Technische Universität München

Fakultät für Chemie

Lehrstuhl für Biomolekulare NMR-Spektroskopie

Development and Application of Novel NMR Methods for Biological Macromolecules

Christoph Hartlmüller

Vollständiger Abdruck der von der Fakultät für Chemie der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Bernd Reif

Prüfer der Dissertation:

1. Prof. Dr. Tobias Madl
2. Prof. Dr. Michael Sattler
3. Prof. Dr. Dierk Niessing

Die Dissertation wurde am 27.03.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Chemie am 07.09.2020 angenommen.

Never stop being curious.

Acknowledgement

I want to acknowledge the Technische Universität München and in particular the Department of Chemistry as well as the Helmholtz Zentrum München for providing an excellent environment throughout the course of the PhD thesis.

This thesis would not have been possible without the teaching, support, mentoring and trust of several people, and I want to say “Thank you” to all those people!

First of all, I want to thank Prof. Dr. Tobias Madl for funding and supervising my PhD thesis. The engagement and continuous support I received throughout the entire project was outstanding. You supported me by giving guidance and providing orientation. You trusted in me and gave me room to develop and grow towards my desired goals. Thank you, Tobias!

I also want to thank Dr. Christoph Göbl. Your amazing support had so many facets, ranging from knowledge-sharing in the wet lab, to deep and theoretical discussions about NMR, to being a role model in regards of scientific working. Thank you, Christoph, for your support and for being a great mentor.

I want to thank Dr. Sandra Scanu for the great support and awesome time we had! You not only shared your experience in the lab, but more importantly, were always open for discussion and gave advice about any topic. Thank you, Sandra!

I want to thank all Madls, for an amazing time with lots of great and unforgettable moments. It's been a great time and I am very glad to have worked together with such great colleagues and friends! Thank you, Tobias, Christoph, Sandra, Martin, Gesa, Ben, Hannah, Emil and Sarah!

I also want to thank Prof. Dr. Michael Sattler for hosting our group and providing an excellent place for studying and outstanding facilities for research. You not only shared your passion about science but also dedicated yourself to teaching and were always available for discussions and giving valuable advice. Thank you, Michael!

I want to thank Dr. Rainer Haeßner! With your dedication and motivation, you are a great role model to me! Thanks for all your support, your excellent teaching, your trust as well as making me think outside the box. Thank you, Rainer!

A huge thank you to the entire Sattler Group! We have always had a great time working together. Thanks for all the great events, as well as all the trips and conferences we organized and went to together!

I also want to thank Prof. Dr. Bernd Reif and Prof. Dr. Dierk Niessing. Thank you for great collaborations on several projects, sharing your lab as well as your support during my PhD thesis. Thank you, Bernd and Dierk!

Above all, I want to thank my family:

Karoline, Martin and Georg, thank you for always being there for each other, no matter what the situation requires. I am so thankful for all the small and big moments we have shared together and I am so looking forward for many more to come. Thank you!

Irene and Peter, I am full of thank that is hard to phrase into words. For sure, without you, this thesis and so many other things would not be possible. Your love and unconditional dedication as parents have been the foundations for my entire life. You have always given me the room and freedom to choose which way I want to go. At the same time, you have always inspired and motivated me to pursue my goals, and when needed, you are always here to support me with advice and guidance. Thank you so much!

Daniela, I am so thankful to have you in my life. You are always here for me and always believed in me. You have supported me in every project and step I took and helped me face any challenge. I know that I can always rely on you! With your warmth, heartiness and dedication you are an inspiration every single day. I am so glad for all the amazing moments we have shared and I am thankful to always have you by my side. Thank you, Dani!

Zusammenfassung

Die Kernspinresonanzspektroskopie (Nuclear magnetic resonance, NMR spectroscopy) ist ein wichtiges Instrument um die Struktur und Dynamiken von Biomolekülen auf einer atomaren Ebene zu enthüllen. Als solches hat diese Methode einen großen Beitrag auf dem Gebiet der Strukturbiologie geleistet, welche zum Ziel hat biologische Prozesse auf molekularer Ebene aufzuklären und zu verstehen. Aktuelle Trends in der Strukturbiologie zeigen ein wachsendes Interesse hin zu anspruchsvolleren Systemen wie zum Beispiel großen RNAs und Proteinen, intrinsisch ungeordneten Proteinen (IDP) sowie biomolekularen Komplexen. Weiterhin hat die herausragende Entwicklung der Computertechnologie die Basis für ein großes Arsenal an neuen computergestützten Rechenmethoden gelegt. Aktuelle Rechenmethoden verwenden genauere Modelle, erzielen Ergebnisse auch mit unvollständigen experimentellen Datenmengen und kombinieren die Daten von unterschiedlichen experimentellen Methoden. Das Ziel dieser Arbeit ist die Entwicklung und Etablierung von neuen NMR-basierten Ansätzen durch das Ausnutzen der neuesten Entwicklungen von experimentellen und computergestützten Methoden. Diese Arbeit fokussiert dabei insbesondere auf Daten zur Lösungsmittelzugänglichkeit, die mit Hilfe des sogenannten solvent paramagnetic relaxation enhancement (solvent PRE)-Effektes gemessen werden können. Durch die enthaltene Information über den Abstand zur Moleküloberfläche sowie durch experimentelle Vorteile sind solvent PRE-Daten vor allem für die strukturelle Analyse von großen und flexiblen Biomolekülen interessant. Diese Arbeit zeigt, dass das Verbessern der solvent PRE-Experimente und der dazugehörigen Datenanalyse sowie das Kombinieren der gemessenen Daten mit aktuellen Rechenmethoden, wie zum Beispiel mit dem Rosetta Framework, ein vielversprechender Ansatz für zukünftige Arbeiten auf dem Gebiet der Strukturbiologie ist.

Abstract

Nuclear magnetic resonance (NMR) spectroscopy is a powerful method to unveil structure and dynamics of biomolecules on an atomic level. As such, the method has significantly contributed to the field of structural biology that aims to understand biological processes on a molecular level. Current trends in structural biology show an increasing interest towards more challenging systems such as large RNAs and proteins, intrinsically disordered proteins (IDPs) as well as biomolecular complexes. Moreover, the tremendous development of computer technology has set the basis for a large arsenal of computational methods. State-of-the-art computational methods use more accurate models, obtain results with limited sets of experimental data and combine data sets of different experimental methods. The goal of this thesis is the development and establishment of new NMR-based approaches by exploiting these new advancements of experimental and computational methods. In particular, this thesis focuses on solvent accessibility data that can be obtained by recording solvent paramagnetic relaxation enhancement (solvent PRE) data. By encoding distance-to-surface information and providing experimental advantages, solvent PRE data are particularly interesting for structural characterization of large and flexible biomolecules. This thesis shows that improving solvent PRE experiments and the corresponding data analysis as well as combining the acquired data with state-of-the-art computational approaches, such as the Rosetta framework, provides a promising approach for future studies in the field of structural biology.

List of Abbreviations

CD	Circular dichroism
cryo-EM	Cryo-electron microscopy
CSA	Chemical shift anisotropies
CSP	Chemical shift perturbation
DNA	Desoxyribonucleic acid
DNP	Dynamic nuclear polarization
FFT	Fast Fourier transform
Flop	Floating point operations
Flops	Floating point operations per second
IDP	Intrinsically disordered protein
INEPT	Insensitive nuclei enhanced by polarization transfer
NMR	Nuclear magnetic resonance
NOE	Nuclear Overhauser effect
NOESY	Nuclear Overhauser enhancement spectroscopy
NUS	Non-uniform sampling
PCS	Pseudo contact shifts
PRE	Paramagnetic relaxation enhancement
QM/MM	Quantum mechanics/molecular mechanics
RDC	Residual dipolar coupling
rf	Radiofrequency
RNA	Ribonucleic acid
SAXS	Small angle scattering
sPRE, solvent PRE	Solvent paramagnetic relaxation enhancement
UV/Vis	Ultraviolet/visible

Table of Contents

I Introduction

I.1 The Field of Structural Biology

Structural biology is an important research focus that aims for a detailed understanding of biology by investigating the structure of biological macromolecules on an atomic level. Biological molecules of interest include proteins, desoxyribonucleic acids (DNAs), ribonucleic acids (RNAs), lipids and membranes as well as complexes thereof[1]. From a classical point of view, the significance of structural biology is justified by the fact that biological macromolecules adopt a specific three-dimensional conformation on an atomic level which in turn is essential for the function of the molecule. In the case of proteins, the three-dimensional atomic structure depends on the sequence of the amino acids, as postulated by the famous Anfinsen's dogma[2]. In the last decades, the range of target molecules was broadened, in particular, understanding the role of intrinsically disordered proteins (IDPs) became a new focus[3, 4]. In addition to studying an ever-increasing set of molecules, the scope of structural biology goes beyond static protein structures. A deep understanding of biological processes requires the studies of dynamic processes ranging from slow protein folding, domain movements and binding events to fast rotational and vibrational flexibility of specific protein sidechains. The combined knowledge about structure and dynamics is ultimately connected to biological functions enabling detailed insights into a variety of cellular processes and paving the way towards a comprehensive biochemical understanding of living cells and organisms, characterization of diseases and a rational structure-based drug design[5].

A major milestone in structural biology was the discovery of the double helix conformation of DNA in the 1950s by X-ray crystallography[6], followed by the first three-dimensional models of proteins obtained by X-ray crystallography in the late 1950s[7], electron microscopy in the 1970s[8, 9] and nuclear magnetic resonance (NMR) spectroscopy in the 1980s[10, 11]. Since these milestones, tremendous efforts were made to determine three-dimensional models of biological macromolecules and complexes thereof. These efforts are best reflected in the growth of the Protein Data Bank (PDB)[12], a combined database for structural biology which was established in the beginning of the 1970s and which today contains more than 110,000 entries.

I.1.1 Overview of Experimental Methods

The most commonly used experimental methods to obtain structural models of biological macromolecules are X-ray crystallography[13, 14], single-particle cryo-transmission electron microscopy[15] (often only referred to as cryo-electron microscopy or cryo-EM), NMR spectroscopy[16-18] and different combinations thereof. Today, the majority of structures found in the PDB are determined using X-ray crystallography (89.56 %), followed by NMR spectroscopy (9.45 %) and electron microscopy (0.99 %), as illustrated in Figure I-1.

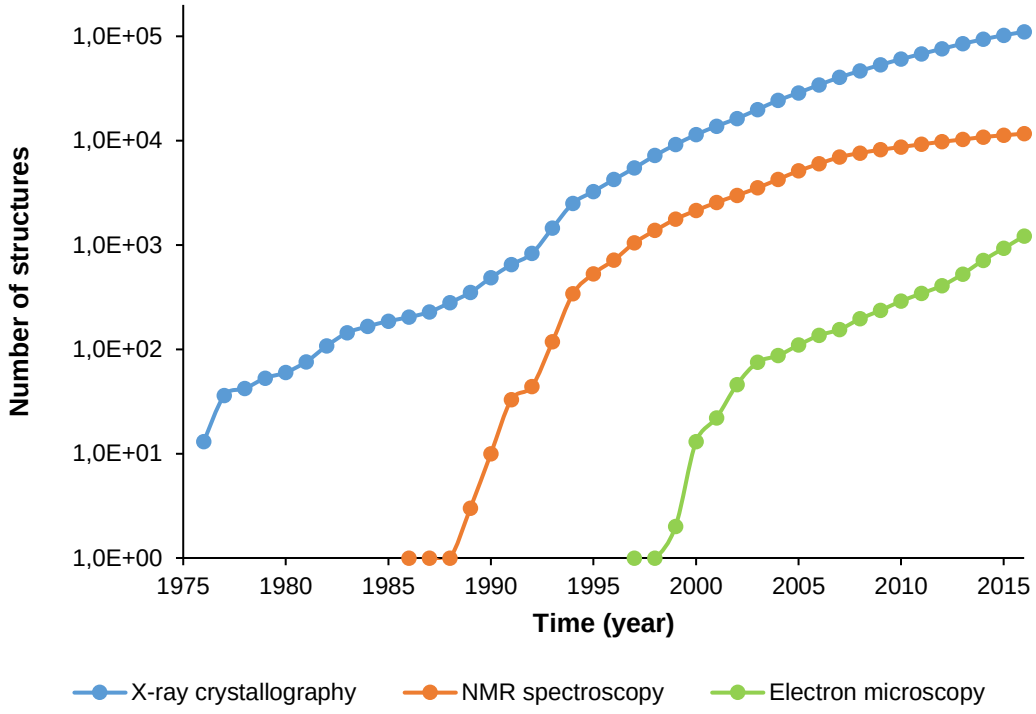


Figure I-1: Number of entries in the PDB by experimental method
Data was obtained from http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html (as of October 2016).

The method of X-ray crystallography[13, 19] can provide structural models with a high spatial resolution. By optimizing sample preparation as well as the procedures of measurement and analysis, X-ray crystallography studies became the driving force for structural biology allowing to unveil thousands of protein structures. However, as X-ray crystallography depends on the preparation of proteins in the crystalline state[13], the availability of appropriate samples is often the bottleneck of this methodology[20]. Although the crystallization of proteins was vastly improved during the last decades, a detailed understanding of the process of crystallization is still missing and finding the optimal crystallization conditions in most cases involves highly-optimized empirical trial-and-error

screenings[21]. The process of crystallization can give rise to artifacts by altering the structure as a consequence of crystal contacts, preparation protocols or changing the primary sequence of biomolecules. Moreover, flexible parts of the molecule or imperfections in the crystal alignment result in a loss of signal and therefore those regions are not observable by X-ray crystallography.

With the very recent development and availability of direct electron detectors, cryo-EM underwent a rapid development, often referred to as "resolution revolution" and today, the structure of large biomolecular complexes can be determined with nearly atomic resolution[15, 22, 23]. To honor these major advantages in technology, the current Nobel Prize in Chemistry was awarded to pioneer researchers in the field of cryo-EM. Structure determination by cryo-EM involves the collection of millions of two-dimensional images of single-particles obtained from a thin layer of quick-frozen solution containing the biomolecule of interest. The complex analysis pipeline includes a classification of the images and angular assignment of the images as well as an iterative three-dimensional reconstruction using the average image for every class of image[24, 25].

In contrast to X-ray crystallography, the crystallization of protein samples is not required in the case of cryo-EM and as a consequence both methods can be considered complementary with regards to sample preparation. However the basic concepts of both methods are very similar[26]. Data obtained from both methods are direct representations of the three-dimensional molecular structure which in turn can be computed by either solving the phase problem or the three-dimensional reconstruction. Moreover, flexibility and dynamics of the target molecule lead to missing electron density in the case of X-ray crystallography[13] and, depending on the quality of the image classification, to blurred-out or low-resolution regions of the molecule in the case of cryo-EM[27].

Compared to X-ray crystallography and cryo-EM, biomolecular NMR spectroscopy is based on a completely different physical concept[16-18]. Instead of recording a direct representation of the global three-dimensional structure of the biomolecule, NMR spectroscopy allows to extract site-specific molecular properties including geometric properties such as distance, angles and orientations, physical properties such as magnetic shielding as well as dynamic properties such as amplitudes and timescales of different molecular motions. NMR spectroscopy is based on measuring the time evolution of the magnetic properties of the

sample. To this end, a precisely-orchestrated sequence of short radio frequency (rf) pulses are required to excite the system. These excitation schemes are often referred to as pulse program and are crucial for the success of the experiment. The nature of the pulse program determines which molecular property is encoded in the detected signal. By carefully selecting and optimizing the NMR pulse program, the obtained molecular parameter is selected and thus the experiment can be tailored to answer a specific question. This approach allows to create a large set of highly-optimized and purpose-bound pulse programs that render NMR spectroscopy a powerful method which is complementary to X-ray crystallography and cryo-EM (see chapter I.2 for a more detailed introduction of biomolecular NMR spectroscopy).

The life time of excited states created in any type of spectroscopy is limited. In the case of biomolecular NMR spectroscopy, the finite lifetimes lead to a limitation of the molecular size of the target molecule. While classical NMR approaches are limited to molecular sizes of 20-30 kDa, advances in the field of biomolecular NMR spectroscopy pushed the molecular size limit well above 100 kDa [28-32]. Despite its intrinsic molecular weight limitations, NMR spectroscopy is a key technique in the field of structural biology for several reasons. It allows to unveil conformations and dynamics of structured biomolecules as well as to probe parameters of highly-flexible targets, such as RNAs or intrinsically disordered proteins (IDPs). In addition, measurements are performed using samples in solution, which closely resembles physiological conditions and greatly simplifies sample preparation. Considering the example of RNAs, which are often highly flexible and more challenging to crystallize[33], about 40 % of all structures found in the PDB were solved by NMR spectroscopy compared to 8.9 % in the case of proteins[1].

In the last decades, the field of structural biology underwent a rapid development and resulted in the determination of more than 100,000 structures of biological macromolecules. With this tremendous amount of knowledge, research in the field of structural biology shifted towards studying more complicated and challenging systems, such as large biomolecular complexes or flexible and dynamic targets of interest[3]. This development is accompanied and greatly accelerated by the increasing importance of hybrid methods. In these hybrid approaches, two

[1] Numbers as of October 2016 and according to http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html.

or more complementary methods are combined to address a biological question that could not be answered by a single method only. As an example, structural insights can be obtained using X-ray crystallography while information about dynamics or binding behavior can be obtained by NMR spectroscopy. In the last years, this concept of integrative structural biology was greatly extended to combine a variety of methods, including small angle X-ray and neutron scattering, mass spectroscopy or even biochemical methods[34-39].

The development of the techniques introduced in this chapter is closely linked to the development of computational methods. This linkage becomes even more important in the case of hybrid methods or integrative structural biology. The following chapter describes the impact and common challenges of computational methods.

I.1.2 The Role of Computational Methods

In its early beginnings, the field of structural biology has been driven by the advances of experimental methods. With the tremendous development of computer technology starting in the second half of the 20th century, computational methods began to play an important role in the development of structural biology. The exponential increase of available transistors in computing units was first described in the well-known Moore's law[40] in 1965 and is still valid 50 years later (see Figure I.2). The impact of computer technology on NMR spectroscopy was impressively demonstrated when the Fast Fourier Transform (FFT) algorithm was developed and computational resources became broadly available. This development set the basis of Fourier Transform-based NMR (FT-NMR) and the development of two-dimensional NMR spectroscopy[41]. Moreover, the computational resources available today enabled the wide applicability of non-uniform sampling techniques and paved the way towards time-efficient high-dimensional NMR spectroscopy[42, 43]. And for a third example, high-performance computing and the development of efficient algorithms are the essential basis of the reconstruction pipeline required for high resolution cryo-EM[15, 22, 23]. These examples show that the advances of computer technology not only improved the arsenal of methods in the field of structural biology but also allowed the creation of new research focus areas such as multidimensional NMR spectroscopy.

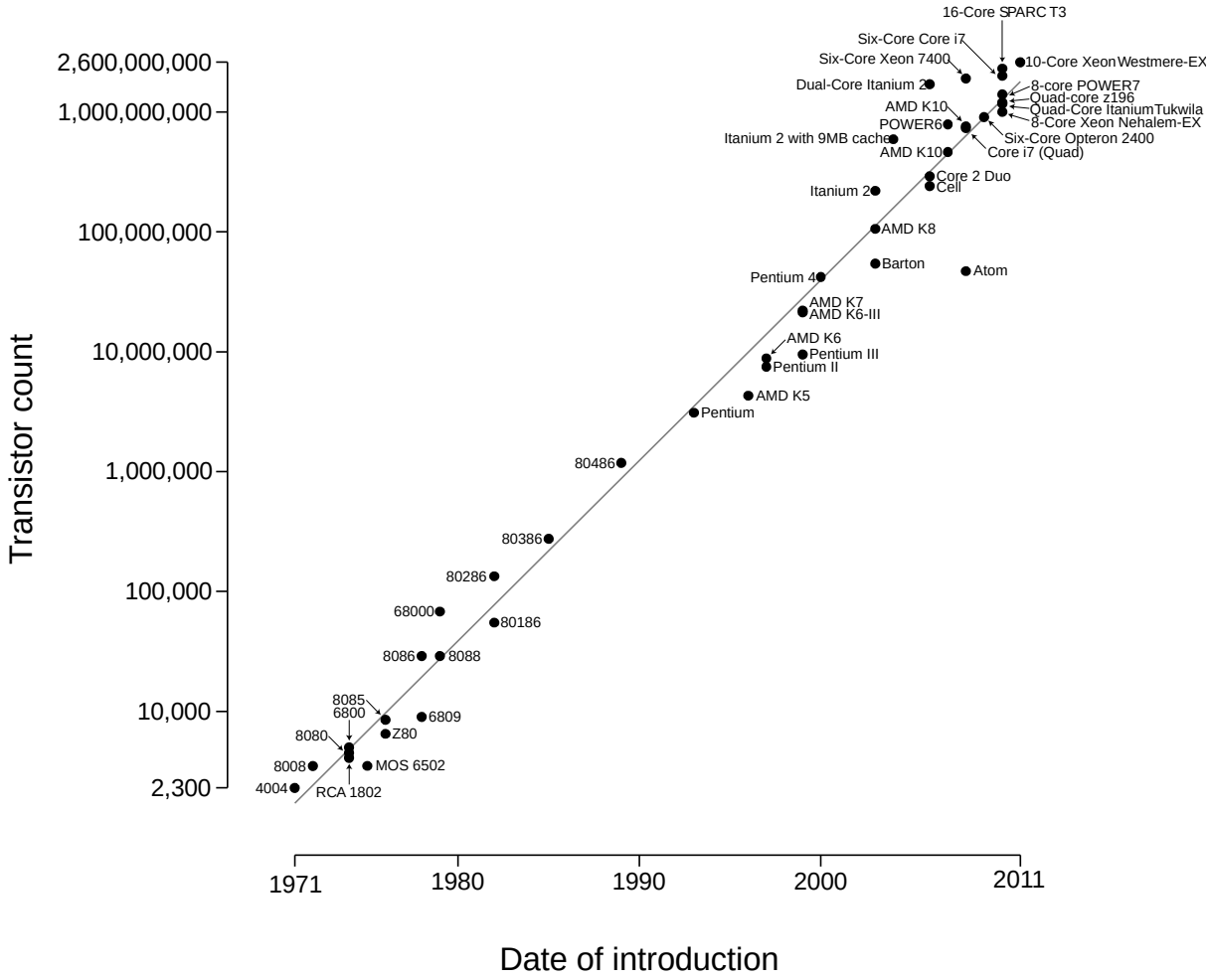


Figure I-2: Increase of the number of transistors during the last decades
The increase in the number of transistors over the last decades is shown on a logarithmic scale. The properties of several common processors introduced between 1971 and 2011 are shown as dots. The solid line corresponds to a doubling of the transistor count every two years. Figure was obtained from https://en.wikipedia.org/wiki/File:Transistor_Count_and_Moore%27s_Law_-_2011.svg (Author Wgsimon, Creative Commons Attribution-Share Alike 3.0 Unported license)

The previous examples clearly demonstrate how the development of computer technology drives the development of experimental methods. Beyond improving existing experimental methods, the increase of computational power sets the cornerstones of a complementary approach in the field of structural biology which solely relies on computational methods. This approach is often referred to as computational structure biology and aims to understand biomolecules based purely on the results of computations and simulations of the target molecule. After studying the first protein in the late 1970s[44], these computational studies and simulations of biological macromolecules have seen a tremendous development[45-51] with the 2013 Nobel Prize in Chemistry underlining the importance of molecular dynamics simulations for the field of structural biology. Computational methods open up unique possibilities to study biological systems and processes in an entirety that is not or only in a

limited extend accessible by experiments. Experimental methods such as NMR spectroscopy allow to investigate dynamic processes including protein folding as well as motions of sidechains or entire domains of the protein. While experimental studies are typically limited to specific sites or motions, computational methods provide detailed insights of the entire molecule on an atomic level[52-58]. Moreover, computational screening methods complement experimental high-throughput approaches providing powerful tools for efficient protein design and drug discovery pipelines[59-62].

It should be noted that although computational methods have extended the arsenal of methods in structural biology, the development of most computational methods required a considerable set of experimental data. Since the computational efforts of pure quantum mechanics-based simulations scale tremendously with the number of electrons of the target molecule[63, 64], the simulations of biological systems require simplified models such as force fields or hybrid quantum mechanics/molecular mechanics (QM/MM) methods[65, 66]. The simplicity of force fields comes at the costs of determining a large set of parameters[63, 64] that greatly influences the outcome and accuracy of the simulation, emphasizing the dependency of computational methods on experimental data[67-71]. In the last years, the combination of computational and experimental methods, often referred to as integrative modeling, was shown to be a powerful approach: Experimental data from different sources including X-ray crystallography, NMR spectroscopy or biochemical methods are used to set up, drive and confirm molecular simulations and at the same time computational methods unveil structure and dynamics of biological process at an atomic level[72-76].

The previous chapters gave a brief overview of the methods involved in structural biology and showed that the development of experimental and computational methods is closely connected. Computational methods enabled significant advances of experimental methods, and on the other hand, experimental methods provided essential parameters for biomolecular simulations.

I.2 NMR Spectroscopy

NMR Spectroscopy is a key technology in structural biology. It opens up unique possibilities as it allows to probe structure and dynamics of biomolecules on an atomic level in solution and thus under near-native conditions. In this chapter, the general concept of NMR

spectroscopy is briefly introduced. In the later parts of this chapter, the effect of paramagnetic relaxation and its usage for structure determination is briefly outlined. For a more detailed introduction into NMR spectroscopy the reader is referred to reviews [77-81] and textbooks[16-18]. It should also be noted that this chapter focuses on biomolecular NMR spectroscopy in solution. Therefore, topics specific to NMR on small molecules or solid-state NMR spectroscopy are not covered in this work.

I.2.1 The Concept of NMR Spectroscopy

The very basic concept of NMR spectroscopy relies on taking advantage of the nuclear spin. In quantum mechanics, the spin is a fundamental property of every elementary particle, such as protons, neutrons and electrons. Similar to assigning a certain mass or a charge to a given particle, the spin of a particle depends on the type of the particular particle. In contrast to the mass of an elementary particle, the spin is quantized and as such only takes discrete values. It is quantized by the spin quantum number $I = {}^{n}/_{2}$ where n is a positive integer number. For the purpose of NMR spectroscopy, the spin of the nuclei of the atoms is essential. In particular, nuclei with a spin $I = {}^{1}/_{2}$ are most relevant since nuclei with $I = 0$ are not detectable and describing as well as measuring nuclei with spins larger $I = {}^{1}/_{2}$ becomes tedious. Considering the comparatively limited chemical composition of proteins, the most relevant nuclear spins in the context of biomolecular NMR spectroscopy are those of protons ^{1}H, carbon ^{13}C and nitrogen ^{15}N. Although the spin of a single electron is significantly larger than the spin of a nucleus, the electron spin can be neglected in the case of NMR spectroscopy since electrons in most biological molecules form pairs which in turn averages the overall electron spin to zero. Nevertheless, in the special case of an unpaired electron, the spin of the electron strongly influences the observed nuclear spins.

A non-zero spin quantum number of a nucleus gives rise to a nuclear magnetic moment μ which is a quantized vector quantity. Such a nuclear magnetic moment interacts with an external magnetic field as it affects the energy of the nucleus. This energy E is proportional to the negative scalar product of the vector of the field strength of the external magnetic field $\vec{B}$ and the vector of nuclear magnetic moment $\vec{\mu}$:

$$E = -\vec{B}\vec{\mu} \tag{1}$$

Like the spin quantum number, the energy of a single particle is quantized to N distinct levels which are described by the magnetic quantum number $m \in [-I, -I+1, \dots, I-1, I]$ and $N = 2 \cdot I + 1$. Consequently, only two energy levels exist in the case of a spin-½ nucleus, which can be pictured as the nuclear spin being aligned either in the same or opposite direction compared to the external field. Those states are often symbolized by the wave functions $|\alpha\rangle$ and $|\beta\rangle$ corresponding to $m = -½$ and $m = +½$, respectively. In a simplified model, spin-½ nuclei are represented as small bar magnets where the lower energetic state $|\alpha\rangle$ corresponds to the bar magnet being aligned with the external field and the high energy state $|\beta\rangle$ corresponding to the bar magnet being rotated against the direction of the external field.

As mentioned in equation (1), the nuclear magnetic moment $\vec{\mu}$ interacts with an external magnetic field $\vec{B}$. Therefore, the presence of an external magnetic field affects the energy of the system and introduces two distinct energy levels for spin-½ nuclei. To compute these energies, the coordinate system is aligned to the z-axis of the magnetic field $\vec{B}$ such that the external magnetic field can be described by a scalar magnetic field strength B_z. The z-component of the nuclear magnetic moment $\vec{\mu}$ is given by

$$\mu_z = \gamma \frac{h}{2\pi} m = \gamma \hbar m \tag{2}$$

where h is the Planck constant, $\hbar$ is the reduced Planck constant $\frac{h}{2\pi}$ and γ is the gyromagnetic ratio, a constant that depends on the type of nucleus. The energies of both states of a spin-½ nuclei are then given by

$$E_{|\alpha\rangle} = \gamma B_z \hbar \cdot m_{|\alpha\rangle} = -\frac{1}{2} \cdot \gamma B_z \hbar$$

$$E_{|\beta\rangle} = +\frac{1}{2} \cdot \gamma B_z \hbar \tag{3}$$

As in the case of any type of spectroscopy, the presence of energy levels allows the excitation of the system from the low to the high energy level. For such a transition to occur, the energy of the electromagnetic radiation has to match the energy gap between the two states. Assuming an excitation from the $|\alpha\rangle$ to the $|\beta\rangle$ state of a spin-½ nucleus, the energy per photon E_{rf} has to match $E_{\mathrm{rf}} = \Delta E = E_{|\beta\rangle} - E_{|\alpha\rangle} = \gamma B_z \hbar$. With the angular frequency of a photon ω_{rf} given by $\omega_{\mathrm{rf}} = \frac{E_{\mathrm{rf}}}{\hbar}$, the resonating frequency for the excitation reduces to

$$\omega_{\mathrm{rf}} = \gamma B_z \tag{4}$$

This angular frequency is often referred to as the Larmor frequency ω_0. Assuming the field strength of a typical NMR spectrometer, the Larmor frequency of a proton is in the order of several hundred MHz and falls into the frequency band of radio frequency (rf).

In many types of spectroscopy, such as UV/Vis absorption spectroscopy, the sample of interest is not directly observable and therefore changes in the intensity or the frequency of the radiation interacting with the sample are measured. In the case of NMR spectroscopy, the time development of the nuclear magnetic moment is measured directly by detecting the induced voltage upon changes of the magnetic moment. This concept gives rise to the following simplified setup of an NMR experiment. The sample is positioned in a strong external magnetic field, where the field strength is often referred to as B_0 and the direction of the field is defined to be the z-Axis. Coils are positioned next to the sample to introduce excitation pulses as well as to record the change of the magnetic moment in the x/y plane. A NMR experiment is then performed by applying an excitation scheme which is followed by a detection period in which the magnetization of the sample is recorded as a function of time. Compared to other types of spectroscopy, the excited states of nuclear spins have long life times, ranging from milliseconds to seconds. This time frame allows to apply complex excitation schemes, so-called pulse sequences or pulse programs. Fast and precise adjustments of the involved pulses allow to accurately control which nuclei are excited as well as the strength of excitation. Pulse programs offer an extreme variety of tools, such as applying different pulses simultaneously, using time-dependent pulse shapes, including so-called delays during which no pulses are active as well as modifying the external magnetic field which in turn shifts all energy levels during a single NMR experiment. Exploiting these tools, a tremendous set of pulse sequences can be designed which in turn allows to tailor the outcome of the experiment such that the detected signal contains the desired information. At the end of the NMR experiment, the oscillating magnetization in the x/z plane is recorded and Fourier-transformed using Fast Fourier transform to unveil the desired NMR spectrum which contains all frequencies and their respective amplitudes. It should be noted that the interpretation of a NMR spectrum requires a complete knowledge and understanding of the applied pulse program.

It should also be noted that in equilibrium the energy levels of an ensemble of nuclei are populated according to the Boltzmann distribution. Since the magnetic interaction energy between the external field and a nucleus is relatively small compared to the thermal energy, the energy levels are nearly equally populated. Since the detected signal in an NMR experiment is generated by a tremendous number of protons, the bulk signal results in a small but detectable net magnetization. As an example, a typical signal of protons is reduced by a factor of about 10^5 compared to the theoretical state of all protons being in the lower energy state (assuming a 11.7 T magnetic field; see [16] for a detailed derivation). This reduced signal is typically compensated by using concentrated samples (concentrations in the upper μM to mM range and volumes between 0.5 and 1.5 ml) and averaging of repetitive measurements. Another approach to boost the signal to noise ratio involves dynamic nuclear polarization (DNP)[82] and can be applied to solid state NMR. In this approach, molecules with unpaired electrons are introduced and excited using a strong external microwave generator, the so-called gyrotron. The significantly larger gyromagnetic ratio of electrons gives rise to a strongly increased initial magnetization which is subsequently transferred to the nuclei of the molecule of interest. Although this technique requires low cryo temperatures, it is subject of current research as it allows to enhance the magnetization by up to two orders of magnitude[83].

To make use of the full potential of NMR spectroscopy, the physics of nuclear spin have to be understood and described mathematically. In particular the detected time evolution of the nuclear magnetic moment has to be described for a given excitation. To this end, the theoretical description of NMR spectroscopy is split into two types of considerations. First, quantum mechanics are used to derive a model that describes the state of the magnetization during the course of an NMR experiment. Eventually, this model can be simplified to so-called Product Operator formalism that can be applied in a straight-forward manner. The second part of the theoretical treatment considers the intensity of the NMR signal and is based on the concept of relaxation. Although both theoretical aspects are connected, product operators and the concept of relaxation are, to a large extent, applied independently of one another. The following chapters give a brief introduction into both concepts.

I.2.1 Product Operator Formalism

Performing and analyzing an NMR experiment requires a complete understanding of the time evolution of the magnetization, starting from the initial equilibrium state until the end of detection. To this end, a quantum mechanical analysis of the nuclear spin is required. As for the previous chapter, the reader is referred to textbooks[16-18] for derivations and a more detailed background.

For an isolated spin-½ system, such as a ^{1}H proton, the nuclear spin has two energy levels corresponding to its two states. These states can be represented using two wave functions, $|\alpha\rangle$ and $|\beta\rangle$. The actual state of a given system is a superposition of those two states with two complex coefficients describing the contribution of each wave function. For example, the wave $\psi(t)$ function of an isolated spin-½ system at time t can be written as

$$\psi(t) = c_\alpha(t) \cdot |\alpha\rangle + c_\beta(t) \cdot |\beta\rangle \tag{5}$$

where c_α and c_β are the coefficients of the wave functions of the two states $|\alpha\rangle$ and $|\beta\rangle$. The squares of the coefficients describe the probability to find the system in the corresponding state and, more importantly, these coefficients determine the expectation value of a given measurement. In quantum mechanics, a measurement corresponds to applying a mathematical operator on the wave function $\psi(t)$ and then computing the scalar product to obtain the expected value of the measured quantity. Assuming a system with n states, these operators can be represented by square $n \times n$ Hermitian matrices which guarantees to obtain real expectation values due to the real eigenvalues of the matrices. For an isolated spin-½ system, these operators can be written as square 2×2 matrices. In the case of a NMR experiment, the x/y magnetization is measured and the corresponding operator can be formulated using its representative matrix. So therefore, to determine the time evolution of the magnetization, the time evolution of the complex coefficients of both states needs to be determined. Note that no exact mathematical formulation of the wave functions themselves is needed, as long as the result of applying an operator to the wave function is known.

Knowing the coefficients of the different states at any point of time is sufficient to fully describe the current state of the magnetization and thus to understand and interpret the result of the NMR experiment. For an efficient computation, all relevant coefficients of the system are typically combined in a so-called density matrix. The density matrix is the representation of

the projection operator and, for a system with n states, is a squared $n \times n$ matrix where the matrix element $\sigma_{ij} = \overline{c_\iota^* \cdot c_J}$ is the product of the coefficients of state i and state j ($*$ referres to the complex conjugate and $\bar{}$ represents the ensemble average of the bulk in solution). For example, the 2×2 density matrix of an isolated spin-½ system at time t can be written as

$$\sigma(t) = \begin{pmatrix} \overline{c_\alpha(t)} \cdot \overline{c_\alpha(t)^*} & \overline{c_\alpha(t)} \cdot \overline{c_\beta(t)^*} \\ \overline{c_\beta(t)} \cdot \overline{c_\alpha(t)^*} & \overline{c_\beta(t)} \cdot \overline{c_\beta(t)^*} \end{pmatrix} = \begin{pmatrix} \overline{|c_\alpha(t)|^2} & \overline{c_\alpha(t)} \cdot \overline{c_\beta(t)^*} \\ \overline{c_\beta(t)} \cdot \overline{c_\alpha(t)^*} & \overline{|c_\beta(t)|^2} \end{pmatrix} \tag{6}$$

The usage of the density matrix is motivated by two favorable properties of the matrix. First, multiplying the density matrix with the operator representing the measurement and then computing the trace of the product yields the expected measurement. And secondly, the evolution of the density matrix can be described using distinct rules. Both aspects are discussed in the following.

As shown above, the diagonal of the density matrix $\sigma_{ii} = \overline{c_\iota^* \cdot c_\iota}$ contains the probability for every state i. On the other hand, the off-diagonal elements $\sigma_{ij} = \overline{c_\iota^* \cdot c_J}$ correspond to interactions between the two states i and j (also referred to as coherences or correlations). The complex coefficient c_i of a state consists of a real and imaginary contribution which can be converted into an amplitude and a phase. Coherence describes a correlation between the phases of two states. More precisely, a coherence is present when the distribution of the phases across all molecules in solution does not average to zero. In the simple case of an isolated spin-½, the 2×2 density matrix contains two diagonal elements representing the population of the two states and two off-diagonal elements representing the coherence between the two states. Since these coherences connect two states that require to flip a single spin to transform one state to the other, these coherences are referred to as single quantum coherences. These single quantum coherences can be pictured as the magnetization being aligned on the x- or y-axis.

With this understanding of the density matrix, the initial density matrix of the equilibrium can be constructed. Since no coherences are presented in equilibrium, all off-diagonal elements are set to zero. The diagonal elements correspond to the population of the states which in turn is described by the Boltzmann distribution. Knowing the strength of the external field allows to derive the energy levels with in turn allows to compute the populations for a given

temperature. Combining these assumptions, the initial density matrix of an equilibrated system can be formulated.

Formally, the density matrix is the matrix representation of the density operator $P = |\psi\rangle\langle\psi|$ where ψ is an arbitrary wave function as defined in equation (5). Knowing that the density matrix is the matrix representation of an operator, it becomes intuitive to relate snapshots of the density matrix $\sigma(t)$ (that is, the density matrix at a given point in time t) to other, time-constant operators I. In particular, the density matrices of a spin-½ nucleus with its spin being aligned along the x-, y- or z-axis correspond to the operators I_x, I_y and I_z, respectively. These matrices are also referred to as Pauli matrices and are defined according to

$$I_x = \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \; I_y = \frac{1}{2}\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \text{ and } I_z = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{7}$$

In addition to these operators, operators that alter the state of the system and therefore correspond to a transition of the spin include the lowering and raising operators I^+ and I^-, respectively. With the raising operator being defined as

$$I^+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \tag{8}$$

applying it to the $|\beta\rangle$ state results in a $|\alpha\rangle$ state, while the lowering operator performs the reverse transition. These operators are at the core of the product operator formalism as they are used to describe the state of the system as well as the Hamiltonian operators acting on the system.

As stated above, the density matrix allows to compute the outcome of an NMR experiment. Due to the nature of the detection in an NMR spectrometer, applying the measurement operator (often denoted as $I^+ = I_x + iI_y$) to the density matrix and then computing the trace corresponds to extracting the single quantum coherences. In other words, the single quantum coherence in the density matrix gives rise to the detectable signal of the NMR experiment. Nevertheless, the entire density matrix is relevant since the diagonal and off-diagonal elements are mixed under the influence of the external magnetic field and by applying rf pulses.

As described above, the density matrix allows to easily describe the initial state of the system as well as to compute the outcome of an experiment. Secondly, the usage of the density matrix

is advantageous as the evolution of the density matrix can be derived from the Schrödinger equation. The so-called Liouville-von Neumann equation

$$\frac{d\sigma(t)}{dt} = -i\left[\hat{H}, \sigma(t)\right] \tag{9}$$

describes the evolution of the density matrix where $\sigma(t)$ is the density matrix at time t, $\hat{H}$ is the Hamiltonian acting on the system and $[\cdot, \cdot]$ corresponds to the commutator $[A, B] = AB - BA$ (for a derivation see [16]). The Hamiltonian $\hat{H}$ is the operator that describes the energy of the system and as such depends on the nature of the system, the surrounding as well as external influences such as rf pulses. In the absence of any excitation rf pulse, the Hamiltonian is constant over time and the equation (9) can be solved to yield the desired description of the evolution of the density matrix over time

$$\sigma(t) = e^{-i\hat{H}t}\sigma(0)e^{i\hat{H}t} \tag{10}$$

To account for non-constant Hamiltonians during rf pulses, a transformation of the basis of the system $U = e^{i\omega I_z t}$ can be formulated such that the transformed Hamiltonian becomes constant (ω refers to the frequency of the radiation, t is the time and I_z is the product operator as defined in equation (7)). This so-called rotating-frame transformation U corresponds to the x and y axis rotating around the z-axis with the frequency of the radiation which in turn corresponds to the energy gap of the two states. As shown in equation (4) this frequency is given by the Larmor frequency ω_0 which, after normalization, is referred to as the chemical shift of the nuclei. It can be shown that in this rotating frame, the constant Hamiltonian corresponds to an effective magnetic field in the direction of the rf pulse.

Using the Taylor series expansion of an exponential, the evolution of the system described in equation (10) can be simplified to the basic scheme

$$\sigma(t) = e^{-iAt}\sigma(0)e^{iAt} = \sigma(0) \cdot \cos(\omega_A t) + B \cdot \sin(\omega_A t) \tag{11}$$

where A corresponds to the active Hamiltonian, ω_A is a frequency associated with A and B is the new state of the system. This equation shows the general idea of the product operator formalism, where the state $\sigma(t)$ of the system oscillates over time t between the initial state $\sigma(0)$ and a new state B. Knowing the frequency ω_A, a given system can be directed into a desired state by activating the corresponding Hamiltonian A for a specific time, which in turn requires the application of a rf pulse. For example, consider an isolated spin-½ system with

the states being defined by the Pauli matrices I_x, I_y and I_z. An excitation pulse along the x-axis leads to an oscillation of the initial state $\sigma(0) = I_z$ corresponding to a rotation around the x-axis with the Larmor frequency ω_0

$$\sigma(t) = I_z \cdot \cos(\omega_0 t) - I_y \cdot \sin(\omega_0 t) \tag{12}$$

Stopping the rf pulse after a period of $\tau = {}^{1}/_{2\omega_0}$ leaves the system in the state $-I_y$, corresponding to the magnetization being aligned along the negative y-axis. This example demonstrates how the state of the magnetization can be precisely controlled by choosing the direction (also referred to as phase) as well as the duration of the rf pulse.

The concept of product operator formalism can be extended in order to describe the evolution of multiple interacting spins. A typical source of interaction between spins are chemical bonds and interacting spins are often referred to as a spin system. To adequately describe a spin system of m interacting spin-½ nuclei, a total of $n = 2^m$ states are required. The density matrix of the system is constructed using the Pauli matrices (I_x, I_y and I_z) and applying the Kronecker product $\otimes$. For example, the density matrix of a 3-spin system with spin 1 being in state I_z, spin 2 being in state I_x and spin 3 being in state $-I_y$, is constructed by computing the Kronecker product $I_z \otimes I_x \otimes -I_y$. By extending the Hamiltonians in the same manner, the concept of oscillating density matrices can be transformed to any multi-spin system.

With the product operator formalism, a set of rules can be derived that describes how the system evolves over time and how the application of precisely-timed rf pulses alter this time evolution. The correct rules depend on the type of the spin system, the active Hamiltonian and the state of the system before applying the Hamiltonian. Since every Hamiltonian is connected to a given energy, the corresponding frequency can be used to determine the required pulse length. Applying the set of product operator rules allows to transfer the magnetization from one state to another which ultimately enables the transfer of magnetization between interacting spins. Consequently, the set of product operator rules is the basis for the development of a tremendous arsenal of NMR pulse programs. Pulse programs are typically constructed using common building blocks, such as for example the insensitive nuclei enhanced by polarization transfer (INEPT). While the magnetization is transferred across different nuclei of a spin system, the detected NMR signal encodes different physical

properties of the corresponding nuclei. Therefore, a given pulse program serves a given purpose as it allows to unveil a specific type of information[84, 85].

I.2.2 Basics of Spin Relaxation

Relaxation is omnipresent in every type of spectroscopy since the life times of excited states are limited. In the case of NMR spectroscopy, the life time of an excited state depends on the life time of the coherence between the corresponding states. In the previous section, the concept of product operators was introduced. While product operators describe which states of the system will be populated and thus which states will give rise to a signal, relaxation theory focuses on the life times of excited states and thus on the strength and decay of the signal. Although these two aspects are typically addressed independently, relaxation times set an upper limit on the available time for an NMR experiment. On the other hand, the exact result of the relaxation theory depends on the states of the spin system. During the development of an NMR experiment, both theories are combined to find the optimal solution. However, from a theoretical perspective, relaxation-specific approaches and ansatzes were developed and applied successfully. In the following, a brief description of some concepts and the consequences for NMR spectroscopy are discussed. A complete and detailed description of relaxation[16, 18, 86] involves a significant portion of mathematical treatment and is beyond the scope for this thesis.

Relaxation is the process of an excited system returning to its ground or equilibrium state and therefore determines the life time of an excited state. In the context of NMR spectroscopy, the excited or non-equilibrium states that give rise to the NMR signal are characterized by a coherence between different states within the solution ensemble. As a consequence, the life time of an excited state depends on the life time of the coherence between the corresponding states. Understanding the process of relaxation is tightly coupled with models describing the loss of coherence or correlation between states across a spin ensemble. As described in the previous chapter, the detection of x- and y-magnetization requires non-zero off-diagonal elements in the density matrix which in turn corresponds to a correlation between states in the solution ensemble. Such a correlation can be pictured as follows: In solution, the nuclear spins of the protons of the molecules are either aligned with or against the external field according to the Boltzmann distribution (compare equation (19)). Averaging the nuclear spins across all molecules leads to a non-zero net magnetization aligned with the external field since more

nuclear spins are in the energetically lower state. Moreover, no net magnetization perpendicular to the external field (that is along the x- or y-axis) is observed, since every spin rotates around the vector of the external field, but with the ensemble being in equilibrium, this rotation is randomly distrusted over different phases (which correspond to different angular positions). Applying a 90° rf pulse rotates all spins by 90° and thus leads to a rotation of the net magnetization such that no net magnetization is present along the external field, but instead, a net magnetization perpendicular to the external field is observed. Note that a net magnetization along the field is the result of equally-populated $|\alpha\rangle$ and $|\beta\rangle$ states. After the rf pulse, every spin rotates around the axis of the external field in the same manner as before applying the pulse (the rotation itself is the same since the active Hamiltonian before and after the pulse is identical). Besides the equally-populated $|\alpha\rangle$ and $|\beta\rangle$ states, the difference between the situation before and after the pulse, is a correlation within the ensemble. This can be regarded as a correlated spin rotation around the axis of the external field which in turn leads to a net magnetization in the x/y-plane, and thus the observed signal. The loss of this correlated rotation leads to a random distribution of the phase angles and therefore leads to a loss of the detect signal. To fully relax to the equilibrium state, the population difference between the $|\alpha\rangle$ and $|\beta\rangle$ state is restored as well. Since the loss of correlation as well as changes of the population difference between the two states are the main aspects of relaxation, this chapter will briefly classify and discuss the relevant molecular processes of relaxation.

