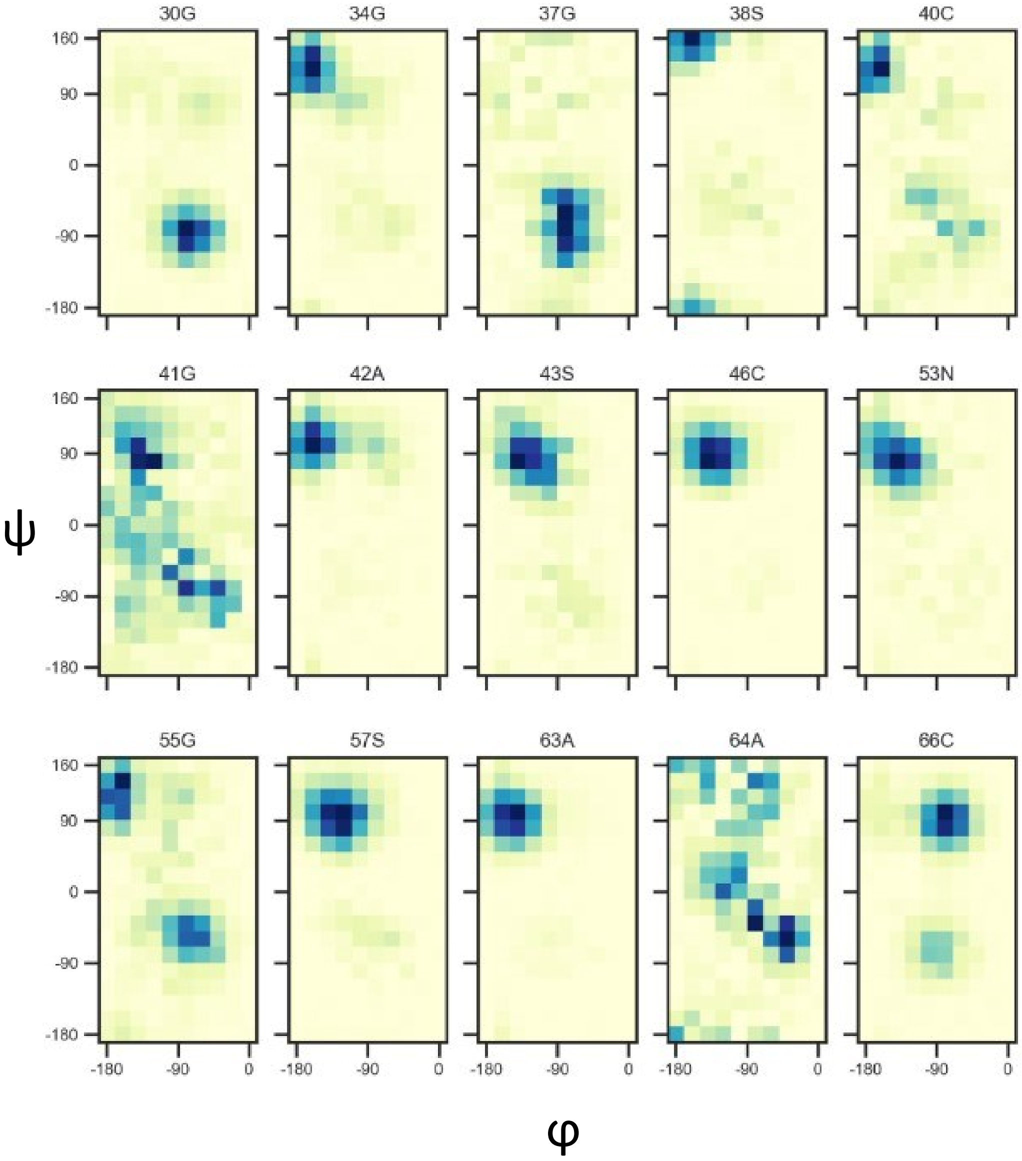


Figure S13. Folded collective TALOS-N (φ, ψ)-maps for $EAS_{\Delta15}$ rodlets. Folding allows to merge the predictions for right- and left-wound structures.

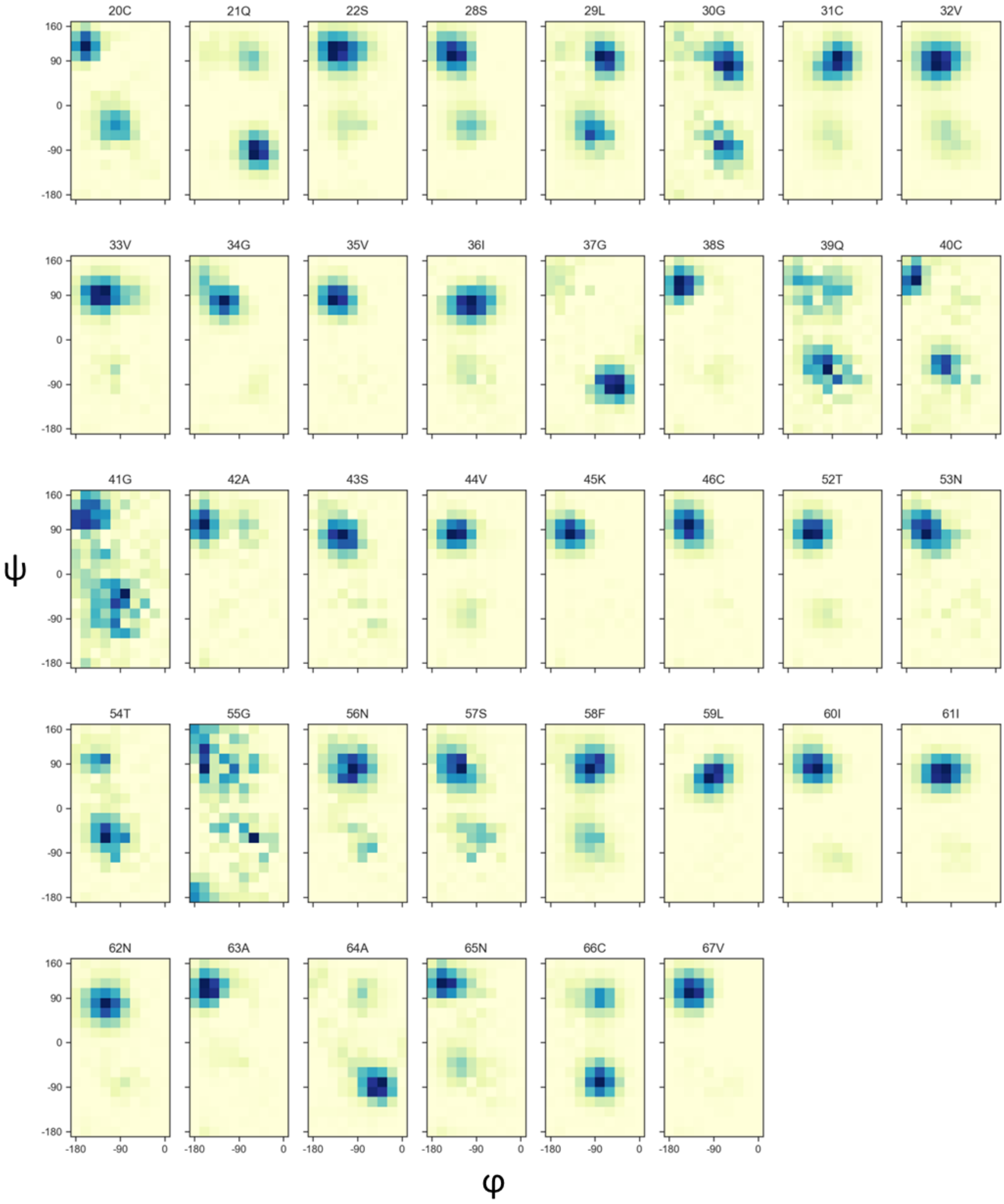


Figure S14. Folded (φ, ψ) maps predicted by regular TALOS-N analysis (only considering peak maxima) of EAS$_{\Delta 15}$ rodlets.

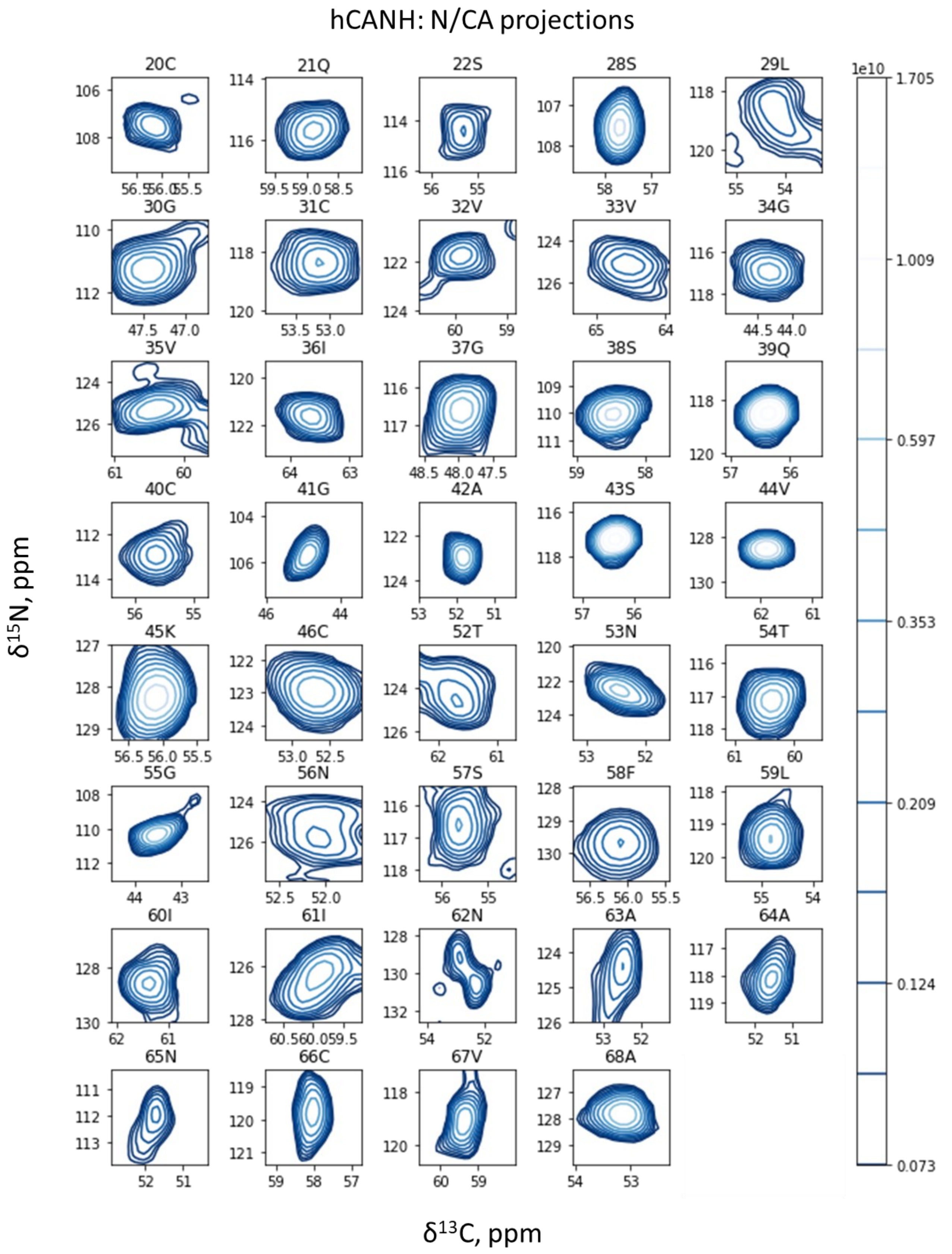


Figure S15. ^{15}N/^{13}Cα projections of the 3D hCANH peaks of the EAS$_{Δ15}$ rodlet sample. The peak boxes were extracted from the 3D spectrum as described in *Section M.3.5.* Contours start at 0.15 of the absolute intensity of each individual peak.

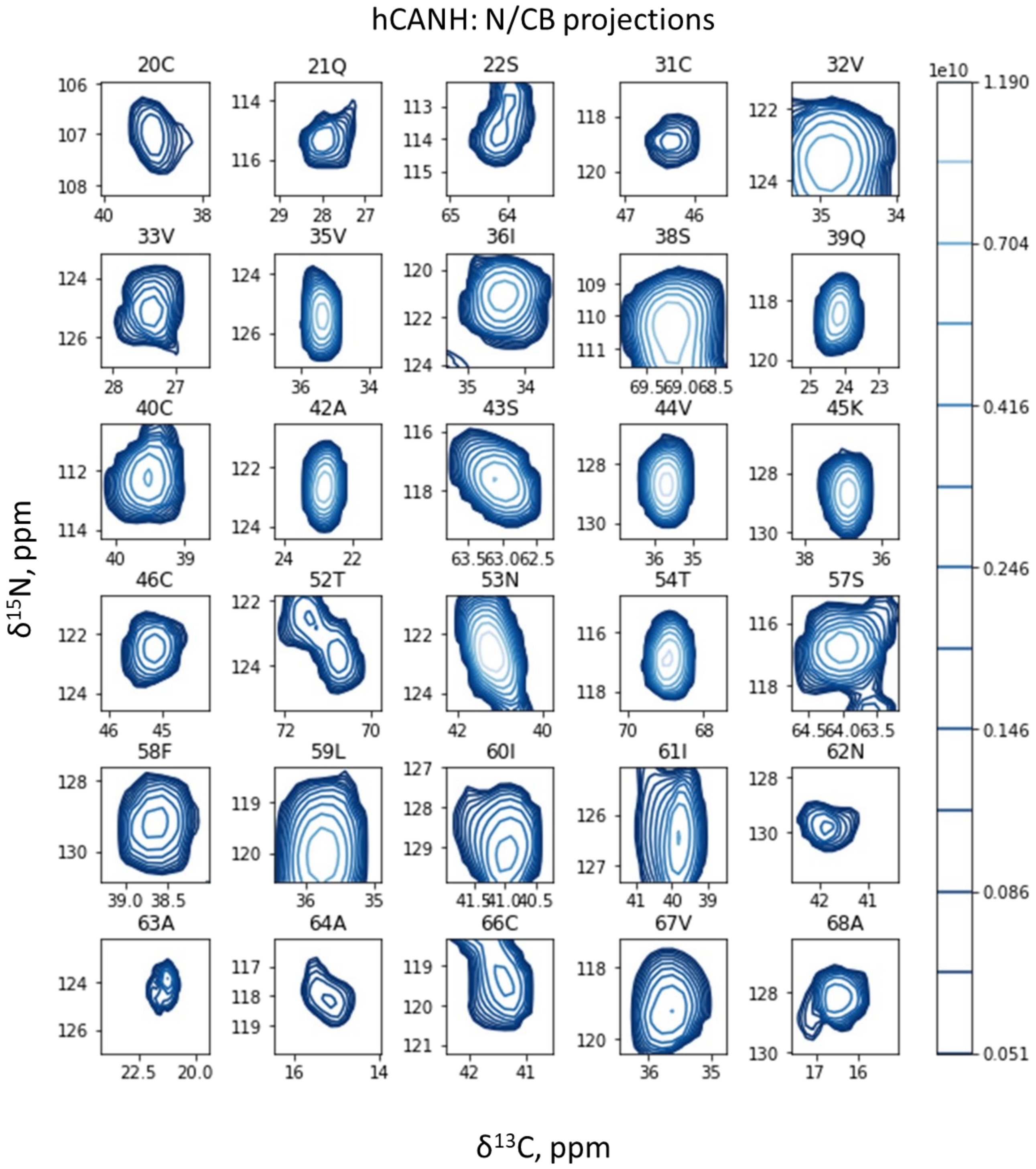


Figure S16. $^{15}N/^{13}C\beta$ projections of the 3D hcaCBcaNH peaks of the $EAS_{\Delta 15}$ rodlet sample. The peak boxes were extracted from the 3D spectrum as described in *Section M.3.5.* Contours start at 0.15 of the absolute intensity of each individual peak.