In NMR spectroscopy, the energy gap between two states of a nuclear spin is relatively small and therefore spontaneous emission becomes unlikely and can be neglected. The dominant mechanisms of relaxation are caused by interactions of the spin with other spins in the surrounding, which are also referred to as lattice. The fact that relaxation is dominated by the interactions with the surrounding spins has several consequences. Firstly, NMR spectroscopy is characterized by long life times of the excited states ranging from several milliseconds up to several seconds. Secondly, describing the rate of relaxation requires a complex theoretical approach which in turn includes various parameters such as the nature of the coupling between the spins, the states of all spins in the surrounding, the molecular structure as well as the timescale and amplitudes of molecular dynamics such as the Brownian motion and rotation.

In the context of the density matrix, relaxation can be regarded in a more distinct manner. The equilibrium state is characterized by a Boltzmann distribution of the states which are directly connected to the diagonal of the density matrix. Relaxation processes that affect the population of the states directly affect the magnetization along the z-axis and are therefore referred to as longitudinal relaxation with the rate constant R_1. In equilibrium, states show no coherences and consequently relaxation leads to all off-diagonal elements of the density matrix becoming zero. Since the magnetization in the x/y-plane is always caused by a coherence between states, the loss of the coherences is also referred to as transverse relaxation and is characterized by the rate constant R_2. In summary, longitudinal R_1 relaxation is the process of the amplitudes of the states converging to the equilibrium while transverse R_2 relaxation causes the phases of the states across the bulk of the solution to average out at zero.

This distinction between longitudinal and transverse relaxation is particularly relevant as the consequences for NMR experiments are different in both cases. At the start of an NMR experiment, only the diagonal elements are non-zero and therefore any changes within the density matrix originate from these diagonal elements. Therefore, longitudinal relaxation is required to be completed before the start of any NMR experiment as it recovers the initial density matrix during the so-called recovery delay. In other words, R_1 relaxation affects NMR experiments by setting an upper limit to the maximum number of NMR experiments that can be performed in a given amount of time. On the other hand, coherences are not present in the beginning of an experiment but are essential for the detection of any signal since NMR spectrometers detect magnetization in the x/y-plane. Therefore, transverse R_2 relaxation reduces the detectable signal, and thus, sets an upper limit to the length in time of the NMR experiment. Moreover, R_2 relaxation continues during the detection period and directly affects the spectral resolution. Due to the nature of the Fourier transformation, fast-decaying signals result in broad peaks in the frequency domain. Fast R_2 relaxation sets an upper limit to the physically possible spectral resolution and might render it impossible to distinguish between similar frequencies due to spectral overlap.

So far, the consequences of spin relaxation were categorized and described in the context of the density matrix. In the following, the causes of relaxation are discussed. Spin relaxation traces back to interactions between a spin and other spins in its surrounding. The magnetic field which is experienced by a single spin is the sum of the constant external magnetic field

and many small magnetic fields that are generated by spins in close vicinity. Since the spins of the surrounding are dynamic, the local magnetic field acting on a given nucleus undergoes steady fluctuations. These fluctuations of the magnetic field can be split into two components, namely fluctuations changing the magnetic field in the z-direction and those changing the magnetization in the x/y plane. This distinction allows the following simplified view of relaxation mechanisms. Fluctuations along the z-axis modulate the strength of the external magnetic field. Since the energy level directly depends on the strength of the magnetic field, these fluctuations translate into a steady variation of the energy levels of the system. As discussed in the case of product operators, the energy levels directly determine the frequency of the oscillations between states and coherences in the density matrix. Therefore, fluctuations along the z-axis lead to some spins evolving faster and others slower and, on the level of the bulk in solution, leads to a loss of coherence and therefore contributes to the transverse relaxation rate R_2. Since the populations of the states are not changed, this mechanism of relaxation is referred to as adiabatic relaxation. In a simplified random-phase model, it can be derived that the transverse relaxation rate R_2 increases with the correlation time of the variations as well as with the mean square magnitude of the fluctuations (see [16] for the derivation and assumptions).

The second type of random fluctuations caused by the surrounding spins are those that affect the field in the x/y plane. Such fluctuating magnetic fields can be considered as electromagnetic pulses. Depending on the frequency of the fluctuations, these pulses can introduce transitions in the density matrix comparable to those induced by the pulses of the NMR spectrometer. Since the populations of the spin systems are changed, this relaxation is often referred to as non-adiabatic relaxation. Fluctuations of the local magnetic field with a smaller frequency can be regarded as low-energy rf pulses. In such cases a transition is stimulated that involves two spins, with one spin going from a high to a low and the second spin going from a low to a high energy state. Since the net energy required for this coupled transition is comparatively low, these transitions are effectively triggered by fluctuations of the local magnetic field with a smaller frequency. Such relaxation processes are referred to as cross-relaxation and require a coupling between the two spins, mediated either through chemical bonds or through space. It is important to note, that spins of the lattice (e.g. protons in the solvent) are also involved in these processes. As a consequence, relaxation stimulates an energy transfer between the

molecule and its surrounding. The surrounding of the molecule can be regarded as an energy bath capable of dissipating the energy of heat transfers and thus acting as an energy sink. Over time, this will eventually lead to an equilibrium distribution of all states in the spin system. The result of this spin-lattice relaxation is the build-up of the equilibrium net magnetization along the axis of the external field. Besides the direct effects of changing the spin population, the steady transitions induced by the fluctuating field leads to uncertainties in the energy levels as the average life time of a given state is shortened. As described in the previous chapter, the spin evolution of a given spin depends on the energy of the state. Constantly swapping the energy levels of a given spin leads to an uncertainty of the energy level which will eventually lead to a loss of coherences. In summary, non-adiabatic relaxation affects both, the population of the states as well as the coherence, and therefore contributes to the longitudinal R_1 as well as to the transverse R_2 relaxation rate.

To exactly describe relaxation, the idea of perturbation Hamiltonians was introduced. While the derivations are out of scope for this thesis (as with the previous section, a profound introduction can be found in [16]), the resulting concept of spectra density functions will be briefly described. A spectra density function describes the strength of a certain frequency within a given fluctuation. Following the Wiener-Chintschin-Theorem, the spectra density function can be derived by Fourier transforming the autocorrelation function of the fluctuating signal. This spectrum of the fluctuations can then be interpreted in the context of relaxation processes. Non-adiabatic and cross-relaxation processes require specific energies and thus only certain frequency bands of the spectra density function are relevant for triggering these relaxation processes. So, for example, slower motions are characterized by a broad autocorrelation function which in turn corresponds to small frequencies dominating the signal. On the other hand, cross-relaxation in a 2-spin system between the states $|\alpha\beta\rangle$ and $|\beta\alpha\rangle$ involves one spin changing from a low to a high energy state while the other spin is excited from the low to the high energy state. The required net energy is comparatively small and thus small frequencies in the spectral density facilitate the transition. As a consequence, the slow tumbling of large molecules (which is characterized by a large correlation time τ_c) leads to fast cross-relaxation.

In biomolecular NMR spectroscopy, the underlying mechanism of relaxation depends on the concrete spin system, and more precisely, on the interactions or couplings between spins

which give rise to fluctuating magnetic fields. Due to the defined chemical structure of biomolecules, the following sources of coupling are relevant in the context of relaxation. The influence of the magnetic field of spin on another spin is referred to as a through-space or direct dipolar interaction. The strength of this interaction depends on the distance between both spins and on the angle between the connecting line and the external field. As a consequence, dipolar interactions are affected by changes of the molecular structure as well as the orientation of the entire molecule, and therefore, molecular tumbling gives rise to fluctuating magnetic fields. Another cause of fluctuating magnetic fields traces back to chemical shift anisotropy (CSA). CSA denotes the asymmetry of the chemical shift which in turn is significantly affected by the shielding of the electron density. Since electron orbitals such as those forming chemical bonds are asymmetric with regard to the center of a nucleus, the chemical shift is not a scalar but instead is direction-dependent. As a consequence, molecular tumbling causes variation of the chemical shift which in turn leads to variations of the energy levels. Another potential source of fluctuating magnetic fields are scalar J-couplings between bonded nuclei. However in biomolecules, the most dominant mechanisms of relaxation trace back to through-space dipolar interactions as well as CSA[16].

Due to the different underlying molecular mechanism, the observed relaxation rates depend on various structural and dynamic properties of the molecule. This allows to utilize relaxation to obtain insights into the molecule of interest. The following section will provide a brief overview about the structural properties that can be observed for biomolecules.

I.2.3 NMR Spectroscopy on Proteins

NMR spectroscopy is a powerful tool to examine the structure and dynamics of proteins and other biological macromolecules on an atomic level. In the previous chapter, the basics of product operators and relaxation have been introduced. In this introduction, several quantities have been described that depend on the molecular structure and dynamics. These dependencies are essential for NMR spectroscopy as they allow to derive insights on an atomic level. The power of NMR spectroscopy lies in particular within the large set of accessible structural and dynamic information. However, this thesis focuses on structure determination aspects and therefore NMR methods for studying dynamics are not discussed and the reader is referred to textbooks and reviews[16-18, 80, 87]. The following section gives an overview over typical structural information that can be unveiled using NMR spectroscopy.

I.2.3.1 Direct Measurement of Geometric Properties

A prominent approach to obtain structural information exploits the scalar J-coupling between spins which is mediated by the electrons involved in the chemical bond. The scalar J-coupling constant between chemically-bonded nuclei depends on the rotational angle of the bond. This dependency is described in the Karplus equations and can be approximated by the use of cosine functions[16, 88-90] such as

$$J(\theta) = A \cdot \cos^2(\theta) + B \cdot \cos(\theta) + C \tag{13}$$

where J is the observed coupling constant, θ is the rotational angle of the bond and A, B and C are coefficients that are used to fit the model.

The most prominent example of structural information derived from NMR spectroscopy are distances based on the Nuclear Overhauser Effect (NOE) [16, 79, 91, 92]. The NOE is a cross-relaxation effect which is utilized in Nuclear Overhauser enhancement spectroscopy (NOESY) pulse programs to transfer magnetization from a given spin to nearby spins. In biological molecules, through-space dipolar couplings are the most relevant contribution to cross-relaxation. Considering two protons, I and S, that are in close proximity such that a relevant dipolar coupling is present, this system possesses four energetic states, with either none, only spin I, only spin S or both spins being in the high energy $|\beta\rangle$ state. As described in previous chapters, through-space dipolar coupling between spin I and S allows for relaxation processes in which both spins change their states in a correlated manner. An example for such a relaxation process is the simultaneous transition of spin I from the $|\alpha\rangle$ to the $|\beta\rangle$ state and vice versa for spin S. In a two-dimensional NOESY experiment, the NOE is typically detected according to the following scheme. All protons are excited non-selectively to obtain a net x/y magnetization and their respective Larmor frequencies are encoded by incrementing the so-called t_1 delay between successive NMR experiments. After this delay, all spins are rotated back into the z-direction. During the subsequent NOE mixing time, cross-relaxation of the z-magnetization can occur between adjacent protons. Eventually, the magnitude of the magnetization after the mixing time is measured using a third rf pulse followed by the direct acquisition of the frequencies of all protons (t_2). In the case of cross-relaxation, signals are obtained that encode the frequencies of the source and the target spins of the transition in t_1 and t_2, respectively. These signals give rise to off-diagonal peaks in the spectrum and the

strength of the through-space dipolar interaction between the two nuclei can be related to the distance using the relationship

$$r_{ij} = r_{\text{ref}} \left(\frac{S_{ij}}{S_{\text{ref}}} \right)^6 \tag{14}$$

where r_{ij} is the distance between the interaction spins i and j, r_{ref} is the known distance between two reference nuclei, S_{ij} is the signal observed from the NOE cross-peak of spins i and j and S_{ref} is the signal observed from the NOE cross-peak of the reference spin pair. In contrast to computing the distance as an absolute quantity, the usage of the reference spin allows to account for multiple factors that are typically more complicated to be determined. It should be noted, that the strong dependency on the distance (6^{th} power) renders the NOE very sensitive to minor changes of the distances and thus allows to obtain useful distance information. On the other hand, for long range distances, the NOE becomes undetectable and therefore limits the structural information of NOE data to distances below 5-6 Å, although in some cases magnetization transfers can be observed for distances up to about 10 Å.

Another type of restraints is obtained by measuring residual dipolar couplings (RDCs) [16, 93]. RDCs are observed by slightly reducing the rotational degrees of freedom. This rotational alignment of the molecules is achieved by the addition of components that form elongated structures. Molecules, such as virus particles, lipids or DNA nanotubes, are used to form regularly-ordered macromolecular structures which lead to a partial molecular alignment of the target molecule. It should be noted that no direct interaction between this matrix and the sample of interest may be formed since such an interaction leads to a strong alignment which in turn leads a significant reduction of the signal [93, 94]. The concept of RDCs, relies on the previously described dipolar couplings between spins. As the interaction between two spins depends on the angle between the direction of the external field and the connecting line between the spins, rotations of the molecule lead to variations of the coupling. In solution the rotation is random and equally distributed across all angles and therefore the dipolar coupling on average has no net effect. In contrast, bond-mediated scalar J-couplings are not affected by molecular tumbling, and therefore lead to distinct energy levels and thus different signals, such as duplets or triplets. Introducing a partial alignment using the above mentioned additives, reduced molecular rotation results in an imperfect averaging of dipolar coupling interactions which in turn is observed by a variation of the energy levels and thus as a variation

of the coupling constants. This variation of the coupling constant contains structural information since the variation depends on the angle between the external field and the connecting line between the interacting spins. This angular dependency can be expressed as

$$D = D_{\mathrm{a}}\left(3 \cdot cos^2(\theta) + \frac{3}{2}R \cdot \sin^2(\theta) \cdot \cos(2 \cdot \phi) - 1\right) \tag{15}$$

where D is the observed residual coupling and θ as well as ϕ are the angles of the spherical coordinate of the vector connecting the interaction spins in the alignment tensor frame. The rhombicity R and alignment order D_{a} depend on the nature of the alignment and need to be determined for every sample and alignment medium. In practice, the effort of preparation of partially aligned samples varies significantly. For many metal-binding proteins, the external magnetic field is sufficient to obtain an alignment without the need of any additional alignment agent. For most proteins however, sample preparation can become tedious as different alignment media need to be tested and the obtained alignment has to be tuned to find the optimal compromise between signal-to-noise ratio and an accurately measurable change of coupling constants. Nevertheless, RDC data are often recorded due to the excellent agreement with structural data. Moreover, since RDCs depend on the direction of the external field, all angles share the same reference direction. As such, RDC data that contain long-range orientation information often complement short distance data sets obtained from NOEs.

Another type of long range distance information can be obtained by means of paramagnetic relaxation enhancement (PRE) as well as pseudo contact shifts (PCS) [95-98]. Both effects are caused by the introduction of unpaired electrons into the system of interest. Typical sources of unpaired electrons include lanthanide ions such as Gadolinium, Dysprosium or Ytterbium as well as stable radicals such as (2,2,6,6-Tetramethylpiperidin-1-yl)oxyl (TEMPO)[32, 99]. While the latter can be directly attached to protein sidechains, the introduction of metal ions can be achieved by using natural metal binding sites or by covalently attaching a metal-binding spin label. These spin labels chelate metal ions with unpaired electrons and, ideally, neither interfere with the protein nor undergo significant motions. Due to the large gyromagnetic ratio and thus large magnetic field induced by electrons, the introduction of unpaired electrons leads to a paramagnetic relaxation enhancement, that is, a distance-dependent increase of the relaxation rates R_1 and R_2. This electron-mediated PRE relies on dipolar interactions between unpaired electrons and the nuclear spins. As previously

described, dipolar interactions of surrounding spins cause fluctuating magnetic fields. The same holds for unpaired electrons of the attached spin labels which, compared to the field induced by nuclear spins, induce fluctuating magnetic fields that are three orders of magnitude stronger due to the significantly larger gyromagnetic moment. In addition, metal ions with nuclear spin quantum numbers $S \geq 1$ introduce a quadrupole moment. A quadrupole moment as well as the strong dipolar coupling lead to a strong enhancement of spin relaxation in a distance dependent manner. A detailed description of the PRE is provided in the Solomon–Bloembergen–Morgan theory [97, 100-102].

Based on theoretical descriptions, several approaches were developed to utilize the PRE to obtain structural information of the target molecule. A commonly-used approach is described in [95] and involves the determination of the paramagnetic contribution to the observed R_2 relaxation rate. To this end, a paramagnetic spin label is attached to the protein and, in a second step, the spin label is transformed from its paramagnetic state to its PRE-inactive diamagnetic state. This transition is accomplished by chemically reducing the radical of the spin label. The R_2 relaxation rate is then estimated in the absence of any PRE using the peak width. The relaxation rate in presence of paramagnetic relaxation ${R_2}^{\text{para}}$, and thus the contribution of the PRE, is determined via the signal intensities

$$\frac{I_{\text{para}}}{I_{\text{ref}}} = \frac{R_2 \cdot e^{-{R_2}^{\text{para}} \cdot \tau}}{R_2 + {R_2}^{\text{para}}} \tag{16}$$

where I_{para} and I_{ref} are the observed peak intensities in presence of a paramagnetic and a diamagnetic (and thus non-relaxation-enhancing) spin label and τ is the total INEPT duration of the pulse program. The obtained PRE rates ${R_2}^{\text{para}}$ are then converted to distances using the relationship (see [95, 97] for details)

$$r = \sqrt[6]{\frac{K}{{R_2}^{\text{para}}} \cdot \left(4 \cdot \tau_c + \frac{3 \cdot \tau_c}{1 + {\omega_h}^2 {\tau_c}^2}\right)} \tag{17}$$

where ${R_2}^{\text{para}}$ is the previously determined relaxation rate of equation (16), τ_c is the correlation time for interaction of the electron with the proton and ω_h is the Larmor frequency of the proton. The factor K in equation (17) contains several constants (see [95, 103]) and is computed using

$$K = \frac{1}{15} \cdot S(S+1)\gamma^2 g^2 \beta^2 = 1.23 \cdot 10^{-32} \frac{\mathrm{cm}^6}{\mathrm{s}^2} \quad (18)$$

where $S = \frac{1}{2}$ is the spin quantum number, γ is the gyromagnetic ratio for protons, g is the dimensionless magnetic moment (g-factor) of the electron and β is the Bohr magneton. Following this procedure, a distance to the paramagnetic center is obtained for every nucleus. By repeating the experiments with spin labels located at different protein sites, multiple sets of distances can be obtained. Another advantage of this method is the possibility to observe long range distances of up to 20 Å. As a result, PRE-based relaxation studies based on spin labels gained popularity.

For the sake of completeness, it should be noted that PCS data allow to complement PRE studies. Metal ions such as Lanthanides possess electrons in outer electron orbitals which in turn lead to a specific electron distribution. Besides a relaxation-enhancing effect, the spatial electron distribution of Lanthanide ions, lead to the pseudo chemical shift (PCS), an alternation of the chemical shift of surrounding nuclei. As a consequence measuring these PCSs, unveils structural information such as orientations and distances[99, 104].

I.2.3.2 Structural Information of the Local Environment

The previously described quantities allow to obtain a direct measurement of defined structural information, including short and long-range distances, dihedral angles as well as orientations. Besides these quantities, NMR spectroscopy offers additional possibilities to acquire data that are indirectly related to the structural properties of the sample. The most prominent type of information is the chemical shift. The broad spectrum of different Larmor frequencies and varying chemical shifts of nuclei in a biomolecule traces back to local distribution of electrons which in turn leads to site-specific shielding of the external magnetic field[16-18, 105]. As a consequence, the set of chemical shifts of a molecule can be regarded as a fingerprint that strongly depends on the local environment of every nuclei and as such on the conformation of the entire molecule. As a consequence, using chemical shifts to obtain structural information is subject to active research[106-108]. For example, the prediction of backbone angles and secondary structure elements based on chemical shifts is a commonly-used approach which significantly facilitates the structure determination process[109-111]. Using the full potential of chemical shift requires an efficient prediction[112, 113]. Based on such a prediction, structures of biomolecules can be obtained and validated by comparing the observed chemical

shift data with predicted data that were computed based upon candidate structure models[114-116].

Another commonly-used and chemical shift-based approach involves observing the change of the chemical shift while changing the conditions of the sample. The chemical shift depends on the magnetic shielding of the surrounding environment, which in turn depends strongly on the structural and dynamic arrangement of all neighboring nuclei. As a consequence, any induced change in the surrounding of a nucleus, leads to a perturbation of the chemical shift. As a consequence, measuring modifications of chemical shifts is a very sensitive method to detect and localize changes in the surroundings of a given moiety of the molecule. For example, titrating a protein of interest with a ligand results in frequency shifts of all sites involved in the binding. Observing these chemical shift perturbations (CSPs) is a powerful application of NMR spectroscopy and gives structural insights regarding the ligand binding site[117]. Besides determining the location of bindings sites, CSP data allow to unveil kinetic and thermodynamic insights into the binding event.

The chemical shift depends strongly on the local environment, and certainly, combining the local information of all sites of the protein allows to obtain the overall fold. However, keeping the global fold and solely changing the properties of a single site will lead to changes of the chemical shifts in the proximity of the altered site. Therefore, chemical shift data indirectly depends on the global fold, but is, to a larger extent, directly dependent on the chemical properties of the local environment. Consequently, chemical shift data are powerful to investigate local conformations, but typically fail to provide insights into the overall fold of biomolecules.

I.2.3.3 Structural Information of the Global Fold

To complement chemical shift data, structural data that are sensitive to the global fold of biomolecules and at the same time insensitive to the local environment need to be obtained. This challenge can be addressed by acquiring solvent accessibility data which encodes structural information depending on the solvent exposure and thus on the fold of the biomolecule[118]. Solvent accessibility data can be obtained by NMR in a straight-forward manner by observing the (NOE-mediated) magnetization transfer from solvent water to the protein[119]. However, this approach suffers from site-specific artifacts due to an additional

chemical exchange contribution. To overcome this limitation, solvent accessibility data can be obtained by applying the concept of solvent paramagnetic relaxation enhancement (solvent PRE or sPRE) [97, 118-122] as indicated in Figure I-3a. Instead of covalently attaching a spin label to the protein, solvent PRE extends the idea of PRE by titrating the sample with a soluble paramagnetic compound. Dipolar interactions between unpaired electrons of the soluble compound and the nuclei of the molecule of interest result in an additional relaxation contribution. The strength of this contribution depends on the interaction between the soluble compound and the biomolecule, which can be described by a solely-diffusional outer sphere interaction model or by the formation of a transient complex between the compound and the biomolecule[123]. In the absence of specific binding and in the case of paramagnetic compounds with a fast electron spin relaxation, such as Lanthanide ions, the strength of the dipolar interaction depends mainly on the solvent accessibly and both models lead to the same results: Nuclei positioned at the outside of the biomolecule are more exposed to the solvent and the paramagnetic compound and thus experience a strong paramagnetic effect which in turn increases the observed relaxation rate. The effect of solvent PRE can be quantified by measuring the increase of the relaxation rate as a function of the concentration of the paramagnetic compound. To this end, the soluble compound TEMPOL (4-Hydroxy-2,2,6,6-tetramethylpiperidinyloxyl) was used since it is a derivative of the commonly-used paramagnetic spin label TEMPO (2,2,6,6-Tetramethylpiperidinyloxyl)[118, 124, 125]. Using TEMPOL, the increase of the relaxation rates was successfully correlated to the helical secondary structure of RNAs[126]. However due to its tendency to form H-bonds, TEMPOL shows an enhanced interaction with certain moieties of biomolecules such as exposed amides[127]. The specific binding and the resulting biasing of specific moieties of the target molecule was circumvented by using Gd(DTPA-BMA) as a soluble paramagnetic compound. Gd(DTPA-BMA) is a Gadolinium-chelate (also referred to as Omniscan) and a contrast agent developed for MRI applications. Studies have successfully proven the correlation between the structure of biomolecules with the solvent-accessibility data obtained using Gd(DTPA-BMA)[122, 127].

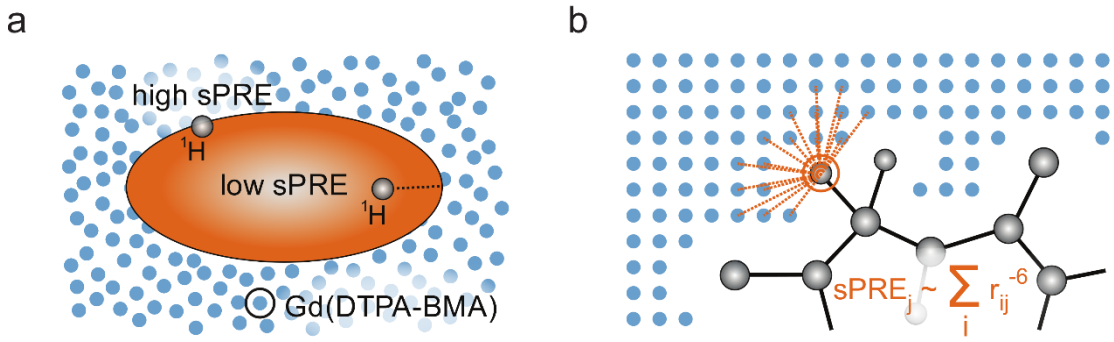


Figure I-3: Concept of solvent PRE
(**a**) Addition of the paramagnetic agent Gd(DTPA-BMA) (blue) to the solution leads to a bleaching of NMR signals of protons (grey) positioned at solvent-exposed sites of the protein (orange). Signals of the buried protein core are significantly less affected and signal intensity is maintained (indicated as low sPRE). (**b**) The numerical model to compute the solvent PRE of a given nuclei relies on constructing a regularly-spaced grid (blue) representing the solvent. The solvent PRE of a nucleus is computed as the sum of the contribution from all grid positions where every individual contribution is proportional to r^{-6}. Figure was obtained from [128].

Similar to the usage of chemical shifts, solvent PRE data was used to unveil structural features and to facilitate structure prediction of proteins[118, 120, 121, 129]. The usage of sPRE data in these studies relies on the possibility to numerically predict solvent PRE data from a given structural model[122]. The concept of the numerical approach is depicted in Figure I-3b and can be regarded as a numerical integration of the paramagnetic effect acting on a specific site of the protein. To this end, the structural model of the protein is positioned in an evenly-spaced three-dimensional grid and all grid positions overlapping with the protein are removed. Since the van-der-Waals radius of all protein nuclei is extended by the radius of a single Gd(DTPA-BMA) molecule, the remaining grid represents the accessible volume of the paramagnetic agent. The solvent PRE of a given nuclei is then computed by summing all contributions of the remaining grid position where every contribution vanishes with r^{-6} where r is the distance between the grid and the respective nucleus. This numerical algorithm allows to predict the solvent PRE given a structure of the target molecule.

I.3 Computational Structural Biology

The term computational structural biology is used in many fields of biological research. In this thesis, computational structural biology refers to the efforts of understanding the structure and dynamics of biomolecules on an atomic level by an extensive usage of computational methods. In this context, molecular dynamics simulations have become the most prominent example[50, 130, 131] and in 2013 the Nobel Prize in Chemistry was awarded to Martin Karplus, Michael Levitt and Arieh Warshel for their work in the field of computational biology.

At its core, molecular dynamics simulations aim to determine molecular motions by computing molecular forces of every nucleus of the system and then following the trajectory in time given by these forces. Depending on the setup of the simulations, different aspects of the system can be studied including dynamics as well as thermodynamic properties[49, 132]. Molecular dynamics simulations studying the dynamics of biomolecules, are complemented by structure prediction studies that aim to determine the equilibrium structure of the target protein. These methods focus on efficiently sampling the conformational space of a given molecule. With the use of an energy function, the conformation with the lowest energy, corresponding to the equilibrium structure, is determined. Several approaches have been developed that are optimized for predicting the fold of proteins, including highly-efficient Monte Carlo sampling-based methods[49, 133-135].

In the case of molecular dynamics simulations as well as structure prediction computations, the energy or force field describing the system is a key factor for obtaining reliable results. The following sections provide insights into these force fields. Later, molecular dynamics and Monte Carlo-based sampling methods for proteins are described in more detail.

I.3.1 Force Fields

The challenge of studying biomolecules traces back to the large molecular size in combination with the tremendously complex quantum mechanical calculations. In the field of computational chemistry, several methods were developed that are capable of describing molecules by means of quantum chemistry. However, the computational resources required for advanced Post-Hartree Fock methods increase drastically with the number of basis functions which in turn is linked to the number of electrons (n_b) in the system. The computational complexity rises with $O({n_b}^4)$ in the case of Hartree Fock, $O({n_b}^5)$ for simple Møller–Plesset methods, $O({n_b}^6)$ for Coupled Cluster methods and up to $O({n_b}^{10})$ for high-order methods[63, 64]. For example, computing the energy of a short DNA strand on a multi-core machine requires over a month of computational time[136]. As a result, even less computationally-demanding approaches such as density functional theory $O({n_b}^3)$ are not capable of computing large biomolecules[63, 64, 73].

To enable computational methods on large biomolecules, computations were developed that consider chemical bonds as non-breakable. This allows to simplify the computations by not

treating electrons explicitly, but instead, only consider the nuclei in combinations with empirical energy functions that mimic the observed physical properties of chemical bonds. Since deriving these energy functions allows to obtain forces, these empirical approaches are also referred to as force fields[63, 64, 131, 137]. Energies that are typically considered in force fields, include the stretches, angles and rotations of chemical bonds, torsion angles, planarity constraints as well as van-der-Waal and electrostatic interactions. These energies are then described by simple sine, parabola or reciprocal functions that only depend on the positions of the nuclei[63, 64]. The simplicity of the force fields leads to a dramatically improved scalability of $O({n_a}^2)$ where n_a is the number of atoms. The squared complexity originates from the van-der-Waals and electrostatic interactions that have to be computed pairwise. However, current state-of-the-art methods exploit the vanishing of these interactions over the distance which further reduces the complexity to $O(n_a)$ [131].

It should be noted that the chemical structure of the molecule has to be provided and is considered to be constant. In particular, the usage of force fields prevents the formation and breakage of chemical bonds. As a consequence, hybrid methods combining molecular dynamics and quantum mechanical computations were developed[63, 64, 138]. These approaches are in particular useful in the case of studies on enzymes as they allow to simulate electrons and bonds for a comparatively small but chemically active site of the target molecule.

I.3.2 Computational Methods

Computational studies of biomolecules require the usage of some type of simplified force fields. Depending on the aim of the studies, different computational approaches can be chosen that eventually apply the force field to answer a given question. Molecular dynamics simulations as well as Monte Carlo-based sampling approaches are two commonly-used computational methods which are discussed in the following sections.

I.3.2.1 Molecular Dynamics Simulations

The basic idea of molecular dynamics simulations can be regarded as a numerical integration. That is, starting from an initial structure, the force field is used to compute the forces acting on every nucleus. With a given integration or time step, the new positions are computed. With these new positions, a new set of forces is obtained by re-evaluating the force field. This numerical integration is repeated until the desired time span is computed. Here, a simplified

Euler method for integration is described, while in most cases a leapfrog or velocity verlet integration is used[63, 64]. It should also be noted that depending on the setup of the simulation, the positions of the nuclei as well as the velocities are adjusted to keep the volume, the pressure or the temperature constant.

Molecular dynamics simulation are primarily used to study the dynamics of proteins, but are also capable of finding new intermediate conformations, refining structures or even finding the folded structure at the first place[45, 57, 139, 140]. In all cases, a common challenge for applying molecular dynamics is the validation that the simulation has sufficiently sampled the conformational space and thus converged to the correct results. Improving the sampling of simulations has been a major research focus and several techniques were developed[132, 141]. Limited sampling of a molecular dynamics simulation is the result of the system being trapped in low-energy states. Since the system will preferable populate these states, only a minor fraction of the simulation time can typically be used to study the desired states of interest. Therefore, the principle idea of advanced sampling approaches relies on pushing the simulated system out of low-energy traps. This can be accomplished by increasing the temperature (replica-exchange approach) or by the introduction of an additional energy potential (Umbrella sampling and metadynamics approaches) [142-147].

I.3.2.2 Monte Carlo-based Methods

Molecular dynamics simulations have a broad range of applications in which cases the kinetics and dynamics of the system are the subject of interest. For studies that solely aim to find the folded structure of a given biomolecule or biomolecular complex, the usage of a Monte Carlo-based sampling method in combination with the Metropolis algorithm is a powerful tool[49, 148, 149]. Markov chain Monte Carlo methods are efficient algorithms to generate ensembles for a given energy distribution which in turn can be used to find the optimum in a high-dimensional system.

Similar to molecular dynamics, the Markov chain Monte Carlo algorithm starts off with an initial structure. In contrast to molecular dynamics, this initial structure is used to compute the energy, rather than the molecular forces. Next, a random variation is introduced into the structure which corresponds to a random jump within the conformational space. For this new structure, the energy is computed and compared to the old energy by applying the Metropolis

criterion: If the energy is lower, the new structure is accepted and replaces the old structure. In case the energy is higher, the new structure is accepted only with a certain probability P that is proportional to a Boltzmann factor

$$P(E_{\text{new}}) = e^{\frac{E_{\text{old}} - E_{\text{new}}}{kT}} \tag{19}$$

where E_{new} and E_{old} refer to the energies of the new and old structure, respectively, k is the Boltzmann constant and T is the temperature of the system. In case the structure was accepted (declined), the new (old) structure is then varied again in the next iteration. It can be shown that the ensemble generated by this algorithm corresponds to a Boltzmann distribution[149, 150].

The Rosetta software framework applies the concept of the Metropolis algorithm to different challenges such as structure determination of proteins and RNA, docking as well as protein design[151-153]. The Rosetta framework provides efficient and highly optimized sampling strategies which in the case of structure determination is used to efficiently fold an elongated protein chain. The high sampling efficiency of the algorithm is a result of several optimizations and adjustments for biomolecules. These optimizations are described in more detail in the following section, in particular the usage of statistical knowledge, the reduction of the conformational space as well as an efficient representation of the biomolecules.

In most applications of the Rosetta framework, the underlying energy function is extended to not only include physical interaction energies but also statistical information. Statistical information refers to using existing structural knowledge and to facilitate structural computations. In the past decades, the number of structures submitted to the Protein Database is growing at a tremendous rate (see Figure I-1), providing a valuable data source. At the same time, the number of unique protein folds increased as well, but started to level off at the beginning of the 21st century (see Figure I-4). In agreement to this observation, different studies have estimated the total number of protein folds to be between 1,000 and 10,000[154-156]. These studies also unveiled an uneven distribution of folds. A comparatively small number of so-called superfolds cover the majority of folds for most proteins families[154, 157]. This bias towards a limited set of folds as well as the upper bound for the total number of folds strongly suggest to include statistical analysis in the structure determination process. In the described algorithm of the Rosetta framework, statistical knowledge about known proteins folds is used

to generate a library of folds. This library is then used to facilitate structure determination as it allows to effectively construct candidate structures and to efficiently sample the structural landscape of the target protein.

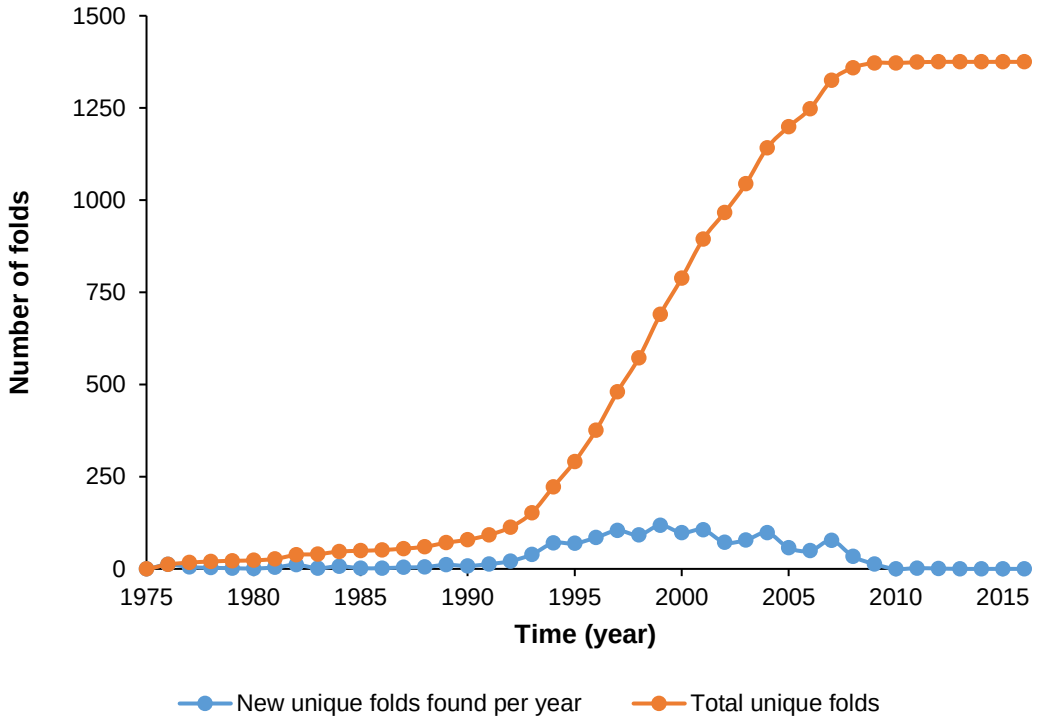


Figure I-4: Number of unique protein folds in the PDB as defined by number of CATH (Class, Architecture, Topology/fold, Homologous superfamily; v4.0.0) levels
While the total number of structures submitted to the Protein Database increased rapidly during the last decades, the number of unique folds stagnated in the last years. Data was obtained from http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=fold-cath (as of October 2016).

Another essential optimization of the Rosetta folding algorithm traces back to the reduction of the high-dimensional conformational space. This is accomplished by using a simplified representation of the protein in the early stages of the folding backbone and switching to a full-atom representation at a later stage when the algorithm is close to the native structure. In addition, the dimensionality of the folding problem is reduced through the construction of new candidate structures by combining 3- and 9-mer peptide fragments. The conformations of these short fragments are obtained in a preparation step which performs sequence alignments to choose optimal conformations for every fragment of the target protein. With this approach, random alteration of protein during the Monte Carlo algorithm is reduced to randomly choosing a fragment of the protein and then randomly selecting a conformation from the pre-selected list of conformations. These optimizations are the key factor for the efficient *ab initio* protein structure prediction algorithm.

I.4 Computational Approaches in NMR Spectroscopy

Computational methods and NMR spectroscopy are closely connected. On one hand, NMR spectroscopy is well suited to deliver parameters for computational studies such as the distances and angles as well as time amplitudes and timescales of dynamics processes. On the other hand, many applications of NMR spectroscopy require the usage of computational methods to extract the full potential of NMR data set, for example by computing structure ensembles that fit the observed NMR data. The following section will describe computational methods that are particularly used in the field of NMR spectroscopy.

I.4.1 Common Methods

Commonly-used software programs to determine biomolecular structures include for example Xplor-NIH[158, 159] and CNS[160]. Both packages are capable of using a variety of experimental NMR data, such as angles obtained from scalar J-couplings, distances derived from NOEs or PREs and solvent accessibility data. Moreover, data from other methods can be imported such as scattering curves obtained by small angle scattering (SAXS). Determining the structure of a protein by means of NMR is typically accomplished using simulated annealing protocols. In this approach, a molecular dynamics simulation of the protein is performed using an extended or random conformation as the initial model. Starting at high temperatures, the simulated annealing protocol gradually reduces the temperature of the system and eventually, after removing the thermal energy of the system, the protein is found in its native state. These simulations are characterized by a large set of experimental restraints that drive the simulation towards a structure that agrees with the experimentally observed data. The most relevant restraints for this purpose are NOE-derived distances, angles obtained from J-couplings or predicted from chemical shifts as well as orientations obtained from RDC data.

Another aspect of protein structure determination using NMR data involves the automated assignment of spectral signals to the nuclei of the protein. In a typical workflow, considerable time is required to assign peaks of the NMR spectra to the correct spin systems of the protein. In most cases, the assignment process is a multi-stage process with the sequential assignment of the protein backbone being followed by the assignment of the amino acid sidechains as well as the assignment of NOE spectra. It should be noted, that all steps are subject to

automation[161-164] and two particularly popular software packages for the assignment of NOE spectra are Aria[165] and CYANA[166].

Besides the molecular dynamics-based approaches, the efficient sampling-based *ab initio* structure prediction capabilities of the Rosetta framework have been applied to the field of NMR spectroscopy. To this end, the protocol CS-Rosetta was developed combining the Rosetta framework and NMR chemical shift[114, 115]. The experimental NMR data are used in two stages of the structure prediction protocol, as described in the following sections.

The Rosetta structure prediction protocol requires the creation of a fragment library that contains candidate structures for every fragment of the primary sequence (for details see previous sections). This fragment library is used to limit the conformational space of the Metropolis optimization algorithm. This concept significantly increases the efficiency of the algorithm. However, at the same time, the quality of the algorithm depends on the selection of fragment confirmations. To improve the crucial selection of the conformations of the fragment, the CS-Rosetta protocol makes use of NMR chemical shift data. As described in previous sections, NMR chemical shift data correlate well with backbone angles and in particular with secondary structure elements such as α-helices or β-sheets. This property allows to select highly accurate fragment confirmations. It was shown, that considering the chemical shift data in the selection of fragment conformations significantly improves the accuracy of the structure prediction.

The *ab initio* sampling approach of the Rosetta framework, relies on optimizing a large number of random initial structures resulting in a large ensemble of folded candidate structures. To find the model closest to the native fold of the protein, energy functions are used to score and cluster the ensemble. The quality of this selection process relies on the accuracy of the energy function. Back-predicting the chemical shift data for every candidate structure and comparing this prediction to the experimental values provides an additional energy function. The CS-Rosetta protocol makes use of this chemical shift-based energy function to filter the final ensemble which leads to an improved accuracy of the structure prediction.

I.4.2 Challenges

Recent trends in structural biology show an increasing importance of integrative hybrid methods which combine data derived from different experimental approaches[34, 167-169]. At

the same time, the focus of recent research in the field of structural biology is moving from studying single biomolecules to the investigation of large biomolecules as well as biomolecular interactions and complexes[28, 31, 32, 170-173]. Moreover, new challenges in the field of structural biology arise from studying highly flexible target proteins such as IDPs[169, 174-176] and RNAs[177, 178]. These trends lead to a variety of new challenges that have to be overcome. In the following these challenges will be discussed in more detail.

In the late 1960s, Levinthal formulated a famous paradox. Considering the flexibility of the protein backbone, the complexity of finding the native fold increases exponentially with the number of amino acids. By randomly sampling these conformations, the folding of a protein becomes impossible[179, 180]. Although this paradox was solved by showing that the folding is a directed and biased process[180], the concept of connecting the size of the biomolecular system to the complexity of the associated conformational space remains essential: The difficulties and complexity of studying large systems arise from the dimension of the associated conformational space. For example, investigating large proteins and complexes of proteins increases the number of degrees of freedoms and thus significantly expands the conformational space of the system[73]. A similar consideration can be done regarding highly flexible systems. Rather than expanding the conformational space, flexibility and dynamics increase the relevant portion of the conformational space that the system can occupy[175].

Integrative hybrid methods aim to combine different types of experimental data in order to provide more complete and complementary experimental data for a given biomolecular system[147, 174, 181]. In addition to studying large and dynamic systems, the trend towards integrative hybrid methods increases the complexity of computational methods as well. Using experimental data in computational methods typically involves comparing measured experimental data to the corresponding back-predicted data based on the structural model. Therefore, including different types of experimental data requires the addition of the corresponding back-prediction in the energy or force field function. Moreover, energy terms that are based on different experimental observables need to be normalized in order to be combined in total energy. In summary, integrative hybrid methods lead to the development of more complex energy functions and algorithms.

From a methodical point of view, these challenges have several implications. First, computational methods became computational-expensive and thus more resource-

demanding. Until the beginning of this century, the development of computer technology lead to an increase of computational resources. Equally important, this newly-gained computational power directly transformed into a reduction of run times and more time-efficient algorithms. In the last decade, computational resources still increased, however the performance per computing unit stagnated. Therefore, the increase of computing power development came at the cost of large distributed systems (see Figure I-5). This trend of distributed systems introduced new requirements for computational methods as the possibility to parallelize the work load becomes essential. In addition to the tremendous development of off-the-shelf processors and accelerators such as graphics cards, the field of computational structure biology also benefits from the development of custom build technology. As an example, the application-specific integrated circuit Anton [182, 183] enabled to study longer time frames of molecular motions which in turn allowed to study the folding of certain proteins[57].

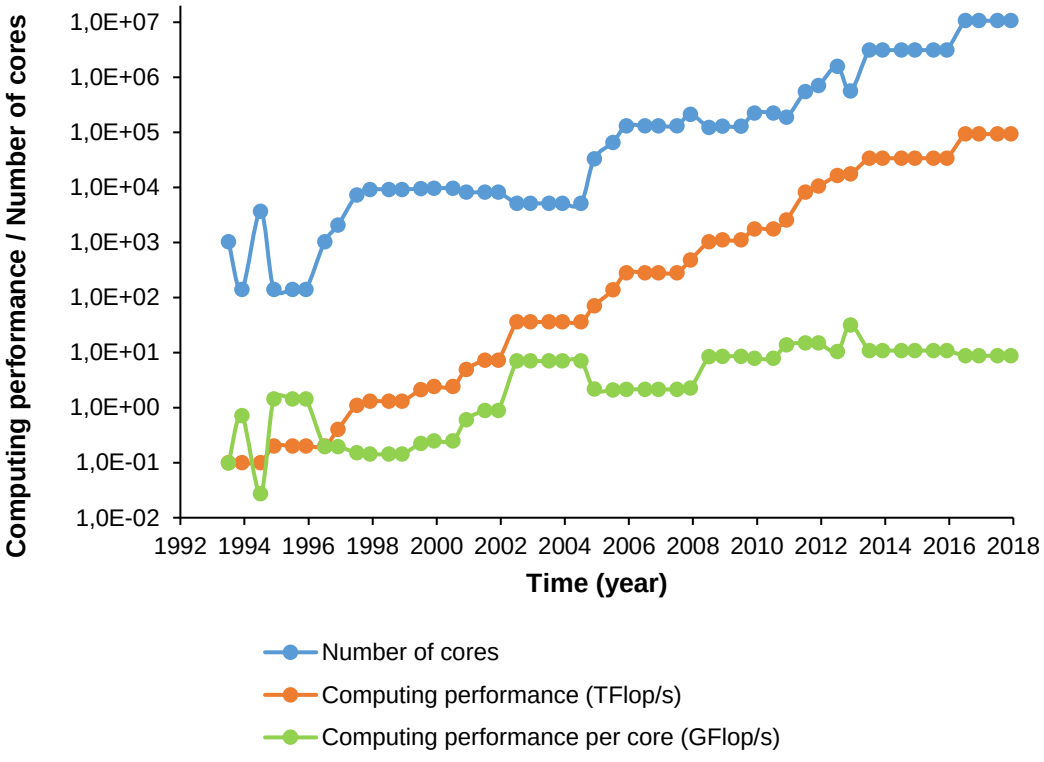


Figure I-5: Development of the computational power of the world's fastest supercomputer
The computational performance of the fastest super computer on the Top 500 list according to the LINPACK benchmark (R_{max} value) is plotted over time (in TFlops; orange). The number of computing cores is plotted in blue and the computational power per core is shown in green (values in GFlops). While the computational performance increases exponential, the performance per computing unit stagnated. As a consequence, modern supercomputers consist of a tremendous number of computing cores and are highly distributed systems. For comparison, modern quad core desktop machines have a typical performance of about 50 GFlop/s. Data was obtained from https://www.top500.org/statistics/sublist/ (as of January 2018).

Current trends of studying large biomolecules and biomolecular complexes as well as investigating dynamic systems, introduced new requirements on the methodology of biomolecular NMR spectroscopy. Overcoming these challenges has been a major concern of recent research. For example, NMR studies on large RNAs suffer from a significant signal overlap as well as from a reduced density of protons[184, 185]. As a second example, studies on large proteins and complexes suffer from low signal-to-noise ratios and significant spectral overlaps[28, 31, 32, 172]. Thirdly, the increasing interest in IDPs leads to a new paradigm of describing the structure of highly flexible systems by means of structural ensembles[174, 175]. Although tremendous advances have been made regarding the described issues, extending the methodology of biomolecular NMR spectroscopy is an essential research focus. Constantly improving the existing techniques drives the field of structural biology and allows to study more and more complex systems.

II Aim of this Thesis

Current methods in biomolecular NMR spectroscopy rely on the local distance derived from NOEs, long range distances derived from PREs, dihedral angles and orientations of bond vectors as well as information regarding secondary structure elements and hydrogen bonding. These structural restraints form a powerful set of information that is eventually combined by computational methods.

As the interest of structural biologist shifts towards larger and more complex systems, the existing methods have to be extended. Challenging systems such as intrinsically disordered proteins (IDPs), RNAs, large proteins and protein complexes dramatically increase the conformational space that needs to be considered. Therefore, the number of structural restraints must not only be increased, but instead new types of restraints are required that are well-suited to analyze systems with a complex conformational space.

The goal of this thesis is to address this challenge by developing and establishing NMR methods that push the limits of NMR structure determination. In particular, solvent accessibility data obtained from solvent PRE experiments are to be investigated and novel NMR methods are to be developed, tested and established by exploiting the full potential of solvent accessibility data.

To this end, the thesis covers the following key aspects. First, NMR experiments for the acquisition of solvent PRE data are to be extended and established not only for proteins but also for IDPs and RNAs in particular. Secondly, data processing and analysis need to be improved and streamlined to provide an accurate and efficient data processing pipeline. Thirdly, current state-of-the-art computational methods are to be extended to allow the inclusion of solvent PRE data. These new methods are to be optimized, tested and established to provide powerful tools that allow to tackle challenging system by means of biomolecular NMR spectroscopy.

III Published Results: Novel Methods based on Paramagnetic NMR Spectroscopy

This chapter presents different papers demonstrating the applicability of solvent PRE data to overcome the challenges of a variety of studies, including structure prediction of proteins, structure determination and refinement of RNAs, characterization of IDPs as well as solid state NMR studies on protein complexes. Eventually, a paper is presented that shows how paramagnetic spin labels can be used to enhance NMR studies on IDPs.

III.1 Prediction of Protein Structure Using Surface Accessibility Data

Christoph Hartlmüller, Christoph Göbl and Tobias Madl

Angew Chem Int Ed Engl. 2016 Sep 19; 55(39):11970-11974
doi: 10.1002/anie.201604788

In this work, a novel approach for structure prediction using solvent PRE data in combination with the CS-Rosetta framework is presented. Chemical shift-based structure prediction using CS-Rosetta is a popular approach in structural studies. In the extended approach presented in this work, a new energy function was developed that allows to use solvent accessibility data as an input to the structure prediction algorithm (see Figure III-1).

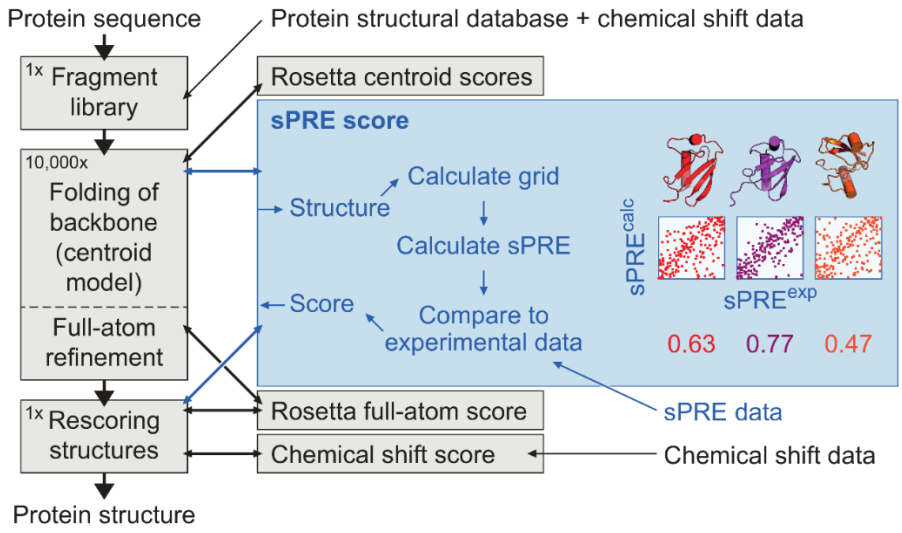


Figure III-1: Concept of Rosetta and the new solvent PRE energy function
Starting from the protein sequence, the CS-Rosetta protocol involves a preparation step to find the conformations of the fragments ("Fragment library"). Next, the protein is folded using the Monte Carlo algorithm and the obtained ensemble is eventually rescored and analyzed to find the native structure. During these stages different scores are used (boxes on the right side). A new energy function for solvent accessibility data was developed (blue box) which can be applied to the simplified centroid as well as to the full-atom model of the protein. Figure was extracted from the presented paper. For reprint permission see Chapter VII .

Based upon an efficient and optimized energy score for solvent accessibility data, the new approach significantly improved the convergence and accuracy of the structure prediction. In particular, the solvent PRE data were shown to drive the algorithm in the early stages of folding the protein. The Rosetta energy function was found to be the dominating driving force at later stages of the folding process in which local adjustments are performed and the refined full-atom structure is obtained. Using the experimental as well as the simulated solvent PRE

dataset, the benefits of the new approach were demonstrated for several proteins (see Figure III-2) and are a promising approach for future NMR studies.

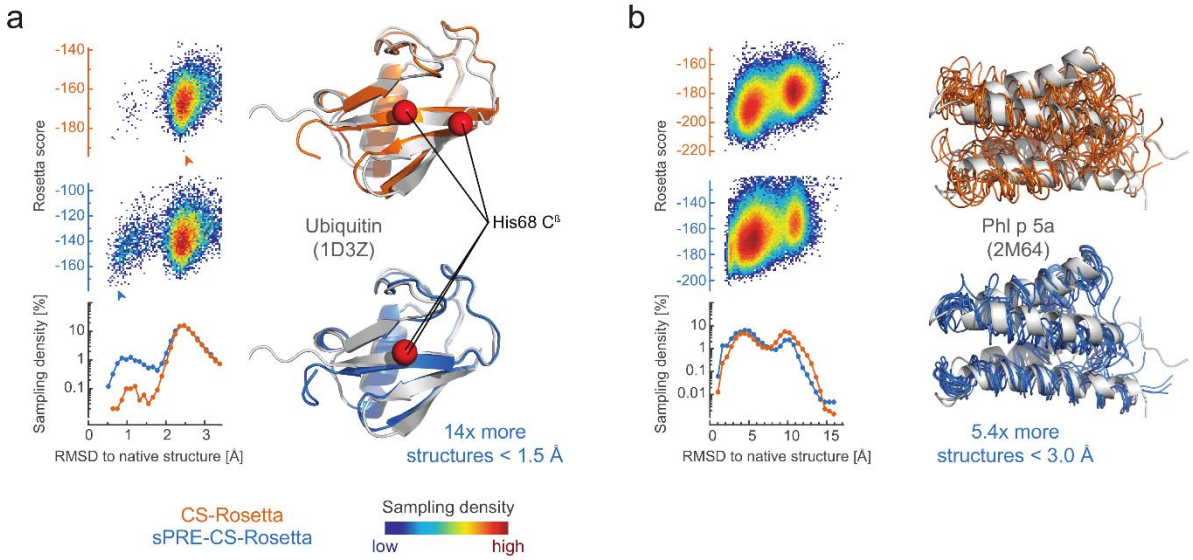


Figure III-2: Benefits of using solvent PRE data
Solvent PRE data facilitates the structure determination process of Ubiquitin (a) and the C-terminal domain of Phl p 5a (b). The obtained structures in the absence (orange) and presence (blue) of solvent PRE data are plotted as a heat map showing the distribution of the Rosetta score over the quality (RMSD) of the models. The improved sampling is further shown in the histogram below the heat maps. (**a**) The Ubiquitin models of the best scored structures in the absence (orange) and presence (blue) of solvent PRE are depicted and compared to the native structure deposited in the PDB (1D3Z, gray). (**b**) The 10 best scored models of Phl p 5a in the absence (orange) and presence (blue) of solvent PRE are depicted and compared to the native structure deposited in the PDB (2M64, gray). Figure was extracted from the presented paper. For reprint permission see Chapter VII .

C.H. and T.M. planned this study. C.G. prepared the protein samples and all authors performed NMR experiments. C.H. developed the analysis of the solvent PRE data, implemented and optimized the solvent PRE energy function for Rosetta, performed the computations for testing the new approach. C.H. analyzed and visualized the test results and, together with T.M., wrote the manuscript.

III.2 RNA Structure Refinement using NMR Solvent Accessibility Data

Christoph Hartlmüller*, Johannes C. Günther*, Antje C. Wolter, Jens Wöhnert, Michael Sattler and Tobias Madl

Sci Rep. 2017 Jul 14;7(1):5393.
doi: 10.1038/s41598-017-05821-z
* Shared first author

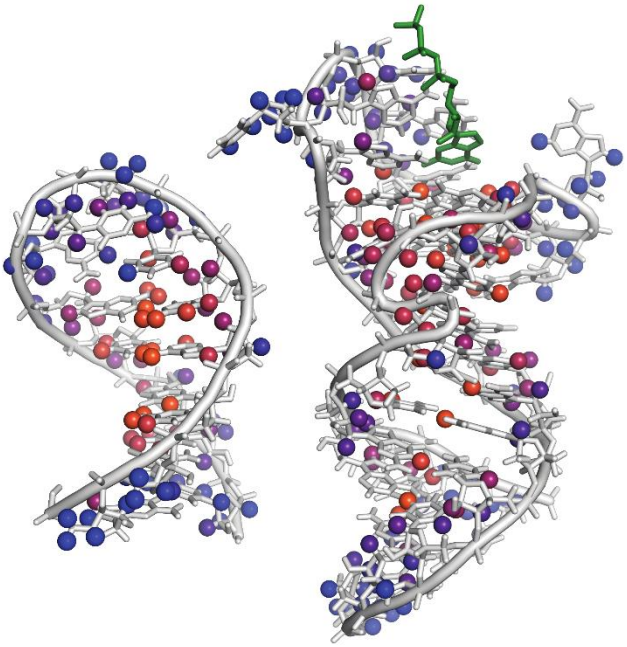


Figure III-3: Solvent PRE data of RNA
The observed ^{1}H proton solvent PRE data is shown as spheres for the UUCG tetraloop (left) and the GTP-bound GTP aptamer (right) where blue corresponds to large and orange to small solvent PRE values. Figure was extracted from the presented paper. For reprint permission see Chapter VII .

In this work, the benefits of solvent PRE data for RNA structure determination were demonstrated. Determination of RNA structure remains a major challenge in NMR spectroscopy which is mainly caused by severe signal overlap as well as a small density of protons leading to a reduced number of experimental restraints. To increase the potential number of restraints, the applicability of solvent PRE data to structural NMR studies on RNA was investigated. To this end, NMR experiments for measuring solvent PRE data of proteins were adapted and optimized for RNA and enabled the acquisition of accurate solvent PRE data for ^{13}C and ^{15}N bound protons (see Figure III-3).

To demonstrate the benefits of solvent PRE data, structure determination protocols using Xplor-NIH were performed. To thoroughly test the performance of the solvent PRE data, additional NMR restraints such as NOE, J-couplings, H-bonds and RDCs were obtained from literature. With these sets of data, several test cases ranging from sparse data to strongly-constrained computations were constructed and executed. The impact of solvent PRE data was examined for all cases and thus provided a clear picture of the benefits of solvent accessibility data. In particular, strong benefits were observed in the case of sparse data

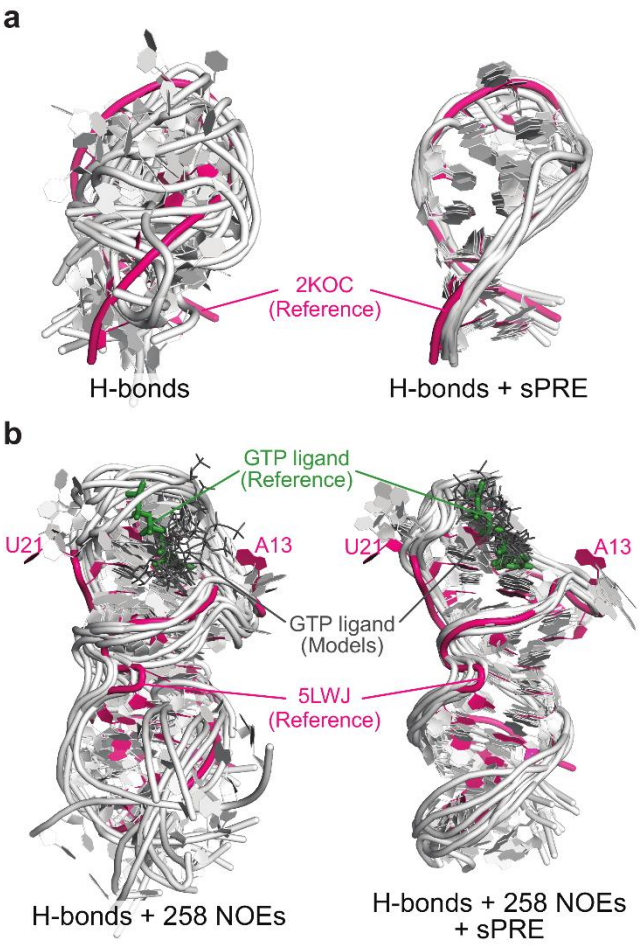


Figure III-4: Benefits of solvent PRE data for RNA structure determination

The 10 best structural models obtained for the UUCG tetraloop (a) and the GTP-bound GTP aptamer (b) in the absence (left) and presence (right) of solvent PRE data are shown. Reference s structures are shown in magenta and the set of constraint used for the computations is indicated below the models. Figure was extracted from the presented paper. For reprint permission see Chapter VII .

examples. For example, using only H-bond information in combination with solvent PRE data was sufficient to determine the structure of the UUCG tetraloop (see Figure III-4a). Moreover, significantly improved models were obtained in the case of the GTP-bound GTP aptamer which demonstrated that solvent PRE enhances the structure prediction algorithm also in the presence of other NMR restraints such as NOEs (see Figure III-4b).

C. H. and T. M. designed the studies. J. G., A. W., J. W. and M. S. prepared the RNA samples. C. H. and T. M. established and tested the NMR experiments. C. H. and J. G. performed the NMR titrations. C. H. processed and analyzed the NMR data, performed the structure prediction computation and analyzed the obtained models. C. H. and T. M. performed the prediction of sPRE data. C. H. wrote the manuscript and all authors have given critical feedback and approved the final version of the manuscript.

III.3 NMR Characterization of Solvent Accessibility and Transient Structure in Intrinsically Disordered Proteins

Christoph Hartlmüller*, Emil Spreitzer*, Christoph Göbl, Fabio Falsone and Tobias Madl

J Biomol NMR. 2019 Jul;73(6-7):305-317
doi: 10.1007/s10858-019-00248-2
* Shared first author

NMR spectroscopy is a powerful method to study highly dynamic IDPs. Typical applications include the detection of residual structure elements, identifying the role of post-translation modifications as well as characterizing biomolecular interactions. The detection of residual secondary structural elements can be efficiently performed by chemical shift analysis and dynamical properties of IDPs can be analyzed using relaxation data. However, the detection of residual tertiary structure and long-range interactions requires chemically-linked spin labels. This paper describes an alternative method to study residual tertiary structure of IDPs by leveraging NMR solvent PRE data.

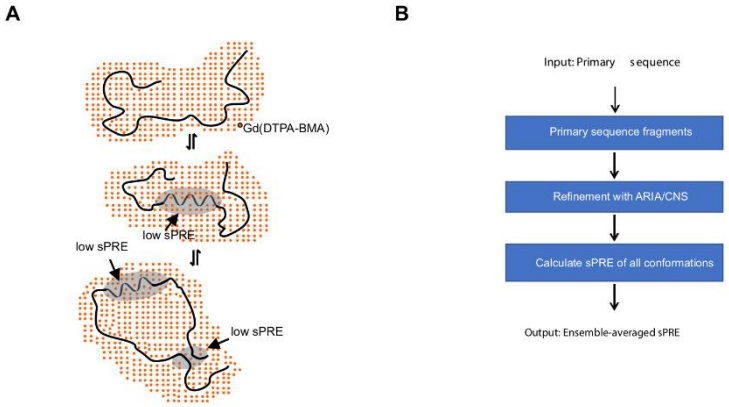


Figure III-5: Detecting residual structure elements of IDPs using solvent PRE data
(**A**) Transient secondary structure elements as well as intramolecular protein-protein contacts can be detected by solvent PRE data as those sites show a reduced solvent accessibility. Using the soluble paramagnetic compound Gd(DTPA-BMA), the reduced solvent accessibility can be observed by a reduced solvent PRE. (**B**) Prediction of the solvent PRE for the target IDP is performed using a fragment-based ensemble approach (see text for details). Figure was extracted from the presented paper. For reprint permission see Chapter VII .

Solvent PRE data encode distance-to-surface information and therefore allow to detect residual structural elements as the solvent accessibility is reduced at those sites of the IDP (see Figure III-5A). Briefly, the approach presented in this work relies on computing the solvent PRE of the target IDP assuming the absence of any residual structural elements and eventually comparing the predicted reference data to the experimental solvent PRE data. Computing the reference solvent PRE data for the target IDP is done using a fragment-based ensemble approach: structural ensembles are computed using ARIA/CNS for every 5-mer, flanked by triple-Alanine on both termini. The solvent PRE is then obtained by computing the solvent PRE for every single conformation and averaging the data across the ensemble (see [186] for details).

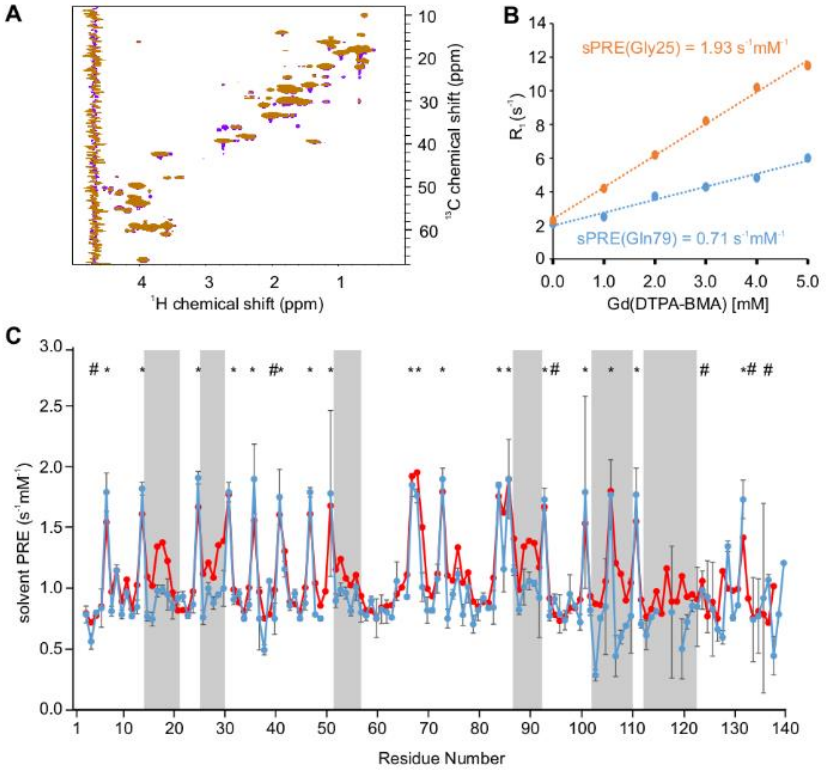


Figure III-6: Solvent PRE data of α-synuclein
(**A**) ^{15}N HSQC spectra in the absence (magenta) and presence (orange) of Gd(DTPA-BMA) show no interaction between the IDP and the paramagnetic agent. (**B**) Linear fits to obtain solvent PRE data are shown for Gly^{25} (orange) and Gln^{79} (blue), representing sites with strong and weak solvent PRE, respectively. (**C**) Predicted (red) and experimental (blue) solvent PRE data of aliphatic protons are plotted with error bars. Gray boxes indicate sites with reduced experimental solvent PRE. Bulky sidechains are indicated with # and Glycine residues are marked with *. Figure was extracted from the presented paper. For reprint permission see Chapter VII .

To demonstrate the effectiveness of the proposed method, measured and predicted solvent PRE data were analysed for three IDP model systems (FOXO4, p53 and α-synuclein; results for α-synuclein as shown in Figure III-6). Solvent PRE data predicted for disordered regions of all three IDPs correlate well with experimental data, indicating that the fragment-based ensemble approach used for prediction is well suited for IDPs. Next, proteins sites were analysed for which a reduced solvent PRE was observed compared to the prediction (compare Figure III-6C). In the case of p53, sites identified by solvent PRE data correspond to sites with a α-helical propensity or a reduced flexibility. In the case of α-synuclein, previously described intramolecular interaction align with the sites identified in this study.

Contributions of C. H. include preparing and performing the NMR experiments, implementing and testing the ensemble-based approach for predicting the solvent PRE as well as analyzing the predicted and experimental data.

III.4 Characterization of Protein-Protein Interfaces in Large Complexes by Solid-State NMR Solvent Paramagnetic Relaxation Enhancements

Carl Öster, Simone Kosol, **Christoph Hartlmüller**, Jonathan M. Lamley, Dinu Iuga, Andres Oss, Mai-Liis Org, Kalju Vanatalu, Ago Samoson, Tobias Madl and Józef R. Lewandowski

J Am Chem Soc. 2017 Sep 6;139(35):12165-12174.
doi: 10.1021/jacs.7b03875

In this work, the application of solvent PRE data in studies using solid state NMR is demonstrated. The system of interest is the small protein GB1 and its complex formed with immunoglobulin G (IgG). In particular, the protein-protein interaction was investigated since IgG was known to contain two binding sites, positioned in the Fc fragment as well as on the Fab fragment (see Figure III-7).

Figure III-7: Model of the complex of GB1 (blue) with IgG
In this model, the Fc fragment of Fc binds to the α-helix of GB1 (left), whereas the Fab fragments binds to the β-sheets of GB1 (right). Figure was extracted from the presented paper. For reprint permission see Chapter VII .

To unveil the structure of the complex, solvent accessibility data obtained from solid state NMR was used. To this end, solvent PRE data was recorded from the isolated GB1 in solution and in the crystal, from the precipitated GB1:IgG complex. To complete the data sets, solvent PRE data were predicted based on different alternative models of the GB1:IgG complex.

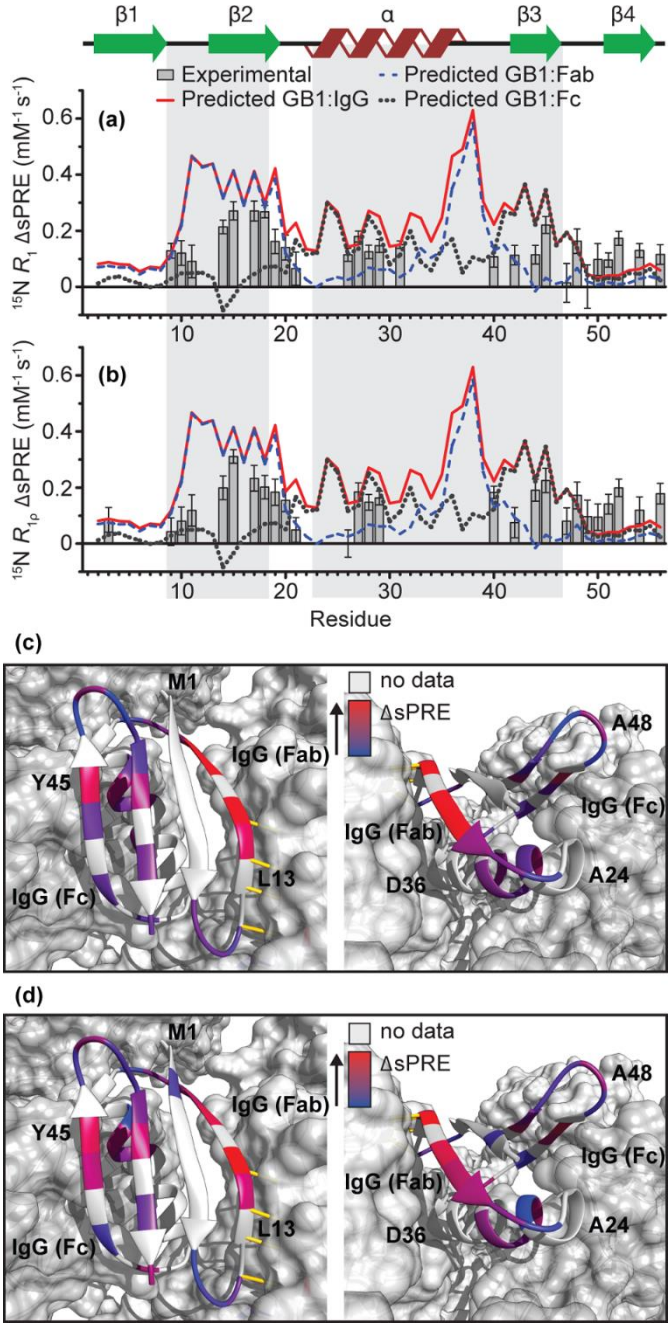


Figure III-8: Using solvent PRE data to unveil the structure of the complex
^{15}N solvent PRE data derived from R_1 (a and c) and $R_{1\rho}$ (b and d) relaxation rates are compared to predicted solvent PRE based on different models of the complex. The difference of the solvent PRE (ΔsPRE) between the complex and the free form is shown in a diagram (a and b) as was plotted onto the structure of GB1 (c and d). Figure was extracted from the presented paper. For reprint permission see Chapter VII .

Based upon previous studies that investigated interactions between GB1 and single fragments of IgG, potential candidate structures of the GB1:IgG complex were constructed. These models were then used to predict the solvent PRE pattern of GB1. These patterns were then compared to the observed data (see Figure III-8) and a final model of the complex was suggested.

Together with C. Ö., S. K., T. M. and J. L., C. H. developed the strategy how to make efficient use of the solvent PRE data. Together with C. Ö. and S. K., C. H. constructed the candidate models used to predict the solvent PRE data. C. H. wrote the program for computing the solvent PRE data, performed the computation and wrote the corresponding paragraphs in the manuscript regarding the prediction of the solvent PRE data.

III.5 Increasing the Chemical-Shift Dispersion of Unstructured Proteins with a Covalent Lanthanide Shift Reagent

Christoph Göbl, Moritz Resch, Madeleine Strickland, **Christoph Hartlmüller**, Martin Viertler, Nico Tjandra and Tobias Madl

Angew Chem Int Ed Engl. 2016 Nov 14;55(47):14847-14851
doi: 10.1002/anie.201607261

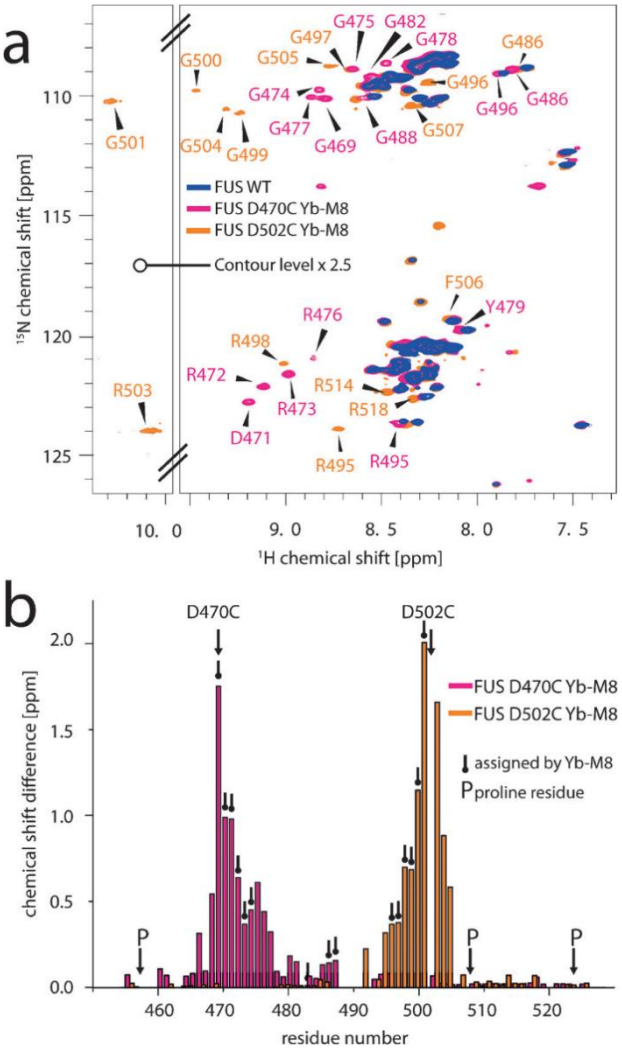


Figure III-9: Ytterbium spin labels increase chemical shift dispersion
NMR spectra (a) of the unstructured protein FUS are shown in the absence (blue) and the presence of a spin label attached to Fus^{470} (magenta) or Fus^{502} (orange). Shifted residues due to PCSs are indicated in the spectra and plotted as a bar chart (b). Figure was extracted from the presented paper. For reprint permission see Chapter VII .

This work demonstrates an approach to enhance the studies of IDPs by means of NMR spectroscopy. In the context of NMR, IDPs are advantages as they give rise to good signal-to-noise ratios. However, studies are often hindered by severe signal overlaps. To address this issue, pseudo chemical shifts (PCSs) induced by covalently attaching Lanthanide spin labels to the protein were exploited to enhance chemical shift dispersion which in turn facilitates assignments and quantification of NMR signals. Attaching the spin label was enabled by introducing Cysteine residues in the proximity of the sites of interest. In particular, commonly-occurring repetitions of Glycine residues were resolved and assigned individually.

In this study, C. H. assisted in the recording and processing of NMR experiments as well as performed the processing of relaxation data.

IV Published Results: Structural Biology Studies

This chapter presents different published studies in the field of structural biology. In all of these studies, NMR spectroscopy as well as computational methods were utilized to leverage the understanding of the corresponding system of interest.

IV.1 Structural Basis of Nucleic-Acid Recognition and Double-Strand Unwinding by the Essential Neuronal Protein Pur-alpha

Janine Weber, Han Bao, **Christoph Hartlmüller**, Zhiqin Wang, Almut Windhager, Robert Janowski, Tobias Madl, Peng Jin and Dierk Niessing

Elife. 2016 Jan 8;5. pii: e11297
doi: 10.7554/eLife.11297

This structural and functional study revealed the molecular details of the Purine-rich element-binding protein A (Pur-alpha) with regard to RNA/DNA binding. Pur-alpha is an essential regulator of mRNA localization that is associated with neurodegenerative diseases. It consists of three domains where domain I and II are involved in the binding of nucleotides and domain III is associated with the homo-dimerization of the protein. Using X-ray crystallography, the ssDNA-bound structure of Pur-alpha was revealed. NMR titration experiments showed that RNA is recognized in the same manner as DNA.

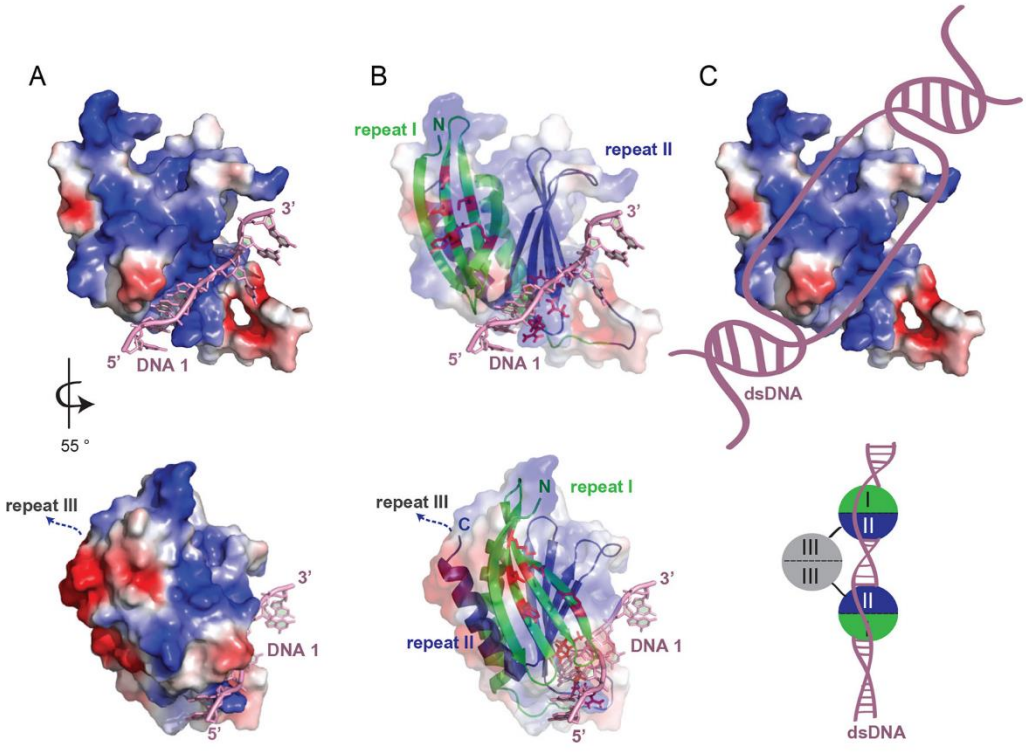


Figure IV-1: Model of Pur-alpha's DNA unwinding mechanism
(**A** and **B**) Crystal structures of Pur-alpha domain I and II (electrostatic surface shown) bound to ssDNA (pink) are depicted. (**C**) Models of the binding of dsDNA (top) and dimerization of Pur-alpha in the dsDNA-bound state. Figure was extracted from the presented paper. For reprint permission see Chapter VII .

Together with J. W., C. H. performed NMR titration of Pur-alpha domain I and II with different construct of ssDNA and RNA. C. H. processed the spectra, created the figures associated with the NMR titration and wrote the corresponding method section in the manuscript.

IV.2 The Redox Environment Triggers Conformational Changes and Aggregation of hIAPP in Type II Diabetes

Diana C. Rodriguez Camargo, Konstantinos Tripsianes, Katalin Buday, Andras Franko, Christoph Göbl, **Christoph Hartlmüller**, Riddhiman Sarkar, Michaela Aichler, Gabriele Mettenleiter, Michael Schulz, Annett Böddrich, Christian Erck, Henrik Martens, Axel Karl Walch, Tobias Madl, Erich E. Wanker, Marcus Conrad, Martin Hrabě de Angelis and Bernd Reif

Sci Rep. 2017 Mar 13;7:44041
doi: 10.1038/srep44041

In this interdisciplinary study, the structure of human islet amyloid polypeptide (hIAPP) was investigated and a model regarding the formation of cell-toxic fibrils was established. Due to its Cysteine residues positioned at the N-terminus, hIAPP can be found in an oxidized and reduced state. The redox state is crucial as it affects the tendency of hIAPP to form fibrils with the reduced form being more likely to do so (see Figure IV-2). *In vivo* experiments with mice carrying the human hIAPP showed fibril formations in the pancreatic islets as well as deficient cell morphology. These findings suggest an important role of hIAPP and therefore *in vitro* studies were performed in which the formation of fibrils was observed with several methods including cryo-EM as well as fluorescence assays.

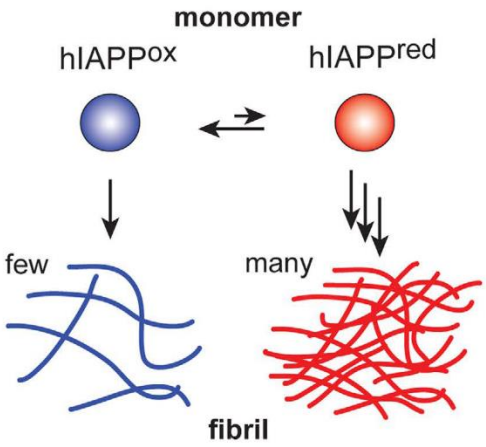


Figure IV-2: Model of fibril formation
The reduced form of hIAPP was shown to possess a higher tendency to form cell-toxic fibrils. Figure was extracted from the presented paper. For reprint permission see Chapter VII .

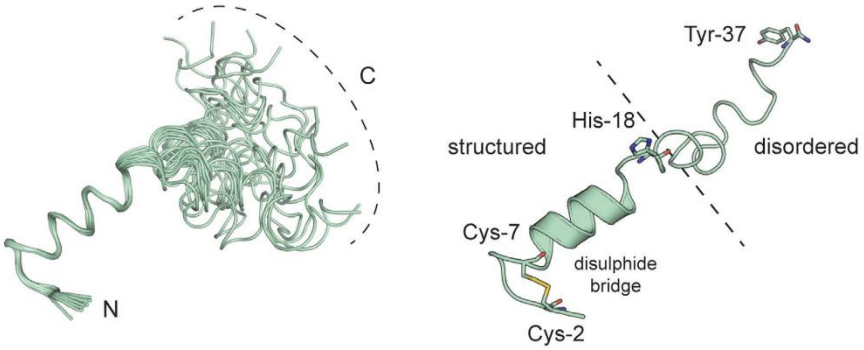


Figure IV-3: Structure model of oxidized hIAPP
The rigid and α-helical N-terminus and the flexible C-terminus are indicated. Figure was extracted from the presented paper. For reprint permission see Chapter VII .

To unveil the structural details of the aggregation process, NMR studies of the oxidized hIAPP were performed. The obtained structure model showed the presence of an N-terminal α-helix (see Figure IV-3) while the C-terminus remains flexible. These findings are further supported by NMR relaxation and hetNOE as well as circular dichroism (CD) measurements. The structural model explains the reduced tendency of the oxidized hIAPP to form fibrils since the presence of the rigid N-terminal helix is in contradiction with the formation of β-sheet-containing fibrils.

Together with D. R. and C. G., C. H. performed NMR experiments and analyzed the data. C. H. assisted in the NMR assignment, the analysis of relaxation data and the computation of the structure.

IV.3 Molecular Basis for Asymmetry Sensing of siRNAs by the Drosophila Loqs-PD/Dcr-2 Complex in RNA Interference

Jan-Niklas Tants, Stephanie Fesser, Thomas Kern, Ralf Stehle, Arie Geerlof Christoph Wunderlich, Michael Juen, **Christoph Hartlmüller**, Romy Böttcher, Stefan Kunzelmann, Oliver Lange, Christoph Kreutz, Klaus Förstemann and Michael Sattler

Nucleic Acids Res. 2017 Dec 1;45(21):12536-12550
doi: 10.1093/nar/gkx886

In this study, the interaction between siRNA and an alternative loading complex (RLC) formed by Loquacious-PD (Loqs-PD) and Dicer-2 (Dcr-2) are investigated. RLCs are an essential part of the RNA interference pathway since they detect the presence of dsRNA and initiate the formation of the RNA induced silencing complex (RISC). The binding of the complex to RNA is performed by two domains double-stranded RNA binding domains (dsRBDs) of Loqs-PD.

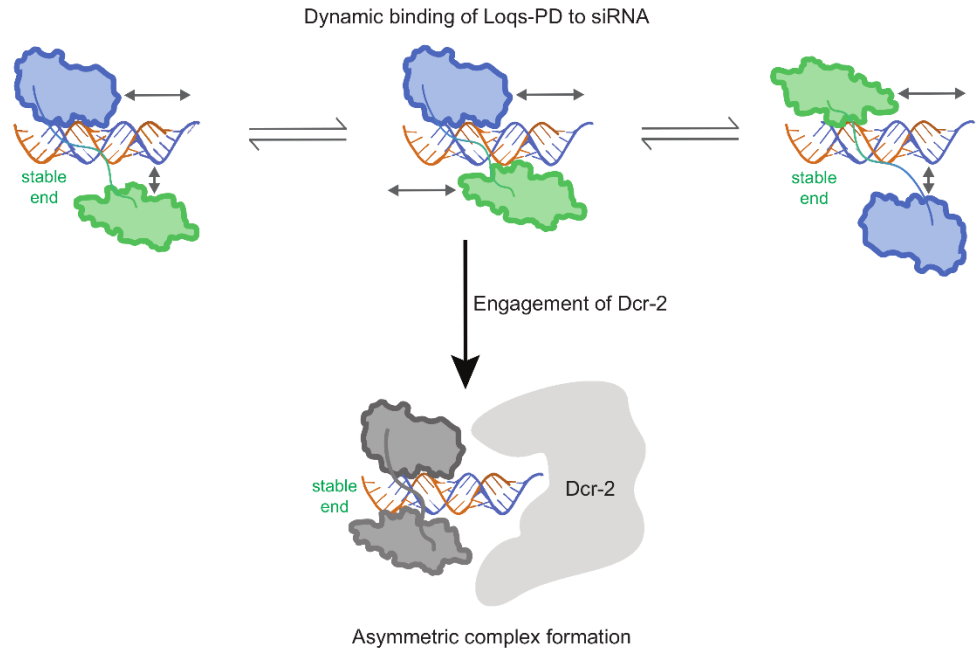


Figure IV-4: Model of the siRNA recognition and binding of the Loqs-PD:Dcr-2 complex
Both dsRBD of Loqs-PD are indicated (blue and green) and the siRNA is shown as helix (blue and orange strands). Both dsRBD scan the siRNA and bind independently. The discrimination of the RNA strand for the RISC complex is associated with the thermodynamic siRNA asymmetry. As both dsRBD slide along the siRNA, a slight preference for binding to the more stable end of the RNA, is amplified by the binding of Dcr-2 to the weak end. Figure was extracted from the presented paper. For reprint permission see Chapter VII .

Using chemical shift and NOE data of the isolated domains the structures of both domains were determined using the Rosetta framework. With additional CSD and PRE data of the complex a model of the RNA binding of the entire complex is proposed (see Figure IV-4).

C. H. contributed to the study by performing the structure calculation of the dsRBDs.

V Unpublished Research: Solvent PRE Data of Exchangeable Protons

Solvent PRE data provide valuable information for structural characterization of biomolecules[118, 121, 122]. In Chapter III , several studies were presented demonstrating the application of solvent PRE to various challenges in structural biology. As shown in these studies, solvent PRE data can be obtained and analyzed very efficiently. Nevertheless, the established methods for measuring and analyzing solvent PRE data neglect the effects of chemical exchange processes, in particular the exchange of water protons with the biomolecule of interest. As a consequence, solvent PRE data of exchangeable protons often show a systematic deviation. This aspect has been discussed in the context of imino protons in RNAs [187]. This chapter presents unpublished work aiming for an improved understanding of exchange processes in the context of solvent PRE. Moreover, different strategies will be suggested to leverage experimental methods to account for exchange processes.

V.1 Relaxation Rates Obtained by Saturation Recovery

Measuring solvent PREs most-commonly involves the acquisition of proton R_1 relaxation rates in the presence of different concentrations of a paramagnetic agent[118, 128, 187]. Proton R_1 relaxation rates are measured using a saturation recovery scheme (see Figure V-1). In short, the magnetization of all protons is dephased during the initial saturation block. During the subsequent recovery delay, the system is relaxed to build-up z-magnetization for a specific time period, which is exponentially incremented between experiments. After the recovery delay, the z-magnetization is measured by common read-out experiments such as ^{13}C or ^{15}N-HSQCs. By incrementing the recovery delay, a build-up over time of the z-magnetization is obtained for every proton and the corresponding R_1 relaxation rates are extracted by fitting a mono-exponential relaxation model

$$I_z\ (t) = I_{z,eq} \cdot \left(1 - e^{-t \cdot R_1}\right) \tag{20}$$

where t is the duration of the recovery delay, R_1 is the desired proton R_1 relaxation rate, $I_{z,eq}$ is the z-magnetization of the system in equilibrium and $I_z\ (t)$ corresponds to the measure data point for a given recovery delay t. It should be noted, that the measured signal intensity of the

saturation recovery experiment is proportional to the proton z-magnetization right before the start of the read-out block. Therefore, equation (20) and all following equations refer to the z-magnetization I_z rather than the measured signal.

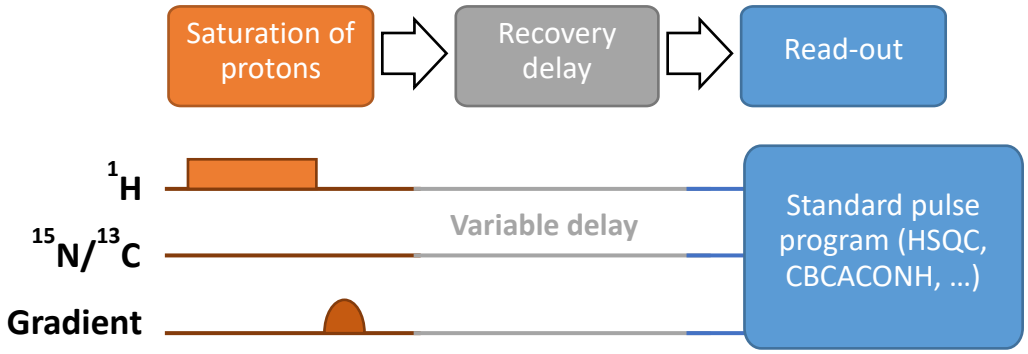


Figure V-1: Saturation recovery scheme to measure proton T_1 relaxation
The saturation recovery scheme allows to measure T_1 relaxation times. The saturation block (orange), here consisting of a trim pulse and a magnetic field gradient, is used to saturate proton magnetization. In the recovery delay phase (gray) the system builds up z-magnetization as it relaxes towards its equilibrium state. Eventually, the proton z-magnetization right after the recovery delay is determined by running a read-out sequence such as an HSQC.

To obtain the solvent PRE, a series of saturation recovery experiments is acquired while increasing the concentration of the soluble paramagnetic agent. In the case of non-exchangeable protons (such as H^{α} and H^{β} in the case of proteins or H6 and H5′ in the case of RNAs) the proton R_1 relaxation rate increases linearly with the concentration (see top row in Figure V-2). The solvent PRE can therefore directly be obtained from the slope of the plot.

V.2 Solvent Exchange in the Context of Solvent PRE

Solvent PRE data for protons that are subject to exchange processes with solvent protons (such as amide H^N of the protein backbone or imino and amino protons of RNAs) show a deviation of the linear dependency at low concentrations of the paramagnetic compound (see bottom row in Figure V-2). In the following section, a model will be derived to understand the nature of this deviation.

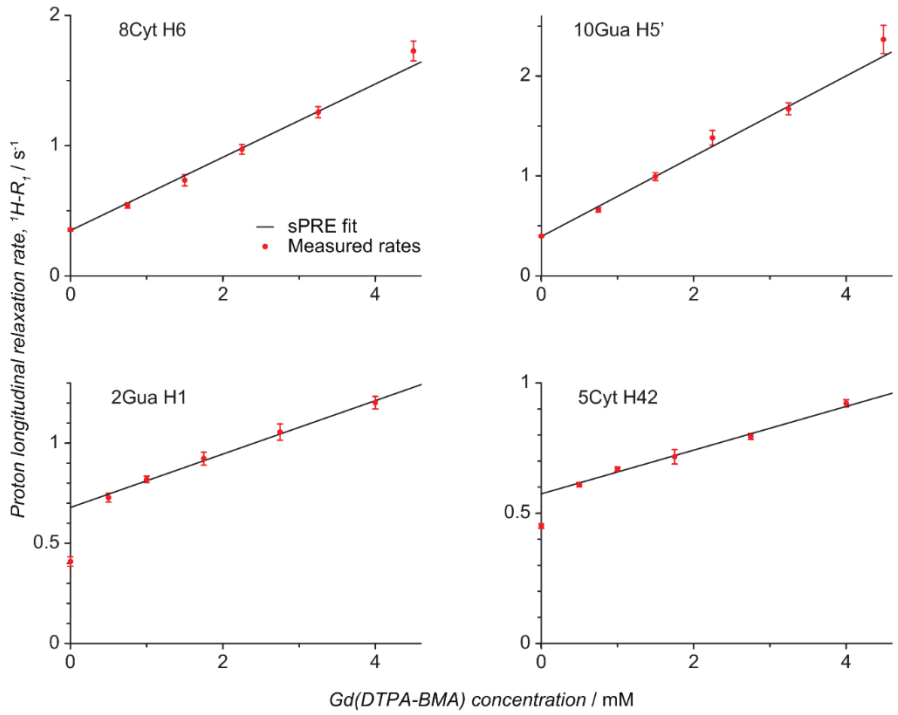


Figure V-2: Typical solvent PRE data for protons in RNA
Proton R1 relaxation rates are plotted as a function of the concentration of the soluble paramagnetic compound Gd(DTPA-BMA). Data for non-exchangeable carbon-bound protons are shown in the top row, while the lower plots show results for exchangeable nitrogen-bound protons. Figure was extracted from [187]. For reprint permission see Chapter VII .