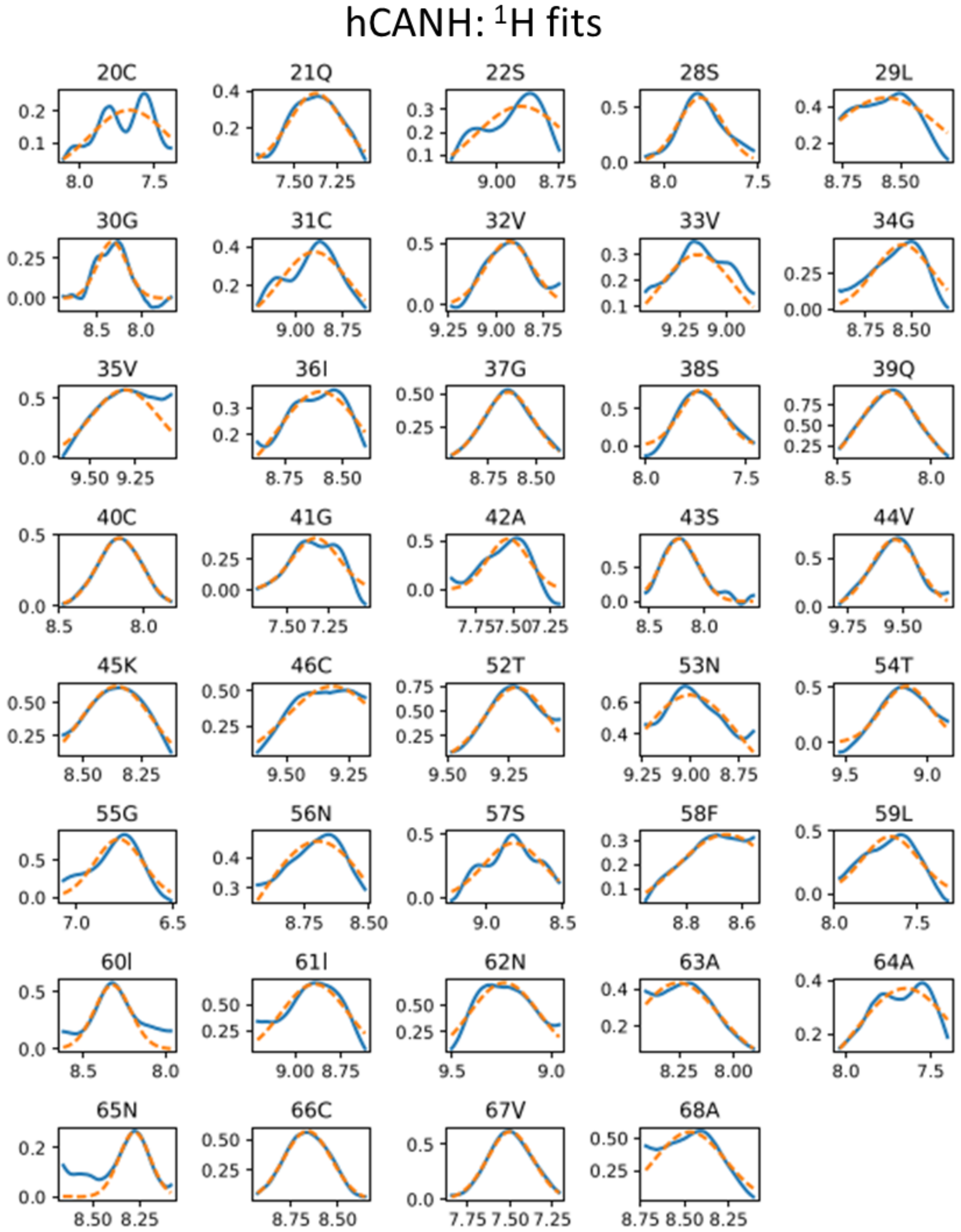


Figure S17.1D projections onto the ^{1}H axis of the 3D hCANH peaks of the EAS$_{\Delta15}$ rodlet sample. Blue lines: peaks; dashed orange lines: fits. Extracted linewidths are shown in *Results, Section 2.3.4*, Fig. 2.3.5B. The scale of the vertical axis is consistent throughout the whole set of fits.

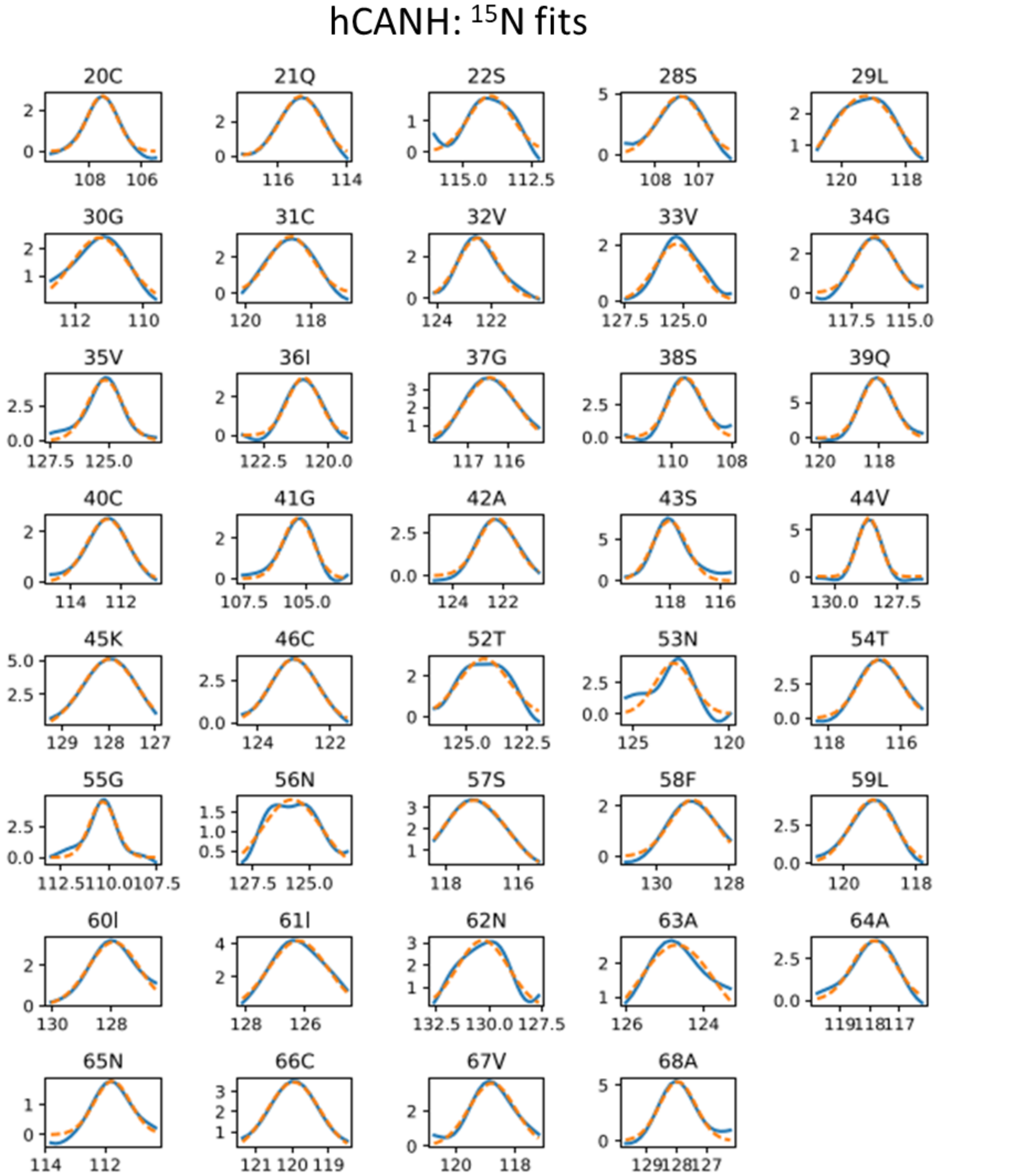


Figure S18. 1D projections onto the ^{15}N axis of the 3D hCANH peaks of the EAS$_{\Delta 15}$ rodlet sample. Blue lines: peaks; dashed orange lines: fits. Extracted linewidths are shown in *Results, Section 2.3.4*, Fig. 2.3.5B. The scale of the vertical axis is consistent throughout the whole set of fits.