To understand the observed deviation for exchangeable protons, the impact of exchange processes on the measurement of solvent PREs needs to be derived. Based on the presented saturation recovery scheme, a model was formulated to simulate the influence of the exchange with solvent protons. In the following, the z-magnetization after the recovery delay t of the solvent protons $I_z^{H_2O}(t)$ and of exchangeable protein amide protons $I_z^{HN}(t)$ are considered. To this end, differential equations for the evolution of the magnetization over time, $I_z^{H_2O}(t)$ and $I_z^{HN}(t)$, are formulated. A linear increase of the relaxation R_1 with the concentration of the paramagnetic agent c_{Gd} is assumed

$$R_1(c_{Gd}) = R_{1,0} + c_{Gd} \cdot \Gamma_{sPRE} \tag{21}$$

where Γ_{sPRE} is the solvent PRE of the corresponding proton and $R_{1,0}$ is the R1 relaxation rate in the absence of the paramagnetic compound. The differential equation for water can therefore be written as

$$\frac{dI_z^{H_2O}(t)}{dt} = \left(I_{z,eq}^{H_2O} - I_z^{H_2O}(t)\right) \cdot R_1^{H_2O}(c_{Gd}) = \left(I_{z,eq}^{H_2O} - I_z^{H_2O}(t)\right) \cdot \left(R_{1,0}^{H_2O} + c_{Gd} \cdot \Gamma_{sPRE}^{H_2O}\right) \tag{22}$$

which can be solved to obtain the mono-exponential build-up of the magnetization of water

$$I_z^{H_2O}(t) = I_{z,eq}^{H_2O} \cdot \left(1 - e^{-t \cdot R_1^{H_2O}(c_{Gd})}\right)$$
$$= I_{z,eq}^{H_2O} \cdot \left(1 - e^{-t \cdot \left(R_{1,0}^{H_2O} + c_{Gd} \cdot \Gamma_{sPRE}^{H_2O}\right)}\right) \quad (23)$$

Consequently, the fraction of un-saturated solvent protons can be written as

$$\frac{I_z^{H_2O}(t)}{I_{z,eq}^{H_2O}} = 1 - e^{-t \cdot R_1^{H_2O}(c_{Gd})} = 1 - e^{-t \cdot \left(R_{1,0}^{H_2O} + c_{Gd} \cdot \Gamma_{sPRE}^{H_2O}\right)} \quad (24)$$

The differential equation for $I_z^{HN}(t)$ is more complicated as it includes additional processes and can be written as

$$\frac{dI_z^{HN}(t)}{dt} =$$
$$\left(I_{z,\,eq}^{HN} - I_z^{HN}(t)\right) \cdot \left(R_1^{HN}(c_{Gd}) + k_{ex} \cdot \frac{I_z^{H_2O}(t)}{I_{z,eq}^{H_2O}}\right) - I_z^{HN}(t) \cdot k_{ex} \cdot \left(1 - \frac{I_z^{H_2O}(t)}{I_{z,eq}^{H_2O}}\right) \quad (25)$$

where k_{ex} is the exchange rate of the amide proton and the surrounding water. The color code marks the different processes:

- $\left(I_{z,\,eq}^{HN} - I_z^{HN}(t)\right) \cdot R_1^{HN}(c_{Gd})$

  The red term describes the R_1 relaxation of the amide proton with the rate depending on the concentration of the paramagnetic agent.

- $\left(I_{z,\,eq}^{HN} - I_z^{HN}(t)\right) \cdot k_{ex} \cdot \frac{I_z^{H_2O}(t)}{I_{z,eq}^{H_2O}}$

  The orange term accounts for the build-up of z-magnetization by chemical exchange with unsaturated water.

- $-I_z^{HN}(t) \cdot k_{ex} \cdot \left(1 - \frac{I_z^{H_2O}(t)}{I_{z,eq}^{H_2O}}\right)$

  The blue term describes the reduction of z-magnetization due to chemical exchange with a saturated solvent proton. It should be noted that the term $1 - \frac{I_z^{H_2O}(t)}{I_{z,eq}^{H_2O}}$ describes the fraction of saturated water.

Assuming that the water magnetization is not affected by the protein, equation (24) can be directly used to solve equation (25). Using $I_z^{HN}(t = 0) = 0$, this results in

$$I_z^{HN}(t) =$$

$$= I_{z,\,eq}^{HN} \cdot \frac{k_{ex} + R_1^{HN}(c_{Gd}) - R_1^{H_2O}(c_{Gd}) - k_{ex} \cdot e^{-R_1^{H_2O}(c_{Gd})\cdot t} + \left(R_1^{H_2O}(c_{Gd}) - R_1^{HN}(c_{Gd})\right) \cdot e^{-R_1^{HN}(c_{Gd})\cdot t - k_{ex}\cdot t}}{k_{ex} + R_1^{HN}(c_{Gd}) - R_1^{H_2O}(c_{Gd})}$$

$$= I_{z,\,eq}^{HN} \cdot \left(1 - \frac{k_{ex} \cdot e^{-R_1^{H_2O}(c_{Gd})\cdot t} + \left(R_1^{HN}(c_{Gd}) - R_1^{H_2O}(c_{Gd})\right) \cdot e^{-R_1^{HN}(c_{Gd})\cdot t - k_{ex}\cdot t}}{k_{ex} + R_1^{HN}(c_{Gd}) - R_1^{H_2O}(c_{Gd})}\right) \quad (26)$$

Equation (26) describes the build-up of z-magnetization over time t during the recovery delay in the presence of exchange with the solvent. It should be noted that in the absence of exchange ($k_{ex} = 0$) the equation simplifies to

$$I_z^{HN}(t) =$$

$$= I_{z,\,eq}^{HN} \cdot \left(1 - \frac{k_{ex} \cdot e^{-R_1^{H_2O}(c_{Gd})\cdot t} + \left(R_1^{HN}(c_{Gd}) - R_1^{H_2O}(c_{Gd})\right) \cdot e^{-R_1^{HN}(c_{Gd})\cdot t - k_{ex}\cdot t}}{k_{ex} + R_1^{HN}(c_{Gd}) - R_1^{H_2O}(c_{Gd})}\right)$$

$$= I_{z,\,eq}^{HN} \cdot \left(1 - \frac{\left(R_1^{HN}(c_{Gd}) - R_1^{H_2O}(c_{Gd})\right) \cdot e^{-R_1^{HN}(c_{Gd})\cdot t}}{R_1^{HN}(c_{Gd}) - R_1^{H_2O}(c_{Gd})}\right)$$

$$= I_{z,\,eq}^{HN} \cdot \left(1 - e^{-R_1^{HN}(c_{Gd})\cdot t}\right) \quad (27)$$

This corresponds to the observed mono-exponential R_1 relaxation that is observed for non-exchangeable protons.

The build-up of the z-magnetization as described by equation (26), is governed by three parameters, the relaxation rate of the exchangeable proton $R_1^{HN}(c_{Gd})$, the relaxation rate of the bulk water $R_1^{H_2O}(c_{Gd})$ as well as the water exchange rate k_{ex}. As shown in Figure V-3 (left), in the case of $R_1^{HN}(c_{Gd}) < R_1^{H_2O}(c_{Gd})$ the build-up of magnetization is characterized by an initial lag phase (due to the slow R_1^{HN}) before a faster exponential build-up occurs (corresponding to the fast $R_1^{H_2O}$). With an increasing exchange rate k_{ex}, the build-up converges from a slow mono-exponential build-up to the faster kinetics of $R_1^{H_2O}$. In the case of $R_1^{HN} > R_1^{H_2O}$, the magnetization is build-up rapidly due to a fast R_1^{HN} and is then followed by a slower relaxation characterized by $R_1^{H_2O}$ (see Figure V-3, right).

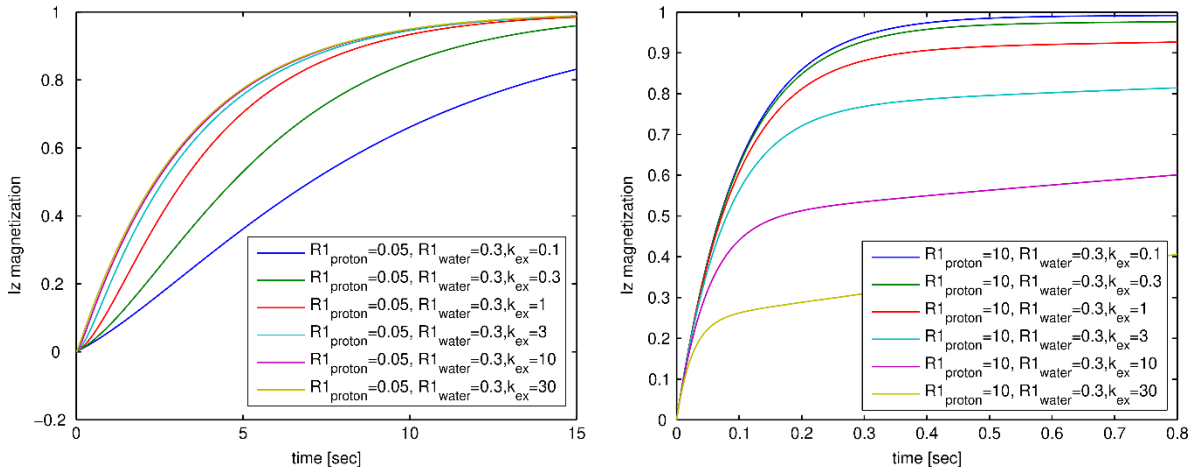


Figure V-3: Recovery of z-magnetization of exchangeable protons
Predicted recovery of the z-magnetization of protons in exchange with water according to equation (26) are shown for different parameters ($R1_{proton}$ is the R_1 relaxation rate of the exchangeable proton, $R1_{water}$ is the R_1 relaxation rate of the bulk water and k_{ex} is the water exchange rate). The left plot shows $R1_{proton} < R1_{water}$ and the right plot shows the build-up of z-magnetization for $R1_{proton} > R1_{water}$. For both regimes, different exchange rates are simulated.

To confirm the model, the predicted build-up of magnetization as a function of the delay time is compared to the measured recovery (see Figure V-4). Strikingly, the observed recovery of magnetization of imino protons fits to the predicted model. In particular, an initial lag-phase of the relaxation process is observed, while at larger times t, the observed relaxation process can be approximated by a mono-exponential kinetic.

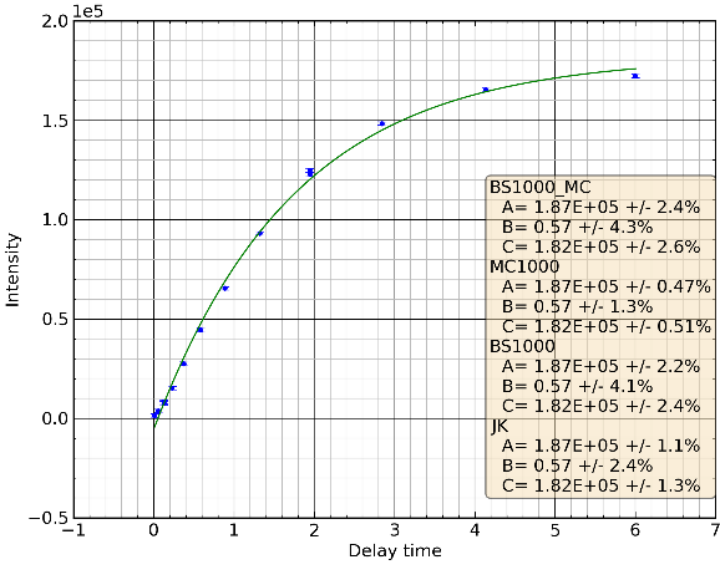


Figure V-4: Observed recovery of z-magnetization in the case of RNA
The observed recovery of z-magnetization over the delay time t is shown for the H1 imino proton of Gua^{12} of the UUCG. The observed rates are shown as blue dots and a mono-exponential fit (green) is shown for comparison (box shows parameters of the fit). For experimental details see [187].

Next, the effect of solvent exchange on the commonly-used process[128, 187] to measure and analyze solvent PRE data is simulated. To this end, the previously derived model is used to simulate the build-up of z-magnetization in the absence and presence of chemical exchange, corresponding to a non-exchangeable (H_C) and an exchangeable proton (H_N) with the same solvent PRE. The predicted recovery curves are then fitted with a mono-exponential equation, simulating the commonly-performed fitting procedure (see Figure V-5). The obtained mono-exponential rates are then plotted against the concentration. Interestingly, the simulation predicts that the fitted relaxation rates converge to a linear slope with an increasing concentration. Strikingly, the data obtained for all imino protons of RNA agree well with the prediction of the model (see [187] and Figure V-5, bottom right).

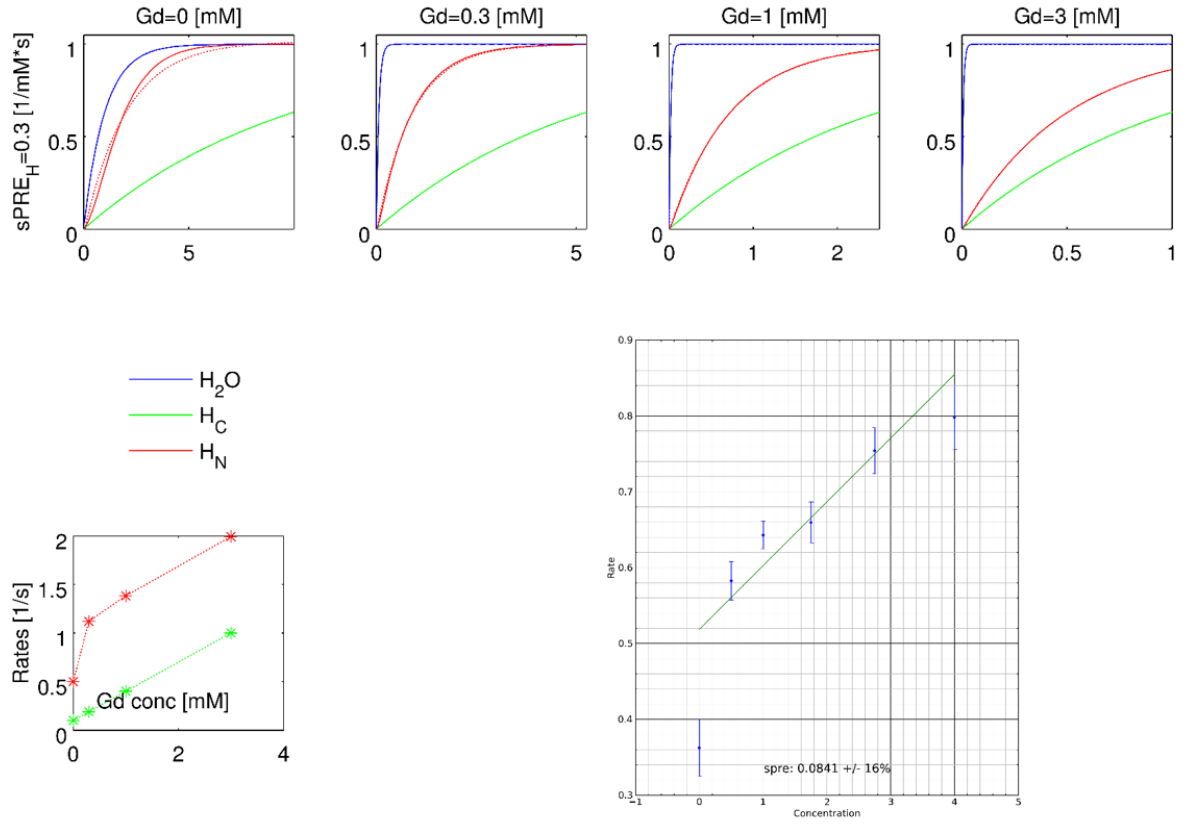


Figure V-5: Predicted solvent PRE measurements in comparison with observed data
(**Top row**) The recovery of z-magnetization during the recovery delay is simulated for water (blue), exchangeable (red) and non-exchangeable (green) protons. The concentration of the paramagnetic compound used in the prediction is indicated above every box. The intrinsic R_1 relaxation rates of the exchangeable and non-exchangeable proton are set to the same rate. Every recovery curve is fitted with a mono-exponential model and the best fit is shown with dotted lines using the same colors (blue for water, red for exchangeable and green for non-exchangeable protons). Note the deviation between the mono-exponential fit and the predicted recovery in the case of H_N in the top left plot. (**Bottom left**) The mono-exponentially fitted rates for exchangeable (red) and non-exchangeable (green) protons are plotted over the concentration of the paramagnetic agent. (**Bottom right**) For comparison, the mono-exponentially fitted relaxation rates of H1 imino proton of Ura[11] of the UUCG tetraloop is plotted over the concentration of paramagnetic compound (blue dots).

Interestingly, the slope of the exchangeable proton H_N converges to the same slope as the slope of the non-exchangeable proton H_C (see Figure V-5, bottom left). This is a crucial aspect, as it suggests that the solvent PRE can be obtained by fitting the linear part of the plot, even in the case of exchangeable protons.

The shown simulations demonstrate the applicability of the model to understand the influence of the exchange on the solvent PRE measurement. In the following, the underlying equations are analyzed to support the results of the simulations: First, equation (25) can be re-written as

$$\frac{dI_z^{HN}(t)}{dt} =$$

$$= k_{ex} \cdot \left(I_{z,\,eq}^{HN} \cdot \frac{I_z^{H_2O}(t)}{I_{z,eq}^{H_2O}} - I_z^{HN}(t) \right) + \left(I_{z,\,eq}^{HN} - I_z^{HN}(t) \right) \cdot (R_1^{HN}(0) + c_{Gd} \cdot \Gamma_{sPRE}^{HN}) \tag{28}$$

where $R_1^{HN}(0)$ is the relaxation rate of the proton in the absence of the paramagnetic agent. From this equation, it can be derived that if $k_{ex} \ll R_1^{HN}(0) + c_{Gd} \cdot \Gamma_{sPRE}^{HN}$ holds, the recovery of the z-magnetization follows a regular, mono-exponential R_1 relaxation model with the rate

$$R_1^{HN}(c_{Gd}) = R_1^{HN}(0) + c_{Gd} \cdot \Gamma_{sPRE}^{HN} \tag{29}$$

Next, the limit of a high concentration of the paramagnetic agent is considered ($c_{Gd} \gg 0$). Moreover, the solvent PRE of the bulk water is assumed to be larger than the solvent PRE of the exchangeable proton ($\Gamma_{sPRE}^{H_2O} > \Gamma_{sPRE}^{HN}$). This assumption is suitable for the commonly-used paramagnetic compound Gd(DTPA-BMA)) as it contains a coordination site for bulk water, and thus, strongly enhances R_1 relaxation of bulk water[122, 127, 187]. In this situation, the water can be assumed to relax extremely rapidly and thus $\frac{I_z^{H_2O}(t)}{I_{z,eq}^{H_2O}} \approx 1$. This leads to

$$\frac{dI_z^{HN}(t)}{dt} \approx \left(I_{z,\,eq}^{HN} - I_z^{HN}(t) \right) \cdot (R_1^{HN}(0) + c_{Gd} \cdot \Gamma_{sPRE}^{HN} + k_{ex}) \tag{30}$$

which results in a mono-exponential recovery of the z-magnetization with the rate

$$R_1^{HN}(c_{Gd}) = R_1^{HN}(0) + k_{ex} + c_{Gd} \cdot \Gamma_{sPRE}^{HN} \tag{31}$$

The observed rate corresponds to the sum of the solvent exchange rate k_{ex} and the R_1 relaxation rate of the proton of interest ($R_1^{HN}(0) + c_{Gd} \cdot \Gamma_{sPRE}^{HN}$). Therefore, in the case of a high concentration of the paramagnetic agent and thus a strong relaxation enhancement of the

water, the target solvent PRE Γ_{sPRE}^{HN} depends linearly on the observed relaxation rate $R_1^{HN}(c_{Gd})$. This is in agreement with the simulations presented above and demonstrates that the solvent PRE of an exchangeable proton can be directly obtained by plotting the observed relaxation rates against the concentration of the paramagnetic agent. This is a crucial implication, since it suggests to measure relaxation rates at high concentrations of the paramagnetic agent, in which case the correct solvent PRE can then be obtained directly obtained from the slope of the plot. Moreover, the linearity of the plot can be used as an indication of the presence of solvent exchange processes.

In summary, the presented model suggests a refined analysis of solvent PRE data for exchangeable protons: The established linear fitting procedure can still be applied to exchangeable protons. However, in case a non-linear increase of rates is observed, the results presented here suggest to limit the fit to the linear high-concentration range of the plot (compare discussion in [187] as well as Figure V-5, bottom right). Moreover, the linearity of the plot can serve as an indication to identify exchange processes.

V.3 Advanced Experimental Methods for Detecting Solvent PREs of Exchangeable Protons

The previous section suggested a refined analysis of the established experimental methods to obtain solvent PRE data for exchangeable protons. In this section, potential enhancements of the experimental NMR experiments are suggested.

V.3.1 Optimization of the Paramagnetic Compound

The previously suggested procedure relies on the assumption that the relaxation-enhancing effect of the paramagnetic compound on the bulk water is large compared to the protons of interest. This suggests Gd(DTPA-BMA) to be a particularly efficient compound for biomolecular solvent PRE studies. Gd(DTPA-BMA) is mainly used as a water-relaxing contrast agent for MRI applications. Nevertheless, its water coordinating site is beneficial also in the case of studying exchangeable protons as it increases the solvent PRE of the solvent.

It should be noted, that the presented findings are in contradiction to recent research efforts. A previously-suggested compound[188] fully coordinates the Gadolinium ion to reduce the solvent PRE of the solvent. Based on the finding of this thesis, an optimization of the paramagnetic compound should rather focus on providing additional coordination sites for

bulk water. Alternatively, an additional paramagnetic compound could be added to the NMR sample, solely to enhance the R_1 relaxation rate of the solvent.

V.3.2 Variation of Saturation Schemes

Another approach to obtain solvent PRE data for exchangeable protons, is to utilize the presented exchange model and including it in the fitting procedure. Instead of fitting a mono-exponential model to the observed data, the data can be fitted using the equation (26). While the mono-exponential fit only relies on a single rate ($R_1^{HN}(c_{Gd})$), the new model requires to fit three parameters (R_1^{HN}, $R_1^{H_2O}$ and k_{ex}) and thus dramatically increases the degree of freedoms. In the following an experimental procedure is outlined to address this issue.

The parameter k_{ex} can be assumed to remain constant upon titrating the paramagnetic compound. Moreover, the solvent-related parameters $R_1^{H_2O}(c_{Gd} = 0)$ and $\Gamma_{sPRE}^{H_2O}$ can be considered to be identical for all protons of interest throughout the titration. Therefore, instead of performing single mono-exponential fits, a global fit needs to be performed that simultaneously fits all protons of interest and all concentrations of the paramagnetic compound.

To further increase the goodness of the fitting procedure, different saturation schemes can be applied to obtain different relaxation curves. By using selective pulses different sets of protons can be saturated before the recovery delay (see Table 1).

Name of saturation scheme	Amides/Iminos saturated	Water saturated	Aliphatic/Sugar protons saturated
sat-all-but-amide	no	yes	yes
sat-all	yes	yes	yes
sat-water-only	no	yes	no
sat-water-and-amide	yes	yes	no

Table 1: Possible saturation schemes of protons
Suggested saturation schemes are shown for proteins. For proteins, saturation is selectively applied to amide and aliphatic protons. For RNA, saturation is selectively applied to iminos and sugar protons.

V.3.2.1 Preliminary Results

Preliminary data using the different saturation schemes were recorded for the IDP FoxP2. Representative data are shown for Asp281 in Figure V-6.

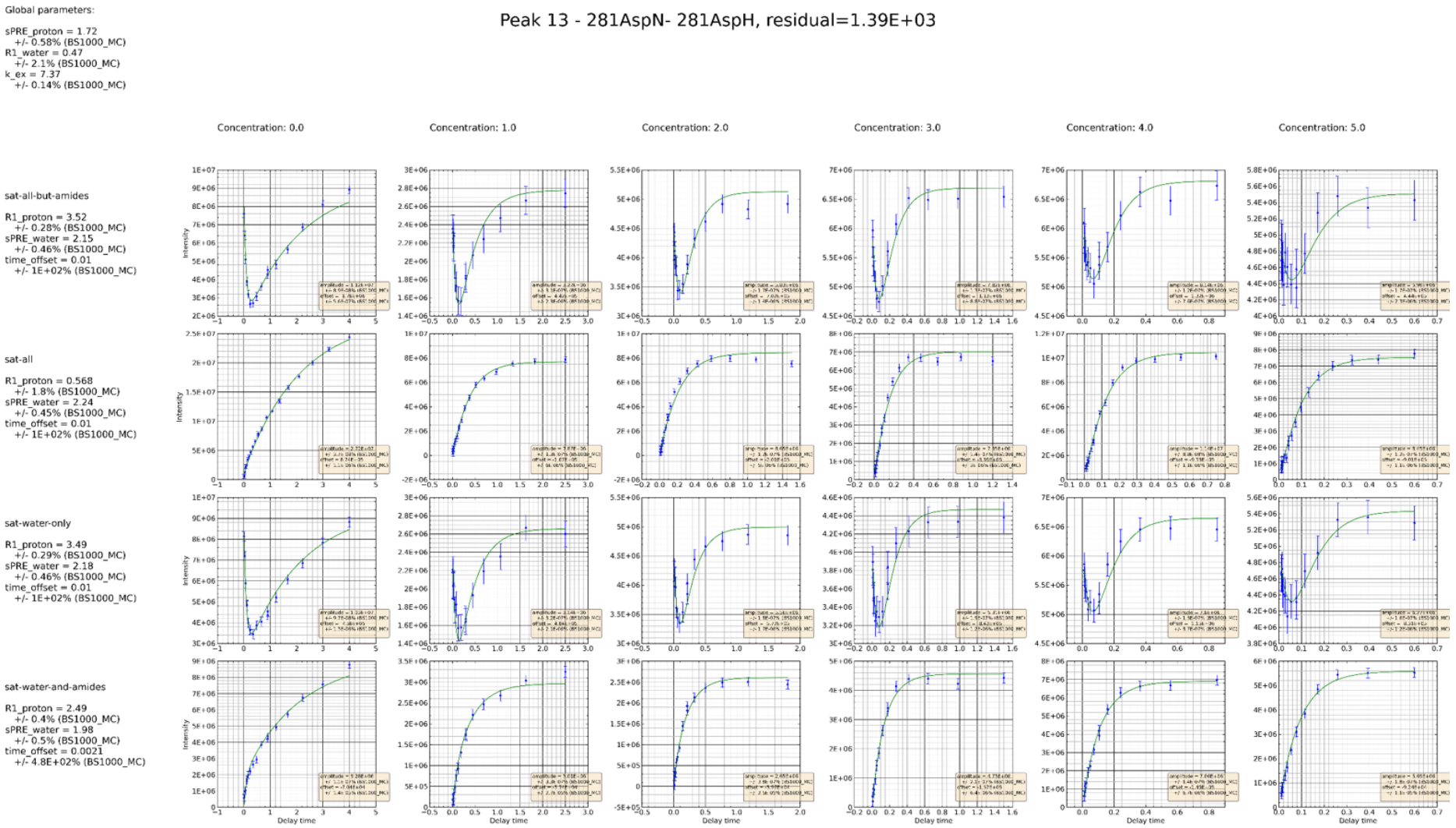


Figure V-6: Preliminary results of applying alternative saturation schemes

Experimental data for the amide proton of Asp281 of the unstructured Forkhead-Box-Protein P2 (FoxP2). All plots show the recovery of magnetization during the recovery delay. Columns from left to right correspond to the increasing concentration of Gd(DTPA-BMA), while different rows correspond to different saturation schemes (from top to bottom: sat-all-but-amide, sat-all, sat-water-only and sat-water-and-amide; see Table 1). The fitted model is indicated as green lines.

The total data set covers four different saturations schemes (as described in Table 1) and six concentrations of Gd(DTPA-BMA). For all peaks of the ^{15}N-HSQC read-out, a single global fit is performed, covering all concentrations and sharing a common $R_1^{H_2O}(c_{Gd} = 0)$ and $\Gamma_{sPRE}^{H_2O}$ for all peaks. For every residue, a specific exchange rate k_{ex}^{residue} is fitted. A similar series of experiments was performed for RNA samples. It should be noted that the data as well as the fitting procedure still require optimization and still need to be considered as preliminary. Therefore, the error of the fits is still considered too large to compare the obtained solvent PRE data to previously obtained solvent PRE data. Interestingly, the preliminary data suggest that the solvent PRE of the water (which is determined by the fitting procedure) is significantly smaller than the solvent PRE obtained when directly measuring the bulk water. In the proximity of Asp281, the solvent PRE of water was found to be 2.14 $s^{-1}mM^{-1}$, compared to 3,97 $s^{-1}mM^{-1}$ in the bulk water. This can be explained by a reduced solvent PRE of the solvation shell. Therefore, further evaluating and optimizing the described procedure is not only interesting to determine exact solvent PRE data, but also to get insight into the rate constant of the exchange as well as the water solvent PRE of the solvation shell.

V.3.2.2 Materials and Methods

Protein and RNA samples have been purified and prepared as previously described[128, 186, 187]. NMR experiments were based on the previously described saturation recovery schemes with a ^{15}N-HSQC[128, 187]. For the presented data of FoxP2, a sample of 365 µl containing uniformly ^{15}N labeled 450µM FoxP2 (20 mM Tris Buffer, pH 6.7, 50 mM NaCl) was titrated with six concentrations (0.0, 1.0, 2.0, 3.0, 4.0 and 5.0 mM) of Gd(DTPA-BMA) (Omniscan, GE Healthcare). NMR experiments were performed on a 750 MHz magnet equipped with a TXI probe head (Bruker). For every titration step, four different saturation recovery experiments were recorded, each using a different saturation scheme as shown in Table 2. Besides the adjusted saturation scheme, the saturation recovery-based NMR experiments were performed as described in [128].

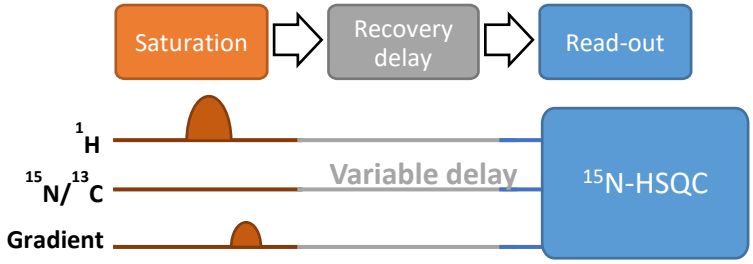
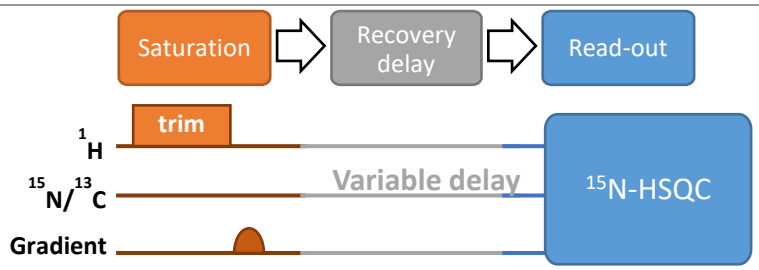
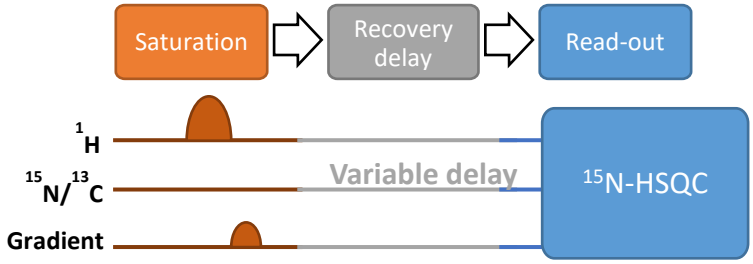
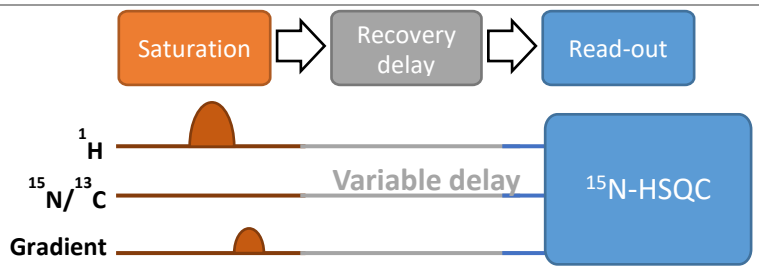
Name of saturation scheme	Amides/Iminos saturated
sat-all-but-amide	 Saturation block contains a gradient preceded by a selective pulse covering the water signal as well as all aliphatic protons
sat-all	 Saturation block contains a gradient preceded by a 2500 ms trim pulse saturating all protons
sat-water-only	 Saturation block contains a gradient preceded by a selective pulse covering only the water signal
sat-water-and-amide	 Saturation block contains a gradient preceded by a selective pulse covering the water signal as well as all amide protons

Table 2: Different saturation recovery experiments with varying saturation schemes used to obtain an extended set of relaxation recovery data

Processing of NMR data and extracting peak intensities were performed as described in [128]. Global fits were performed using SciPy. In the fitting procedure, two different exchange models were used. First, data obtained from the experiments in which amide protons were saturated (sat-water-and-amide and sat-all) show a continuous build-up of magnetization and were fitted based on equation (26). Therefore, for every concentration c_{Gd}, every recovery delay t, every signal i in the ^{15}N-HSQC and every saturation scheme $s \in$ {sat_water_and_amide, sat_all} the following fit was performed

$$I_z^{HN,s,i}(t, c_{Gd}) =$$

$$= \text{off}_{s,i,c_{Gd}} + \text{amp}_{s,i,c_{Gd}} \cdot \left(1 - \frac{k_{ex}^i \cdot e^{-R_1^{H_2O,s}(c_{Gd})\cdot t} - \left(R_1^{H_2O,s}(c_{Gd}) - R_1^{HN,s,i}(c_{Gd}) \right) \cdot e^{-R_1^{HN,s,i}(c_{Gd})\cdot t - k_{ex}^i \cdot t}}{k_{ex}^i + R_1^{HN,s,i}(c_{Gd}) - R_1^{H_2O,s}(c_{Gd})} \right) \quad (32)$$

where $\text{off}_{s,i,c_{Gd}}$ and $\text{amp}_{s,i,c_{Gd}}$ are the offset and the amplitude for the relaxation curve of signal i recorded with saturation scheme s for concentration c_{Gd}. Moreover, $R_1^{HN,s,i}(c_{Gd}) = R_1^{HN,s,i}(0) + c_{Gd} \cdot \Gamma_{sPRE}^{HN,i}$ and $R_1^{H_2O,s}(c_{Gd}) = R_1^{H_2O}(0) + c_{Gd} \cdot \Gamma_{sPRE}^{H_2O,s}$. For all fits, a time offset was allowed to be performed, such that $\hat{t} = t + \Delta t_{s,i}$.

Secondly, data obtained from experiments in which amide protons were not saturated (sat-water-only and sat-all-but-amide) show a decrease of magnetization followed by an increase of magnetization. To obtain a model for fitting the data, equation (25) was solved using equation (24) and $I_z^{HN}(0) = I_{z,\,eq}^{HN}$. Based on the analytical solution, the following fit was performed for every concentration c_{Gd}, every recovery delay t, every signal i in the ^{15}N-HSQC and every saturation scheme $s \in$ {sat_water_only, sat_all_but_aminde}

$$I_z^{HN,s,i}(t, c_{Gd}) =$$

$$= \text{off}_{s,i,c_{Gd}} + \text{amp}_{s,i,c_{Gd}} \cdot \left(1 - \frac{k_{ex}^i \cdot \left(e^{-R_1^{H_2O,s}(c_{Gd})\cdot t} - e^{-R_1^{HN,s,i}(c_{Gd})\cdot t - k_{ex}^i \cdot t} \right)}{k_{ex}^i + R_1^{HN,s,i}(c_{Gd}) - R_1^{H_2O,s}(c_{Gd})} \right) \quad (33)$$

where $\text{off}_{s,i,c_{Gd}}$, $\text{amp}_{s,i,c_{Gd}}$, $R_1^{HN,s,i}(c_{Gd})$ and $R_1^{H_2O,s}(c_{Gd})$ are defined corresponding to equation (32) and a time offset was included in the fit as well.

After performing the fits for all four saturation schemes (sat-all-but-amide, sat-all, sat-water-only and sat-water-and-amide) several parameters are obtained by a single fitting procedure as shown for $i = HN$, Asp281 in Table 3.

Parameter	Parameter description	sat-all-but-amide	sat-all	sat-water-only	sat-water-and-amide
$\Gamma_{sPRE}^{HN,i}$	Solvent PRE of the amide proton [$s^{-1}mM^{-1}$]	1.72			
$R_1^{HN,s,i}(0)$	Proton R_1 relaxation rate at $c_{Gd} = 0$ [s^{-1}]	3.52	0.568	3.49	2.49
$\Gamma_{sPRE}^{H_2O,s}$	Solvent PRE of water [$s^{-1}mM^{-1}$]	2.15	2.24	2.18	1.98
$R_1^{H_2O}(0)$	Solvent R_1 relaxation rate at $c_{Gd} = 0$ [$s^{-1}mM^{-1}$]	0.47			
k_{ex}^{i}	Solvent exchange rate [s^{-1}]	7.37			

Table 3: Parameter values obtained for $i = HN$, Asp281

VI Conclusion and Outlook

The aim of this thesis is the testing and establishment of new NMR methods. As described in Chapter I.4.2 , the current trends of biomolecular NMR spectroscopy are studying large proteins and protein complexes as well as highly dynamic biomolecules such as IDPs and RNAs. These trends lead to new challenges since large as well as dynamic systems need to be represented in a larger conformational space. These challenges are often addressed by using hybrid methods that combine different data sets of complementary experimental methods. This trend imposes new requirements for computational methods as the problems become more complex. Moreover, parallelization of the computational methods becomes an essential aspect, due to the strong trend of distributing computational power across many thousands of computing nodes. To this end, this thesis focuses on improving and further establishing the utilization of solvent PRE data and the distance-to-surface information which it encodes.

The choice of solvent PRE data is motivated by several aspects. First , acquiring a set of solvent PRE data does not involve any specific sample preparation and the sample can be regenerated. Most notably this is also true in the case of RNAs which have been shown to be susceptible to hydrolysis in the presence of lanthanides.

Moreover, including additional information such as solvent accessibility data in the structure prediction algorithm is a straight-forward approach to increase the number of restraints. Increasing the number of restraints is an essential key to accurately determine the structure. To a certain extent, a limited number of restraints can be compensated by extensive sampling of the conformational space. However, an increased sampling also broadens the population of wrong conformations and therefore increases the probability to sample wrong models with good scores. Therefore, increasing the number of restraints is essential for structure prediction. This is particularly important in the case of flexible proteins, RNAs, large proteins or biomolecular complexes. In these cases, obtaining a sufficient number of restraints for structure determination becomes a challenge. Including solvent PRE as an additional type of restraint is a straight-forward approach to enhance structure prediction in these cases.

Moreover, the acquisition of solvent PREs does not require a particular set of NMR experiments. This can be useful in challenging cases that require specific NMR methods to

obtain decent spectra. Solvent PRE data can still be recorded in such cases, since most NMR pulse programs can be expanded and re-used to record solvent PRE data. Moreover, the evaluation of solvent PRE data does not depend on a complete assignment which is required for example in the case of NOEs.

Including solvent PRE data in the structure determination process is not only increasing the number of restraints. The nature of the data is beneficial due to several aspects. Results presented in this thesis show that solvent PRE data are orthogonal to other commonly-used NMR restraints. Distances derived from NOEs and angles derived from J-couplings or predicted from chemical shifts provide local short-range information. Long-range distance information becomes particularly important for larger systems and RDCs, PREs and PCSs are often used in these cases. Solvent PREs are a promising tool to complete the list of long-range restraints since no additional preparations are required which in some studies hinder the acquisition of other long-range restraints. The nature of solvent PRE comes with another advantage as it primarily depends on the global fold of the biomolecule. Therefore, the data can drive structure determination algorithms towards the native fold even in the early stage of the calculation in which structures far from the native fold are sampled. Moreover, solvent PRE data depend only to a limited extent on the local environment such as adjacent residues or even the type of the residue itself. This property enables its application to simplified models such as the centroid model in Rosetta. In the case of chemical shift or RDC data the evaluation requires the presence of certain atoms in the representation of the protein. In summary, solvent PRE data offer a unique set of advantages that might not be relevant for some studies, but can certainly be crucial for others.

In the research presented in this thesis, solvent PRE was successfully applied in several scenarios, such as protein structure prediction, RNA structure determination, characterization of IDPs as well as complexes in solid state NMR (see Chapter III). Moreover, solvent PRE data were also applied using different approaches. For studies aiming to find the correct model out of a limited list of candidate models, sPRE data can be predicted for all candidates and these data sets can then be compared to the observed data. Furthermore, the presented research in this thesis demonstrates the usage as a restraint in a Monte Carlo-based sampling algorithm as well as in a molecular dynamics simulated annealing approach. These findings suggest that solvent PRE data are universal restraints and applicable to different methods.

A crucial aspect for the successful establishment of solvent PRE in NMR studies is an efficient pipeline for measurement and analyzing the data. In the presented research, the NMR experiments to obtain solvent PRE data were adapted for studies on RNA. Moreover, the processing of the data was streamlined and the error handling was improved. It was further shown, that solvent PRE data were recorded using non-uniform sampling which is particularly useful for unstructured proteins. Finally, by utilizing the solvent PRE of water as a reference, a robust approach was suggested to account for batch-to-batch differences and to improve the comparison of solvent PRE recorded at different spectrometers.

The research of this thesis focused on the Rosetta framework and in particular on the usage of Monte Carlo-based sampling approaches. This choice was motivated by the excellent scaling capabilities of the sampling methods. This property will be essential in the near future as it allows to efficiently use highly-distributed computer systems and accelerator hardware such as graphics processing units (GPUs). Compared to molecular dynamics simulations which are based on force fields and thus require differentiable energy functions, Monte Carlo-methods allow to combine any type of energy function. This is particularly useful for hybrid methods that aim to combine data sets of many different experimental methods.

In recent years, highly dynamic IDPs are became an important aspect in many structural biology research projects. To properly characterize IDPs, new methods are required. Therefore, research presented in this thesis investigated how solvent PRE data can be used to study IDPs. To this end, an approach to predict solvent PRE data for IDP was present, allowing to characterize residual structural elements such as residual structural elements as well as intramolecular protein-protein contacts. This approach is particularly interesting since the usage of NUS allows to record solvent PRE data in a time-efficient manner. Compared to the commonly-used spin label-based PRE experiments, recording solvent PRE data allows to quickly scan the entire IDP for potential binding sites without the need of introducing mutations for the attachment of the spin labels. The method presented in this thesis, can therefore be used to efficiently screen an IDP for residual structural elements. Focusing on the protein regions identified by solvent PREs, more laborious methods such as attaching spin labels can be performed in a second phase. For future studies, chemical shift and relaxation data can be extended by solvent PRE data to obtain a powerful set of experimental data for screening IDPs for residual structural elements.

Solvent PRE data are often obtained for exchangeable nitrogen-bound protons, such as amide protons in the case of proteins or amide as well as amino protons in the case of RNA. Nevertheless, the effect of exchange in the context of solvent PREs is not fully understood. In this thesis, a theoretical model was derived that describes the chemical exchange of nitrogen-bound protons with solvent water (see Chapter V). The model correctly predicts the results of saturation recovery experiments as well as the effects on the measurements of solvent PREs. Moreover, the solvent PRE obtained by fitting the linear increase of the relaxation rate over the concentration of the paramagnetic agent is suggested to be equivalent with the solvent PRE in the absence of chemical exchange. Since this assumption is valid in the high concentration limit, the model suggests that paramagnetic compounds that strongly relax bulk water are better suited to measure solvent PRE data. Nevertheless, exact measurements of the solvent PRE data require more efforts. An approach using alternative saturation schemes and a global fitting procedure was presented but is still considered preliminary. However, optimizing this approach is a promising future step towards a robust and exact measurement of solvent PRE data of solvent-exchangeable protons.

Using solvent PRE data is a powerful tool to tackle future challenges in NMR spectroscopy. A promising approach is the usage of RasRec-Rosetta[189] in combination with solvent accessibility data. RasRec-Rosetta makes use of advanced sampling methods and was shown to be more efficient for larger proteins. In this study, solvent PRE data was shown to efficiently direct optimization algorithms in the early stages of the folding. Therefore, combining solvent PRE data with RasRec-Rosetta is a promising tool for future NMR studies on large proteins.

In future studies, solvent accessibility data can potentially be used in an even broader sense. For example, solvent accessibility information can be obtained by bioinformatics as well as by experimental methods such as mass spectroscopy-based surface exposure. Since the implementation of the Rosetta energy function developed in the course of this thesis is not limited to solvent accessibility data, other sources of solvent accessibly data can be included in the Rosetta framework. Another powerful future application of solvent accessibility data, is to use unassigned solvent PRE data sets. For example, measuring the distribution of the solvent PRE data could be used to detect the fold of a protein. To this end, a solvent PRE pattern could be used to filter a list containing all possible folds for a given total length of the

sequence. Although such an analysis is computationally expensive, testing different candidate folds can be easily parallelized and can therefore become feasible in the near future.

VII Reprint Permissions

VII.1 Prediction of Protein Structure Using Surface Accessibility Data

Title:	Prediction of Protein Structure Using Surface Accessibility Data
Author:	Christoph Hartlmüller, Christoph Göbl, Tobias Madl
Publication:	Angewandte Chemie International Edition
Publisher:	John Wiley and Sons
Date:	Aug 25, 2016
Available at:	http://onlinelibrary.wiley.com.eaccess.ub.tum.de/doi/10.1002/anie.201604788/abstract
Copyright	© 2016 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA.
License	This article is available under the terms of the Creative Commons Attribution License (CC BY) (which may be updated from time to time) and permits use, distribution and reproduction in any medium, provided that the Contribution is properly cited. For an understanding of what is meant by the terms of the Creative Commons License, please refer to Wiley's Open Access Terms and Conditions (http://olabout.wiley.com/WileyCDA/Section/id-410895.html). Permission is not required for this type of reuse.

VII.2 RNA Structure Refinement using NMR Solvent Accessibility Data

Title:	RNA Structure Refinement using NMR Solvent Accessibility Data
Author:	Christoph Hartlmüller, Johannes C. Günther, Antje C. Wolter, Jens Wöhnert, Michael Sattler and Tobias Madl
Publication:	Scientific Reports
Publisher:	Springer Nature
Date:	Jul 14, 2017
Available at:	https://www.nature.com/articles/s41598-017-05821-z
Copyright	© 2017, Springer Nature
License	Creative Commons This is an open access article distributed under the terms of the Creative Commons CC BY license (https://creativecommons.org/licenses/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. You are not required to obtain permission to reuse this article.

VII.3 NMR Characterization of Solvent Accessibility and Transient Structure in Intrinsically Disordered Proteins

Title:	NMR Characterization of Solvent Accessibility and Transient Structure in Intrinsically Disordered Proteins
Author:	Christoph Hartlmüller, Emil Spreitzer, Christoph Göbl et al
Publication:	Journal of Biomolecular NMR
Publisher:	Springer Nature
Date:	Jan 1, 2019
Available at:	https://link.springer.com/article/10.1007%2Fs10858-019-00248-2
Copyright	© 2019, The Author(s)
License	Creative Commons This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. You are not required to obtain permission to reuse this article. To request permission for a type of use not listed, please contact Springer Nature

VII.4 Characterization of Protein-Protein Interfaces in Large Complexes by Solid-State NMR Solvent Paramagnetic Relaxation Enhancements

Title:	Characterization of Protein-Protein Interfaces in Large Complexes by Solid-State NMR Solvent Paramagnetic Relaxation Enhancements
Author:	Carl Öster, Simone Kosol, Christoph Hartlmüller, Jonathan M. Lamley, Dinu Iuga, Andres Oss, Mai-Liis Org, Kalju Vanatalu, Ago Samoson, Tobias Madl, Józef R. Lewandowski
Publication:	Journal of the American Chemical Society
Publisher:	ACS Publications
Available at:	https://pubs.acs.org/doi/10.1021/jacs.7b03875
License	ACS AuthorChoice/Editors' Choice via Creative Commons CC-BY Usage Agreement

This ACS article is provided to You under the terms of this ACS AuthorChoice/Editors' Choice via Creative Commons CC-BY agreement between You and the American Chemical Society ("ACS"), a federally-chartered nonprofit located at 1155 16th Street NW, Washington DC 20036. Your access and use of this ACS article means that you have accepted and agreed to the Terms and Conditions of this Agreement. ACS and You are collectively referred to in this Agreement as "the Parties").

1. SCOPE OF GRANT

ACS grants You a non-exclusive and nontransferable permission to access and use this ACS article subject to the terms and conditions set forth in this Agreement.
2. PERMITTED USES

a. ACS grants You the rights in the attached Creative Commons Attribution 4.0 International license. Consistent with the Creative Commons Attribution 4.0 license we note that any use of the article is subject to the following conditions:

i. The authors' moral right to the integrity of their work under the Berne Convention (Article 6bis) is not compromised.

ii. Where content in the article is identified as belonging to a third party, it is your responsibility to ensure that any reuse complies with copyright policies of the owner.
3. TERMINATION

ACS reserves the right to limit, suspend, or terminate your access to and use of the ACS Publications Division website and/or all ACS articles immediately upon detecting a breach of this License.
4. COPYRIGHTS; OTHER INTELLECTUAL PROPERTY RIGHTS

Except as otherwise specifically noted, ACS is the owner of all right, title and interest in the content of this ACS article, including, without limitations, graphs, charts, tables illustrations, and copyrightable supporting information. This ACS article is protected under the Copyright Laws of the United States Codified in Title 17 of the U.S. Code and subject to the Universal Copyright Convention and the Berne Copyright Convention. You agree not to remove or obscure copyright notices. You acknowledge that You have no claim to ownership of any part of this ACS article or other proprietary information accessed under this Agreement.The names “American Chemical Society,” “ACS” and the titles of the journals and other ACS products are trademarks of ACS.
5. DISCLAIMER OF WARRANTIES; LIMITATION OF LIABILITY

ACS warrants that it is entitled to grant this Agreement.

EXCEPT AS SET FORTH IN THE PRECEDING SENTENCE, ACS MAKES NO WARRANTY OR REPRESENTATION OF ANY KIND, EXPRESS OR IMPLIED, WITH RESPECT TO THIS ACS ARTICLE INCLUDING, BUT NOT LIMITED TO WARRANTIES AS TO THE ACCURACY OR COMPLETENESS OF THE ACS ARTICLE, ITS QUALITY, ORIGINALITY, SUITABILITY, SEARCHABILITY, OPERATION, PERFORMANCE, COMPLIANCE WITH ANY COMPUTATIONAL PROCESS, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

ACS SHALL NOT BE LIABLE FOR: EXEMPLARY, SPECIAL, INDIRECT, INCIDENTAL, CONSEQUENTIAL OR OTHER DAMAGES ARISING OUT OF OR IN CONNECTION WITH THE AGREEMENT GRANTED HEREUNDER, THE USE OR INABILITY TO USE ANY ACS PRODUCT, ACS’S PERFORMANCE UNDER THIS AGREEMENT, TERMINATION OF THIS AGREEMENT BY ACS OR THE LOSS OF DATA, BUSINESS OR GOODWILL EVEN IF ACS IS ADVISED OR AWARE OF THE POSSIBILITY OF SUCH DAMAGES. IN NO EVENT SHALL THE TOTAL AGGREGATE LIABILITY OF ACS OUT OF ANY BREACH OR TERMINATION OF THIS AGREEMENT EXCEED THE TOTAL AMOUNT PAID BY YOU TO ACS FOR ACCESS TO THIS ACS ARTICLE FOR THE CURRENT YEAR IN WHICH SUCH CLAIM, LOSS OR DAMAGE OCCURRED, WHETHER IN CONTRACT, TORT OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, DUE TO NEGLIGENCE.

The foregoing limitations and exclusions of certain damages shall apply regardless of the success or effectiveness of other remedies. No claim may be made against ACS unless suit is filed within one (1) year after the event giving rise to the claim.
6. GENERAL

This Agreement sets forth the entire understanding of the Parties. The validity, construction and performance of this Agreement shall be governed by and construed in accordance with the laws of the District of Columbia, USA without reference to its conflicts of laws principles. You acknowledge that the delivery of the ACS article will occur in the District of Columbia, USA. You shall pay any taxes lawfully due from it, other than taxes on ACS's net income, arising out of your use of this ACS article and/or other rights granted under this Agreement.
7. ACCEPTANCE

You warrant that You have read, understand, and accept the terms and conditions of this Agreement. ACS reserves the right to modify this Agreement at any time by posting the modified terms and conditions on the ACS Publications Web site. Any use of this ACS article after such posting shall constitute acceptance of the terms and conditions as modified.

Creative Commons Attribution 4.0 International Public License

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution 4.0 International Public License ("Public License"). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.

Section 1 – Definitions

a. Adapted Material means material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor. For purposes of this Public License, where the Licensed Material is a musical work, performance, or sound recording, Adapted Material is always produced where the Licensed Material is synched in timed relation with a moving image.

b. Adapter's License means the license You apply to Your Copyright and Similar Rights in Your contributions to Adapted Material in accordance with the terms and conditions of this Public License.

c. Copyright and Similar Rights means copyright and/or similar rights closely related to copyright including, without limitation, performance, broadcast, sound recording, and Sui Generis Database Rights, without regard to how the rights are labeled or categorized. For purposes of this Public License, the rights specified in Section 2(b)(1)-(2) are not Copyright and Similar Rights.

d. Effective Technological Measures means those measures that, in the absence of proper authority, may not be circumvented under laws fulfilling obligations under Article 11 of the WIPO Copyright Treaty adopted on December 20, 1996, and/or similar international agreements.

e. Exceptions and Limitations means fair use, fair dealing, and/or any other exception or limitation to Copyright and Similar Rights that applies to Your use of the Licensed Material.

f. Licensed Material means the artistic or literary work, database, or other material to which the Licensor applied this Public License.

g. Licensed Rights means the rights granted to You subject to the terms and conditions of this Public License, which are limited to all Copyright and Similar Rights that apply to Your use of the Licensed Material and that the Licensor has authority to license.

h. Licensor means the individual(s) or entity(ies) granting rights under this Public License.

i. Share means to provide material to the public by any means or process that requires permission under the Licensed Rights, such as reproduction, public display, public performance, distribution, dissemination, communication, or importation, and to make material available to the public including in ways that members of the public may access the material from a place and at a time individually chosen by them.

j. Sui Generis Database Rights means rights other than copyright resulting from Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, as amended and/or succeeded, as well as other essentially equivalent rights anywhere in the world.

k. You means the individual or entity exercising the Licensed Rights under this Public License. Your has a corresponding meaning.

Section 2 – Scope

a. License grant.

1. Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to:

A. reproduce and Share the Licensed Material, in whole or in part; and
B. produce, reproduce, and Share Adapted Material.

2. Exceptions and Limitations. For the avoidance of doubt, where Exceptions and Limitations apply to Your use, this Public License does not apply, and You do not need to comply with its terms and conditions.

3. Term. The term of this Public License is specified in Section 6(a).

4. Media and formats; technical modifications allowed. The Licensor authorizes You to exercise the Licensed Rights in all media and formats whether now known or hereafter created, and to make technical modifications necessary to do so. The Licensor waives and/or agrees not to assert any right or authority to forbid You from making technical modifications necessary to exercise the Licensed Rights, including technical modifications necessary to circumvent Effective Technological Measures. For purposes of this Public License, simply making modifications authorized by this Section 2(a)(4) never produces Adapted Material.

5. Downstream recipients

A. Offer from the Licensor – Licensed Material. Every recipient of the Licensed Material automatically receives an offer from the Licensor to exercise the Licensed Rights under the terms and conditions of this Public License.

B. No downstream restrictions. You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, the Licensed Material if doing so restricts exercise of the Licensed Rights by any recipient of the Licensed Material.

6. No endorsement. Nothing in this Public License constitutes or may be construed as permission to assert or imply that You are, or that Your use of the Licensed Material is, connected with, or sponsored, endorsed, or granted official status by, the Licensor or others designated to receive attribution as provided in Section 3(a)(1)(A)(i).

b. Other rights.

1. Moral rights, such as the right of integrity, are not licensed under this Public License, nor are publicity, privacy, and/or other similar personality rights; however, to the extent possible, the Licensor waives and/or agrees not to assert any such rights held by the Licensor to the limited extent necessary to allow You to exercise the Licensed Rights, but not otherwise.

2. Patent and trademark rights are not licensed under this Public License.

3. To the extent possible, the Licensor waives any right to collect royalties from You for the exercise of the Licensed Rights, whether directly or through a collecting society under any voluntary or waivable statutory or compulsory licensing scheme. In all other cases the Licensor expressly reserves any right to collect such royalties.
Section 3 – License Conditions

Your exercise of the Licensed Rights is expressly made subject to the following conditions.

a. Attribution

1. If You Share the Licensed Material (including in modified form), You must:

A. retain the following if it is supplied by the Licensor with the Licensed Material:

i. identification of the creator(s) of the Licensed Material and any others designated to receive attribution, in any reasonable manner requested by the Licensor (including by pseudonym if designated);

ii. a copyright notice;

iii. a notice that refers to this Public License;

iv. a notice that refers to the disclaimer of warranties;

v. a URI or hyperlink to the Licensed Material to the extent reasonably practicable;

B. indicate if You modified the Licensed Material and retain an indication of any previous modifications; and

C. indicate the Licensed Material is licensed under this Public License, and include the text of, or the URI or hyperlink to, this Public License.

2. You may satisfy the conditions in Section 3(a)(1) in any reasonable manner based on the medium, means, and context in which You Share the Licensed Material. For example, it may be reasonable to satisfy the conditions by providing a URI or hyperlink to a resource that includes the required information.

3. If requested by the Licensor, You must remove any of the information required by Section 3(a)(1)(A) to the extent reasonably practicable.

4. If You Share Adapted Material You produce, the Adapter's License You apply must not prevent recipients of the Adapted Material from complying with this Public License.
Section 4 – Sui Generis Database Rights

Where the Licensed Rights include Sui Generis Database Rights that apply to Your use of the Licensed Material:

a. for the avoidance of doubt, Section 2(a)(1) grants You the right to extract, reuse, reproduce, and Share all or a substantial portion of the contents of the database;

b. if You include all or a substantial portion of the database contents in a database in which You have Sui Generis Database Rights, then the database in which You have Sui Generis Database Rights (but not its individual contents) is Adapted Material; and

c. You must comply with the conditions in Section 3(a) if You Share all or a substantial portion of the contents of the database.

For the avoidance of doubt, this Section 4 supplements and does not replace Your obligations under this Public License where the Licensed Rights include other Copyright and Similar Rights.
Section 5 – Disclaimer of Warranties and Limitation of Liability

a. Unless otherwise separately undertaken by the Licensor, to the extent possible, the Licensor offers the Licensed Material as-is and as-available, and makes no representations or warranties of any kind concerning the Licensed Material, whether express, implied, statutory, or other. This includes, without limitation, warranties of title, merchantability, fitness for a particular purpose, non-infringement, absence of latent or other defects, accuracy, or the presence or absence of errors, whether or not known or discoverable. Where disclaimers of warranties are not allowed in full or in part, this disclaimer may not apply to You.

b. To the extent possible, in no event will the Licensor be liable to You on any legal theory (including, without limitation, negligence) or otherwise for any direct, special, indirect, incidental, consequential, punitive, exemplary, or other losses, costs, expenses, or damages arising out of this Public License or use of the Licensed Material, even if the Licensor has been advised of the possibility of such losses, costs, expenses, or damages. Where a limitation of liability is not allowed in full or in part, this limitation may not apply to You.

c. The disclaimer of warranties and limitation of liability provided above shall be interpreted in a manner that, to the extent possible, most closely approximates an absolute disclaimer and waiver of all liability.
Section 6 – Term and Termination

a. This Public License applies for the term of the Copyright and Similar Rights licensed here. However, if You fail to comply with this Public License, then Your rights under this Public License terminate automatically.

b. Where Your right to use the Licensed Material has terminated under Section 6(a), it reinstates:

1. automatically as of the date the violation is cured, provided it is cured within 30 days of Your discovery of the violation; or

2. upon express reinstatement by the Licensor.For the avoidance of doubt, this Section 6(b) does not affect any right the Licensor may have to seek remedies for Your violations of this Public License.

c. For the avoidance of doubt, the Licensor may also offer the Licensed Material under separate terms or conditions or stop distributing the Licensed Material at any time; however, doing so will not terminate this Public License.

d. Sections 1, 5, 6, 7, and 8 survive termination of this Public License.
Section 7 – Other Terms and Conditions

a. The Licensor shall not be bound by any additional or different terms or conditions communicated by You unless expressly agreed.

b. Any arrangements, understandings, or agreements regarding the Licensed Material not stated herein are separate from and independent of the terms and conditions of this Public License.
Section 8 – Interpretation

a. For the avoidance of doubt, this Public License does not, and shall not be interpreted to, reduce, limit, restrict, or impose conditions on any use of the Licensed Material that could lawfully be made without permission under this Public License.

b. To the extent possible, if any provision of this Public License is deemed unenforceable, it shall be automatically reformed to the minimum extent necessary to make it enforceable. If the provision cannot be reformed, it shall be severed from this Public License without affecting the enforceability of the remaining terms and conditions.

c. No term or condition of this Public License will be waived and no failure to comply consented to unless expressly agreed to by the Licensor.

d. Nothing in this Public License constitutes or may be interpreted as a limitation upon, or waiver of, any privileges and immunities that apply to the Licensor or You, including from the legal processes of any jurisdiction or authority.

Back to ACS Publishing Policies

Posted: 03/06/2014

VII.5 Increasing the Chemical-Shift Dispersion of Unstructured Proteins with a Covalent Lanthanide Shift Reagent

Title: Increasing the Chemical-Shift Dispersion of Unstructured Proteins with a Covalent Lanthanide Shift Reagent

Author: Dr. Christoph Göbl, Moritz Resch, Dr. Madeleine Strickland, Christoph Hartlmüller, Martin Viertler, Dr. Nico Tjandra, Prof. Dr. Tobias Madl

Publication: Angewandte Chemie

Publisher: Wiley-VCH

Available at: https://onlinelibrary.wiley.com/doi/full/10.1002/anie.201607261

© 2016 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA.

License This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

VII.6 Structural Basis of Nucleic-Acid Recognition and Double-Strand Unwinding by the Essential Neuronal Protein Pur-alpha

DOI 10.7554/eLife.11297
Cite this as eLife 2016;5:e11297
Authors Janine Weber, Han Bao, Christoph Hartlmüller, Zhiqin Wang, Almut Windhager, Robert Janowski, Tobias Madl, Peng Jin and Dierk Niessing
Publication history Received September 03, 2015.
Accepted January 07, 2016.
Published January 08, 2016.
Reviewing editor Karsten Weis, Reviewing editor, ETH Zürich, Switzerland
Copyright © 2016, Weber et al
This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

eLife is an open-access publication. All results are available to a worldwide audience, for free, immediately on publication. Everyone has the right to use the results without restriction (using a Creative Commons Attribution license). Citing, downloading, re-using, reproducing, re-purposing, and building on material published in eLife can be done without requesting permission from the authors or the publisher, provided full attribution to the authors is given. We also deposit all content into PubMed Central, as well as sites such as Mendeley, scribd, and Github, and we make article data openly available through an API.

VII.7 The Redox Environment Triggers Conformational Changes and Aggregation of hIAPP in Type II Diabetes

Title: The redox environment triggers conformational changes and aggregation of hIAPP in Type II Diabetes

Author: Diana C. Rodriguez Camargo, Konstantinos Tripsianes, Katalin Buday, Andras Franko, Christoph Göbl et al.

Publication: Scientific Reports
Publisher: Nature Publishing Group
Date: Mar 13, 2017
Copyright © 2017, Rights Managed by Nature Publishing Group

Creative Commons
The article for which you have requested permission has been distributed under a Creative Commons CC-BY license (please see the article itself for the license version number). You may reuse this material without obtaining permission from Nature Publishing Group, providing that the author and the original source of publication are fully acknowledged, as per the terms of the license.
For license terms, please see http://creativecommons.org/

VII.8 Molecular Basis for Asymmetry Sensing of siRNAs by the Drosophila Loqs-PD/Dcr-2 Complex in RNA Interference

Title: Molecular Basis for Asymmetry Sensing of siRNAs by the Drosophila Loqs-PD/Dcr-2 Complex in RNA Interference
Author: Jan-Niklas Tants, Stephanie Fesser, Thomas Kern, Ralf Stehle, Arie Geerlof, Christoph Wunderlich, Michael Juen, Christoph Hartlmüller, Romy Böttcher, Stefan Kunzelmann, Oliver Lange, Christoph Kreutz, Klaus Förstemann, Michael Sattler
Publication: Nucleic Acids Research
Publisher: Oxford University Press
Available at: https://academic.oup.com/nar/article/45/21/12536/4508871
License See below

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS
Nov 10, 2019

This Agreement between Christoph Hartlmüller ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number 4705380439916
License date Nov 10, 2019

Licensed content publisher	Oxford University Press
Licensed content publication	Nucleic Acids Research
Licensed content title	Molecular basis for asymmetry sensing of siRNAs by the *Drosophila* Loqs-PD/Dcr-2 complex in RNA interference
Licensed content author	Tants, Jan-Niklas; Fesser, Stephanie
Licensed content date	Oct 13, 2017
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	Development and Application of Novel NMR Methods for Biological Macromolecules
Publisher of your work	Technische Universität München
Expected publication date	Mar 2020
Permissions cost	0.00 EUR
Value added tax	0.00 EUR
Total	0.00 EUR
Title	Development and Application of Novel NMR Methods for Biological Macromolecules
Institution name	Technische Universität München
Expected presentation date	Mar 2020
Portions	Figure 6
Publisher Tax ID	GB125506730
Total	0.00 EUR
Terms and Conditions	

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

12. Other Terms and Conditions:

v1.4

VII.9 Other Reprint Permissions

Figure I-2 was created and provided by Wgsimon (https://commons.wikimedia.org/wiki/User:Wgsimon) and was obtained from

https://en.wikipedia.org/wiki/File:Transistor_Count_and_Moore%27s_Law_-_2011.svg . It is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license, which is available online (https://creativecommons.org/licenses/by-sa/3.0/deed.en).

Figure I-3 was obtained from the paper Prediction of Protein Structure Using Surface Accessibility Data. For the reprint permission, see Chapter VII.1

VIII References

1. Campbell, I.D., *Timeline: the march of structural biology.* Nat Rev Mol Cell Biol, 2002. **3**(5): p. 377-81.
2. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(4096): p. 223-30.
3. Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions.* Nat Rev Mol Cell Biol, 2005. **6**(3): p. 197-208.
4. Wright, P.E. and H.J. Dyson, *Intrinsically disordered proteins in cellular signalling and regulation.* Nat Rev Mol Cell Biol, 2015. **16**(1): p. 18-29.
5. Mandal, S., M. Moudgil, and S.K. Mandal, *Rational drug design.* Eur J Pharmacol, 2009. **625**(1-3): p. 90-100.
6. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.* Nature, 1953. **171**(4356): p. 737-8.
7. Kendrew, J.C., et al., *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.* Nature, 1958. **181**(4610): p. 662-6.
8. Henderson, R. and P.N. Unwin, *Three-dimensional model of purple membrane obtained by electron microscopy.* Nature, 1975. **257**(5521): p. 28-32.
9. Unwin, P.N. and R. Henderson, *Molecular structure determination by electron microscopy of unstained crystalline specimens.* J Mol Biol, 1975. **94**(3): p. 425-40.
10. Williamson, M.P., T.F. Havel, and K. Wuthrich, *Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry.* J Mol Biol, 1985. **182**(2): p. 295-315.
11. Kline, A.D., W. Braun, and K. Wuthrich, *Determination of the complete three-dimensional structure of the alpha-amylase inhibitor tendamistat in aqueous solution by nuclear magnetic resonance and distance geometry.* J Mol Biol, 1988. **204**(3): p. 675-724.
12. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res., 2000. **28**(1): p. 235-42.
13. Drenth, J. and J. Mesters, *Principles of protein x-ray crystallography*. 3rd ed. 2007, New York: Springer. xiv, 332 p.
14. Jaskolski, M., Z. Dauter, and A. Wlodawer, *A brief history of macromolecular crystallography, illustrated by a family tree and its Nobel fruits.* FEBS J, 2014. **281**(18): p. 3985-4009.
15. Fernandez-Leiro, R. and S.H. Scheres, *Unravelling biological macromolecules with cryo-electron microscopy.* Nature, 2016. **537**(7620): p. 339-46.
16. Cavanagh, J., *Protein NMR Spectroscopy: Principles and Practice*. 2nd ed. 2007, Amsterdam ; Boston: Academic Press. xxv, 885 p.
17. Keeler, J., *Understanding NMR spectroscopy*. 2nd ed. 2010, Chichester, U.K.: John Wiley and Sons. xiii, 511 p.
18. Levitt, M.H., *Spin dynamics : basics of nuclear magnetic resonance*. 2nd ed. 2008, Chichester, England ; Hoboken, NJ: John Wiley & Sons. xxv, 714 p., 7 p. of plates.
19. Shi, Y., *A glimpse of structural biology through X-ray crystallography.* Cell, 2014. **159**(5): p. 995-1014.
20. Dale, G.E., C. Oefner, and A. D'Arcy, *The protein as a variable in protein crystallization.* J Struct Biol, 2003. **142**(1): p. 88-97.

21. McPherson, A. and J.A. Gavira, *Introduction to protein crystallization.* Acta Crystallogr F Struct Biol Commun, 2014. **70**(Pt 1): p. 2-20.
22. Binshtein, E. and M.D. Ohi, *Cryo-electron microscopy and the amazing race to atomic resolution.* Biochemistry, 2015. **54**(20): p. 3133-41.
23. Cheng, Y., *Single-Particle Cryo-EM at Crystallographic Resolution.* Cell, 2015. **161**(3): p. 450-7.
24. Llorca, O., *Introduction to 3D reconstruction of macromolecules using single particle electron microscopy.* Acta Pharmacol Sin, 2005. **26**(10): p. 1153-64.
25. Carazo, J.M., et al., *Three-dimensional reconstruction methods in Single Particle Analysis from transmission electron microscopy data.* Arch Biochem Biophys, 2015. **581**: p. 39-48.
26. Wang, H.W. and J.W. Wang, *How cryo-electron microscopy and X-ray crystallography complement each other.* Protein Sci, 2016.
27. Leschziner, A.E. and E. Nogales, *Visualizing flexibility at molecular resolution: analysis of heterogeneity in single-particle electron microscopy reconstructions.* Annu Rev Biophys Biomol Struct, 2007. **36**: p. 43-62.
28. Frueh, D.P., *Practical aspects of NMR signal assignment in larger and challenging proteins.* Prog Nucl Magn Reson Spectrosc, 2014. **78**: p. 47-75.
29. Pervushin, K., et al., *Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution.* Proc Natl Acad Sci U S A, 1997. **94**(23): p. 12366-71.
30. Chiliveri, S.C. and M.V. Deshmukh, *Recent excitements in protein NMR: Large proteins and biologically relevant dynamics.* J Biosci, 2016. **41**(4): p. 787-803.
31. Jiang, Y. and C.G. Kalodimos, *NMR Studies of Large Proteins.* J Mol Biol, 2017. **429**(17): p. 2667-2676.
32. Gobl, C., et al., *NMR approaches for structural analysis of multidomain proteins and complexes in solution.* Prog Nucl Magn Reson Spectrosc, 2014. **80**: p. 26-63.
33. Mooers, B.H., *Crystallographic studies of DNA and RNA.* Methods, 2009. **47**(3): p. 168-76.
34. van den Bedem, H. and J.S. Fraser, *Integrative, dynamic structural biology at atomic resolution--it's about time.* Nat Methods, 2015. **12**(4): p. 307-18.
35. Ward, A.B., A. Sali, and I.A. Wilson, *Biochemistry. Integrative structural biology.* Science, 2013. **339**(6122): p. 913-5.
36. Schlundt, A., J.N. Tants, and M. Sattler, *Integrated structural biology to unravel molecular mechanisms of protein-RNA recognition.* Methods, 2017.
37. Faini, M., F. Stengel, and R. Aebersold, *The Evolving Contribution of Mass Spectrometry to Integrative Structural Biology.* J Am Soc Mass Spectrom, 2016. **27**(6): p. 966-74.
38. Konermann, L., S. Vahidi, and M.A. Sowole, *Mass spectrometry methods for studying structure and dynamics of biological macromolecules.* Anal Chem, 2014. **86**(1): p. 213-32.
39. Hennig, J. and M. Sattler, *The dynamic duo: combining NMR and small angle scattering in structural biology.* Protein Sci, 2014. **23**(6): p. 669-82.
40. Moore, G.E., *Cramming More Components onto Integrated Circuits.* Electronics, 1965: p. 114-117.
41. Aue, W.P., E. Bartholdi, and R.R. Ernst, *Two - dimensional spectroscopy. Application to nuclear magnetic resonance.* The Journal of Chemical Physics, 1976. **64**(5): p. 2229-2246.
42. Hyberts, S.G., et al., *Perspectives in magnetic resonance: NMR in the post-FFT era.* J Magn Reson, 2014. **241**: p. 60-73.
43. Nowakowski, M., et al., *Applications of high dimensionality experiments to biomolecular NMR.* Prog Nucl Magn Reson Spectrosc, 2015. **90-91**: p. 49-73.

44. McCammon, J.A., B.R. Gelin, and M. Karplus, *Dynamics of folded proteins.* Nature, 1977. **267**(5612): p. 585-90.
45. Hospital, A., et al., *Molecular dynamics simulations: advances and applications.* Adv Appl Bioinform Chem, 2015. **8**: p. 37-47.
46. Orozco, M., et al., *Coarse-grained representation of protein flexibility. Foundations, successes, and shortcomings.* Adv Protein Chem Struct Biol, 2011. **85**: p. 183-215.
47. Kamerlin, S.C., et al., *Coarse-grained (multiscale) simulations in studies of biophysical and chemical systems.* Annu Rev Phys Chem, 2011. **62**: p. 41-64.
48. Kamerlin, S.C. and A. Warshel, *Multiscale modeling of biological functions.* Phys Chem Chem Phys, 2011. **13**(22): p. 10401-11.
49. Maximova, T., et al., *Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics.* PLoS Comput Biol, 2016. **12**(4): p. e1004619.
50. Karplus, M. and J.A. McCammon, *Molecular dynamics simulations of biomolecules.* Nat Struct Biol, 2002. **9**(9): p. 646-52.
51. Sim, A.Y., P. Minary, and M. Levitt, *Modeling nucleic acids.* Curr Opin Struct Biol, 2012. **22**(3): p. 273-8.
52. Bohm, G., *New approaches in molecular structure prediction.* Biophys Chem, 1996. **59**(1-2): p. 1-32.
53. Schonbrun, J., W.J. Wedemeyer, and D. Baker, *Protein structure prediction in 2002.* Curr Opin Struct Biol, 2002. **12**(3): p. 348-54.
54. Bonneau, R., et al., *De novo prediction of three-dimensional structures for major protein families.* J Mol Biol, 2002. **322**(1): p. 65-78.
55. Bonneau, R. and D. Baker, *Ab initio protein structure prediction: progress and prospects.* Annu Rev Biophys Biomol Struct, 2001. **30**: p. 173-89.
56. Perez, A., et al., *Advances in free-energy-based simulations of protein folding and ligand binding.* Curr Opin Struct Biol, 2016. **36**: p. 25-31.
57. Lindorff-Larsen, K., et al., *How fast-folding proteins fold.* Science, 2011. **334**(6055): p. 517-20.
58. Dill, K.A. and J.L. MacCallum, *The protein-folding problem, 50 years on.* Science, 2012. **338**(6110): p. 1042-6.
59. Leelananda, S.P. and S. Lindert, *Computational methods in drug discovery.* Beilstein J Org Chem, 2016. **12**: p. 2694-2718.
60. Hernandez-Rodriguez, M., et al., *Current Tools and Methods in Molecular Dynamics (MD) Simulations for Drug Design.* Curr Med Chem, 2016. **23**(34): p. 3909-3924.
61. Ghemtio, L., et al., *Recent trends and applications in 3D virtual screening.* Comb Chem High Throughput Screen, 2012. **15**(9): p. 749-69.
62. Pantazes, R.J., M.J. Grisewood, and C.D. Maranas, *Recent advances in computational protein design.* Curr Opin Struct Biol, 2011. **21**(4): p. 467-72.
63. Cramer, C.J., *Essentials of computational chemistry : theories and models*. 2nd ed. 2004, Chichester, West Sussex, England ; Hoboken, NJ: Wiley. xx, 596 p.
64. Jensen, F., *Introduction to computational chemistry*. Third edition. ed. 2017, Chichester, UK ; Hoboken, NJ: John Wiley & Sons. pages cm.
65. Quesne, M.G., T. Borowski, and S.P. de Visser, *Quantum Mechanics/Molecular Mechanics Modeling of Enzymatic Processes: Caveats and Breakthroughs.* Chemistry, 2016. **22**(8): p. 2562-81.
66. Groenhof, G., *Introduction to QM/MM simulations.* Methods Mol Biol, 2013. **924**: p. 43-66.

67. Piana, S., K. Lindorff-Larsen, and D.E. Shaw, *How robust are protein folding simulations with respect to force field parameterization?* Biophys J, 2011. **100**(9): p. L47-9.
68. Piana, S., J.L. Klepeis, and D.E. Shaw, *Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations.* Curr Opin Struct Biol, 2014. **24**: p. 98-105.
69. Vanommeslaeghe, K., O. Guvench, and A.D. MacKerell, Jr., *Molecular mechanics.* Curr Pharm Des, 2014. **20**(20): p. 3281-92.
70. Beauchamp, K.A., et al., *Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements.* J Chem Theory Comput, 2012. **8**(4): p. 1409-1414.
71. Chen, W., et al., *Conformational Dynamics of Two Natively Unfolded Fragment Peptides: Comparison of the AMBER and CHARMM Force Fields.* J Phys Chem B, 2015. **119**(25): p. 7902-10.
72. Bermudez, M., et al., *More than a look into a crystal ball: protein structure elucidation guided by molecular dynamics simulations.* Drug Discov Today, 2016. **21**(11): p. 1799-1805.
73. Perilla, J.R., et al., *Molecular dynamics simulations of large macromolecular complexes.* Curr Opin Struct Biol, 2015. **31**: p. 64-74.
74. Salvatella, X., *Understanding protein dynamics using conformational ensembles.* Adv Exp Med Biol, 2014. **805**: p. 67-85.
75. Karaca, E. and A.M. Bonvin, *Advances in integrative modeling of biomolecular complexes.* Methods, 2013. **59**(3): p. 372-81.
76. Schneidman-Duhovny, D., R. Pellarin, and A. Sali, *Uncertainty in integrative structural modeling.* Curr Opin Struct Biol, 2014. **28**: p. 96-104.
77. Bothwell, J.H. and J.L. Griffin, *An introduction to biological nuclear magnetic resonance spectroscopy.* Biol Rev Camb Philos Soc, 2011. **86**(2): p. 493-510.
78. Kwan, A.H., et al., *Macromolecular NMR spectroscopy for the non-spectroscopist.* FEBS J, 2011. **278**(5): p. 687-703.
79. Marion, D., *An introduction to biological NMR spectroscopy.* Mol Cell Proteomics, 2013. **12**(11): p. 3006-25.
80. Kleckner, I.R. and M.P. Foster, *An introduction to NMR-based approaches for measuring protein dynamics.* Biochim Biophys Acta, 2011. **1814**(8): p. 942-68.
81. Kovermann, M., P. Rogne, and M. Wolf-Watz, *Protein dynamics and function from solution state NMR spectroscopy.* Q Rev Biophys, 2016. **49**: p. e6.
82. Carver, T.R. and C.P. Slichter, *Experimental Verification of the Overhauser Nuclear Polarization Effect.* Physical Review, 1956. **102**(4): p. 975-980.
83. Wenk, P., et al., *Dynamic nuclear polarization of nucleic acid with endogenously bound manganese.* J Biomol NMR, 2015. **63**(1): p. 97-109.
84. Sattler, M., *Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients.* Progress in Nuclear Magnetic Resonance Spectroscopy, 1999. **34**(2): p. 93-158.
85. Kay, L.E., et al., *Three-dimensional triple-resonance NMR Spectroscopy of isotopically enriched proteins. 1990.* J Magn Reson, 2011. **213**(2): p. 423-41.
86. Kowalewski, J.z. and L. Mäler, *Nuclear spin relaxation in liquids : theory, experiments, and applications*. Series in chemical physics. 2006, New York: Taylor & Francis. 426 p.
87. Kempf, J.G. and J.P. Loria, *Protein Dynamics from Solution NMR.* Cell Biochemistry and Biophysics, 2002. **37**(3): p. 187-212.
88. Karplus, M., *Contact Electron - Spin Coupling of Nuclear Magnetic Moments.* The Journal of Chemical Physics, 1959. **30**(1): p. 11-15.

89. Pardi, A., M. Billeter, and K. Wuthrich, *Calibration of the angular dependence of the amide proton-C alpha proton coupling constants, 3JHN alpha, in a globular protein. Use of 3JHN alpha for identification of helical secondary structure.* J Mol Biol, 1984. **180**(3): p. 741-51.
90. Li, F., et al., *High accuracy of Karplus equations for relating three-bond J couplings to protein backbone torsion angles.* Chemphyschem, 2015. **16**(3): p. 572-8.
91. Overhauser, A.W., *Polarization of Nuclei in Metals.* Physical Review, 1953. **92**(2): p. 411-415.
92. Vogeli, B., *The nuclear Overhauser effect from a quantitative perspective.* Prog Nucl Magn Reson Spectrosc, 2014. **78**: p. 1-46.
93. Tjandra, N. and A. Bax, *Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium.* Science, 1997. **278**(5340): p. 1111-4.
94. Chen, K. and N. Tjandra, *The use of residual dipolar coupling in studying proteins by NMR.* Top Curr Chem, 2012. **326**: p. 47-67.
95. Battiste, J.L. and G. Wagner, *Utilization of site-directed spin labeling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited nuclear overhauser effect data.* Biochemistry, 2000. **39**(18): p. 5355-65.
96. Kosen, P.A., *[5] Spin labeling of proteins.* 1989. **177**: p. 86-121.
97. Clore, G.M. and J. Iwahara, *Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes.* Chem Rev, 2009. **109**(9): p. 4108-39.
98. Bertini, I., C. Luchinat, and G. Parigi, *Paramagnetic constraints: An aid for quick solution structure determination of paramagnetic metalloproteins.* Concepts in Magnetic Resonance, 2002. **14**(4): p. 259-286.
99. Otting, G., *Protein NMR using paramagnetic ions.* Annu Rev Biophys, 2010. **39**: p. 387-405.
100. Solomon, I., *Relaxation Processes in a System of Two Spins.* Physical Review, 1955. **99**(2): p. 559-565.
101. Bloembergen, N. and L.O. Morgan, *Proton Relaxation Times in Paramagnetic Solutions. Effects of Electron Spin Relaxation.* The Journal of Chemical Physics, 1961. **34**(3): p. 842-850.
102. Solomon, I. and N. Bloembergen, *Nuclear Magnetic Interactions in the HF Molecule.* The Journal of Chemical Physics, 1956. **25**(2): p. 261-266.
103. Kosen, P.A., *Spin labeling of proteins.* Methods Enzymol, 1989. **177**: p. 86-121.
104. Pintacuda, G., et al., *NMR structure determination of protein-ligand complexes by lanthanide labeling.* Acc Chem Res, 2007. **40**(3): p. 206-12.
105. de Dios, A.C. and E. Oldfield, *Recent progress in understanding chemical shifts.* Solid State Nuclear Magnetic Resonance, 1996. **6**(2): p. 101-125.
106. Wishart, D.S. and B.D. Sykes, *[12] Chemical shifts as a tool for structure determination.* 1994. **239**: p. 363-392.
107. Oldfield, E., *Chemical shifts and three-dimensional protein structures.* J Biomol NMR, 1995. **5**(3): p. 217-25.
108. Wishart, D.S., *Interpreting protein chemical shift data.* Prog Nucl Magn Reson Spectrosc, 2011. **58**(1-2): p. 62-87.
109. Wishart, D.S., B.D. Sykes, and F.M. Richards, *Relationship between nuclear magnetic resonance chemical shift and protein secondary structure.* Journal of Molecular Biology, 1991. **222**(2): p. 311-333.

110. Wishart, D.S. and B.D. Sykes, *The 13C Chemical-Shift Index: A simple method for the identification of protein secondary structure using 13C chemical-shift data.* Journal of Biomolecular NMR, 1994. **4**(2).
111. Shen, Y., et al., *TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts.* J Biomol NMR, 2009. **44**(4): p. 213-23.
112. Han, B., et al., *SHIFTX2: significantly improved protein chemical shift prediction.* J Biomol NMR, 2011. **50**(1): p. 43-57.
113. Shen, Y. and A. Bax, *SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network.* J Biomol NMR, 2010. **48**(1): p. 13-22.
114. Shen, Y., et al., *Consistent blind protein structure generation from NMR chemical shift data.* Proc. Natl. Acad. Sci. USA, 2008. **105**(12): p. 4685-90.
115. Raman, S., et al., *NMR Structure Determination for Larger Proteins Using Backbone-Only Data.* Science, 2010.
116. Lange, O.F., et al., *Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples.* Proc Natl Acad Sci U S A, 2012. **109**(27): p. 10873-8.
117. Williamson, M.P., *Using chemical shift perturbation to characterise ligand binding.* Prog Nucl Magn Reson Spectrosc, 2013. **73**: p. 1-16.
118. Hocking, H.G., K. Zangger, and T. Madl, *Studying the structure and dynamics of biomolecules by using soluble paramagnetic probes.* Chemphyschem, 2013. **14**(13): p. 3082-94.
119. Wang, Y., C.D. Schwieters, and N. Tjandra, *Parameterization of solvent-protein interaction and its use on NMR protein structure determination.* J Magn Reson, 2012. **221**: p. 76-84.
120. Madl, T., et al., *Structural analysis of large protein complexes using solvent paramagnetic relaxation enhancements.* Angew Chem Int Ed Engl, 2011. **50**(17): p. 3993-7.
121. Madl, T., W. Bermel, and K. Zangger, *Use of relaxation enhancements in a paramagnetic environment for the structure determination of proteins using NMR spectroscopy.* Angew Chem Int Ed Engl, 2009. **48**(44): p. 8259-62.
122. Pintacuda, G. and G. Otting, *Identification of protein surfaces by NMR measurements with a paramagnetic Gd(III) chelate.* J. Am. Chem. Soc., 2002. **124**(3): p. 372-373.
123. Bernini, A., et al., *Probing protein surface accessibility with solvent and paramagnetic molecules.* Prog. Nucl. Mag. Res. Sp., 2009. **54**(3–4): p. 278-289.
124. Esposito, G., et al., *A 1H NMR study on the interaction of aminoxyl paramagnetic probes with unfolded peptides.* Journal of the Chemical Society, Perkin Transactions 2, 1993(8): p. 1531.
125. Hernández, G., et al., *O2Penetration and Proton Burial Depth in Proteins: Applicability to Fold Family Recognition.* Journal of the American Chemical Society, 2002. **124**(16): p. 4463-4472.
126. Venditti, V., N. Niccolai, and S.E. Butcher, *Measuring the dynamic surface accessibility of RNA with the small paramagnetic molecule TEMPOL.* Nucleic Acids Res, 2008. **36**(4): p. e20.
127. Bernini, A., et al., *NMR studies on the surface accessibility of the archaeal protein Sso7d by using TEMPOL and Gd(III)(DTPA-BMA) as paramagnetic probes.* Biophys Chem, 2008. **137**(2-3): p. 71-5.
128. Hartlmuller, C., C. Gobl, and T. Madl, *Prediction of Protein Structure Using Surface Accessibility Data.* Angew Chem Int Ed Engl, 2016. **55**(39): p. 11970-4.

129. Zangger, K., et al., *Positioning of micelle-bound peptides by paramagnetic relaxation enhancements.* J Phys Chem B, 2009. **113**(13): p. 4400-6.
130. Lupas, A.N., *The long coming of computational structural biology.* J Struct Biol, 2008. **163**(3): p. 254-7.
131. Vlachakis, D., et al., *Current state-of-the-art molecular dynamics methods and applications.* Adv Protein Chem Struct Biol, 2014. **94**: p. 269-313.
132. Bernardi, R.C., M.C. Melo, and K. Schulten, *Enhanced sampling techniques in molecular dynamics simulations of biological systems.* Biochim Biophys Acta, 2015. **1850**(5): p. 872-7.
133. Bradley, P., K.M. Misura, and D. Baker, *Toward high-resolution de novo structure prediction for small proteins.* Science, 2005. **309**(5742): p. 1868-71.
134. Floudas, C.A., et al., *Advances in protein structure prediction and de novo protein design: A review.* Chemical Engineering Science, 2006. **61**(3): p. 966-988.
135. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. **294**(5540): p. 93-6.
136. Maurer, S.A., L. Clin, and C. Ochsenfeld, *Cholesky-decomposed density MP2 with density fitting: accurate MP2 and double-hybrid DFT energies for large systems.* J Chem Phys, 2014. **140**(22): p. 224112.
137. Mackerell, A.D., Jr., *Empirical force fields for biological macromolecules: overview and issues.* J Comput Chem, 2004. **25**(13): p. 1584-604.
138. Ryde, U., *QM/MM Calculations on Proteins.* Methods Enzymol, 2016. **577**: p. 119-58.
139. Rizzuti, B. and V. Daggett, *Using simulations to provide the framework for experimental protein folding studies.* Arch Biochem Biophys, 2013. **531**(1-2): p. 128-35.
140. Gelman, H. and M. Gruebele, *Fast protein folding kinetics.* Q Rev Biophys, 2014. **47**(2): p. 95-142.
141. Doshi, U. and D. Hamelberg, *Towards fast, rigorous and efficient conformational sampling of biomolecules: Advances in accelerated molecular dynamics.* Biochim Biophys Acta, 2015. **1850**(5): p. 878-88.
142. Laio, A. and M. Parrinello, *Escaping free-energy minima.* Proc Natl Acad Sci U S A, 2002. **99**(20): p. 12562-6.
143. Hansen, H.S. and P.H. Hunenberger, *Using the local elevation method to construct optimized umbrella sampling potentials: calculation of the relative free energies and interconversion barriers of glucopyranose ring conformers in water.* J Comput Chem, 2010. **31**(1): p. 1-23.
144. Sugitaa, Y. and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding.* Chem. Phys. Lett., 1999. **314**(1-2): p. 141-151.
145. Bonomi, M., C. Camilloni, and M. Vendruscolo, *Metadynamic metainference: Enhanced sampling of the metainference ensemble using metadynamics.* Sci Rep, 2016. **6**: p. 31232.
146. Camilloni, C., A. Cavalli, and M. Vendruscolo, *Replica-Averaged Metadynamics.* J Chem Theory Comput, 2013. **9**(12): p. 5610-7.
147. Granata, D., et al., *Characterization of the free-energy landscapes of proteins by NMR-guided metadynamics.* Proc Natl Acad Sci U S A, 2013. **110**(17): p. 6817-22.
148. Liwo, A., et al., *Computational techniques for efficient conformational sampling of proteins.* Curr Opin Struct Biol, 2008. **18**(2): p. 134-9.
149. Metropolis, N., et al., *Equation of State Calculations by Fast Computing Machines.* The Journal of Chemical Physics, 1953. **21**(6): p. 1087-1092.
150. Chib, S. and E. Greenberg, *Understanding the Metropolis-Hastings Algorithm.* The American Statistician, 1995. **49**(4): p. 327-335.

151. Kaufmann, K.W., et al., *Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You.* Biochemistry, 2010. **49**(14): p. 2987-2998.
152. Das, R. and D. Baker, *Macromolecular modeling with rosetta.* Annu. Rev. Biochem., 2008. **77**: p. 363-82.
153. Simons, K.T., et al., *Ab initio protein structure prediction of CASP III targets using ROSETTA.* Proteins, 1999. **Suppl 3**: p. 171-6.
154. Grant, A., D. Lee, and C. Orengo, *Progress towards mapping the universe of protein folds.* Genome Biol, 2004. **5**(5): p. 107.
155. Wolf, Y.I., N.V. Grishin, and E.V. Koonin, *Estimating the number of protein folds and families from complete genome data.* J Mol Biol, 2000. **299**(4): p. 897-905.
156. Govindarajan, S., R. Recabarren, and R.A. Goldstein, *Estimating the total number of protein folds.* Proteins, 1999. **35**(4): p. 408-14.
157. Magner, A., W. Szpankowski, and D. Kihara, *On the origin of protein superfamilies and superfolds.* Sci Rep, 2015. **5**: p. 8166.
158. Schwieters, C.D., et al., *The Xplor-NIH NMR molecular structure determination package.* Journal of Magnetic Resonance, 2003. **160**(1): p. 65-73.
159. Schwieters, C.D., G.A. Bermejo, and G.M. Clore, *Xplor-NIH for molecular structure determination from NMR and other data sources.* Protein Sci, 2018. **27**(1): p. 26-40.
160. Brunger, A.T., et al., *Crystallography & NMR system: A new software suite for macromolecular structure determination.* Acta Crystallogr D Biol Crystallogr, 1998. **54**(Pt 5): p. 905-21.
161. Wurz, J.M., et al., *NMR-based automated protein structure determination.* Arch Biochem Biophys, 2017. **628**: p. 24-32.
162. Evangelidis, T., et al., *Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra.* Nat Commun, 2018. **9**(1): p. 384.
163. Jung, Y.S. and M. Zweckstetter, *Mars -- robust automatic backbone assignment of proteins.* J Biomol NMR, 2004. **30**(1): p. 11-23.
164. Leutner, M., et al., *Automated backbone assignment of labeled proteins using the threshold accepting algorithm.* Journal of Biomolecular NMR, 1998. **11**(1): p. 31-43.
165. Rieping, W., et al., *ARIA2: automated NOE assignment and data integration in NMR structure calculation.* Bioinformatics, 2007. **23**(3): p. 381-2.
166. Guntert, P. and L. Buchner, *Combined automated NOE assignment and structure calculation with CYANA.* J Biomol NMR, 2015. **62**(4): p. 453-71.
167. Cuniasse, P., et al., *Structures of biomolecular complexes by combination of NMR and cryoEM methods.* Curr Opin Struct Biol, 2017. **43**: p. 104-113.
168. van Ingen, H. and A.M. Bonvin, *Information-driven modeling of large macromolecular assemblies using NMR data.* J Magn Reson, 2014. **241**: p. 103-14.
169. Sibille, N. and P. Bernado, *Structural characterization of intrinsically disordered proteins by the combined use of NMR and SAXS.* Biochem Soc Trans, 2012. **40**(5): p. 955-62.
170. Mackereth, C.D. and M. Sattler, *Dynamics in multi-domain protein recognition of RNA.* Curr Opin Struct Biol, 2012. **22**(3): p. 287-96.
171. Kerfah, R., et al., *Methyl-specific isotopic labeling: a molecular tool box for solution NMR studies of large proteins.* Curr Opin Struct Biol, 2015. **32**: p. 113-22.
172. Frueh, D.P., et al., *NMR methods for structural studies of large monomeric and multimeric proteins.* Curr Opin Struct Biol, 2013. **23**(5): p. 734-9.
173. Im, W., et al., *Challenges in structural approaches to cell modeling.* J Mol Biol, 2016. **428**(15): p. 2943-64.