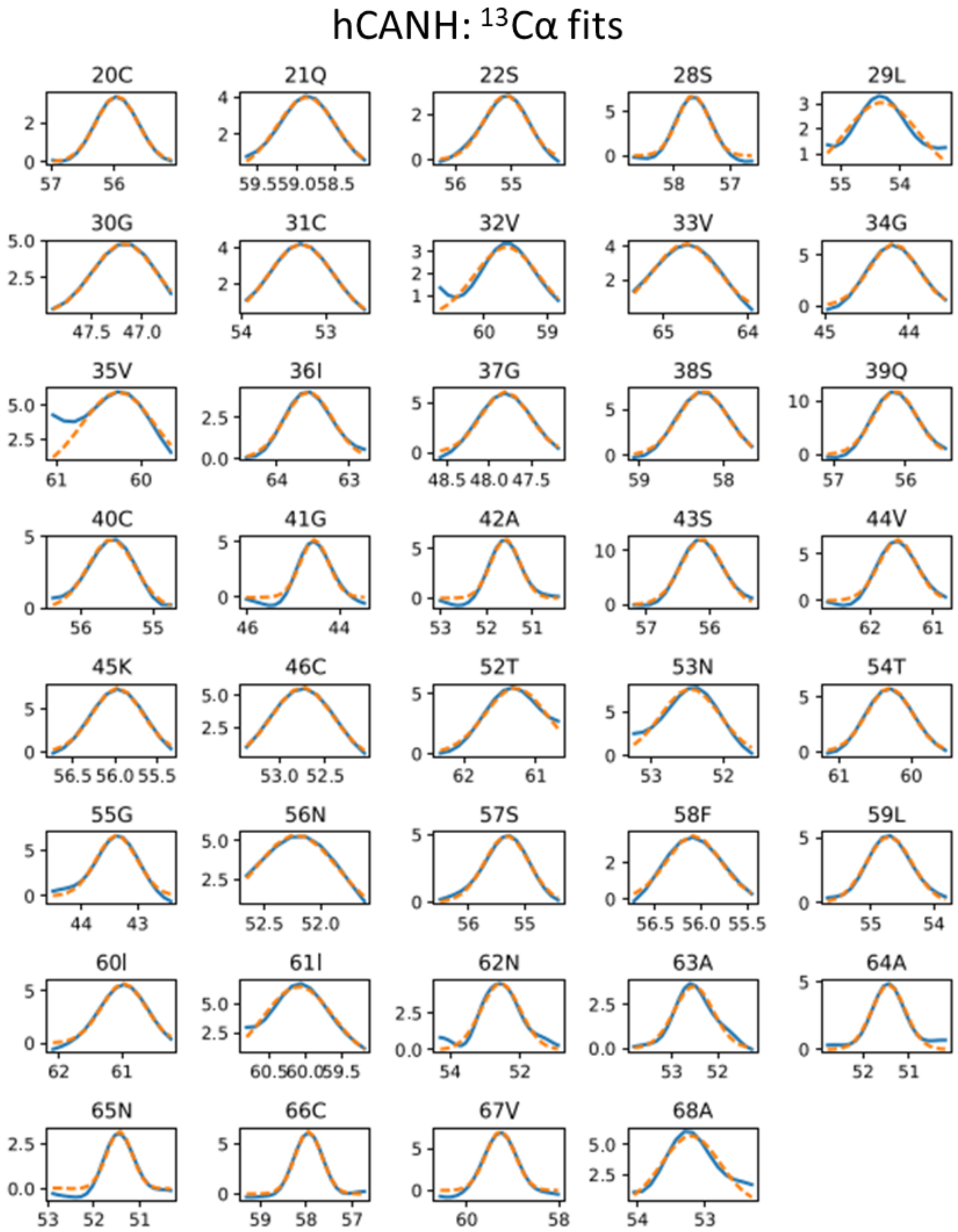


Figure S19. 1D projections onto the ^{13}C axis of the 3D hCANH peaks of the EAS$_{\Delta 15}$ rodlet sample. Blue lines: peaks; dashed orange lines: fits. Extracted linewidths are shown in *Results, Section 2.3.4*, Fig. 2.3.5A. The scale of the vertical axis is consistent throughout the whole set of fits

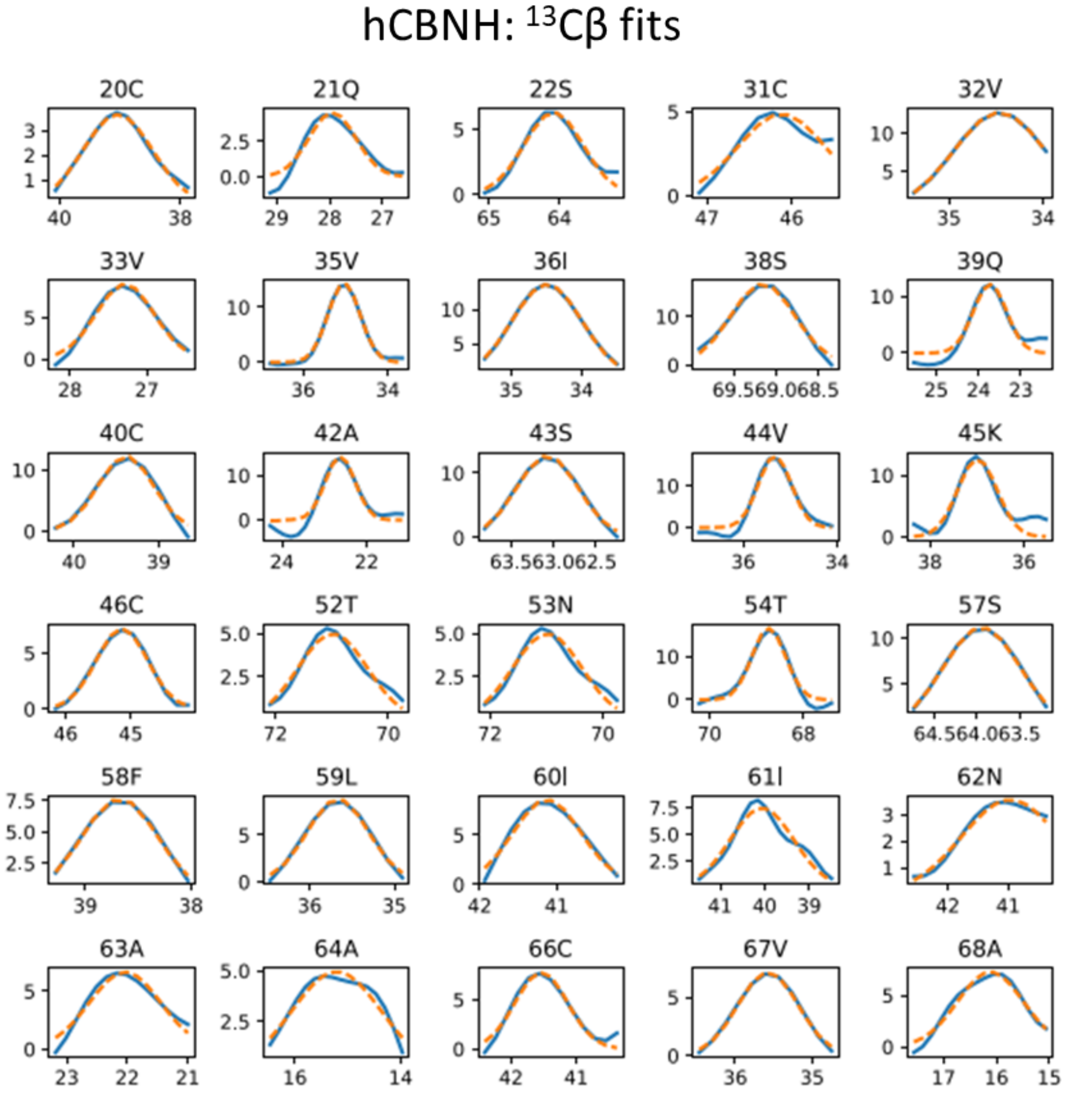


Figure S20.1D projections onto the ^{13}C axis of the 3D hCBNH peaks of the EAS$_{\Delta15}$ rodlet sample. Blue lines: peaks; dashed orange lines: fits. Extracted linewidths are shown in *Results, Section 2.3.4*, Fig. 2.3.5A. The scale of the vertical axis is consistent throughout the whole set of fits.