174. Camilloni, C., et al., *Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts.* Biochemistry, 2012. **51**(11): p. 2224-31.
175. Kragelj, J., et al., *Conformational propensities of intrinsically disordered proteins from NMR chemical shifts.* Chemphyschem, 2013. **14**(13): p. 3034-45.
176. Ambadipudi, S. and M. Zweckstetter, *Targeting intrinsically disordered proteins in rational drug discovery.* Expert Opin Drug Discov, 2016. **11**(1): p. 65-77.
177. Carlomagno, T., *Present and future of NMR for RNA-protein complexes: a perspective of integrated structural biology.* J Magn Reson, 2014. **241**: p. 126-36.
178. Yadav, D.K. and P.J. Lukavsky, *NMR solution structure determination of large RNA-protein complexes.* Prog Nucl Magn Reson Spectrosc, 2016. **97**: p. 57-81.
179. Dill, K.A. and H.S. Chan, *From Levinthal to pathways to funnels.* Nat Struct Biol, 1997. **4**(1): p. 10-9.
180. Zwanzig, R., A. Szabo, and B. Bagchi, *Levinthal's paradox.* Proc Natl Acad Sci U S A, 1992. **89**(1): p. 20-2.
181. Pilla, K.B., K. Gaalswyk, and J.L. MacCallum, *Molecular modeling of biomolecules by paramagnetic NMR and computational hybrid methods.* Biochim Biophys Acta, 2017. **1865**(11 Pt B): p. 1654-1663.
182. Shaw, D.E., et al., *Millisecond-Scale Molecular Dynamics Simulations on Anton.* Proceedings of the ACM/IEEE Conference on Supercomputing (SC09), 2009.
183. Shaw, D.E., et al., *Anton, a special-purpose machine for molecular dynamics simulation.* Communications of the ACM - Web science, 2008. **51**(7): p. 91-97
184. Duss, O., P.J. Lukavsky, and F.H. Allain, *Isotope labeling and segmental labeling of larger RNAs for NMR structural studies.* Adv Exp Med Biol, 2012. **992**: p. 121-44.
185. Hennig, M., et al., *Recent advances in RNA structure determination by NMR.* Curr Protoc Nucleic Acid Chem, 2001. **Chapter 7**: p. Unit 7 7.
186. Hartlmuller, C., et al., *NMR characterization of solvent accessibility and transient structure in intrinsically disordered proteins.* J Biomol NMR, 2019. **73**(6-7): p. 305-317.
187. Hartlmuller, C., et al., *RNA structure refinement using NMR solvent accessibility data.* Sci Rep, 2017. **7**(1): p. 5393.
188. Gu, X.H., et al., *A decadentate Gd(III)-coordinating paramagnetic cosolvent for protein relaxation enhancement measurement.* J Biomol NMR, 2014. **58**(3): p. 149-54.
189. Lange, O.F. and D. Baker, *Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation.* Proteins, 2012. **80**(3): p. 884-95.

IX Appendix: Reprints of Papers

This appendix contains the reprints for the following papers and the corresponding supplementary information:

- **Prediction of Protein Structure Using Surface Accessibility Data**
  Christoph Hartlmüller, Christoph Göbl and Tobias Madl
  Angew Chem Int Ed Engl. 2016 Sep 19; 55(39):11970-11974
  doi: 10.1002/anie.201604788
- **RNA Structure Refinement using NMR Solvent Accessibility Data**
  Christoph Hartlmüller*, Johannes C. Günther*, Antje C. Wolter, Jens Wöhnert, Michael Sattler and Tobias Madl
  Sci Rep. 2017 Jul 14;7(1):5393.
  doi: 10.1038/s41598-017-05821-z
  * Shared first author
- **NMR Characterization of Solvent Accessibility and Transient Structure in Intrinsically Disordered Proteins**
  Christoph Hartlmüller*, Emil Spreitzer*, Christoph Göbl, Fabio Falsone and Tobias Madl
  J Biomol NMR. 2019 Jul;73(6-7):305-317
  doi: 10.1007/s10858-019-00248-2
  * Shared first author

Structural Biology

International Edition: DOI: 10.1002/anie.201604788
German Edition: DOI: 10.1002/ange.201604788

Prediction of Protein Structure Using Surface Accessibility Data

*Christoph Hartlmüller, Christoph Göbl, and Tobias Madl**

Abstract: *An approach to the de novo structure prediction of proteins is described that relies on surface accessibility data from NMR paramagnetic relaxation enhancements by a soluble paramagnetic compound (sPRE). This method exploits the distance-to-surface information encoded in the sPRE data in the chemical shift-based CS-Rosetta de novo structure prediction framework to generate reliable structural models. For several proteins, it is demonstrated that surface accessibility data is an excellent measure of the correct protein fold in the early stages of the computational folding algorithm and significantly improves accuracy and convergence of the standard Rosetta structure prediction approach.*

During the last few decades, NMR spectroscopy has become the method of choice for studying high-resolution protein structures in solution. In the standard NMR-based structure determination approach, structurally relevant data from different sources, such as pair-wise interatomic distances and orientation information, are collected and used as restraints for structure calculation.[1] Very recently, several groups have realized that the growing number of structural data available in the Protein Data Base[2] (PDB) provide a valuable source for NMR-based structure determination, in particular when combined with NMR chemical shifts.[3] In these de novo structure prediction approaches, only the amino acid sequence is needed, and structures are calculated in an often Monte Carlo-based conformation-searching algorithm. The benefits of NMR chemical shift data in fragment selection and evaluation of structural quality have been recognized[4] and impressively demonstrated.[3,5] However, this method is still limited to small proteins owing to computational bottlenecks[6] and requires extensive sets of NMR-based structural data, which are difficult to obtain in case of larger proteins as a result of the increasing complexity of NMR spectra and line broadening of NMR signals because of overall slower protein tumbling.

Herein we describe an approach in which we exploit NMR-based surface accessibility data obtained from measurement of paramagnetic relaxation enhancements induced by a soluble paramagnetic compound for de novo structure prediction in the Rosetta framework.[6,7] The addition of soluble paramagnetic compounds leads to a concentration-dependent increase of relaxation rates, the so-called paramagnetic relaxation enhancement (here denoted as solvent PRE, sPRE; also known as co-solute PRE, Figure 1a). This effect depends on the distance of the spin to the protein surface, with the spins on the surface being affected most, and has been shown to correlate well with protein structure.[8] sPREs have been exploited for structural studies of biomol-

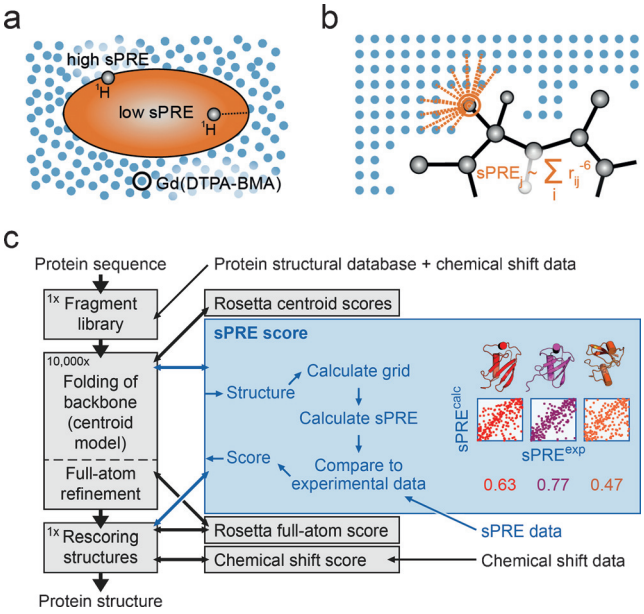


Figure 1. Principle of sPRE-CS-Rosetta. a) NMR sPRE data provides quantitative and residue specific information on the solvent accessibility as the effect of paramagnetic probes such as Gd(DTPA-BMA) is distance dependent. b) Back-calculation of sPRE data relies on placing the protein into equidistantly spaced grid points, while overlapping grid points are removed. The sPRE is approximated by the sum of all contributions of the surrounding grid points. c) The sPRE module is implemented as a scoring function capable of scoring centroid as well as full-atom models. At its core, the experimental sPRE data ($sPRE^{exp}$) is compared to the predicted sPRE data of the current Rosetta model ($sPRE^{calc}$) and a score based on the Spearman correlation coefficient (colored numbers) is computed. In this scheme, the sPRE score is used during the folding of the protein backbone using the simplified centroid model as well as for rescoring the final full-atom models.

[*] C. Hartlmüller, Dr. C. Göbl, Prof. Dr. T. Madl
Center for Integrated Protein Science Munich
Technische Universität München, Department of Chemistry
Lichtenbergstrasse 4, 85748 Garching (Germany)
and
Institute of Structural Biology, Helmholtz Zentrum München
Ingolstädter Landstrasse 1, 85764 Neuherberg (Germany)
Prof. Dr. T. Madl
Institute of Molecular Biology & Biochemistry
Center of Molecular Medicine, Medical University of Graz
8010 Graz (Austria)
E-mail: tobias.madl@medunigraz.at

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under: http://dx.doi.org/10.1002/anie.201604788.

© 2016 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

ecules such as for structure determination of proteins,[8a,9] docking of protein complexes,[10] and detection of dynamics[11] in the recent years.

Although sPRE data has been used to evaluate structural quality, its use in structure calculations has been limited owing to the lack of time-efficient computational methods for back-calculation of sPRE data. This is essential because in Rosetta, every scoring function (that is, the sPRE score) is evaluated several ten thousand times for obtaining a single structure. Furthermore, a typical structural ensemble required for accurate structure prediction contains at least several thousands of such structure models, emphasizing the need for efficient scoring functions. Recently, an approach has been presented for the molecular dynamics software XPLOR-NIH using a structure-based metric including the neighboring heavy atoms.[9] Herein, we use a different approach optimized for high-performance and time-efficiency in which we directly use a model structure and map it onto a bit array (Figure 1b). This simplifies the required computations to simple grid-based operations that are further accelerated by lookup tables. In this approach, the protein is placed in a regularly spaced grid represented by a three-dimensional bit array. Grid positions that overlap with the protein are marked, such that the remaining unmarked grid positions represent the inverted shape of the protein, and can be regarded as a spatial distribution of the paramagnetic agent. The sPRE of a protein atom is then calculated by summing up all contributions of the unmarked grid positions within the integration radius around the atom (Figure 1b).

We then extended the Rosetta de novo structure prediction method to incorporate sPRE data to take advantage of the surface accessibility information in the folding of the protein backbone (Figure 1c). A new scoring function for sPRE data was implemented and is available to the entire Rosetta framework. In short, the sPRE module first back-calculates the sPRE data for a given structure using the grid-based algorithm described above. The back-calculated sPRE data is then compared to the experimental sPRE data using the Spearman correlation coefficient and converted into an energy score (sPRE score).

The suitability of sPRE-based surface-accessibility data as an indicator of structural accuracy was evaluated for the individual CS-Rosetta refinement stages using a set of proteins ranging from of 6.4 to 41 kDa. To this end, we created structural ensembles for the individual stages of the Rosetta AbinitioRelax protocol and compared the sPRE

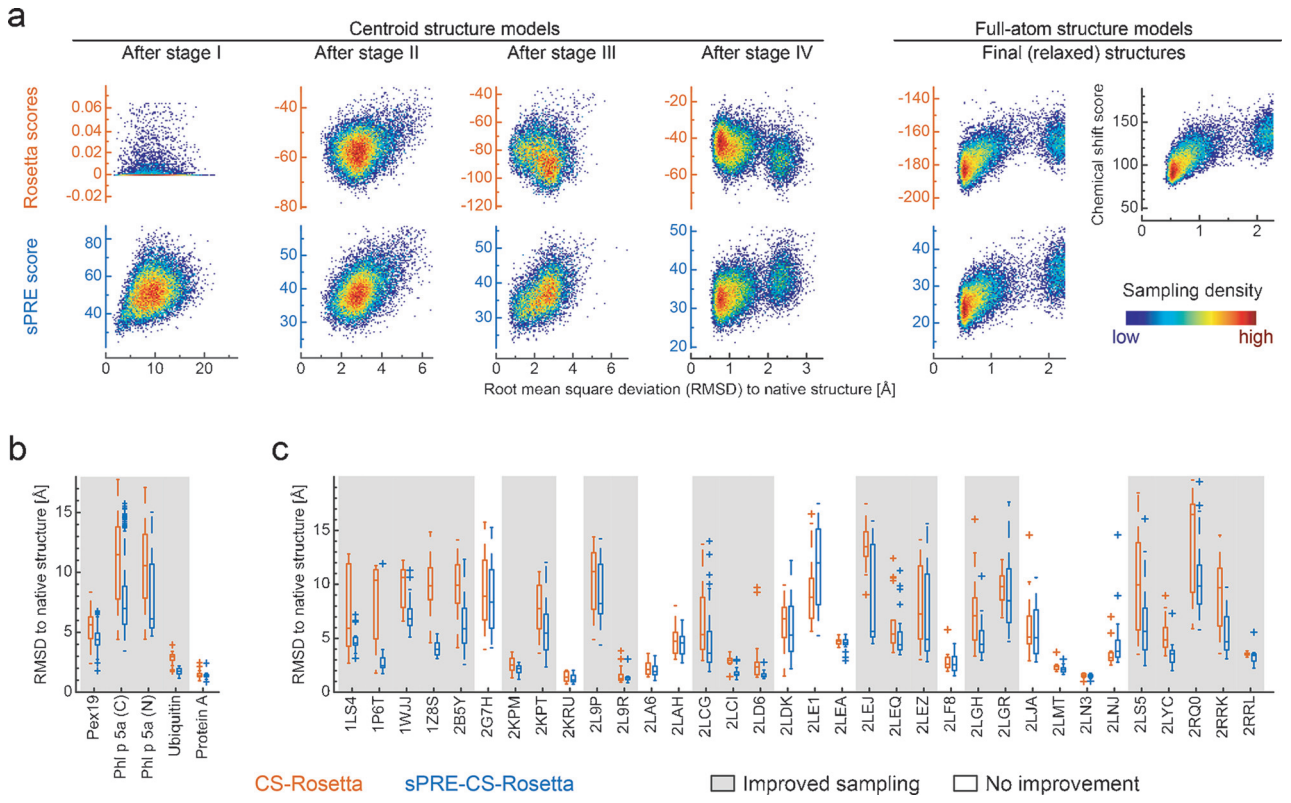


Figure 2. sPRE data is an excellent measure of the correct protein fold and improves protein structure prediction. a) Structural ensembles of ubiquitin representing different stages of the AbinitioRelax protocol were rescored using Rosetta centroid and full-atom scores (orange axis), the sPRE score (blue axis), and the chemical shift score (black axis). Experimental sPRE data for H^N and $H^{aliphatic}$ protons were used as input for the sPRE score. b),c) Box plots showing the average C^α-RMSD to the native structure for models obtained from CS-Rosetta (orange) and sPRE-CS-Rosetta (blue). sPRE data was determined by NMR experiments (b) or back-calculated (c). All obtained structural models were scored according to the sum of the Rosetta, chemical shift and sPRE score (b) or according to the sum of the Rosetta and the chemical shift score (c). For every protein, the best scored 0.2% structures of all models were selected and used to generate the box plots. Proteins for which the sampling was improved by the sPRE module (reduced mean RMSD to native structure compared to CS-Rosetta) are marked with a gray background and proteins for which CS-Rosetta and sPRE-CS-Rosetta failed are not shown (average C^α-RMSD >10 Å in the case of p16, 1CX1, 1F2H, 1GXE, 1IX5, 1ON4, 1RFL, 1XWE, 2KNR, 2LFC, 2LFP, 2LLL, 2PQE, 2RRF, 3ZQD, and 4A5V). All scores are shown in arbitrary units.

 © 2016 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

score to the Rosetta scores. We observed that the sPRE score outperforms the initial scores in the early protein folding stage I, which has been initially optimized to collapse the extended chain but also in the later stages II–IV in which the fold of the backbone is determined (Figure 2a; Supporting Information, Figure S1). Over a wide C^{α}-RMSD range of 3–20 Å, the sPRE score shows a clear correlation with structural accuracy. In the later stages II–IV, the quality of the standard Rosetta scores improves and they cooperate with the sPRE score when combined. This strongly indicates that the sPRE score is capable of guiding the sampling of a Rosetta AbinitioRelax run towards the native structure. Interestingly, for near-native-like structures (C^{α}-RMSD < 2 Å), the Rosetta score shows a better performance compared to the sPRE score. This is probably due to the higher susceptibility of the sPRE to variations on the protein surface where minor conformational changes, for example, side-chain rotations, translate into a large variation of the sPRE. Summarizing, our findings suggest that sPRE data can be valuable for Rosetta-based de novo structure prediction when sampling states far from the native state, since it is able to guide it towards more native-like states. From these more native-like states, the common Rosetta scoring functions are able to drive the sampling to high-resolution, full-atom structures.

To examine the potential of solvent accessibility data for Rosetta de novo structure prediction, we carried out classical CS-Rosetta as well as CS-Rosetta with sPRE scoring (referred to as sPRE-CS-Rosetta) calculations with experimental NMR data (Figure 2b; Supporting Information, Table S1) and back-calculated sPRE data (Figure 2c; Supporting Information, Table S2). For ubiquitin and using experimental amide ($^1H^N$) and aliphatic ($^1H^{aliphatic}$) proton sPRE data, the sPRE-CS-Rosetta approach improved the sampling significantly in a set of about 10000 models (Figure 3). As a result, more structures in the C^{α}-RMSD range up to 1.5 Å were sampled, and subsequently the common Rosetta scores converge to high-resolution structures as close as 0.7 Å C^{α}-RMSD to the native structure. The main structural difference of the ubiquitin ensemble at 2.5 Å compared to the ensemble at about 0.7 Å is a register shift of β-strand 5 (Figure 3a). To evaluate the robustness, we carried out sPRE-CS-Rosetta calculations using only subsets of the experimental sPRE data. Surprisingly, even with restricted sPRE data sets ($^1H^N$, sidechain ^1H, or $^1H^{\alpha}/^1H^{\beta}$) the sampling was not deteriorated (Supporting Information, Figure S2a). This suggests that the surface-accessibility information is already encoded in a low number of sPRE restraints and that scoring the global fold of a protein does not require precise input data as long as the correct trend of the solvent accessibility pattern is present in the data. This is further supported by the observation that a complete set of synthetic ubiquitin sPRE data did not further improve the structural quality (Supporting Information, Figure S2b). Summarizing, this indicates that even in case of sparse and incomplete chemical shift assignments, sPRE data can provide high-quality structural models. Similar results were obtained using experimental data for the C-terminal domain of Phl p 5a, a four helix bundle in which case the sPRE-CS-Rosetta approach significantly improved convergence and accuracy of the structural models in a set of about 100000 models (Figure 3b; Supporting Information, Figure S3).

To further examine the potential of solvent accessibility data for Rosetta structure prediction, we built a benchmark of challenging proteins with sizes up to 170 residues and using synthetic sPRE data (Supporting Information, Table S2). The structure of each protein was subsequently determined using classical CS-Rosetta as well as sPRE-CS-Rosetta. Comparing the average C^{α}-RMSD to the native structure for the best 0.2% models, filtered by the sum of Rosetta and chemical shift score, revealed that for several proteins (22 of 49) the accuracy of the structure prediction was notably improved to models closer to the native structure (Figure 2c). To solely account for the sampling improvement, we additionally filtered 1% of the models that are closest to the native structure and compared the average C^{α}-RMSD of these sets (Supporting Information, Figure S4). These results show that for most of the tested proteins (30 of 49) the sampling is significantly improved. Two proteins of this benchmark, 2LEJ and 1P6T, are illustrated in Figure 3c and 3d, respectively. To further evaluate the robustness of the sPRE scoring module, we determined the structure of four proteins using back-calculated sPRE data with an increasing level of noise, various assignment completeness and different sets of resonances (Supporting Information, Tables S3a–d). Our results for fully assigned proteins and using only sPRE data for H^N, H^{α}, and H^{β} resonances show that the sampling is improved even in the presence of simulated noise with a range of four times the sPRE value (± 2 sPRE value, here denoted as noise level of 200%). Moreover, even for partially assigned proteins and using only amide protons, which corresponds to less than one restraint per residue, the number of models close to the native structure is still enhanced. Therefore, the results of the benchmark showed that solvent accessibility data improves accuracy and convergence even if only sparse data is available.

To further evaluate the performance of sPRE-CS-Rosetta in combination with (sparse) NMR-based structural data, we carried out de novo structure predictions using random subsets of experimental nuclear Overhauser enhancement (NOE)-based distance and residual dipolar coupling (RDC)-based orientation data. Most notably, the addition of experimental sPRE data increases the sampling significantly in all cases (Supporting Information, Figure S5, Tables S4a–b, S5). This confirms that the sPRE data acts as an orthogonal restraint.

Iterative sampling has been shown to improve Rosetta-based de novo structure prediction in some cases. We compared the performance of our approach to the iterative sampling algorithm CS-Rasrec-Rosetta.[12] We find that the performance of the Rasrec-based structure predictions does not improve significantly in terms of sampling (that is, the RMSD of the best structures), but rather excludes the high-RMSD structures during the iteration. In line with this, inclusion of sPRE data in the sPRE-CS-Rosetta shows significantly improved performance (Supporting Information, Figure S6). An explanation for the comparable performance of CS-Rasrec-Rosetta is the fact that the Abinitio part of the classical Rosetta is still an integral part of CS-Rasrec-Rosetta.

 © 2016 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

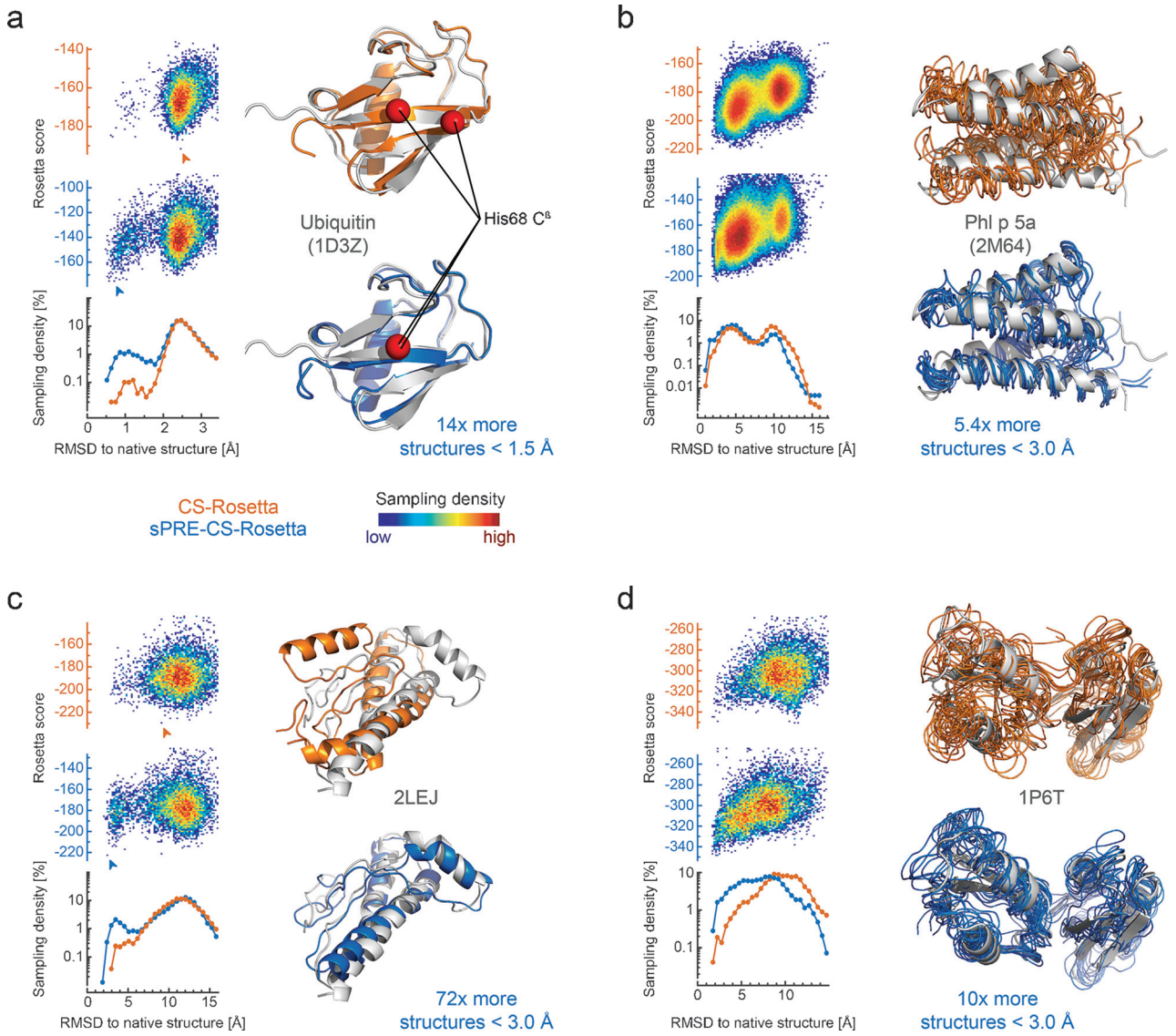


Figure 3. sPRE data enhances accuracy and convergence of CS-Rosetta structure prediction. The lowest-energy models of CS-Rosetta (orange) and sPRE-CS-Rosetta (blue) are compared to the NMR solution structures (gray, PDB code). For both methods, the corresponding Rosetta score (score13_env_hb) is plotted on the left and the distribution of the C^{α}-RMSD of the sampled structures is shown below for both methods in a logarithmic histogram. For ubiquitin (a) and the C-terminal domain of Phl p 5a (b) experimental sPRE data for amide and aliphatic protons is used, and for human prion protein (c) and the P-type ATPase CopA (d) the input sPRE data was back-calculated using the lowest energy model. In (a) and (c), the best scored model according to the Rosetta score is shown (see arrow in score plots), and for (b) and (d) the 10 lowest-energy models are shown. For ubiquitin (a), a red sphere represents the position of the C^{β} atom of His68, indicating the wrong positioning of the β-strand in the CS-Rosetta run. A more detailed picture of the scores is shown in the Supporting Information, Figure S3. All scores are shown in arbitrary units.

Our findings for several model proteins show that sPRE data improves conformational sampling and scoring of CS-Rosetta, subsequently provides more accurate and better converged structural models, and thereby effectively shifts the size limitations of CS-Rosetta. Our observation that a restricted set of sPRE data is sufficient to improve structural quality indicates that this class of restraints will be particularly powerful for de novo structure prediction of larger proteins where complete chemical shift assignments are difficult to obtain. With this respect sPRE data can be used in combination with (sparse) restraints from conventional approaches and offer several benefits over conventional approaches based on NOE-derived distance restraints only: sPRE data can be obtained for any kind of NMR-active nucleus for which chemical shift assignments are available (including for example ^{13}C[8a]), and as long as a NMR spectrum can be obtained. This is independent of the completeness of chemical shift assignments which is essential for NOE-based approaches. Combination of the sPRE-CS-Rosetta approach with recently developed iterative sampling algorithms,[12] or comparative modeling[13] in the future promises further improvements for de novo structure prediction of larger proteins. In these cases, surface accessibility data can be particularly useful as it provides orthogonal information

 © 2016 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

compared to other NMR restraints that often contain local, short-distance information. Furthermore, the sPRE module is open to complementary types of surface accessibility data such as for example bioinformatics and mass spectrometry (cross-linking, radical-mediated protein footprinting) data and will thereby allow integrating different techniques in one program.

Acknowledgements

We thank Dr. Oliver Lange for fruitful discussions and valuable advice. This work was supported by the Bavarian Ministry of Sciences, Research and the Arts (Bavarian Molecular Biosystems Research Network, to T.M.), the Deutsche Forschungsgemeinschaft (Emmy Noether program MA 5703/1-1, to T.M.), the President's International Fellowship Initiative of CAS (No. 2015VBB045, to T.M.), the National Natural Science Foundation of China (No. 31450110423, to T.M.), and the Austrian Science Fund (FWF: P28854). C.G. gratefully acknowledges the postdoctoral fellowship program (PFP) of the Helmholtz Zentrum München. The authors gratefully acknowledge the Leibniz Supercomputing Centre (LRZ, www.lrz.de) for funding this project by providing computing time on the Linux-Cluster (Projects t3623/t3671).

Keywords: CS-Rosetta · NMR spectroscopy · paramagnetic relaxation · protein structure prediction · structural biology

How to cite: *Angew. Chem. Int. Ed.* **2016**, *55*, 11970–11974
Angew. Chem. **2016**, *128*, 12149–12153

[1] a) C. Göbl, T. Madl, B. Simon, M. Sattler, *Prog. Nucl. Magn. Reson. Spectrosc.* **2014**, *80*, 26–63; b) J. Cavanagh, *Protein NMR Spectroscopy: Principles and Practice*, 2nd ed., Academic Press, Amsterdam, Boston, **2007**; c) G. S. Rule, T. K. Hitchens, *Fundamentals of Protein NMR Spectroscopy*, Springer, Dordrecht, **2006**.

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235–242.

[3] a) A. Cavalli, X. Salvatella, C. M. Dobson, M. Vendruscolo, *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 9615–9620; b) Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, A. Ignatchenko, C. H. Arrowsmith, T. Szyperski, G. T. Montelione, D. Baker, A. Bax, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 4685–4690; c) D. S. Wishart, D. Arndt, M. Berjanskii, P. Tang, J. Zhou, G. Lin, *Nucleic Acids Res.* **2008**, *36*, W496–W502.

[4] a) P. M. Bowers, C. E. M. Strauss, D. Baker, *J. Biomol. NMR* **2000**, *18*, 311–318; b) V. Tugarinov, W. Y. Choy, V. Y. Orekhov, L. E. Kay, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 622–627.

[5] S. Raman, O. F. Lange, P. Rossi, M. Tyka, X. Wang, J. Aramini, G. Liu, T. Ramelot, A. Eletsky, T. Szyperski, M. Kennedy, J. Prestegard, G. T. Montelione, D. Baker, *Science* **2010**, *327*, 1014–1018.

[6] R. Das, D. Baker, *Annu. Rev. Biochem.* **2008**, *77*, 363–382.

[7] K. W. Kaufmann, G. H. Lemmon, S. L. DeLuca, J. H. Sheehan, J. Meiler, *Biochemistry* **2010**, *49*, 2987–2998.

[8] a) T. Madl, W. Bermel, K. Zangger, *Angew. Chem. Int. Ed.* **2009**, *48*, 8259–8262; *Angew. Chem.* **2009**, *121*, 8409–8412; b) G. Pintacuda, G. Otting, *J. Am. Chem. Soc.* **2002**, *124*, 372–373; c) A. Bernini, V. Venditti, O. Spiga, N. Niccolai, *Prog. Nucl. Magn. Reson. Spectrosc.* **2009**, *54*, 278–289.

[9] Y. Wang, C. D. Schwieters, N. Tjandra, *J. Magn. Reson.* **2012**, *221*, 76–84.

[10] T. Madl, T. Guttler, D. Gorlich, M. Sattler, *Angew. Chem. Int. Ed.* **2011**, *50*, 3993–3997; *Angew. Chem.* **2011**, *123*, 4079–4083.

[11] a) H. G. Hocking, K. Zangger, T. Madl, *ChemPhysChem* **2013**, *14*, 3082–3094; b) Y. Sun, J. I. Friedman, J. T. Stivers, *Biochemistry* **2011**, *50*, 10724–10731.

[12] O. F. Lange, D. Baker, *Proteins Struct. Funct. Bioinf.* **2012**, *80*, 884–895.

[13] Y. Shen, A. Bax, *Nat. Methods* **2015**, *12*, 747–750.

Received: May 16, 2016
Revised: July 12, 2016
Published online: August 25, 2016

 © 2016 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

Angewandte Chemie
Eine Zeitschrift der Gesellschaft Deutscher Chemiker

Supporting Information

Prediction of Protein Structure Using Surface Accessibility Data

*Christoph Hartlmüller, Christoph Göbl, and Tobias Madl**

anie_201604788_sm_miscellaneous_information.pdf

Contents

Experimental Section

Rosetta framework and the sPRE energy module.

The sPRE module is capable of scoring full-atom models as well as simplified centroid models, which are used by CS-Rosetta during the folding of the backbone in the AbinitioRelax protocol. In this work, we used the classical CS-Rosetta protocol for structure determination consisting of i) a fragment selection using chemical shift data, ii) the AbinitioRelax protocol for generating structures and iii) a final rescoring of the structure ensemble.
It should also be noted, that the AbinitioRelax protocol is a combination of the Abinitio and the Relax protocol. The sPRE score is only used in the Abinitio folding algorithm but not in the Relax stage. The Relax algorithm is a gradient-based energy minimization step and is the last step of the AbinitioRelax protocol. At that late stage, the global protein fold is already determined by the Abinitio protocol and the Relax algorithm performs a full-atom refinement of the sidechains, without introducing extensive changes that could affect the overall fold of the protein. Therefore, the sPRE score is only used in the Abinitio stage where the global fold is determined but will not affect the final energy minimization. For the final rescoring of the full-atom structures, we used Rosetta's full-atom scores, the chemical shift score as well as the sPRE score. The Rosetta scoring functions as well as the chemical shift score of the Rosetta framework were used without any adjustments. As an input for the chemical shift score, we

use the published chemical shift data (the corresponding BMRB codes are listed in supplementary table 1).
Implementing a sPRE scoring function in the Rosetta framework, requires the back-calculation of sPRE data for a given protein model and the comparison of this back-calculated data with experimental data. The sPRE module is implemented as a WholeStructureEnergy and communicates with Rosetta using the common scoring interface of the framework. The module integrates seamlessly into existing Rosetta protocols and can easily be activated in every Rosetta application by assigning a non-zero weight to the identifier "spre" in the score function weighting set. Once the sPRE module is activated, an input file containing the sPRE data in talos file format (NMRPipe Table Format) has to be supplied. For details regarding code availability, setup instructions as well as a tutorial, please refer to http://mbbc.medunigraz.at/en/research/forschungseinheiten-und-gruppen/research-group-prof-madl/.

Code availability

In this work, we tested the sPRE module in conjunction with the AbinitioRelax and score_jd2 protocols of Rosetta 3.2 for structure calculation and rescoring structural ensembles, respectively. The sPRE module is entirely included in the Rosetta framework, requires no additional software or online service and will be made available in the upcoming releases of the Rosetta framework.

Input data

Although the module was mainly tested with proton sPRE data as input, the module supports input data for all carbon and proton atoms. The input data is adjusted as follows. For methyl groups, the sPRE values of all protons and the corresponding center carbon atom (if sPRE data is available) are averaged and assigned to the hetero atom. Similar re-mapping is done for tyrosine and phenylalanine sidechains in which case the protons and carbon atoms at the meta positions are averaged with the para carbon atom and mapped onto the para carbon atoms. For data sets that were not assigned stereo-specifically (non-stereo specific assignment is assumed by default), the sPRE values of prochiral protons are averaged with the corresponding carbon atom and projected onto the center atom (i.e. for serine, data for both H^{β} protons and for the C^{β} carbon atom is averaged and assigned to the C^{β} atom). Missing data for carbons or protons does not affect the averaging. As the paramagnetic effect shows a γ^{-2} dependency, where γ is the gyromagnetic ratio of the corresponding nucleus, all sPRE values are normalized by γ^2 to allow proper averaging of sPRE data of different types of nuclei.
To score centroid models, further mapping needs to be performed since centroid models contain only 7 atoms per residue (H, N, C^{α}, C, O, C^{β} and a sidechain pseudo atom CEN). Consequently, the data for H^{α} is averaged with data for C^{α} (if available) and assigned to C^{α}. In a similar manner, data for H^{β} is assigned to C^{β} and all other sidechain data is merged and assigned to the CEN atom.

Back-calculation of sPRE data

sPRE data for a given structural model is back-calculated by an optimized grid-based approach (compare figure 1b) similar as described in previous studies.[1] In a first step, a uniform grid with a typical spacing of 0.5-2 Å between grid points (default 2 Å) is created around the given structural model. By default, every atom of the protein is at least 10 Å away from the boundaries of the grid.
Next, the atom positions of the protein are discretized onto this grid by replacing the x, y and z coordinates of the atom with the coordinates of the closest grid point. In a third step, grid positions that fall within the van-der-Waals region of the protein are marked as occupied. The atom radius is obtained from the Rosetta database. For centroid pseudo atoms, the radius is approximated by the distance between the centroid atom and the C^{α} atom of the same amino acid. All atom radii are increased by the radius of the paramagnetic agent (default 3.5 Å[1b, c]). This effectively marks all positions of the grid that are not accessible by a paramagnetic agent.
Next, the sPRE value for every atom is approximated by the sum of all grid positions within an integration radius (default 10 Å) that have not been marked as occupied in the previous step:

$$\mathrm{sPRE}_i^{\mathrm{model}} = \sum_{r_{i,j} < d_{\mathrm{int}}}^{N} \frac{1}{{r_{i,j}}^6} \cdot m_j \qquad (1)$$

where i is the index of the protein atom, j is the index of the grid point, N is the number of grid points, $\mathrm{sPRE}_i^{\mathrm{model}}$ is the approximated sPRE value for the i^{th} atom of the given protein structure, $r_{i,j}$ is the

discretized distance on the grid between the ith atom and the jth grid position, d_{int} is the integration radius (default 10 Å), and $m_j = \begin{cases} 0 & \text{if } j\text{-th grid point marked as occupied} \\ 1 & \text{else} \end{cases}$. Since the protein atoms and the grid positions are discretized on the same grid, $\frac{1}{r_{i,j}{}^6}$ is computed beforehand and stored in a lookup table.

Scoring a structural model

After back-calculating the sPRE data for every atom of the protein, the calculated values $\text{sPRE}_i^{\text{model}}$ are compared to the experimental data $\text{sPRE}_i^{\text{exp}}$. By default, the sPRE module uses a robust spearman correlation coefficient to compute a scalar score. To calculate the spearman correlation coefficient, both data sets ($\text{sPRE}_i^{\text{model}}$ and $\text{sPRE}_i^{\text{exp}}$) are ranked independently, generating two new data sets $\text{r}_i^{\text{model}} \in [1, \text{n}]$ and $\text{r}_i^{\text{exp}} \in [1, \text{n}]$. The raw score is then obtained using

$$s_{\text{spearman}} = 1.906 \cdot \left(1 - \text{correlation}_{\text{spearman}}\right) - 0.144$$

$$= 1.906 \cdot \left(1 - \frac{\sum_{i=1}^{\text{n}} \left(\left(\text{r}_i^{\text{model}} - \overline{\text{r}^{\text{model}}} \right) \cdot \left(\text{r}_i^{\text{exp}} - \overline{\text{r}^{\text{exp}}} \right) \right)}{\sqrt{\sum_{i=1}^{\text{n}} \left(\text{r}_i^{\text{model}} - \overline{\text{r}^{\text{model}}} \right)^2 \cdot \sum_{i=1}^{\text{n}} \left(\text{r}_i^{\text{exp}} - \overline{\text{r}^{\text{exp}}} \right)^2}} \right) - 0.144 \qquad (2)$$

where $\text{correlation}_{\text{spearman}}$ is the correlation coefficient of the ranked data sets $\text{r}_i^{\text{model}}$ and r_i^{exp}, i is the index of a protein atom for which experimental sPRE data is available, n is the number of atoms for which experimental sPRE data is available, $\text{r}_i^{\text{model}}$ is the ranked back-calculated sPRE data of the ith atom, r_i^{exp} is the ranked experimental sPRE data of the ith atom and $\overline{\text{r}^{\text{model}}} = \frac{1}{\text{n}} \sum_{i=1}^{\text{n}} \text{r}_i^{\text{model}}$ as well as $\overline{\text{r}^{\text{exp}}} = \frac{1}{\text{n}} \sum_{i=1}^{\text{n}} \text{r}_i^{\text{exp}}$ correspond to the average ranks over all atoms. The scaling of 1.906 and offset of 0.144 were chosen such that for a test set of small proteins (see supplementary table 6) the raw score of most final full-atom structures ranges from 0 to 1.
The spearman correlation was chosen as the default method from a set of several alternative methods. In this study, several other scores were tested and all of these scores were computed according to equation (3)

$$s_{\text{score}_j} = A_j \cdot \tilde{s}_{\text{score}_j} - B_j \qquad (3)$$

where $\tilde{s}_{\text{score}_j}$ denotes the raw score, s_{score_j} is the scaled score, and A_j as well as B_j are constants. s_{score_j} was then normalized with an appropriate power of the average reference sPRE. For every type of score, the values for A_j and B_j as well as the computation of $\tilde{s}_{\text{score}_j}$ are listed in the supplementary table 7.

The scores based on the Pearson correlation (s_{pearson}), Spearman correlation (s_{spearman}) and the quadrant count ratio (s_{quadrant}) are derived from the corresponding correlation coefficients and as a consequence are mathematically bounded. All other scores in the supplementary table 7 are unbounded and normalization of these scores becomes more challenging. Here, we used a test of fully-relaxed, full-atom protein structures for normalization. As structure models in the initial phase of CS-Rosetta are entirely different compared to the optimized final models, the chosen set of constants A_i and B_i can give rise to large score values for these initial structural models. Furthermore, the absolute values of these scores depend on the size of the protein as well as the number of total constraints and can be dominated by outliers. Therefore, we only considered s_{pearson}, s_{spearman} and s_{quadrant} for optimizing the sPRE module, since those scores are based on correlation coefficients and resulted in a stable and robust folding algorithm. Moreover, the correlation-based scores can be utilized in scenarios involving different proteins, different sets of sPRE data or a mix of centroid and full-atom models.
In addition to the fixed A_i and B_i values, the score is scaled and shifted before it is returned to the Rosetta framework. This scaling step can be adjusted by the user and is performed according to

$$\text{sPRE score} = \text{scaling} \cdot s_{\text{score}_i} - \text{offset} \qquad (4)$$

where s_{score_i} is the score as calculated above, scaling is given by the Rosetta command line option -score:spre:scaling (default 67), and offset is given by the option -score:spre:offset (default 0).

Note that the sPRE score (as any other scoring function used in Rosetta) is also scaled according to the Rosetta weight sets. The Rosetta weight for the sPRE score was set to 1.0 throughout the study. Furthermore, it should be noted that the Monte-Carlo algorithm of Abinitio relies on differences between scores. The offset was implemented only for the sake of completeness and chosen to be 0.

Verification of the sPRE back-calculation

The sPRE module uses a discretized, low resolution back-calculation of the sPRE data. To visualize the error of this approximated back-calculation, we compared the back-calculated sPRE data obtained from the Rosetta sPRE with data obtained from a classical grid-based back-calculation (Supplementary figure 7). As expected, the low resolution of the grid results in a weakening of the correlation. However, even with a grid of 2 Å, solvent exposed residues are still predicted to have a high sPRE, indicating that such low-resolution back-calculated data can still be used to guide the Rosetta sampling algorithm towards the native structure. Furthermore, assuming a global protein, the accuracy of the approximated sPRE back-calculation increases as the size of protein increases. For larger proteins, the effect of missing some high resolution structural features can be neglected compared to the large sPRE gradient between the core and the surface of the protein structure.

Optimizing the sPRE module

The sPRE scoring function can be directly adjusted using several parameters most notably, the resolution of the grid, the cut-off radius (integration radius) and the method for comparing the measured and the back-calculated sPRE data. An example showing how these parameters affect the scoring performance is illustrated in supplementary figure 8. In cases where the sPRE score is used in the Abinitio protocol of CS-Rosetta, two additional parameters become crucial, the global weight as well as a stage-specific weighting of the sPRE score. We optimized these parameters using a set of proteins (Supplementary table 6) and the recommended settings are listed in supplementary table 8.
The optimal values for the grid resolution and the cut-off radius can vary depending on the size of the protein, the quality of the sPRE data and the computational resources. Since both parameters affect the accuracy of the back-calculation as well as the computational costs, the optimal value is a trade-off between computational time and accuracy (Improving the resolution by lowering the distances between the grid points or enlarging the integration radius leads to a cubic increase of computational costs. For details see the section Computational costs).
As a rule of thumb, a minimum value of 10 Å is required for the integrational cut-off radius since smaller values lead to a significantly reduced correlation between the sPRE score and the C^{α}-RMSD to the native structure (Supplementary figure 8a). Increasing the integration threshold to more than 10 Å might be beneficial in the case of large proteins. However, given the current size limitations of CS-Rosetta, an integration threshold of 10 Å was sufficient throughout this work. Regarding the resolution of the grid, a grid spacing of 2 Å is sufficient in most cases and allows a fast computation of the sPRE score. Spacings above 2 Å lead to a significantly increased error and a broadening of the scoring correlation (Supplementary figure 8b). Higher resolved grids with a spacing of 1 Å or 0.5 Å increase the scoring performance in case where the score is used to distinguish between similar conformations with different high-resolution features (for example the tilting of the 4 helices in the case of C-terminal phpl5a, see figure 2d). In cases where the sPRE score guides the folding mainly in the early folding stages, an increase in resolution is typically not beneficial but requires more computational resources. In summary, using a grid spacing of 2 Å and an integration of 10 Å is a good compromise between performance and accuracy for most cases. To include high resolution information for near native-like structures, a grid spacing of 0.5 to 1 Å is preferable.
To find the optimal method of comparison, as well as the optimal global and stage-specific weights we used the following approach. We first chose a set of small to medium-sized proteins (Supplementary table 6) and predicted structure models using classical CS-Rosetta and sPRE-CS-Rosetta with different settings of the sPRE score. We then computed the percentage of models close to the experimental NMR structure (below 4 Å for 1Q02, 2JTV, 2JMB and 2CKX, below 1.5 Å for 2OSQ and 2K52) and used it as a measure for convergence for every parameter value. Using this procedure to optimize the global weighting, we found an optimum for the global scaling factor of 67 in our test set (Supplementary table 9). Although this default choice resembles a good compromise for many scenarios, it should be noted that the optimal scaling factor varies between proteins. Over-emphasizing the sPRE score can lead to physically incorrect structures, while a low weighted sPRE score fails to drive the sampling in a significant manner. The scaling of the sPRE score can therefore be adjusted by the user either by changing the score weight set within of the Rosetta framework or by changing the scaling command line option of equation (4).

We then used the same strategy to evaluate how different weightings in the individual Abinitio stages affect the sampling. In our test set, we found that every stage of the Abinitio protocol can benefit from the sPRE score (Supplementary table 10). Furthermore, depending on the protein, the improvement of the sampling can be traced back to different stages of the Abinitio protocol. Therefore, our tests suggest to include the sPRE score in all stages of Abinitio with a constant weight throughout the protocol.
Among the five stages of Abinitio, stage I uses the simplest scoring function, solely based on a van-der-Waals term. We still chose to include the sPRE score in this early stage, as the sPRE score depends on the global fold and favors compact structures. It is therefore suited to collapse the initial extended chain which is the main purpose of stage I.
Eventually we tested different algorithms to compare the experimental and the synthetic data using the same procedure. We found different types of correlation coefficients to perform best and to be the most robust among several common choices such as RMSD, correlation coefficient and Chi values (Supplementary table 7). Chi values and several variations of RMSD performed well only in a few test cases. The classical Pearson correlation coefficient performed considerably better, but was out-performed by the Spearman correlation coefficient which gave the best convergence in most cases (Supplementary table 11). Consequently, we chose the Spearman correlation as the default method to compare experimental and back-calculated sPRE data.
The Spearman correlation coefficient is obtained by ranking both data sets (measured and back-calculated sPRE data) independently, and subsequently calculating a classical Pearson correlation coefficient of the ranks. As the ranks are bounded by the number of data points, the Spearman correlation is robust and less sensitive for outliers, making it well suited to cope with amide proton sPRE data that might contain additional relaxation contribution due to chemical exchange with water. Also note that due to the ranking of the input data, the sPRE module can potentially be used to include solvent accessibility data from different sources such as bioinformatics or other experimental methods such as mass spectrometry.