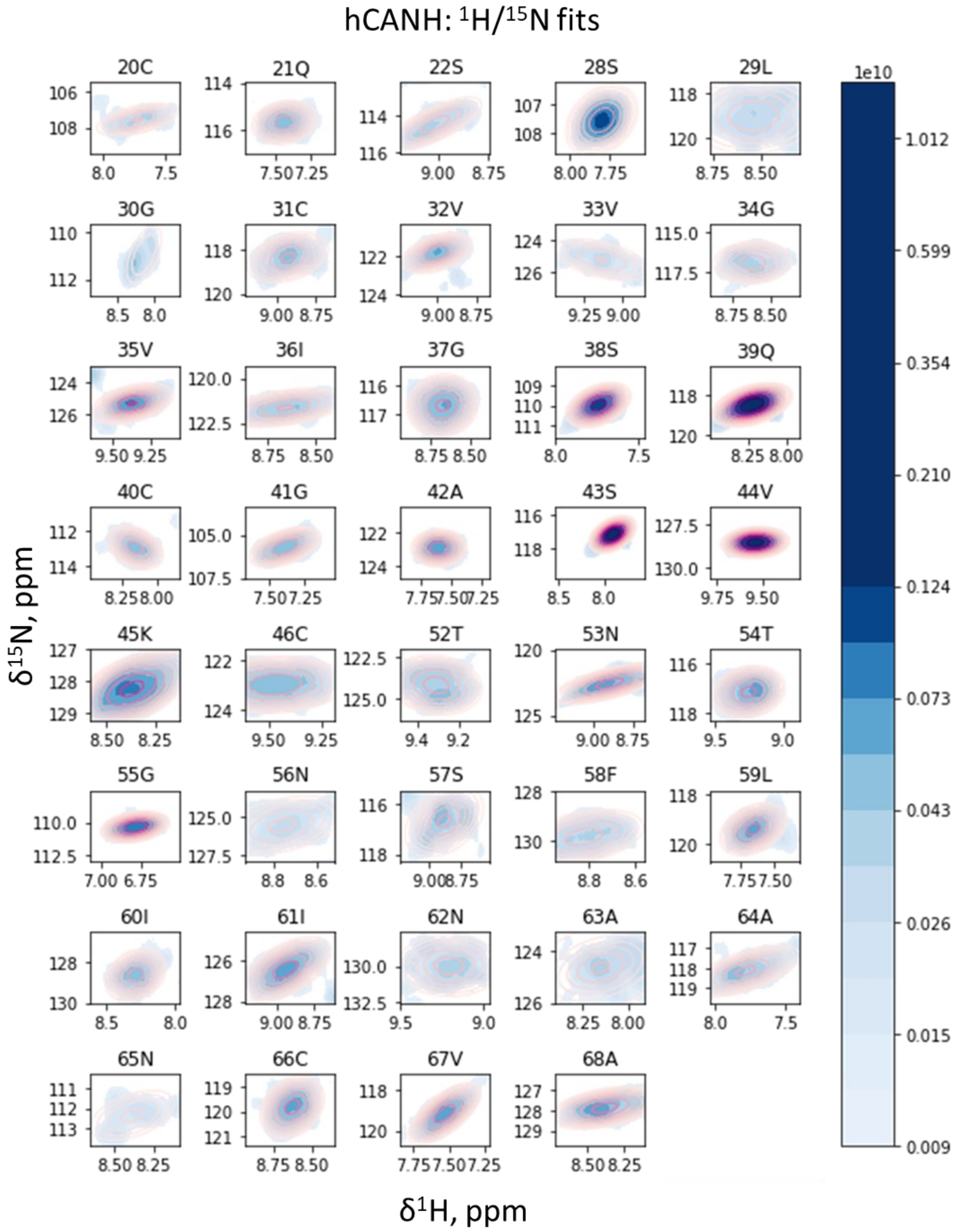


Figure S21. 2D projections onto the ^{13}C axis of the 3D hCANH peaks of the EAS$_{\Delta15}$ rodlet sample. Blue shades: peaks; magenta contours: fits. The peaks were fitted as described in *Section M.3.5*. The projections demonstrate different H/N slopes of the individual peaks (note the different scales of the vertical axis). The slopes (cross-correlation coefficients of the 2D Gaussian) are shown in *Results, Section 2.3.4*, Fig. 2.3.5D.

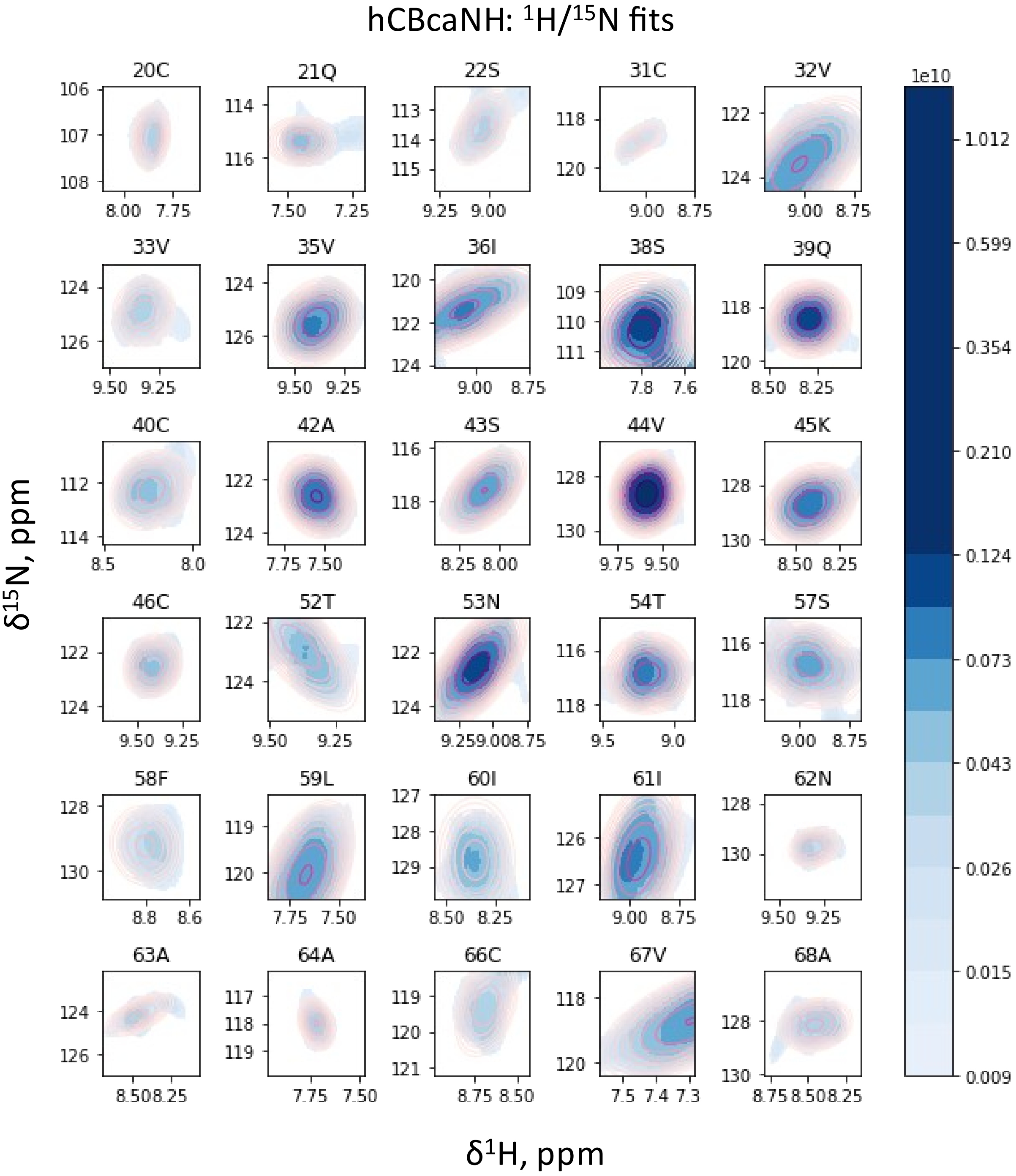


Figure S22. 2D projections onto the ^{13}C axis of the 3D hCANH peaks of the EAS$_{\Delta 15}$ rodlet sample. Blue shades: peaks; magenta contours: fits. The peaks were fit as described in *Section M.3.5*. The projections demonstrate different H/N slopes of the individual peaks (note the different scales of the vertical axis). The slopes (cross-correlation coefficients of the 2D Gaussian) are shown in *Results, Section 2.3.4*, Fig. 2.3.5D.

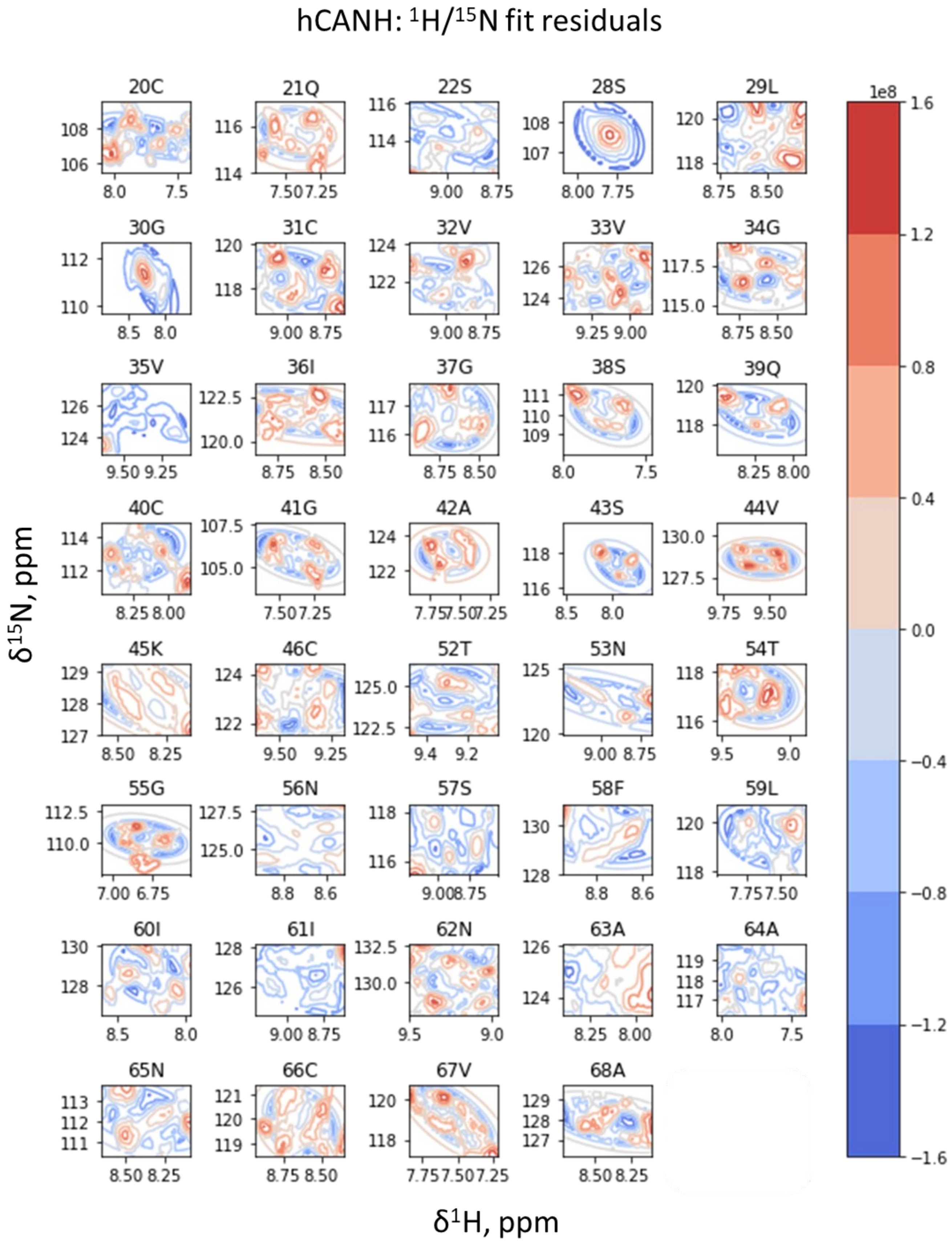


Figure S23. Residuals of the fits shown in Fig. S21.

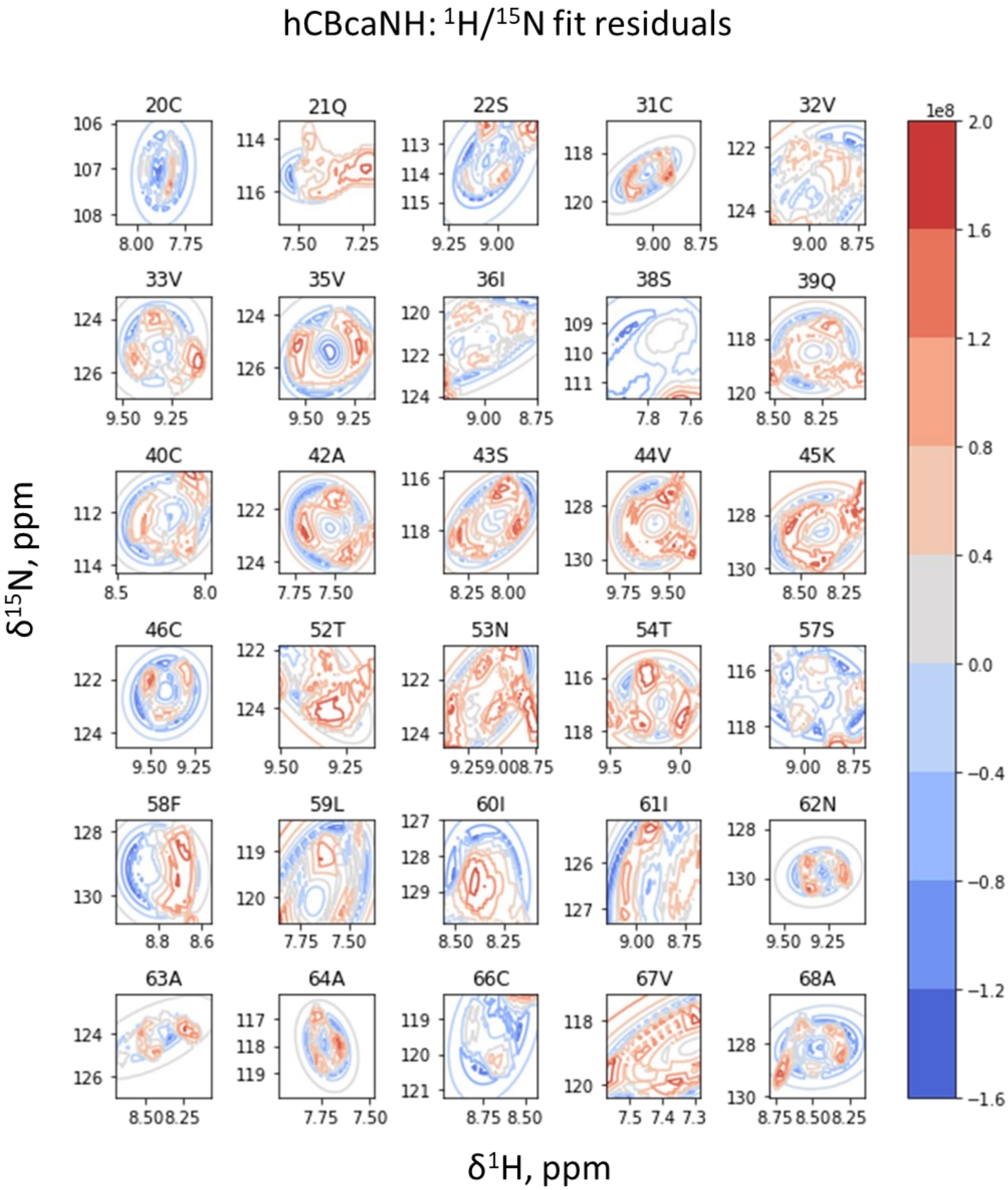


Figure S24. Residuals of the fits shown in Fig. S22.

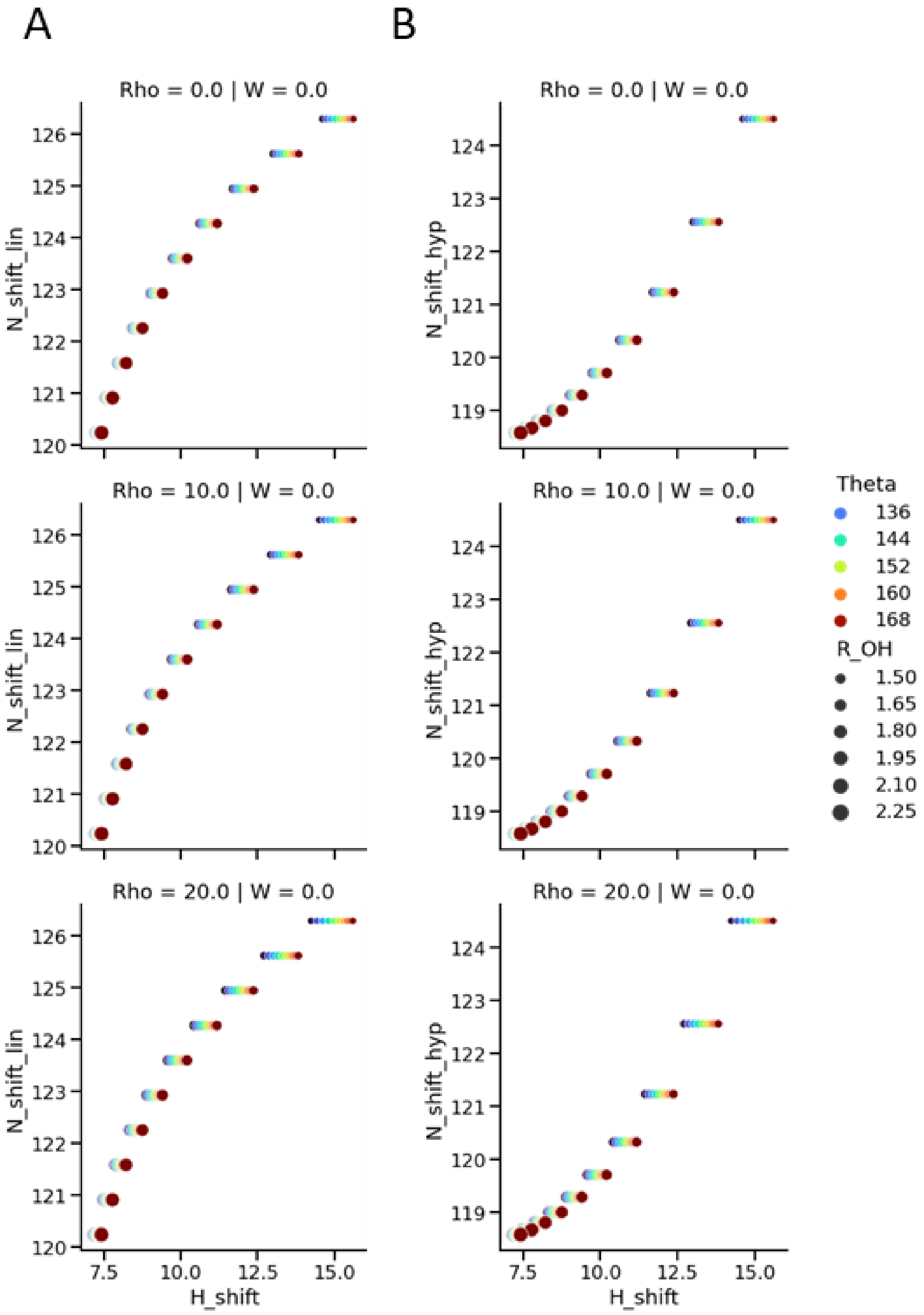


Figure S25. Proton and nitrogen chemical shifts as a function of H-bond length r_{OH} and angle θ. Proton chemical shifts are calculated with the model of Parker *et al.* (2006). Nitrogen shifts are calculated with **A**: linear models of Paramasivam *et al*. (2018), **B**: hyperbolic-exponential model of Xu and Case (2002). See parameter definition in main text, *Section 1.3.2,* Fig, 1.3.4B.