Computational costs

The main contribution of the overall computational effort is the back-calculation of the sPRE data. We therefore approximated the sPRE by discretizing the protein atom positions to positions of the same grid that is used to model the paramagnetic substance (see chapter Back-calculation of sPRE data). This simplifies the required computations to simple grid-based operations that can be accelerated by techniques such as lookup tables. We also aimed to reduce the total amount of memory to improve cache efficiency.
To quantify the computational costs of our sPRE score, we compared the runtime of rescoring an ensemble of ubiquitin structural models using the sPRE score with the runtime of computing the Rosetta centroid scores. As shown in supplementary table 12, the computational costs mainly depend on the resolution of the grid and the radius of integration. As an example, choosing a grid resolution of 1 or 2 Å requires an extra computational cost that is in the same order as required by the efficient centroid Rosetta scores (about 80% more computational time compared to only calculating a Rosetta centroid score). Furthermore, we compared the computational costs of CS-Rosetta and sPRE-CS-Rosetta (Supplementary table 13). The extra costs of the sPRE module do not change the order of magnitude of the total runtime when choosing a grid with 2 Å spacing (computational cost roughly doubles). Reducing the grid to 1 Å requires about 5 to 6 times the computational time compared to a classical CS-Rosetta run. It should be noted that in this comparison, the number of computed structures was kept constant. In practice, using the sPRE score can dramatically speed up the complete procedure, as less models need to be computed to sample near-native conformations. The additional computational costs of the sPRE module become even less important considering that CS-Rosetta can easily by parallelized and the number of computational cores in modern clusters increases rapidly.

Scoring Benchmark

To evaluate the potential of the sPRE score, we performed a comparison between the common Rosetta scoring functions and the sPRE score. In particular, we did not limit the benchmark to fully-relaxed full-atom structures, but we also analyzed how the sPRE score performs in the case of centroid structure models since those simplified models are used to fold the extended chain in the Abinitio protocol.
In a first step, we chose a set of proteins for which the native structure was determined by NMR spectroscopy and experimental sPRE data was either measured or already available in (Supplementary table 1). For every protein, a test ensemble of structures was generated by starting classical CS-Rosetta structure prediction runs and collecting the centroid models at the end of each stage (stage I, II, III and IV) as well as the final full-atom structures.

To ensure that the ensemble covers a broad RMSD range, from only partially-folded proteins to near native structures, we added distance restraints that were derived from the native structure. Gradually improved structures were obtained by running several CS-Rosetta runs and narrowing the distance potential in steps of 10, 6, 4, 3, and 2 Å. In total, we generated 15000 models for every protein and for every stage (3000 per protein, Abinitio stage and distance potential window size).
These ensembles were then scored using the corresponding centroid Rosetta score (score0 for stage I, score1 for stage II, score2 for stage III and score3 for stage IV) and the sPRE score. The chemical shift score was only used for fully-relaxed structures, as the score is only applicable to full-atom models.
The results of the scoring benchmark clearly suggest that the sPRE score can be used to find near-native structures (Supplementary figure 1). In particular in the case of centroid models, the Rosetta scoring function in some cases prefers wrongly folded models over near-native structures. For these cases, we observed that the sPRE score outperforms the Rosetta score (see for example supplementary figure 1a and b). On the other hand, for full-atom models the Rosetta and in particular the chemical shift score are more reliable and the performance of these scores is similar to that of the sPRE score. Although the sPRE score mainly depends on the global fold properties with only minor contributions from local high-resolution structural features, in some cases the sPRE score performs as well as the chemical shift score and outperforms the Rosetta full-atom score even in the low RMSD range (see for example supplementary figure 1c, g and h). Moreover, considering that in some cases only sPRE data for amide protons was used, it is interesting to note that in our test set we never observed the sPRE score to perform worse than the Rosetta centroid score.
In summary, the scoring benchmark revealed the potential of the sPRE score in finding native-like structures and consequently suggests the score to be perfectly suited to improve sampling and thus the overall performance of CS-Rosetta.

Sampling Benchmark

To study the benefit of including solvent accessibility data into the folding algorithm of CS-Rosetta, we built a test set of 49 proteins by randomly selecting protein models of the protein data base[2] (PDB) with a protein core size up to 170 residues and for which chemical shift data is available (Supplementary table 2). A full set of synthetic carbon and proton sPRE data was back-calculated using the lowest-energy model of the submitted structure in the PDB. We then predicted models using classical CS-Rosetta as well as sPRE-CS-Rosetta. For both methods, the obtained structure ensembles were ranked according to the sum of chemical shift score and Rosetta full-atom score (score13_env_hb) and the average C^{α}-RMSD of the best ranked 0.2% was computed (Figure 2c). To solely address the sampling of both methods, we also compared the best 1% by C^{α}-RMSD (Supplementary figure 4b). Proteins, for which both methods fail (average C^{α}-RMSD > 10 Å) where not analyzed. As indicated by the scoring benchmark, the additional solvent accessibility data significantly improved the sampling compared to classical CS-Rosetta.

To evaluate if the observed benefit is also present when using experimental NMR data, we repeated the sampling benchmark using a set of proteins for which experimental sPRE data is available (Supplementary table 1). We again predicted models using both methods and computed the average C^{α}-RMSD of the best ranked models by the sum of the chemical shift score, the Rosetta full-atom score (score13_env_hb) and the sPRE score (Figure 2b) as well as by the C^{α}-RMSD (Supplementary figure 4a). Although both methods failed to predict reasonable folds in the case of MBP and p16, the experimental sPRE data significantly improved sampling in the case of Pex19, Ubiquitin and both domains of Phl p 5a. In the case of Protein A, both methods resulted in high-resolution models.

Since the previous sampling benchmarks clearly showed that sPRE data improves the convergence and accuracy of CS-Rosetta, we used 4 proteins (2LEJ, 1LS4, 1P6T and 1Z8S) to quantify the robustness and applicability of sPRE-CS-Rosetta regarding typical challenges in protein NMR spectroscopy. For this benchmark, we first back-calculated a full sPRE dataset as described before. We then generated different sets of sPRE data by simulating incomplete assignments (40%, 70% and 100% assigned), different noise levels (30%, 60%, 100%, 200% and 400%) and different atom subsets (H^N only, H^N and $H^{methyl\text{-}ILV}$ as well as H^N, H^{α} and H^{β}). For a comparison of the simulated noise with experimental sPRE data see supplementary figure 9. Next, structure ensembles were predicted using sPRE-CS-Rosetta for every sPRE data set and the percentages of models with an C^{α}-RMSD of 5 Å or less to the native structure were computed for every ensemble as well as for the reference CS-Rosetta ensemble (Supplementary tables 3a-d). Interestingly, this sampling benchmark revealed the robustness of the sPRE score and clearly suggests its applicability to sparse and erroneous experimental NMR sPRE data.

Sampling Benchmark using NOE and RDC data

To show the orthogonality of the sPRE score with other experimental NMR data, the structure of ubiquitin was predicted using CS-Rosetta and sPRE-CS-Rosetta in the absence and presence of additional NMR restraints such as RDCs and NOEs (see supplementary table 4a-b and 5 as well as supplementary figure 5).

To this end, experiment NOEs and RDCs for ubiquitin (1 set of H^N-N RDCs recorded in one medium) were obtained from the literature (PDB entry 1D3Z). Next, random subsets of either NOE or RDC data were generated with a varying number of total restrains in the sets (see supplementary table 4a and 5). Ambiguous NOE restraints were counted as a single restraint and the AmbiguousRestraint groups of the Rosetta framework were used to account for the ambiguity. For every RDC subset, the CS-Rosetta toolbox (http://csrosetta.chemistry.ucsc.edu/) was used to prepare the RDC data for the usage in the Rosetta framework.

For every NOE or RDC subset, CS-Rosetta and sPRE-CS-Rosetta runs were used to obtain ensembles of ubiquitin with 5000 models each. For every subset size, 2 to 4 different random subsets were generated and used as input to account for random effects of the selection process (see supplementary tables 4a and 5). To compare the performance of the different input sets, the percentage of models with C^{α}-RMSD of 1.0 Å or less was computed for every ensemble. As the results show, the sPRE module improves the sampling of CS-Rosetta in all cases, even when using large sets of NOEs restraints.

In addition, the percentages of wrong models (C^{α}-RMSD above 4.0 Å) were analyzed using the same data set (see supplementary table 4b). The results show that adding the NOE scoring functions which consist of several thousands of NOEs not only generates more high-resolution structures, but also increases the percentage of models far from the native structure. This can be explained by the fact that such a NOE score containing a large number of restraints can become rather complex and therefore harder to sample efficiently. On the other hand, the sPRE score depends on the global fold and as such is less prone to rapid change upon minor conformational changes. With a smoother energy landscape, the sPRE score can in particularly drive the sampling from far off models to near native-like models. This can be seen by a reduction of models with a C^{α}-RMSD of 4 Å or more.

Comparison with RasRec

To compare the performance of CS-Rosetta and sPRE-CS-Rosetta to the iterative Rosetta protocol RasRec,[3] the Rosetta toolbox (http://csrosetta.chemistry.ucsc.edu/) was used to setup RasRec-Rosetta runs. The corresponding amino acid sequence as well as the chemical shift data as listed in supplementary table 1 were used as input data for the RasRec runs. The pool size of the RasRec protocol was increased to 1000 while all other settings were left as default. The obtained full-atom models were rescored using the same procedure as for ensembles obtained with CS-Rosetta and sPRE-CS-Rosetta. The results are compared in supplementary figure 6.

Protein expression and purification

For the expression of protein A and p16, a pET-M11 vector was modified to express the protein A as a solubility tag for p16 expression. The vector contains an N-terminal hexa-histidine sequence followed by protein A, a TEV (tobacco etch virus) cleavage site and the *E. coli* codon optimized DNA sequence of human p16. After cleavage by TEV protease, the protein A domain includes 16 N-terminal residues (MKHHHHHHPMKQHDEA) and an unstructured C-terminal region of 15 residues including the remaining cleaved TEV-site (MDAGSGSGSENLYFQ). The cleaved p16 protein contains two additional N-terminal residues (GA) followed by its 156 amino acids (canonical sequence, isoform 1). The expression vector was transformed into *E. Coli Bl21 (DE3)* and cells were grown at 37 °C, using 50µg/ml of kanamycin for selection. After inoculation of 150 ml of M9 minimal medium including uniformly ^{13}C labeled glucose (3 g per liter) and uniformly ^{15}N labeled ammonium chloride (1 g per liter), the culture was grown over night while vigorous shaking. In the next morning, the cell suspension was diluted with 850 ml of the same medium and grown to an OD of 0.8 and protein synthesis was induced by addition of IPTG (isopropyl-1-thio-D-galactopyranoside) to a final concentration of 0.5 mM. Then the culture was incubated over night at 19 °C and harvested on the next day. The cell pellets were re-suspended in 30 ml purification buffer (8 M urea, 20 mM TRIS, pH 8.0 and 20 mM Imidazole) and frozen at -20 °C. For purification of the protein, the cell pellet was thawed at room temperature, sonicated and applied to a Ni-NTA agarose (Qiagen) gravity column following the manufacturer's instruction. The gravity column with the bound protein was then washed with 50 ml urea buffer. Afterwards, the buffer was exchanged to HEPES buffer (110 mM potassium acetate, 20 mM HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid), pH 8.0, 2 mM β-mercaptoethanol (BME), 5% (v/v) glycerol and 20 mM

imidazole) by washing of the column with 20 ml. The protein was then eluted with HEPES buffer including 250 mM imidazole and concentrated to 5 ml in a centrifugal filter unit (Amicon Ultra-15 (Millipore, 3kDa molecular weight cut-off) and applied to size exclusion chromatography. After loading on a HiLoad 16/600 Superdex 75 pg (GE Healthcare Life Sciences, 50 mM sodium phosphate, 500 mM NaCl, 2 mM BME, pH 6.0) the target-protein containing fractions were pooled. The sample was dialyzed over night at 4°C against HEPES buffer using a 2 kDa MWCO ZelluTrans V series membrane (Carl Roth) after addition of 400 µl of a 0.1 mg/ml 6xHistidine tagged TEV protease solution. Next day, the solution was applied again to a Ni-NTA agarose column to separate the cleaved p16 while the TEV protease, traces of uncleaved protein and the protein A remained bound to the column. The p16 flow through fraction was buffer exchanged into a HEPES buffer (4 mM HEPES, pH 7.5, 5 mM DTT) by using a 5ml highTrap desalting column (GE Healthcare Life Sciences) and the final concentration for NMR measurements was 150 µM (including 10% D_2O). The protein A fraction was again eluted by HEPES buffer containing 250 mM imidazole and concentrated to 5 ml and a second size exclusion step was performed as described above which allowed the separation of the pure protein. Protein A was concentrated to 500 µl and buffer exchanged into NMR buffer (20 mM potassium phosphate buffer at pH 6.5, 50 mM NaCl) using the desalting column. The protein concentration of the final NMR sample (containing 10% D_2O) was 1 mM.

Expression, purification and assignment of Phl p 5a and Pex19 have been performed using standard methods and are described elsewhere (C. G., Margarete Focke-Tejkl, Evelyne Schrank, T. M., Simone Kosol, Christoph Madritsch, Nazanin Najafi, Sabine Flicker, Rudolf Valenta, Klaus Zangger, Nico Tjandra, *manuscript submitted*; Leonidas Emmanouilidis, Ulrike Schütz, Konstantinos Tripsianes, T. M., Juliane Radke, Robert Rucktäschel, Matthias Wilmanns, Wolfgang Schliebs, Ralf Erdmann, Michael Sattler, *manuscript submitted*). The chemical shifts have been deposited in the biological magnetic resonance data base (BMRB[4], accession code 19107) and NMR samples were prepared in 50 mM KPi buffer, 20 mM NaCl at pH 6.2.

Expression, purification and assignment of MBP and Ubiquitin has been previously described.[1b]

NMR spectroscopy

Recording of sPRE data

To obtain sPRE data by NMR spectroscopy, we used a saturation-based approach as described previously.[1b] Briefly, the R1 relaxation rates are determined by a saturation-recovery scheme followed by a read-out experiment such as a $^{1}H,^{15}N$ HSQC, $^{1}H,^{13}C$ HSQC or a 3D CBCA(CO)NH experiment. For proton saturation, a 7.5 ms ^{1}H trim pulse followed by a gradient was applied. Then, z-magnetization is build up during the recovery delay, ranging between several milliseconds up to several seconds. Iterating through the different recovery delays is done in an interleaved manner, and short and long delays were ordered in an alternating fashion. For every R1 measurement at least 8 delay times were recorded and for error estimation, at least one delay time was recorded as a duplicate.
The measurement of R1 rates was repeated for increasing concentrations of the relaxation-enhancing Omniscan and the sPRE was obtained as the average change of the proton R1 rate per concentration of the paramagnetic agent. After every addition of Omniscan, the recovery delays were shortened such that for the longest delay all NMR signals were still sufficiently recovered. The interscan delay was set to 50 ms, as the saturation-recovery scheme does not rely on an equilibrium z-magnetization at the start of each scan. All NMR samples contained 10% $^{2}H_2O$. Spectra were processed using NMRPipe[5] and analyzed with the NMRView[6] and CcpNmr Analysis[7] software packages.

Required measurement time

To record a full set of sPRE data for H^N and $H^{aliphatic}$ protons about 2 to 5 days of measurement time is required when using $^{1}H,^{15}N$ and $^{1}H,^{13}C$ HSQC-based pseudo-3D relaxation experiments. Acquiring relaxation rates for 4 to 6 different concentrations of the paramagnetic agent is sufficient for most proteins.
For example, using a 400 µM sample of p16 (16.6 kDa) and a 750 MHz magnet equipped with a TXI probe head (Bruker), one set of relaxation rates was measured in 7 to 8 hours (3.5 hours for a pseudo-3D $^{1}H,^{15}N$ HSQC with 8 scans,100 complex points and 12 exponentially-spaced delay points as well as 3 hours for a pseudo-3D $^{1}H,^{13}C$ HSQC with 4 scans, 175 complex points and 12 exponentially-spaced delay points). Using more scans for overnight experiments, relaxation rates for 6 different concentrations of the paramagnetic agent can be acquired in 3 days.
The total measurement time of sPRE data for a 300 µM sample of Phl p 5a (24.1kDa) using pseudo-3D $^{1}H,^{13}C$ HSQC required 22 hours of measurement time. The data was acquired on an Avance III Bruker

700 MHz NMR spectrometer using 4 scans, 128 complex points and 12 exponentially-spaced delay points for 5 different concentrations of the paramagnetic agent.

Measurement of sPRE data used in this study

Details of the assignment and acquisition of sPRE data for MBP and Ubiquitin were published previously[1b] and sPRE data of Pex19 was obtained according to the same protocol.

The assignment of protein A was achieved by transferring the published chemical shift data of BMRB entry 4023[8] and confirmation of the resonance positions by acquiring HNCO, HNCACO and HNCACB experiments. sPRE data of a uniformly ^{13}C, ^{15}N labeled 1 mM sample of the Z domain of protein A was recorded on a 600 MHz magnet (Oxford Instruments) equipped with an AV III console and cryo TCI probe head (Bruker). R1 rates of H^{α} and H^{β} were measured using CBCA(CO)NH read-out spectra at 25 °C in the presence of 0, 0.5, 1, 2, 5 and 10 mM Omniscan (GE Healthcare, Vienna, Austria).

For the assignment of p16, previously reported chemical shifts of p16[9] were obtained from the BMRB[4] (accession code 4086) and the assignment was confirmed by recording backbone HNCA as well as sidechain (H)CCH and H(C)CH tocsy spectra of uniformly ^{13}C, ^{15}N labeled p16 on a Avance III Bruker 900 MHz NMR spectrometer at 25 °C. sPRE data of a uniformly ^{13}C, ^{15}N labeled 400 µM sample of p16 was recorded on a 750 MHz magnet (Bruker) equipped with an AV III console and TXI probe head (Bruker). R1 rates of aliphatic protons and amide protons were measured using ^{1}H,^{13}C HSQC and ^{1}H,^{15}N HSQC read-out spectra, respectively, at 25 °C in the presence of 0, 1, 1.75, 2.5, 3.25, 4 and 4.75 mM Omniscan (GE Healthcare, Vienna, Austria).

sPRE data of uniformly ^{13}C, ^{15}N labeled 0.3 mM Phl p 5a was recorded on an Avance III Bruker 700 MHz NMR spectrometer at 24.8 °C in the absence and after addition of 1, 2, 3 and 5 mM Omniscan (Nycomed, Oslo, Norway). R1 rates of aliphatic protons were measured using ^{1}H,^{13}C HSQC read-out spectra and amide R1 rate were obtained using ^{1}H,^{15}N HSQC read-out spectra.

Analysis of NMR data

Analysis of sPRE data for MBP and Ubiquitin was described previously[1b] and sPRE data for Pex19 was analyzed accordingly. For p16, Phl p 5a and protein A, the sPRE data was analyzed as follows. Peak intensities were extracted using the nmrglue[10] Python package and fitted to a mono-exponential build up curve using the SciPy python package and equation (5)

$$\mathrm{I}(t) = -A \cdot e^{-R_1 \cdot t} + C \tag{5}$$

where $\mathrm{I}(t)$ is the peak intensity of the saturation-recovery experiment, t is the recovery delay, A is the amplitude of the z-magnetization build-up, C is the plateau of the curve and R_1 is the longitudinal relaxation rate. To estimate the error for the fitted rates R_1, the experimental error was estimated using duplicate recovery delays. For every R1 experiment, one absolute error for all peaks $\varepsilon_{\mathrm{exp}}$ was obtained by equation (6)

$$\varepsilon_{\mathrm{exp}} = \sqrt{\frac{1}{2N} \cdot \sum_{i=1}^{N} \delta_i} \tag{6}$$

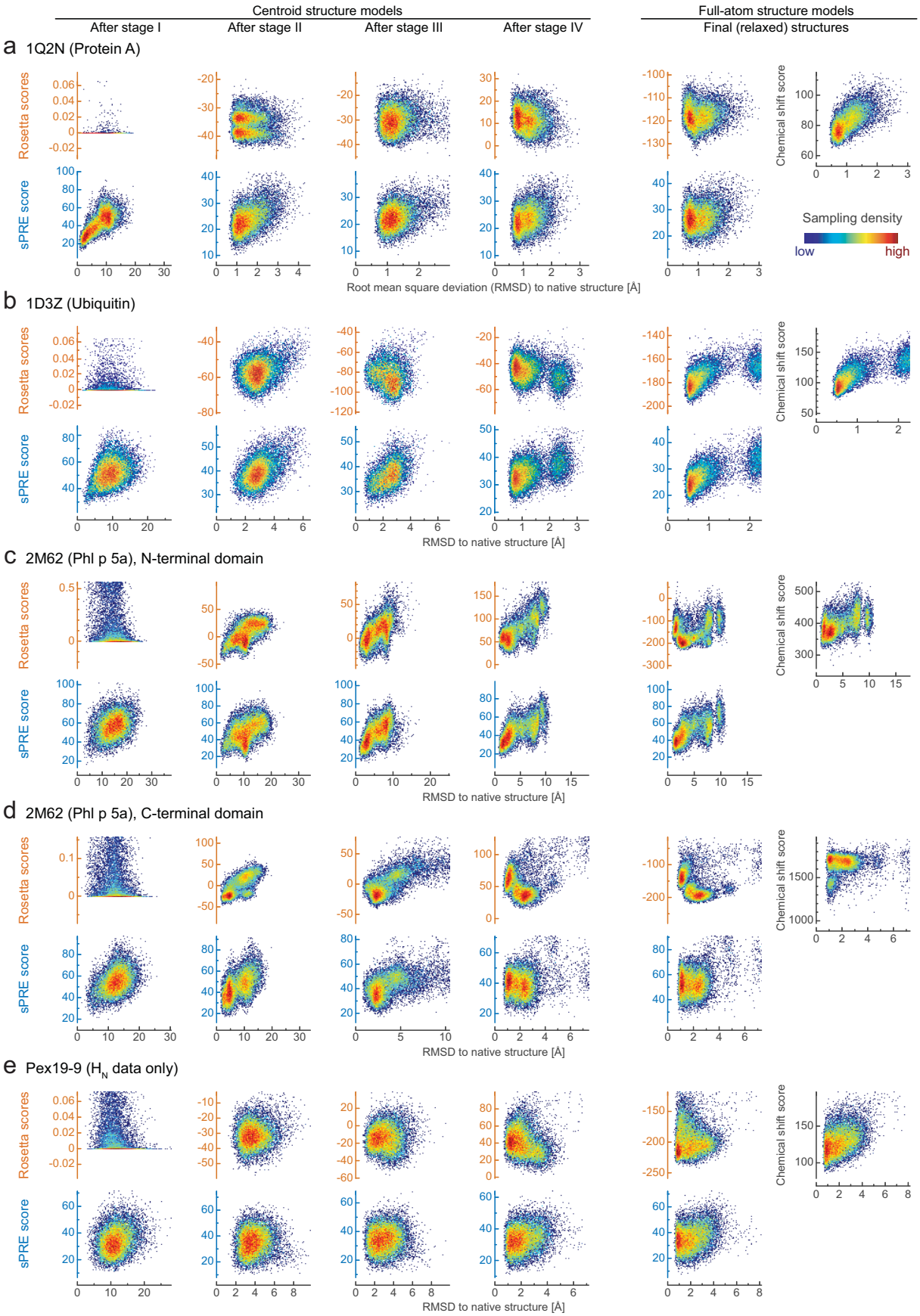
where N is the number of peaks in the spectrum, i is the index of the peak, and δ_i is the difference of the duplicates for the i-th peak. The error of the rates R_1 was then obtained using a Monte Carlo-type resampling strategy. By randomly drawing $3 \cdot n$ from the pool of n unique recovery delays, a new data set was created. Then noise was added to the peak intensities for each of the $3 \cdot n$ data points, according to a normal distribution with a standard deviation of $\varepsilon_{\mathrm{exp}}$. For every peak and saturation-recovery experiment, 1000 of such data sets containing $3 \cdot n$ randomly altered data points were created and fitted to the saturation recovery model as described by equation (5). The standard deviation ΔR_1 of all 1000 fitted R_1 parameters was then used as the error of R_1.
The sPRE is then obtained by performing a weighted linear regression using equation (7)

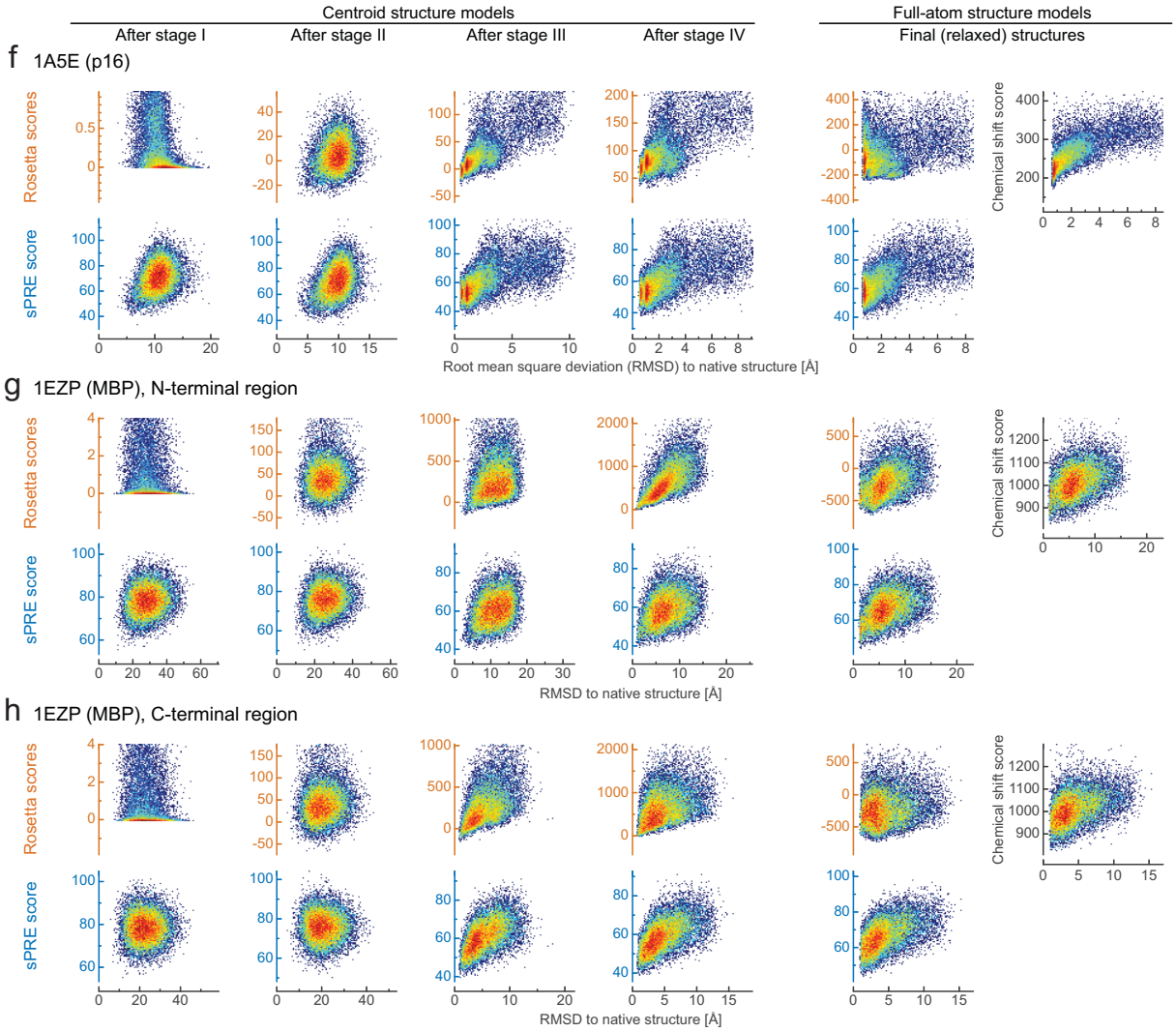
$$R_1(c) = \mathrm{sPRE} \cdot c + R_1^0 \tag{7}$$

where c is the concentration of Omniscan, $R_1(c)$ is the fitted R_1 rate at the present of Omniscan with a concentration c, R_1^0 is the R_1 in the absence of Omniscan and sPRE is the slope and the desired sPRE

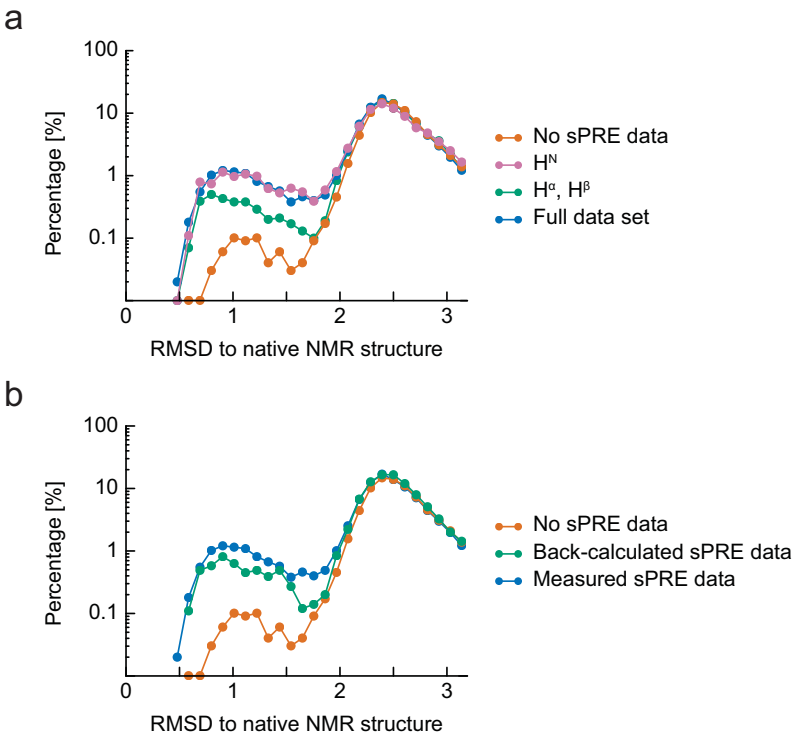
value. For the weighted linear regression, the previously determined errors ΔR_1 for R_1 was used, and the error of the concentration c was neglected.

Supplementary Figures

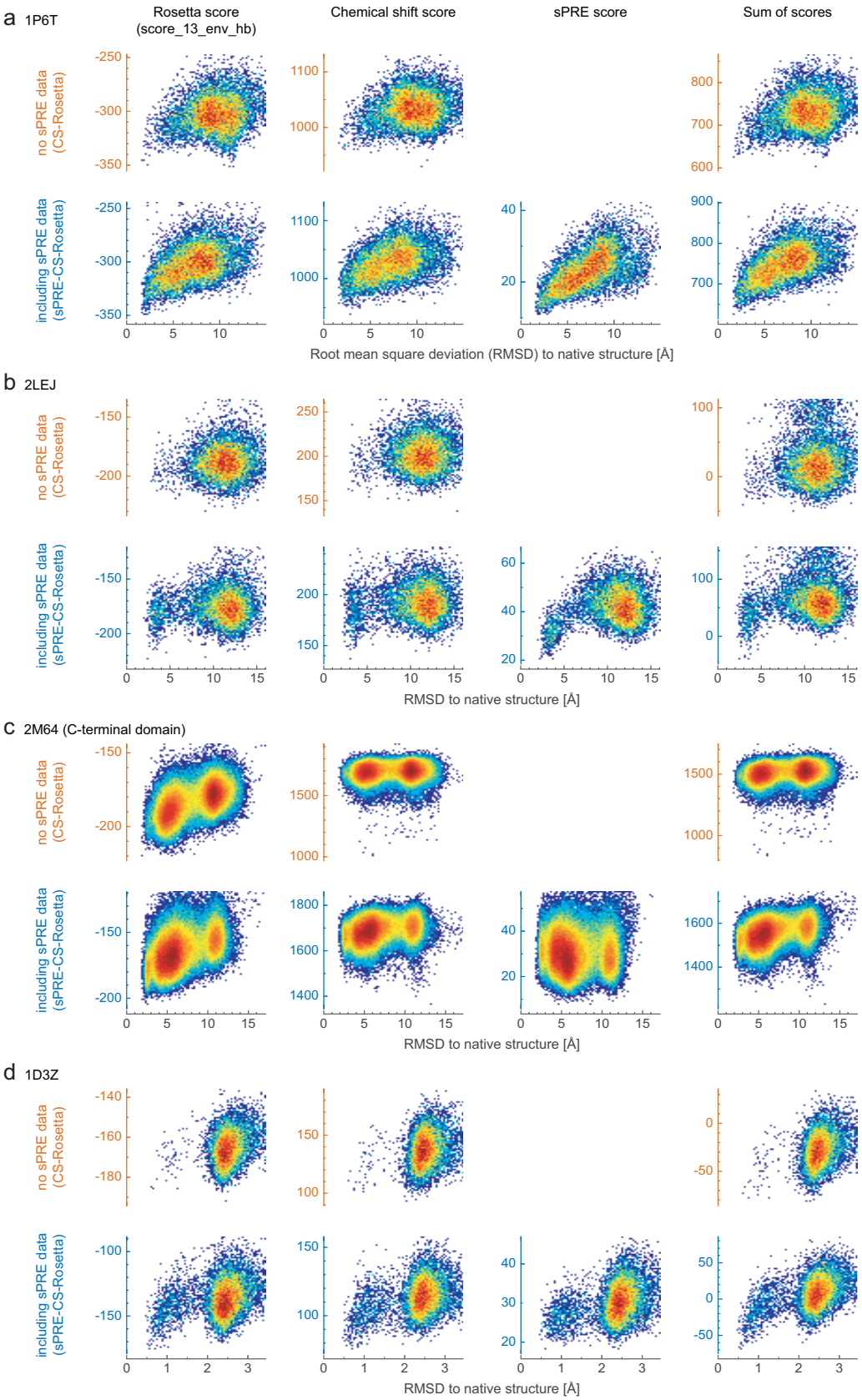




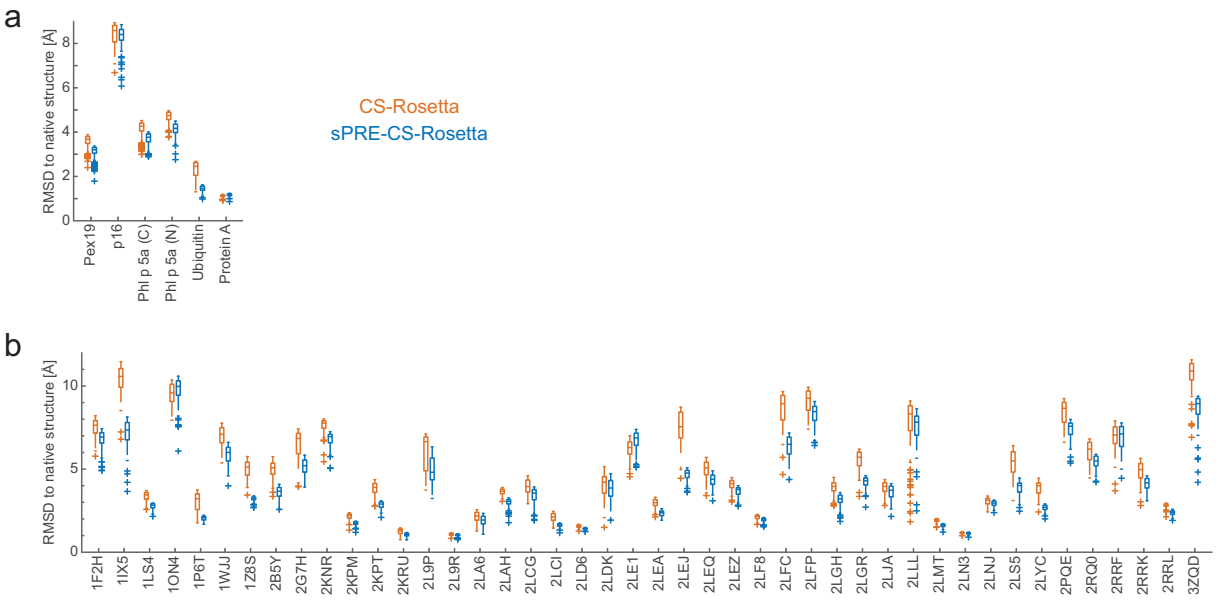
Supplementary figure 1. Scoring performance of the sPRE module. Ensembles of different proteins (a-i) have been generated and rescored using different scoring functions. Plots show 2D-histograms of the score and the C^{α}-RMSD to the native structure, with red corresponding to a high sampling density and dark blue corresponding to single structures. The ensembles contain centroid and full-atom models representing different stages of Rosetta's AbinitioRelax protocol (see column headers). Centroid models for Stages 1-4 were rescored using the corresponding Rosetta centroid score score0-3 (orange axis), the sPRE score (blue axis). Full-atom models were rescored using the Rosetta score score13_env_hb (orange axis), the chemical shift score (black axis) and the sPRE score (blue axis). Experimental sPRE data was used as listed in supplementary table 1.



Supplementary figure 2. Redundancy of sPRE data for Ubiquitin. Plots show histograms of the C^{α}-RMSD distribution of different structure ensembles on a logarithmic scale. (a) Structure ensembles were generated using CS-Rosetta (orange) or sPRE-CS-Rosetta with different sPRE input data: the full experimental data set (blue), sPRE data of amide protons H^N only (magenta) and sPRE data of H^{α} and H^{β} protons only (green). (b) Structure ensembles were generated using CS-Rosetta (orange), sPRE-CS-Rosetta with 193 experimental sPRE values for H^N and $H^{aliphatic}$ protons (blue) and sPRE-CS-Rosetta with back-calculated sPRE data for all protons and carbon atoms (green).

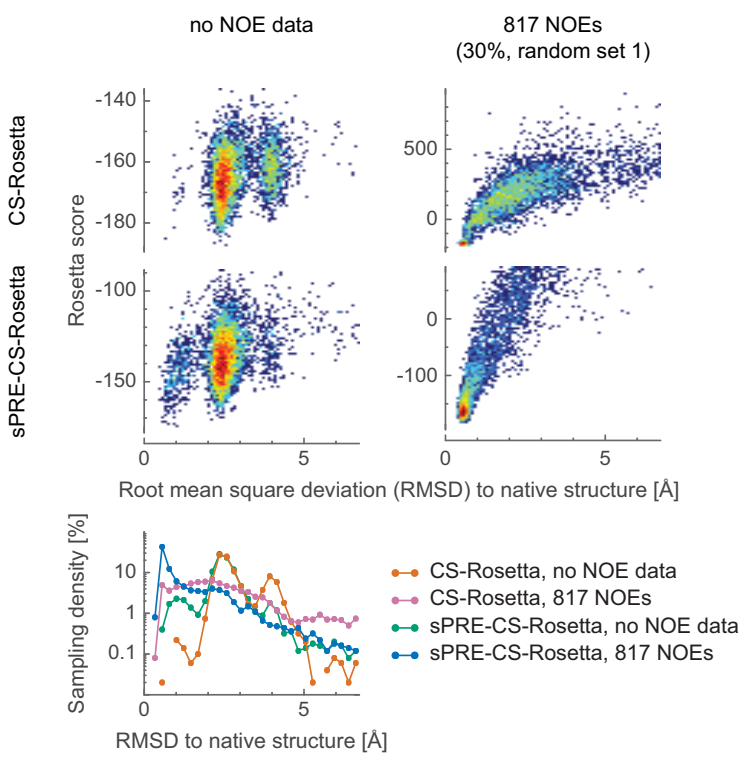


Supplementary figure 3. Performance of sPRE-CS-Rosetta. Comparison of structure ensembles obtained by CS-Rosetta (orange axis) and sPRE-CS-Rosetta (blue axis) for 1P6T (a), 2LEJ (b), 2M64 (c) and 1D3Z (d). Columns show different scores. From left to right: The full-atom Rosetta score (score13_env_hb), the chemical shift score, the sPRE score (for sPRE-CS-Rosetta runs only) and the sum of the first 2 (for CS-Rosetta) or the sum of all 3 scores (sPRE-CS-Rosetta). Plots show 2D-histograms of the respective score and the C^{α}-RMSD to the native structure, with red corresponding to a high sampling density and dark blue corresponding to single structures.

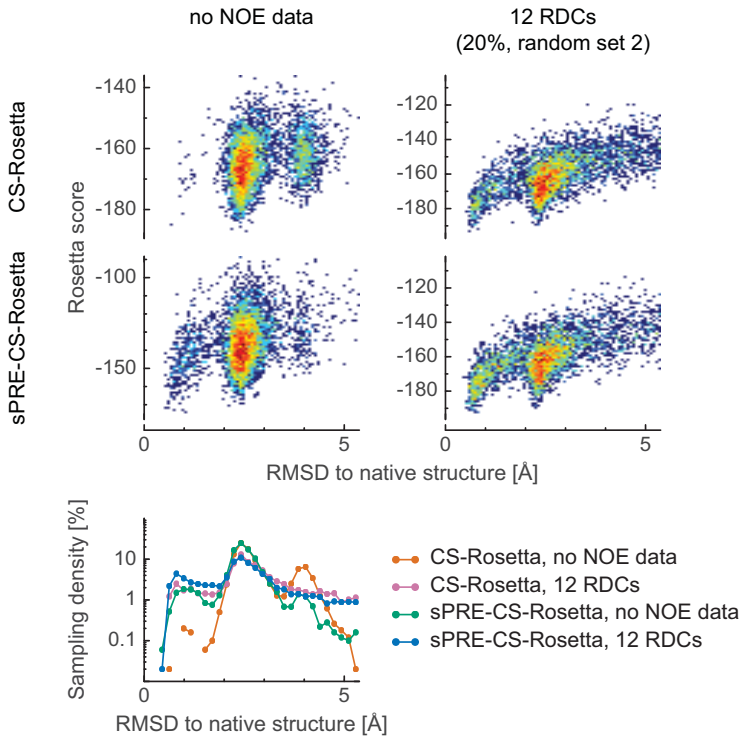


Supplementary figure 4. Effect of sPRE data on sampling efficiency. Structures for several proteins were predicted by CS-Rosetta (orange) and sPRE-CS-Rosetta (blue). For the sPRE module, experimental (a) and back-calculated (b) data was used. The 1% best structures by the C^{α}-RMSD to the native structure were filtered and the average C^{α}-RMSD of the subset was computed and plotted. Proteins for which the average C^{α}-RMSD was above 10 Å for CS-Rosetta and sPRE-CS-Rosetta are not shown (1CX1, 1GXE, 1RFL, 1XWE and 4A5V). All tested proteins are listed in supplementary table 2.

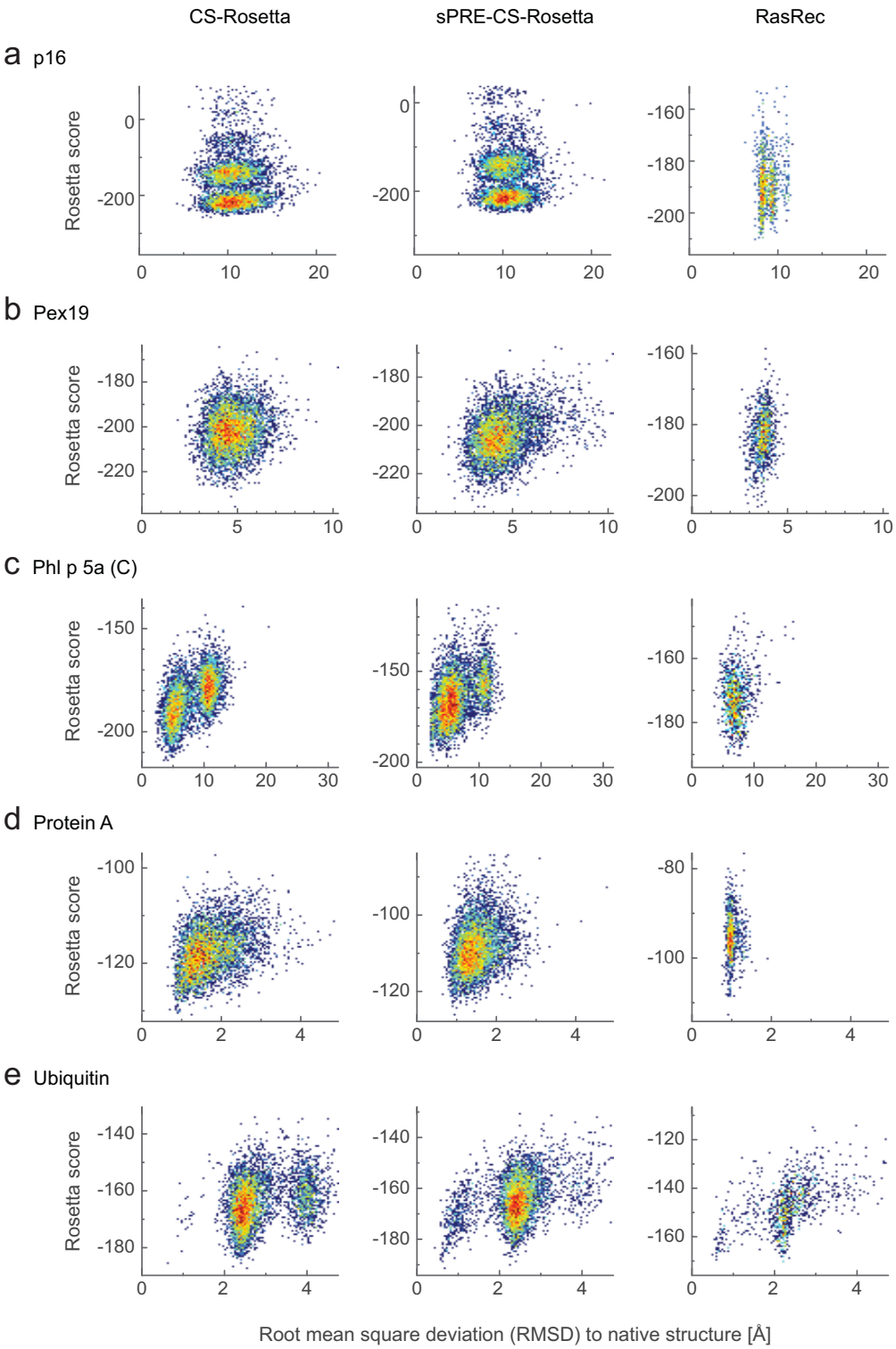
a



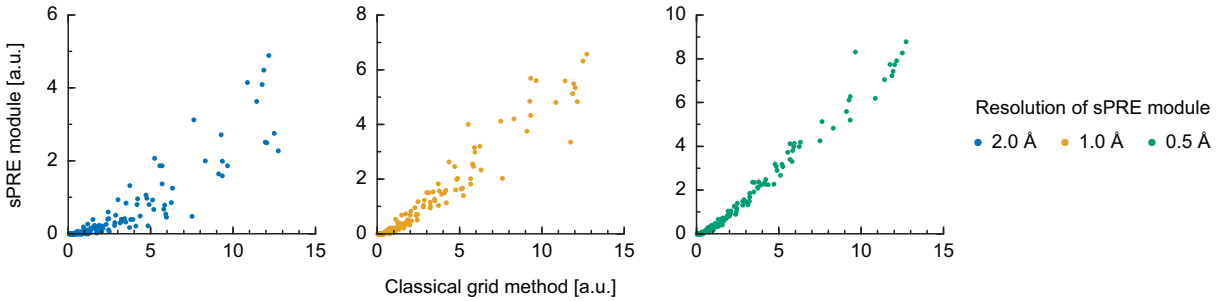
b



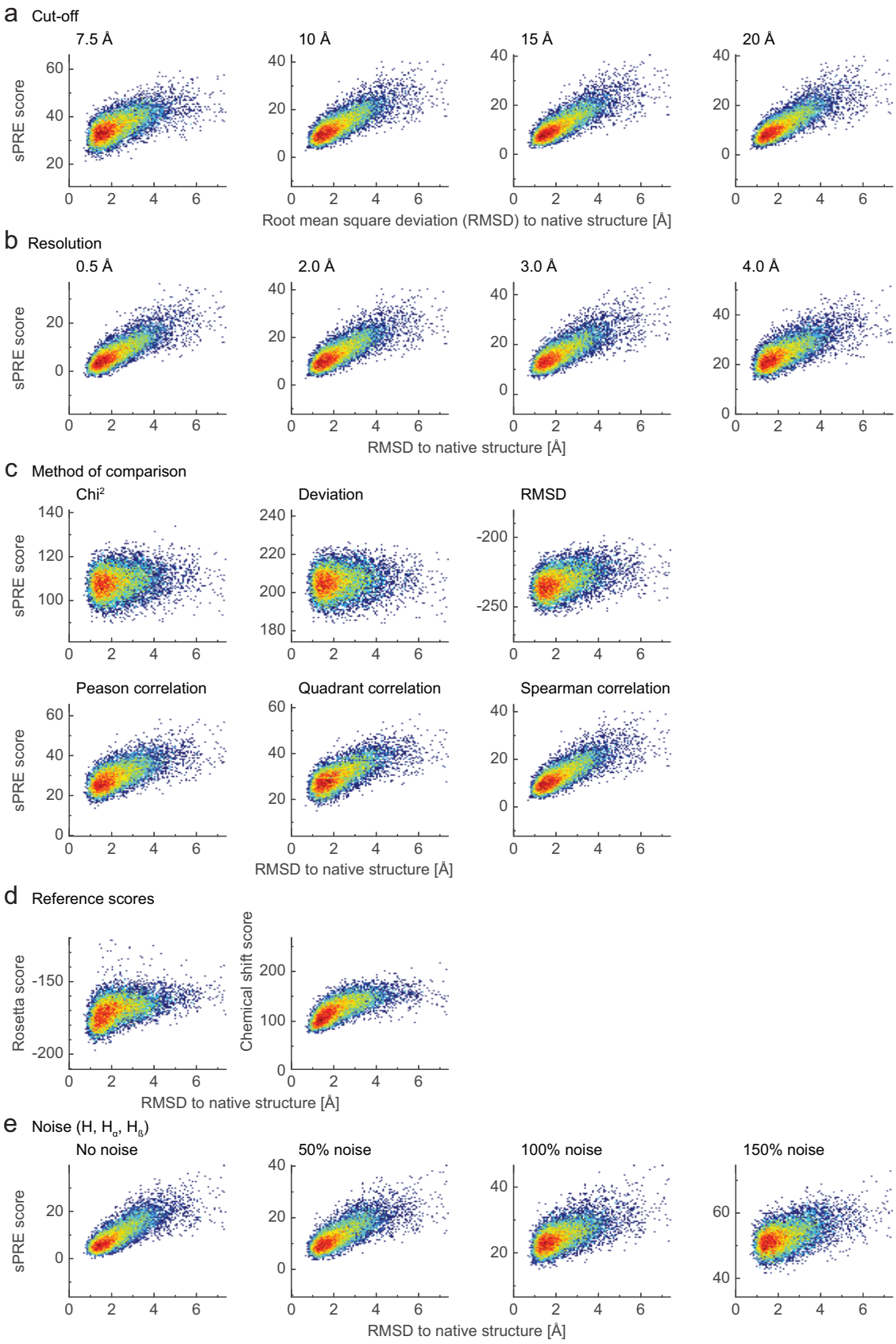
Supplementary figure 5. Sampling performance of sPRE-CS-Rosetta with NOE and RDC data. Structural models of ubiquitin (1D3Z) were predicted using classical CS-Rosetta as well as sPRE-CS-Rosetta in combination with experimental NOE (a) and RDC (b) data. Experimental sPRE data for H^N and $H^{aliphatic}$ protons was used for all computations and NOE and RDC restraints were obtained from the PDB entry 1D3Z. Plots show 2D-histograms of the respective score and the C^α-RMSD to the native structure, with red corresponding to a high sampling density and dark blue corresponding to single structures. Below the 4 plots, a logarithmic histogram shows the distribution the corresponding structures.