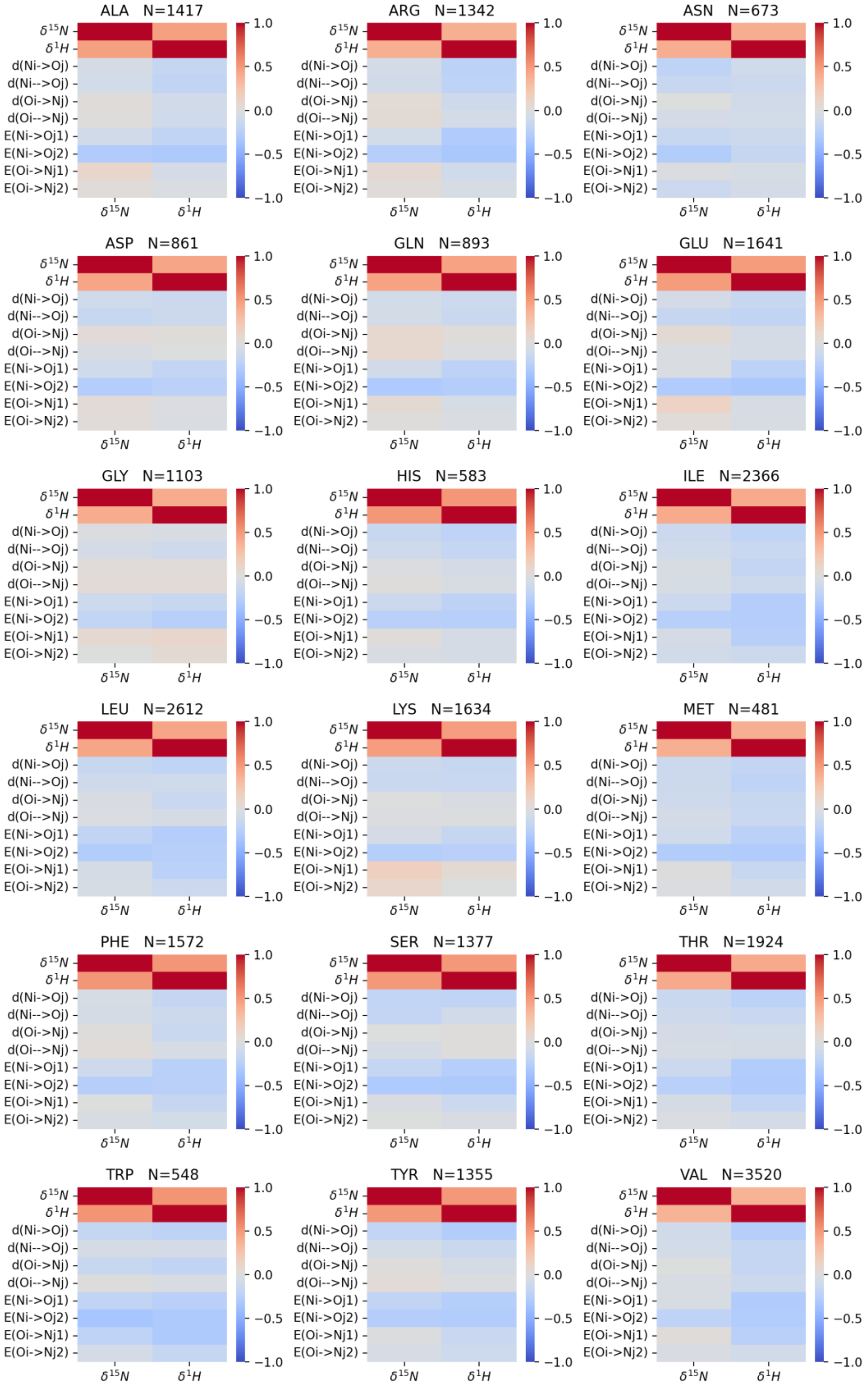


Figure S26. Pearson correlation coefficients for chemical shifts of the backbone amide protons and nitrogens belonging to extended structures ('E') and the hydrogen bond parameters estimated by DSSP (Kabsch and Sander, 1983): d, the lower (->) and the higher(-->) distances between the nitrogen of residue *i* and the oxygens of the bonding partners *j*; E, the lower (1) and the higher (2) H-bond energies where the residue *i* is a donor and an acceptor. N denotes the number of residues. The PDB and BMRB data of 3636 proteins are related by the tables from ReBoxitory (Maciejewski *et al.*, 2017). Compare with Fig. 2.3.6, *Section 2.3.4.*

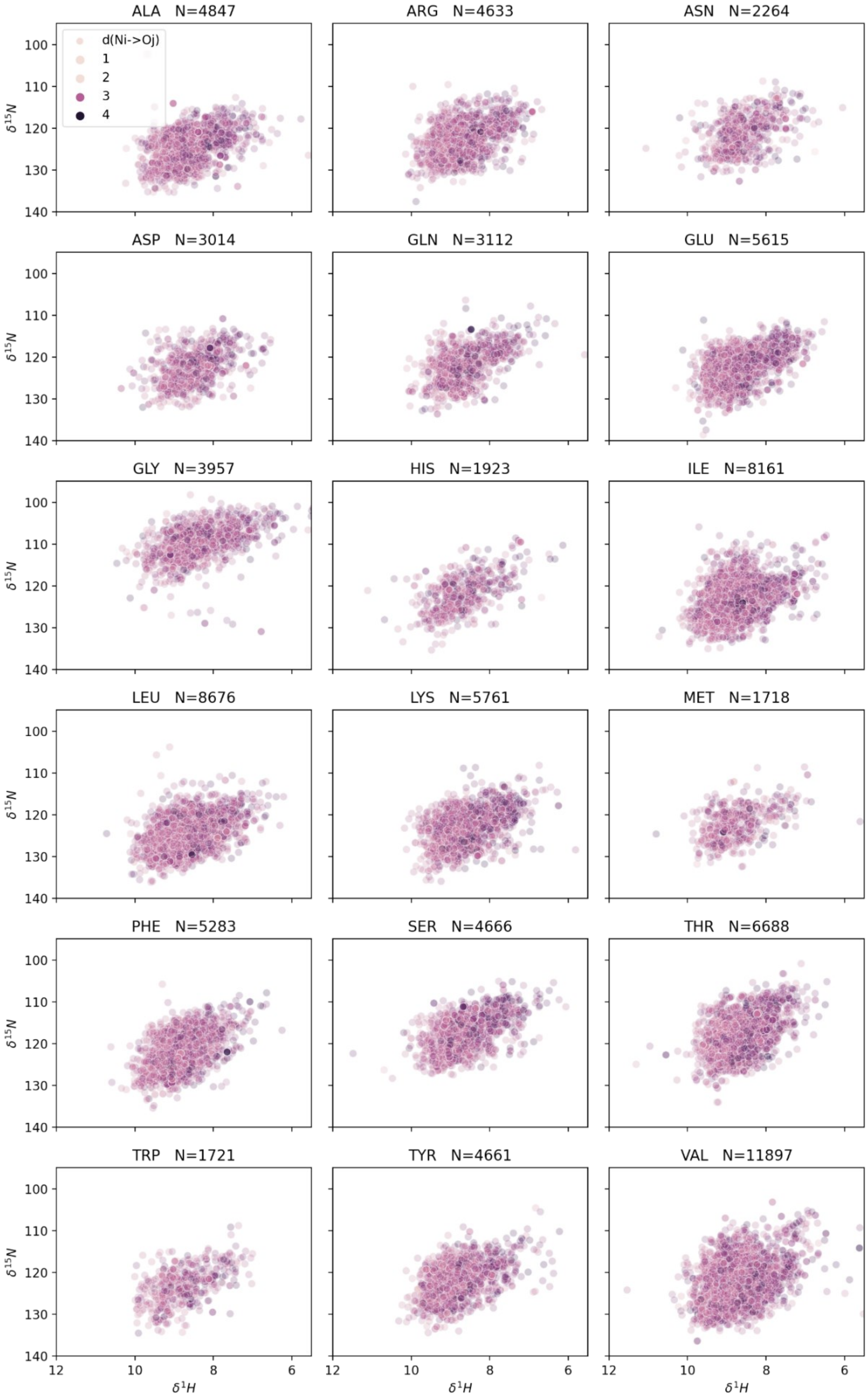


Figure S27. Correlations between the backbone amide chemical shifts and the minimal interatomic N-O distance (denoted by color) formed by a nitrogen of the donor residue *i* and an oxygen from the acceptor residue *j*, estimated by DSSP. N denotes the number of residues.

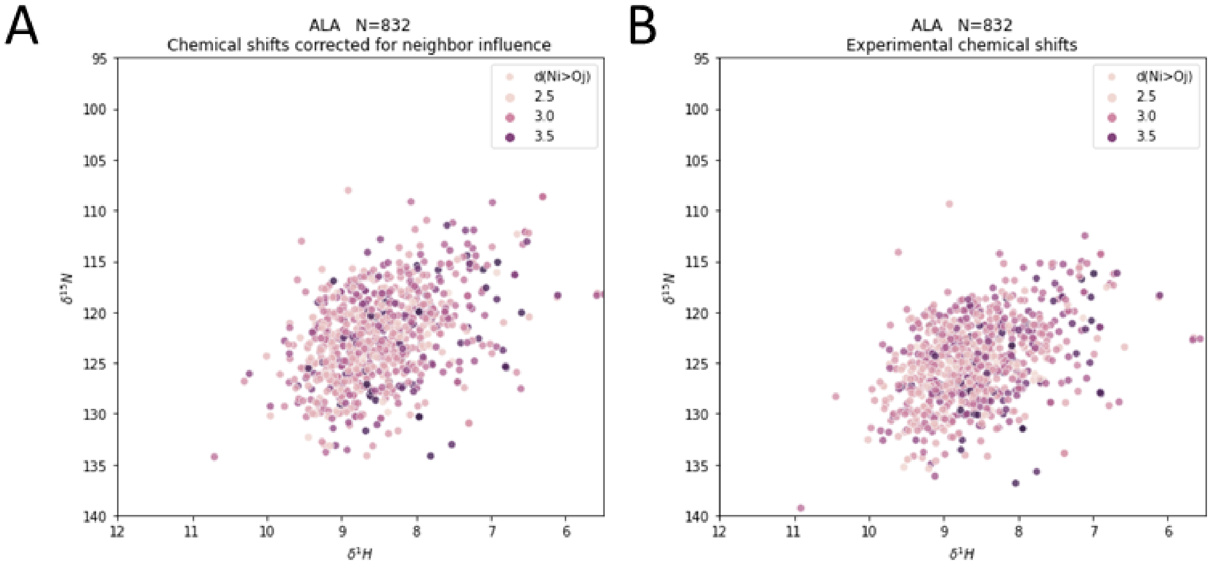


Figure S28. Backbone amide chemical shifts with (**A**) and without (**B**) correction for the neighbor influence and their relationship to the H-bond lengths. The correction increments were taken from Tamiola *et al.* (2010). The data for 838 alanines were taken from the 1707 pairs of PDB-NMRStar files, arbitrary selected from the entire set of all considered pairs (3636) to optimize the computational costs. Structural and NMR data from PDB and BMRB are related with the tables from ReBoxitory (Maciejewski *et al.*, 2017). The H-bond length was estimated by the interatomic distances between nitrogen and most proximate oxygen, identified with `Bio.PDB` module (Hamelryck and Manderick, 2003) of `BioPython` package.

Workshops and conferences

Workshops

2018 *Ampere Biological Solid-State NMR School 2018*, 21-26. Oct. 2018, Palma-de-Mallorca, Spain

2019 *Proteomics in epigenetics*, 12. Apr. 2019, BayBioMS, Munich, Germany

IRTG1309 Summer School, 15-19. Sep. 2019, Innsbruck, Austria

14th Biomolecular NMR: Protein Dynamics. 30 Sep. – 04. Oct. 2019 University of Gothenburg, Sweden

2020 *Academic writing workshop*, 04-05. Mar. 2020, TU Dortmund.

4th G-NMR school, 17-21. Feb. 2020, Göttingen, Germany

2021 15th *Biomolecular NMR: Advanced Tools.* 27. Sep. – 08. Oct. 2021 University of Gothenburg, Sweden

Poster contributions

2018 E. Burakova, E. Akorury, R. Linser. *"The effect of DNA modifications on the stability and mobility of nucleosomes"*, SFB1309 Nikolaus Symposium 2018, 30. Nov. 2018, LMU Munich, Munich, Germany

2019 E. Burakova, J. Dietschreit, R. Linser. *"Reconstitution of protein conformational ensemble from peak shapes in multidimensional solid-state NMR spectra"*. Gordon Research Conference "Computational Aspects of Biomolecular NMR", 09-14. June, 2019 Les Diablerets, Switzerland.

2021 E. Burakova. *"A statistics-driven NMR approach to site-specific analysis of static protein disorder"* Euromar 2021 (online) & 42nd FGMR 2021 (online)

2022 E. Burakova, S. K. Vasa, R. Linser. *"Characterisation of backbone conformational heterogeneity in solid-state protein samples by high-dimensional, proton-detected NMR spectroscopy"* Euromar 2022, Utrecht, The Netherlands.

E. Burakova, S. K. Vasa, R. Linser. *"Characterisation of static disorder of protein backbone by high-dimensional, proton-detected NMR spectroscopy"* 43rd FGMR, Karslruhe, Germany

Oral contributions

2020 *"A closer look at protein conformational disorder by means of solid-state NMR"*. Tag der Chemie 2020 at TU Dortmund, 07. Feb. 2020, Dortmund

2021 "*NMR for Epigenetics*" IRTG 1309 Science Meeting, 12. Feb. 2021 (online)

Eidesstattliche Versicherung (Affidavit)

Burakova, Ekaterina

Name, Vorname
(Surname, first name)

216628

Matrikel-Nr.
(Enrolment number)

Belehrung:

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden, § 63 Abs. 5 Hochschulgesetz NRW.

Die Abgabe einer falschen Versicherung an Eides statt ist strafbar.

Wer vorsätzlich eine falsche Versicherung an Eides statt abgibt, kann mit einer Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft werden, § 156 StGB. Die fahrlässige Abgabe einer falschen Versicherung an Eides statt kann mit einer Freiheitsstrafe bis zu einem Jahr oder Geldstrafe bestraft werden, § 161 StGB.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:

Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offence can be punished with a fine of up to EUR 50,000.00. The competent administrative authority for the pursuit and prosecution of offences of this type is the chancellor of the TU Dortmund University. In the case of multiple or other serious attempts at deception, the candidate can also be unenrolled, Section 63, paragraph 5 of the Universities Act of North Rhine-Westphalia.

The submission of a false affidavit is punishable.

Any person who intentionally submits a false affidavit can be punished with a prison sentence of up to three years or a fine, Section 156 of the Criminal Code. The negligent submission of a false affidavit can be punished with a prison sentence of up to one year or a fine, Section 161 of the Criminal Code.

I have taken note of the above official notification.

Dortmund, 12. July 2023

Ort, Datum
(Place, date)

[redacted]

Unterschrift
(Signature)

Titel der Dissertation:
(Title of the thesis):

Qualitative and quantitative characterization of protein backbone heterogeneity
by solid-state NMR spectroscopy

Ich versichere hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel selbstständig und ohne unzulässige fremde Hilfe angefertigt habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht.
Die Arbeit hat in gegenwärtiger oder in einer anderen Fassung weder der TU Dortmund noch einer anderen Hochschule im Zusammenhang mit einer staatlichen oder akademischen Prüfung vorgelegen.

I hereby swear that I have completed the present dissertation independently and without inadmissible external support. I have not used any sources or tools other than those indicated and have identified literal and analogous quotations.

The thesis in its current version or another version has not been presented to the TU Dortmund University or another university in connection with a state or academic examination.*

*Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the PhD thesis is the official and legally binding version.

Dortmund, 12. July 2023

Ort, Datum
(Place, date)

[redacted]

Unterschrift
(Signature)