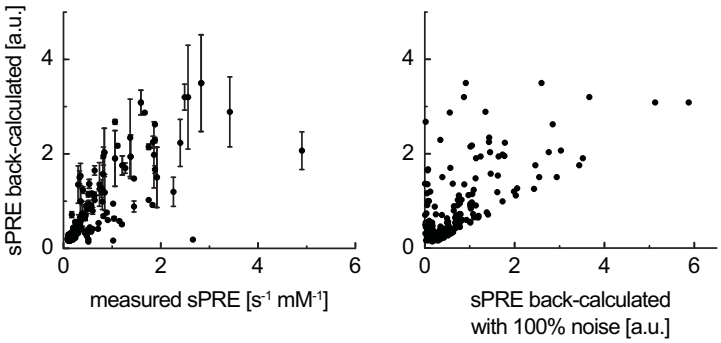
Supplementary figure 6. Comparison with RasRec-Rosetta. Structural ensembles of different proteins have been predicted using CS-Rosetta (left column), sPRE-CS-Rosetta (middle column) and RasRec-Rosetta (right column). Plots show 2D-histograms of the Rosetta score (score13_env_hb, in arbitrary units) and the C^{α}-RMSD to the native structure, with red corresponding to a high sampling density and dark blue corresponding to single structures. 5000 models are shown for CS-Rosetta and sPRE-CS-Rosetta, 1000 models for RasRec-Rosetta. Experimental data was used as input for sPRE-CS-Rosetta.



Supplementary figure 7. Verification of sPRE back-calculation and effect of the grid resolution on the accuracy. Back-calculated sPRE values for 1D3Z (Ubiquitin) using the sPRE module of Rosetta at different resolutions are compared to sPRE values computed using a classical grid-based approach with a resolution of 0.1A



Supplementary figure 8. Effect of different scoring parameters. An ensemble for the DNA-Binding Domain of Ngtrf1 (2CKX) with 15000 full-atom structure models was obtained using classical CS-Rosetta and subsequently rescored using the sPRE score. Plots show 2D-histograms of the respective score and the C^{α}-RMSD to the native structure, with red corresponding to a high sampling density and dark blue corresponding to single structures. (a-c) The sPRE score was computed using back-calculated sPRE data and parameters of the sPRE score module were varied as indicated. (d, e) For comparison, the Rosetta full-atom score (score_13_env_hb) and the chemical shift score (d) as well as the effect of noise (e) are shown.



Supplementary figure 9. Agreement of experimental and back-calculated sPRE data. The correlation between measured sPRE data of ubiquitin (H^N and $H^{aliphatic}$ protons) and the back-calculated sPRE data is shown on the left. sPRE values were calculated using the published NMR ensemble (1D3Z) and error bars indicate the standard deviation between the 10 models of the NMR ensemble. In comparison, the effect of adding 100% noise to the back-calculated data is shown in the scatter plot on the right. Back-calculated sPRE values are shown in arbitrary units and scaled to fit the experimental values.

Supplementary Tables

Supplementary table 1. Proteins with experimental sPRE data used for benchmarking the sPRE module.

Protein	PDB	BMRB	Fold	Fragment used	sPRE data
Maltodextrin-Binding Protein (MBP)	1EZP	4986	α+β	1-370	H^N + $H^{aliphatic}$
Ubiquitin	1D3Z	6457	α+β	1-76	H^N + $H^{aliphatic}$
Z domain of protein A	1Q2N	4023	α	3-58	H^α + H^β
Tumor suppressor p16INK4A	1A5E	4086	α	11-136	H^N + $H^{aliphatic}$
Major Grass Pollen Allergen Phl p 5a	2M64	19107	α	58-171 (N term. domain) 181-284 (C-term. domain)	H^N + $H^{aliphatic}$
Pex19	to be submitted	to be submitted	α	1-110	H^N

Supplementary table 2. Proteins with back-calculated sPRE data for benchmarking the sPRE module.

PDB	BMRB	Protein name	Number of residues	Fold
1CX1	4706	Second N-terminal cellulose-binding domain of β-1,4-Glucanase C	146	β
1F2H	4636	N-terminal domain of the TNFR1 associated protein	170	α+β
1GXE	5014	Central domain of cardiac myosin binding protein C	131	β
1IX5	4668	FK506-binding protein	152	α+β
1LS4	4814	Apolipophorin-III	157	α
1ON4	5742	Soluble domain of Sco1	170	α+β
1P6T	5813	Soluble region of P-type ATPase CopA	142	α+β
1RFL	5861	G-domain of MnmE protein	170	α
1WJJ	10090	Protein F20O9.120	127	α+β
1XWE	6001	C345C (NTR) domain of C5 of complement	148	α+β
1Z8S	6716	DnaB binding domain of DnaG (P16)	101	α
2B5Y	6603	Thioredoxin-like Protein	148	α+β
2G7H	7122	O6-Methylguanine DNA Methyltransferase	159	α+β
2KNR	16476	Ontario Center for Structural Proteomics target ATC0905	122	α+β
2KPM	16560	Northeast Structural Genomics Target NeR103A	83	α+β
2KPT	16569	Northeast Structural Genomics Consortium Target CgR26A	132	α+β
2KRU	16649	PCP_red domain of light-independent protochlorophyllide reductase subunit B	52	α
2L9P	17481	Northeast Structural Genomics Consortium Target SeR147	160	α+β
2L9R	17484	Homeobox domain of Homeobox protein Nkx-3.1	48	α
2LA6	17508	RRM domain of RNA-binding protein FUS	87	α+β
2LAH	17524	N-terminal domain of serine/threonine-protein kinase BUB1	151	α
2LCG	17611	Northeast Structural Genomics Consortium Target CrR115	137	α+β
2LCI	17613	Northeast Structural Genomics Consortium Target OR36	130	α+β
2LD6	17651	Histidine Phosphotransfer Domain of CheA	124	α
2LDK	17670	Northeast Structural Genomics Consortium Target AaR96	161	α+β
2LE1	17688	Northeast Structural Genomics Consortium Target TfR85A	144	α+β
2LEA	17705	RRM domain of Serine/arginine-rich splicing factor 2	88	α+β
2LEJ	17714	Prion protein mutant HuPrP	116	α
2LEQ	17723	Northeast Structural Genomics Consortium Target ChR145	146	α+β
2LEZ	17734	N-terminal domain of Secreted effector protein PipB2	116	α+β
2LF8	17742	ETS Domain of transcription factor ETV6	108	α+β
2LFC	17747	Subunit of fumarate reductase flavoprotein	146	α+β
2LFP	17768	Gp17 protein of bacteriophage SPP1	130	α+β
2LGH	17809	Northeast Structural Genomics Consortium Target AhR99.	139	α+β
2LJA	17927	C-terminal domain of putative thiol-disulfide oxidoreductase	137	α+β
2LLL	18053	C-terminal domain of Lamin-B2	111	β
2LMT	17353	N-terminal domain of Androcam	78	α+β
2LN3	18145	Northeast Structural Genomics Consortium Target OR135	76	α+β
2LNJ	18167	CyanoP subunit of photosystem II	159	α+β
2LS5	18411	Putative protein disulfide isomerase	151	α+β
2LYC	18717	C-terminal domain of Spindle and kinetochore-associated protein 1	129	α+β
2PQE	15232	Proline-free mutant of staphylococcal nuclease	138	α+β
2RQ0	11062	Lipocalin-type Prostaglandin D Synthase	158	α+β
2RRK	11422	E. coli ORF135 protein	136	α+β
2RRL	11423	C-terminal domain of the FliK	97	α+β
2LGR	15593	Human protein C6orf130	141	α+β
2RRF	11251	C-terminal region of Zinc finger FYVE domain-containing protein 21	125	α+β
3ZQD	17701	B. subtilis L,D-transpeptidase	168	α+β
4A5V	18039	N-terminal apple domains of Toxoplasma gondii microneme protein 4	160	α+β

Supplementary table 3a. Effect of assignment completeness and noise level on Rosetta sampling. Structural models of 2LEJ were predicted using simulated sPRE data of different quality. The percentages of models with a C^{α}-RMSD of 5 Å or less to the native structure are shown for different assignment completeness, noise levels and different protein nuclei.

No sPRE data

1.2%

Using simulated H^{N} sPRE data

	100% assignment	70% assignment	40% assignment
Noise +/-30%	6.4%	5.0%	2.3%
Noise +/-60%	6.9%	4.6%	2.2%
Noise +/-100%	4.4%	4.2%	1.6%
Noise +/-200%	1.7%	0.4%	0.9%
Noise +/-400%	0.0%	0.0%	0.0%

Using simulated H^{N}, $H^{methyl\text{-}ILV}$ sPRE data

	100% assignment	70% assignment	40% assignment
Noise +/-30%	4.0%	3.9%	7.0%
Noise +/-60%	4.4%	4.0%	6.2%
Noise +/-100%	1.7%	3.4%	3.9%
Noise +/-200%	2.3%	0.7%	0.1%
Noise +/-400%	0.8%	0.0%	0.3%

Using simulated H^{N}, H^{α}, H^{β} sPRE data

	100% assignment	70% assignment	40% assignment
Noise +/-30%	7.4%	4.4%	2.2%
Noise +/-60%	9.3%	4.7%	1.6%
Noise +/-100%	11.2%	4.8%	1.3%
Noise +/-200%	3.1%	4.7%	0.4%
Noise +/-400%	1.9%	0.9%	0.2%

Supplementary table 3b. Effect of assignment completeness and noise level on Rosetta sampling. Structural models of 1P6T were predicted using simulated sPRE data of different quality. The percentages of models with a C^{α}-RMSD of 5 Å or less to the native structure are shown for different assignment completeness, noise levels and different protein nuclei.

No sPRE data

3.5%

Using simulated H^{N} sPRE data

	100% assignment	70% assignment	40% assignment
Noise +/-30%	25.8%	23.8%	20.7%
Noise +/-60%	26.1%	22.1%	20.7%
Noise +/-100%	27.2%	25.1%	13.9%
Noise +/-200%	15.9%	10.8%	0.1%
Noise +/-400%	12.1%	2.3%	0.6%

Using simulated H^{N}, $H^{methyl\text{-}ILV}$ sPRE data

	100% assignment	70% assignment	40% assignment
Noise +/-30%	28.9%	30.6%	22.3%
Noise +/-60%	29.8%	27.2%	23.2%
Noise +/-100%	30.1%	32.9%	10.4%
Noise +/-200%	16.2%	7.4%	0.9%
Noise +/-400%	2.3%	2.9%	0.4%

Using simulated H^{N}, H^{α}, H^{β} sPRE data

	100% assignment	70% assignment	40% assignment
Noise +/-30%	28.1%	27.5%	18.9%
Noise +/-60%	26.4%	26.9%	17.1%
Noise +/-100%	22.9%	28.7%	14.7%
Noise +/-200%	19.7%	16.7%	12.9%
Noise +/-400%	3.3%	10.9%	3.6%

Supplementary table 3c. Effect of assignment completeness and noise level on Rosetta sampling. Structural models of 1LS4 were predicted using simulated sPRE data of different quality. The percentages of models with a C^{α}-RMSD of 5 Å or less to the native structure are shown for different assignment completeness, noise levels and different protein nuclei.

No sPRE data

5.9%

Using simulated H^{N} sPRE data

	100% assignment	70% assignment	40% assignment
Noise +/-30%	27.3%	21.8%	10.9%
Noise +/-60%	28.4%	14.7%	7.0%
Noise +/-100%	26.3%	7.6%	2.3%
Noise +/-200%	2.0%	2.1%	0.1%
Noise +/-400%	0.3%	0.0%	0.0%

Using simulated H^{N}, $H^{methyl\text{-}ILV}$ sPRE data

	100% assignment	70% assignment	40% assignment
Noise +/-30%	31.3%	40.9%	10.6%
Noise +/-60%	30.8%	41.0%	10.3%
Noise +/-100%	28.0%	41.0%	9.8%
Noise +/-200%	13.7%	43.2%	0.1%
Noise +/-400%	4.4%	2.8%	0.0%

Using simulated H^{N}, H^{α}, H^{β} sPRE data

	100% assignment	70% assignment	40% assignment
Noise +/-30%	44.6%	48.9%	43.7%
Noise +/-60%	46.1%	49.9%	45.4%
Noise +/-100%	46.0%	52.6%	33.1%
Noise +/-200%	29.8%	39.0%	6.8%
Noise +/-400%	2.0%	10.6%	2.4%

Supplementary table 3d. Effect of assignment completeness and noise level on Rosetta sampling. Structural models of 1Z8S were predicted using simulated sPRE data of different quality. The percentages of models with a C^{α}-RMSD of 5 Å or less to the native structure are shown for different assignment completeness, noise levels and different protein nuclei.

No sPRE data

0.6%

Using simulated H^{N} sPRE data

	100% assignment	70% assignment	40% assignment
Noise +/-30%	14.1%	12.3%	12.0%
Noise +/-60%	13.2%	9.5%	6.4%
Noise +/-100%	3.8%	3.7%	1.2%
Noise +/-200%	0.2%	0.1%	0.0%
Noise +/-400%	0.1%	0.3%	0.0%

Using simulated H^{N}, $H^{methyl\text{-}ILV}$ sPRE data

	100% assignment	70% assignment	40% assignment
Noise +/-30%	32.1%	28.9%	6.0%
Noise +/-60%	30.8%	26.5%	5.8%
Noise +/-100%	21.2%	16.2%	2.6%
Noise +/-200%	4.8%	0.7%	0.0%
Noise +/-400%	0.2%	0.0%	0.0%

Using simulated H^{N}, H^{α}, H^{β} sPRE data

	100% assignment	70% assignment	40% assignment
Noise +/-30%	38.6%	28.6%	15.0%
Noise +/-60%	38.2%	26.6%	13.0%
Noise +/-100%	35.7%	25.9%	18.2%
Noise +/-200%	24.4%	7.4%	4.4%
Noise +/-400%	2.6%	3.6%	0.0%

Supplementary table 4a. Sampling performance of sPRE-CS-Rosetta with NOE data. Structural models of ubiquitin (1D3Z) were predicted using classical CS-Rosetta as well as sPRE-CS-Rosetta in combination with different sets of experimental NOE data. Experimental sPRE data for H^N and $H^{aliphatic}$ protons was used for all computations and NOE restraints were obtained from the PDB entry 1D3Z. The percentages of models with a C^α-RMSD of 1 Å or less to the native structure are shown for the full set of NOEs, in the absence of NOE data and with randomly drawn subsets of NOEs varying in size (between 27 and 1363 NOEs). To account for selection effects, 3 to 4 different random sets were generated for every NOE pool size.

		no NOEs	27 NOEs (1%)	54 NOEs (2%)	136 NOEs (5%)	272 NOEs (10%)	817 NOEs (30%)	1363 NOEs (50%)	2726 NOEs (full set)
Set 1	no sPRE	0.2%	0.9%	10.8%	33.4%	60.0%	10.5%	5.0%	31.1%
	sPRE	3.1%	2.3%	24.7%	62.6%	79.9%	58.8%	31.9%	45.3%
Set 2	no sPRE		0.3%	7.4%	12.4%	31.9%	56.7%	61.0%	
	sPRE		1.8%	17.5%	72.6%	77.8%	83.6%	82.4%	
Set 3	no sPRE		1.9%	2.4%	20.2%	35.0%	48.1%	20.0%	
	sPRE		6.7%	7.3%	29.3%	73.3%	79.7%	82.6%	
Set 4	no sPRE		4.3%	23.1%					
	sPRE		9.8%	31.2%					

Supplementary table 4b. Sampling performance of sPRE-CS-Rosetta with NOE data. The same ensembles as in supplementary table 4a were used and the percentages of models with a C^α-RMSD of 4 Å or more to the native structure are shown.

		no NOEs	27 NOEs (1%)	54 NOEs (2%)	136 NOEs (5%)	272 NOEs (10%)	817 NOEs (30%)	1363 NOEs (50%)	2726 NOEs (full set)
Set 1	no sPRE	10.9%	7.6%	8.8%	10.7%	4.0%	31.1%	19.8%	23.4%
	sPRE	4.9%	13.9%	0.5%	2.2%	1.2%	5.2%	5.9%	3.9%
Set 2	no sPRE		4.5%	2.0%	12.3%	6.8%	10.6%	8.4%	
	sPRE		5.1%	1.5%	1.9%	1.8%	1.2%	2.9%	
Set 3	no sPRE		14.6%	5.1%	6.5%	22.3%	7.5%	12.4%	
	sPRE		1.5%	3.7%	3.4%	4.8%	1.8%	2.7%	
Set 4	no sPRE		0.4%	0.3%					
	sPRE		1.9%	0.4%					

Supplementary table 5. Sampling performance of sPRE-CS-Rosetta with RDC data. Structural models of ubiquitin (1D3Z) were predicted using classical CS-Rosetta as well as sPRE-CS-Rosetta in combination with different sets of experimental RDC data. Experimental sPRE data for H^N and $H^{aliphatic}$ protons was used for all computations and H^N-N RDC restraints in one medium were obtained from the PDB entry 1D3Z. The percentages of models with a C^α-RMSD of 1 Å or less to the native structure are shown for the full set of RDCs, in the absence of RDC data and with randomly drawn subsets of RDCs varying in size (between 6 and 47 H^N-N RDCs). To account for selection effects, 2 to 4 different random sets were generated for every RDC pool size.

		no RDC	6 RDCs	12 RDCs	18 RDCs	25 RDCs	31 RDCs	47 RDCs	63 RDCs (full set)
Set 1	no sPRE	0.17%	0.58%	1.70%	0.46%	1.60%	3.52%	2.66%	4.48%
	sPRE	3.12%	2.76%	3.04%	1.02%	2.42%	4.16%	3.32%	4.92%
Set 2	no sPRE		0.10%	4.78%	0.76%	0.88%	2.26%	2.92%	
	sPRE		1.58%	8.70%	1.32%	1.72%	3.06%	3.05%	
Set 3	no sPRE		0.26%	2.06%					
	sPRE		1.70%	4.40%					
Set 4	no sPRE		0.48%	5.42%					
	sPRE		2.04%	10.18%					

Supplementary table 6. Proteins used to optimize the parameters for the sPRE module.

Protein	PDB	Fold	Fragment	sPRE data
UBA domain of p62	1Q02	α	1-52	synthetic
Protein RPA3401	2JTV	α+β	1-64	synthetic
Protein Atu4866	2JMB	β	1-79	synthetic
DNA-Binding Domain of Ngtrf1	2CKX	α	578-660	synthetic
RRM-1 of Yeast NPL3	2OSQ	α+β	9-79	synthetic
Protein MJ1198	2K52	β	1-74	synthetic

Supplementary table 7. Tested candidates for computing the sPRE score according to $s_{\text{score}_j} = A_j \cdot \tilde{s}_{\text{score}_j} - B_j$. Different methods to compare the experimental to the back-calculated sPRE data were implemented according to equation (3). A_j and B_j are the scaling and offset constants for the scoring function score_j, n is the number of atoms for which a reference sPRE value is supplied, i is the index of the atom, $\text{sPRE}_i^{\text{exp}}$ and $\text{sPRE}_i^{\text{model}}$ are the reference and back-calculated sPRE value of the i-th atom, r_i^{exp} and r_i^{model} are the ranks of the reference and back-calculated sPRE value of the i-th atom, $|\cdot|$ denotes the absolute value, $\overline{\cdot}$ is the average of the corresponding quantity over all n atoms, and $\text{sgn}(\cdot)$ is the sign function.

s_{score_j}	$\tilde{s}_{\text{score}_j}$	A_j	B_j
s_{spearman}	$1 - \frac{\sum_{i=1}^{n}\left(\left(r_i^{\text{model}} - \overline{r^{\text{model}}}\right) \cdot \left(r_i^{\text{exp}} - \overline{r^{\text{exp}}}\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(r_i^{\text{model}} - \overline{r^{\text{model}}}\right)^2 \cdot \sum_{i=1}^{n}\left(r_i^{\text{exp}} - \overline{r^{\text{exp}}}\right)^2}}$	1.906	0.144
s_{pearson}	$1 - \frac{\sum_{i=1}^{n}\left(\left(\text{sPRE}_i^{\text{model}} - \overline{\text{sPRE}^{\text{model}}}\right) \cdot \left(\text{sPRE}_i^{\text{exp}} - \overline{\text{sPRE}^{\text{exp}}}\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(\text{sPRE}_i^{\text{model}} - \overline{\text{sPRE}^{\text{model}}}\right)^2 \cdot \sum_{i=1}^{n}\left(\text{sPRE}_i^{\text{exp}} - \overline{\text{sPRE}^{\text{exp}}}\right)^2}}$	1.906	0.144
s_{quadrant}	$1 - \frac{1}{n}\sum_{i=1}^{n}\left(\text{sgn}\left(\text{sPRE}_i^{\text{model}} - \overline{\text{sPRE}^{\text{model}}}\right) \cdot \text{sgn}\left(\text{sPRE}_i^{\text{exp}} - \overline{\text{sPRE}^{\text{exp}}}\right)\right)$	1.906	0.144
s_{chi}	$\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\left(\text{sPRE}_i^{\text{model}} - \text{sPRE}_i^{\text{exp}}\right)^2}{\text{sPRE}_i^{\text{exp}}}}$	95	130
s_{chi2}	$\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\left(\text{sPRE}_i^{\text{model}} - \text{sPRE}_i^{\text{exp}}\right)^2}{\text{sPRE}_i^{\text{exp}}}$	35	67
$s_{\text{meanDeviation}}$	$\frac{1}{n} \cdot \sum_{i=1}^{n} \left\lvert \text{sPRE}_i^{\text{model}} - \text{sPRE}_i^{\text{exp}} \right\rvert$	95	130
$s_{\text{meanDeviation4}}$	$\frac{1}{n} \cdot \sum_{i=1}^{n} \sqrt[4]{\text{sPRE}_i^{\text{model}} - \text{sPRE}_i^{\text{exp}}}$	80	78
s_{rmsd}	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\text{sPRE}_i^{\text{model}} - \text{sPRE}_i^{\text{exp}}\right)^2}$	118	390
s_{msd}	$\frac{1}{n} \cdot \sum_{i=1}^{n}\left(\text{sPRE}_i^{\text{model}} - \text{sPRE}_i^{\text{exp}}\right)^2$	10	93

Supplementary table 8. Critical parameters of the sPRE module and recommended values.

Parameter	Description	Recommended value
Grid resolution	Dimension of each cell of the grid	1 - 2 Å
Integration	Cutoff radius, above which all sPRE contributions are neglected	10 Å
Scaling	Global weight of the sPRE score	67 (scaling of sPRE score in Rosetta framework assumed to be 1.0)
Weights	Stage-specific weight of the sPRE score in the AbinitioRelax protocol	1-1-1-1-1 (equal weights in all stages)
Method of comparison	Algorithm comparing the predicted and the experimental sPRE data	Spearman correlation coefficient

Supplementary table 9. Effect of different weights of the sPRE score on Rosetta sampling. Structure models of different proteins were predicted using different weights of the sPRE score and the corresponding percentages of models with a maximum C^{α}-RMSD of 4 Å (for 1Q02, 2JTV and 2JMB) or 1.5 Å (for 2CKX, 2OSQ and 2K52) to the native structure are shown.

Scaling (a.u.)	1Q02	2JMB	2JTV	2CKX	2K52	2OSQ
0 (Reference)	10%	40%	70%	22%	0.26%	5.9%
3	14%	44%	78%	22%	0.27%	6.2%
10	29%	54%	87%	22%	0.45%	6.2%
33	74%	67%	97%	25%	0.55%	7.3%
67	93%	74%	98%	31%	0.99%	8.5%
100	96%	74%	95%	34%	0.99%	9.5%
150	94%	69%	72%	32%	1.08%	11.0%
300	78%	42%	20%	17%	0.59%	11.1%

Supplementary table 10. Effect of stage-specific weights of the sPRE score on Rosetta sampling. Structure models of different proteins were predicted using varying weights of the sPRE score for the different Abinitio stages and the corresponding percentages of models with an maximum C^{α}-RMSD of 4 Å (for 1Q02, 2JTV and 2JMB) or 1.5 Å (for 2CKX, 2OSQ and 2K52) to the native structure are shown.

Weight in stage										
I	II	III	IIIa	IV	1Q02	2JMB	2JTV	2CKX	2K52	2OSQ
0	0	0	0	0	10%	40%	70%	22%	0.26%	5.9%
1	0	0	0	0	10%	41%	70%	24%	0.29%	7.1%
0	1	0	0	0	11%	42%	72%	31%	0.23%	5.3%
0	0	1	0	0	11%	44%	73%	23%	0.30%	7.3%
0	0	0	1	0	10%	48%	73%	19%	0.35%	7.5%
0	0	0	0	1	92%	60%	98%	28%	1.38%	11.9%
0.3	1	1	1	0.3	53%	63%	94%	26%	0.41%	5.4%
1	1	1	1	1	93%	74%	98%	31%	0.99%	8.5%

Supplementary table 11. Effect of different correlation coefficients on Rosetta sampling. Structure models of different proteins were predicted using different correlation coefficients to compare the back-calculated to the given input sPRE data. The corresponding percentages of models with a maximum C^{α}-RMSD of 4 Å (for 1Q02, 2JTV and 2JMB) or 1.5 Å (for 2CKX, 2OSQ and 2K52) to the native structure are shown.

Correlation coefficient	1Q02	2JMB	2JTV	2CKX	2K52	2OSQ
No sPRE data (Reference)	10%	40%	70%	22%	0.26%	5.9%
Correlation	84%	48%	97%	26%	0.60%	10.7%
Quadrant correlation	63%	52%	94%	28%	0.68%	6.4%
Spearman correlation	93%	74%	98%	31%	0.99%	8.5%

Supplementary table 12. Runtime for scoring a centroid ensemble. 3000 centroid structures of Ubiquitin were scored using a combination of the sPRE score and a Rosetta centroid score (see columns). Settings for the sPRE back-calculation were changed as indicated (res. resolution, int. radius of integration). The user space runtime was recorded and compared to only scoring with the Rosetta centroid score (100%, top row).

Settings for sPRE back-calculation	Rosetta score0	Rosetta score1	Rosetta score2	Rosetta score3
No sPRE score	100%	100%	100%	100%
sPRE score, res. 2 Å, int. 10 Å	168%	166%	166%	166%
sPRE score, res. 1 Å, int. 10 Å	182%	179%	178%	178%
sPRE score, res. 0.5 Å, int. 10 Å	262%	254%	253%	253%
sPRE score, res. 0.2 Å, int. 10 Å	1507%	1439%	1433%	1430%
sPRE score, res. 2 Å, int. 15 Å	173%	170%	170%	169%
sPRE score, res. 2 Å, int. 20 Å	180%	177%	176%	176%
sPRE score, res. 2 Å, int. 30 Å	197%	193%	192%	192%
sPRE score, res. 2 Å, int. 40 Å	227%	220%	219%	219%

Supplementary table 13. Runtime comparison for structure determination. Classical CS-Rosetta AbinitioRelax runs were started for Ubiquitin and Protein A. For the same proteins, CS-Rosetta runs with activated sPRE score were started and the runtime per obtained structure was compared to the reference CS-Rosetta run. Settings for the sPRE back-calculation were changed as indicated (res. resolution, int. radius of integration).

Settings for sPRE back-calculation	Ubiquitin	Protein A
Reference CS-Rosetta	100%	100%
sPRE-CS-Rosetta, res. 2 Å, int. 10 Å	191%	217%
sPRE-CS-Rosetta, res. 1 Å, int. 10 Å	546%	683%
sPRE-CS-Rosetta, res. 0.5 Å, int. 10 Å	2670%	3417%
sPRE-CS-Rosetta, res. 2 Å, int. 20 Å	400%	482%
sPRE-CS-Rosetta, res. 2 Å, int. 30 Å	910%	1086%

References

[1] a) H. G. Hocking, K. Zangger, T. Madl, *Chemphyschem* **2013**, *14*, 3082-3094; b) T. Madl, W. Bermel, K. Zangger, *Angew. Chem. Int. Ed.* **2009**, *48*, 8259-8262; c) G. Pintacuda, G. Otting, *J. Am. Chem. Soc.* **2002**, *124*, 372-373.

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235-242.

[3] O. F. Lange, D. Baker, *Proteins* **2012**, *80*, 884-895.

[4] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. Kent Wenger, H. Yao, J. L. Markley, *Nucleic Acids Res.* **2008**, *36*, D402-408.

[5] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, A. Bax, *J. Biomol. NMR* **1995**, *6*, 277-293.

[6] B. A. Johnson, R. A. Blevins, *J. Biomol. NMR* **1994**, *4*, 603-614.

[7] W. F. Vranken, W. Boucher, T. J. Stevens, R. H. Fogh, A. Pajon, M. Llinas, E. L. Ulrich, J. L. Markley, J. Ionides, E. D. Laue, *Proteins* **2005**, *59*, 687-696.

[8] M. Tashiro, R. Tejero, D. E. Zimmerman, B. Celda, B. Nilsson, G. T. Montelione, *J. Mol. Biol.* **1997**, *272*, 573-590.

[9] A. Tevelev, I. J. Byeon, T. Selby, K. Ericson, H. J. Kim, V. Kraynov, M. D. Tsai, *Biochemistry* **1996**, *35*, 9475-9487.

[10] J. J. Helmus, C. P. Jaroniec, *J. Biomol. NMR* **2013**, *55*, 355-367.

SCIENTIFIC REPORTS

OPEN

RNA structure refinement using NMR solvent accessibility data

Christoph Hartlmüller[1,2], Johannes C. Günther[1,2], Antje C. Wolter[3], Jens Wöhnert[3], Michael Sattler[1,2] & Tobias Madl[1,2,4]

Received: 13 March 2017
Accepted: 2 June 2017
Published online: 14 July 2017

NMR spectroscopy is a powerful technique to study ribonucleic acids (RNAs) which are key players in a plethora of cellular processes. Although the NMR toolbox for structural studies of RNAs expanded during the last decades, they often remain challenging. Here, we show that solvent paramagnetic relaxation enhancements (sPRE) induced by the soluble, paramagnetic compound Gd(DTPA-BMA) provide a quantitative measure for RNA solvent accessibility and encode distance-to-surface information that correlates well with RNA structure and improves accuracy and convergence of RNA structure determination. Moreover, we show that sPRE data can be easily obtained for RNAs with any isotope labeling scheme and is advantageous regarding sample preparation, stability and recovery. sPRE data show a large dynamic range and reflect the global fold of the RNA suggesting that they are well suited to identify interaction surfaces, to score structural models and as restraints in RNA structure determination.

In the last decades, ribonucleic acids (RNAs) have been found to be key regulators for numerous important cellular processes including transcription, translation, splicing, cell differentiation as well as cell death and proliferation[1]. Besides regulatory functions, RNAs are essential for catalysis in ribosomes and spliceosomes, sensing environmental parameters such as temperature or metabolite concentration, as well as storing the genomic information of RNA viruses. Understanding the underlying molecular mechanisms of these processes requires insights into the structure and dynamics at an atomic level. In the last decades, NMR spectroscopy has developed to a powerful technique to study RNA structure and dynamics[2–8]. However, technical challenges are caused by poor chemical shift dispersion leading to spectral overlap as well as by the low proton density and the low number of intramolecular interactions limiting the number of observable distance restraints that can be used to define RNA structure. Moreover, the RNA backbone is defined by several torsion angles, requiring a large set of restraints to obtain an accurate structural model. To overcome these limitations, different labeling strategies, including for example specific or uniformly ^{13}C- and/or ^{15}N-labeling and segmental labelling of RNAs were developed to reduce spectral complexity and signal overlap[9–22]. Large sets of restraints including NOE-based distances, torsion angles and orientational restraints derived from RDCs and J-couplings, respectively, as well as base pairing contacts obtained from J-couplings across hydrogen bonds are typically collected to provide a sufficient number of restraints for high-resolution models of RNAs[2,6,23–26].

In previous NMR studies, solvent accessibility data were obtained by measuring the NOE between water and protein protons, but the data were found to be dominated by chemical exchange[27,28]. Here, we demonstrate that solvent accessibility data derived from solvent paramagnetic relaxation enhancements (sPREs) provide valuable information to characterize the conformation of RNAs and will be useful to enable NMR studies of larger RNAs. sPRE data are a quantitative measure for solvent accessibility and thus readily provide distance-to-surface information[29] (Fig. 1). Directly observing the solvent accessible areas is a promising approach not only for mapping interaction surfaces but also for structure determination, as has been shown for proteins[30–32]. Recently, computational protocols using sPRE data were developed for XplorNIH[27] and the Rosetta framework[33]. The implementation and use of sPRE data has several advantages: i) no covalent labeling of RNAs is required, ii) the sample can be recovered, iii) complete chemical shift assignment is not required and iv) any type of NMR spectrum such as

[1]Center for Integrated Protein Science Munich, Department Chemie, Technical University of Munich, Lichtenbergstr. 4, 85748, Garching, Germany. [2]Institute of Structural Biology, Helmholtz Zentrum München, Ingolstadter Landstr. 1, 85764, Neuherberg, Germany. [3]Institut für Molekulare Biowissenschaften and Zentrum für Biomolekulare Magnetische Resonanz (BMRZ), Goethe-Universität Frankfurt, Max-von-Laue Str. 9, 60438, Frankfurt/M, Germany. [4]Institute of Molecular Biology and Biochemistry, Center of Molecular Medicine, Medical University of Graz, Harrachgasse 21, 8010, Graz, Austria. Christoph Hartlmüller and Johannes C. Günther contributed equally to this work. Correspondence and requests for materials should be addressed to T.M. (email: tobias.madl@medunigraz.at)

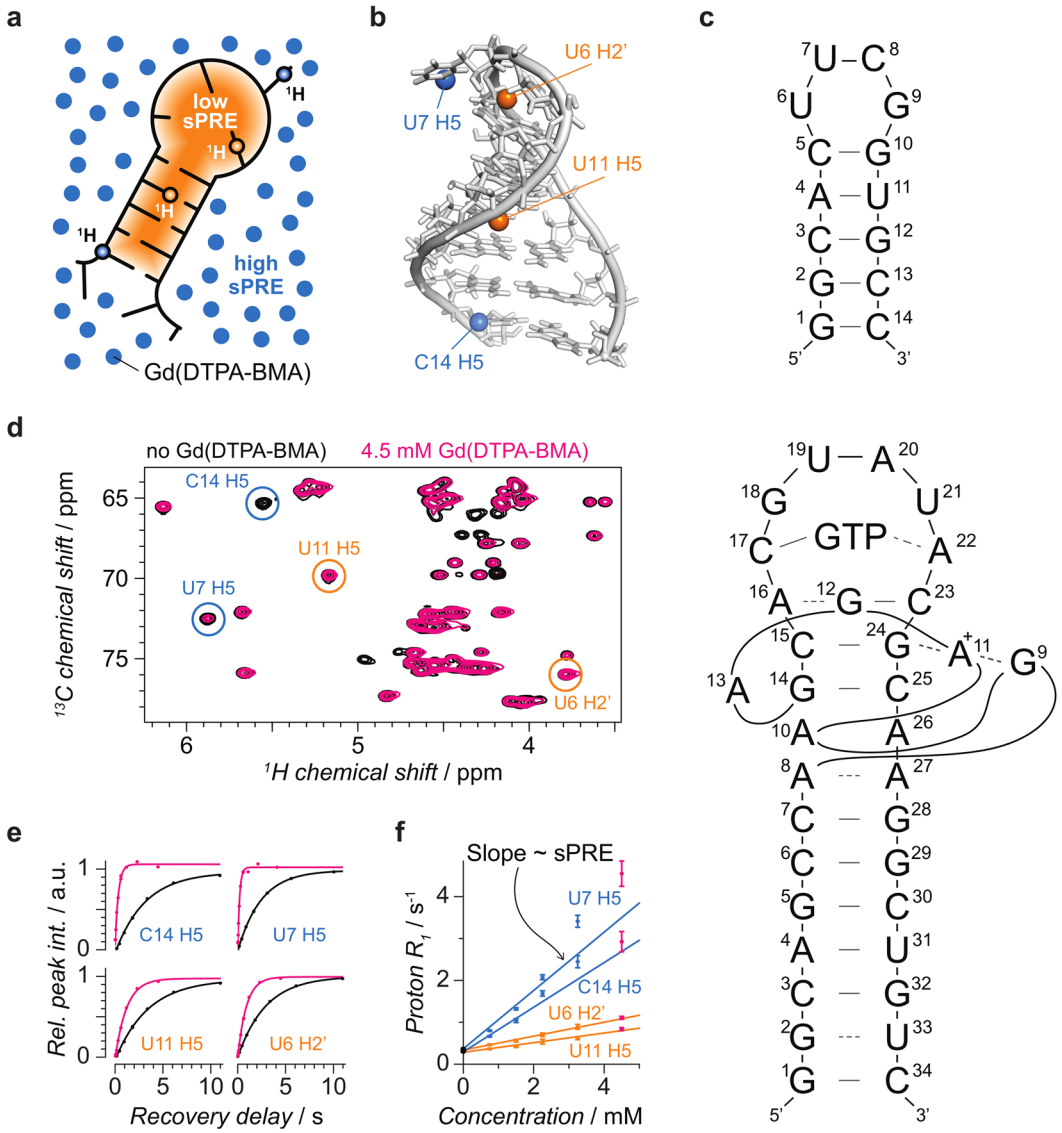


Figure 1. Concept of sPRE and exemplary sPRE data of the UUCG tetraloop. (**a**) sPRE data provide a quantitative measure for solvent accessibility and encode distance-to-surface information. sPRE data are acquired by titrating the RNA with the paramagnetic compound Gd(DTPA-BMA). (**b**) The NMR solution structure of the UUCG tetraloop (PDB code 2KOC) is shown and for illustration purpose, two solvent-exposed protons (blue spheres) and two buried protons (orange spheres) are highlighted. (**c**) Schematic representation of the UUCG tetraloop (top) and the GTP-bound GTP aptamer. (**d**) NMR spectra of the UUCG tetraloop are shown in the absence (black) and presence (magenta) of Gd(DTPA-BMA) with circles around the peaks corresponding to the protons shown in (**b**). (**e**) Quantitative sPRE data are obtained by measuring the longitudinal proton R_1 relaxation rate as a function of the concentration of Gd(DTPA-BMA). (**f**) Proton R_1 rates increase linearly with the concentration of the paramagnetic compound and the slopes correspond to the sPRE values of the respective resonances highlighted in (**b**).

proton 1D, ^{1}H, ^{15}N HSQC or ^{1}H, ^{13}C HSQC experiments can be used. Direct measurements of solvent accessibility data of RNAs have been reported previously using the soluble compound TEMPOL[34]. While the correlation of TEMPOL-induced sPREs agreed with RNA structure for some parts, the data were limited to a qualitative validation of models. Moreover, unexpectedly large sPREs were found for certain nucleotides and it was later suggested that TEMPOL forms intermolecular hydrogen bonds leading to preferred binding to specific sites[35]. Here, sPRE data are obtained by titrating the RNA sample with Gd(DTPA-BMA), a soluble, well-characterized Gd^{3+} chelating paramagnetic contrast agent that was originally developed for MRI imaging[31,36].

Results

Gd(DTPA-BMA) provides quantitative solvent-accessibility data of RNAs. To first assess a potential binding to specific RNA moieties, fingerprint spectra in the absence and presence of the paramagnetic compound were recorded for two RNAs exhibiting distinct folds and structural elements. The resulting NMR spectra

of the UUCG tetraloop[37,38] and the GTP class II aptamer[39,40] are shown in Supplementary Figures 1 and 2, respectively. As expected, the presence of Gd(DTPA-BMA) causes line-broadening of NMR signals. However, even in the presence of the paramagnetic compound, the vast majority of peaks are still well observable and the absence of chemical shift perturbations confirms the absence of specific binding of the compound to the RNA (Fig. 1d and Supplementary Figures 1 and 2). Thus Gd(DTPA-BMA) can be used as a paramagnetic probe that screens the accessible surface area of RNA molecules.

As unchelated lanthanide ions can efficiently hydrolyze RNA[41], we checked for potential cleavage products of the RNAs in the presence of the Gd^{3+} compound. No degradation products were observed for any RNA studied and sPRE data obtained from a single RNA sample were reproducible even after several sPRE measurements, each followed by a removal of Gd(DTPA-BMA) (data not shown). The absence of phosphodiester hydrolysis can be explained by the very high chelating stability of DTPA-BMA[36] and the presence of a slight excess of Ca^{2+}-bound chelator.

The sPRE data were acquired by measuring proton longitudinal relaxation rates (^{1}H-R_1) as a function of increasing concentrations of Gd(DTPA-BMA). As expected, ^{1}H-R_1 rates were found to increase linearly with the concentration of Gd(DTPA-BMA) in the case of carbon-bound protons. For imino and amino protons, the increase of ^{1}H-R_1 rates is larger at small concentrations of the paramagnetic compound (below about 1 mM), and becomes linear at higher concentrations (Supplementary Figure 3). This observation likely reflects an additional exchange contribution that mixes the relaxation mechanisms of water and RNA protons. Nevertheless, a quantitative measure for the solvent accessibility can be obtained for nitrogen-bound exchangeable protons by determining the minimum concentration of Gd(DTPA-BMA) for which a linear response is observed. Only titration points at this concentration or above are used to fit the linear sPRE model (0.5 mM or above for the UUCG tetraloop and 1.29 mM or above for the GTP aptamer; compare Supplementary Tables 1 and 2). For details about data acquisition and analysis, refer to the methods section in the supplementary information.

sPRE data correlate well with RNA structure. Next, we analyzed how sPRE data reflect the structural features of the RNAs studied. To this end, the NMR structures of the UUCG tetraloop and the GTP-bound GTP class II aptamer were used to predict the expected sPRE data using a previously published grid-based approach[30,31] (see method section for details). The experimental sPRE data correlate very well with the predicted data (Fig. 2a and Supplementary Figures 4a and 5). This demonstrates that sPRE data provide a quantitative measure of solvent-accessibility that correlates well with RNA structure. In particular, all buried protons of both RNAs show a small sPRE. Protons for which significant deviations between experimental and back-calculated sPREs are observed are found in loop regions or in terminal nucleotides. These regions are flexible and thus are expected to show a population-weighted average of the sPRE. The structural NMR ensembles used to predict the sPRE data represent the most stable conformations in solution and depend on the algorithm and experimental restraints used for the structure determination. Thus, the NMR ensembles do not necessarily reflect the equilibrium distribution of the different conformations of these solvent-exposed regions of the RNA. Since all conformations of the ensemble were equally weighted in the prediction of the sPRE data, the prediction is expected to deviate for dynamic regions of the RNA. It is further worth noting, that despite the compact fold of the UUCG tetraloop, the experimental sPRE data show a large dynamic range, which renders them a powerful source to characterize the structure of RNAs. The excellent correlation between experimental data and RNA structure in combination with a large dynamic range suggests sPRE data as promising parameters to probe RNA structures by NMR.

As described above, significant deviations between predicted and measured sPRE data are observed for some terminal nucleotides. Among all 258 observed sPRE values (93 for the UUCG tetraloop and 165 for the GTP aptamer), only two protons that are located in non-terminal nucleotides show significant deviations (9Gua H1 and 6Ura H3 in the UUCG tetraloop). Both protons are located in the loop region of the RNA in close spatial proximity to each other (Supplementary Figure 4b) and the corresponding sPRE values are underestimated based on the NMR solution structure, independent of which structural model of the UUCG loop motif (PDB codes 1HLX, 1K2G, 1TLR, 1Z31, 2KHY, 2KOC, 2KZL, 2LHP, 2LUB and 2N6X) was used. Although the sPRE of these protons are expected to have an additional contribution due to chemical exchange with water, the sPRE values of 2.2 and 0.44 $s^{-1}mM^{-1}$ of 6Ura H3 and 9Gua H1, respectively, are significantly larger compared to the average of $0.079 \pm 0.037\ s^{-1}mM^{-1}$ for all other imino protons. This suggests that conformational dynamics in this regions may cause the unusually high experimental sPREs. Indeed, recent molecular dynamics studies that included a large set of NMR relaxation data of the UUCG tetraloop, suggest the loop region to be more flexible than the RNA stem[42–48]. Moreover, intermediate structural states with the bases of the loop region experiencing higher solvent exposure were observed in these studies. This suggests that the increased sPREs are a result of increased flexibility and dynamics of the loop region. However, future studies are required to fully understand the usability of sPRE data to detect dynamics in RNAs.

sPRE data provide orthogonal restraints for RNA structure determination. Next, the potential of sPRE data to facilitate the structure determination of RNAs was investigated (Fig. 3). To this end, structural models of the UUCG tetraloop as well as the GTP-bound aptamer were computed using the XplorNIH framework[27,49–51] in combination with different sets of experimental NOE data (Supplementary Tables 3 and 4, Supplementary Figure 7). The results show that sPRE data significantly improves convergence and accuracy of the structure determination, in particular in cases where only limited sets of NOEs are available. For example, using only 258 experimental NOEs of the GTP-bound aptamer (30% of the all NOEs), the average RMSD of the obtained models was significantly reduced from 7.80 Å to 3.51 Å (Fig. 3b; Random set 2 in Supplementary Table 3). Next, the performance of the sPRE potential in combination with other experimental NMR-based restraints was investigated (Supplementary Table 5). The benchmark revealed that solvent accessibility data is an orthogonal restraint since its usage improves structure determination of the UUCG tetraloop in combination with NOE, RDC and

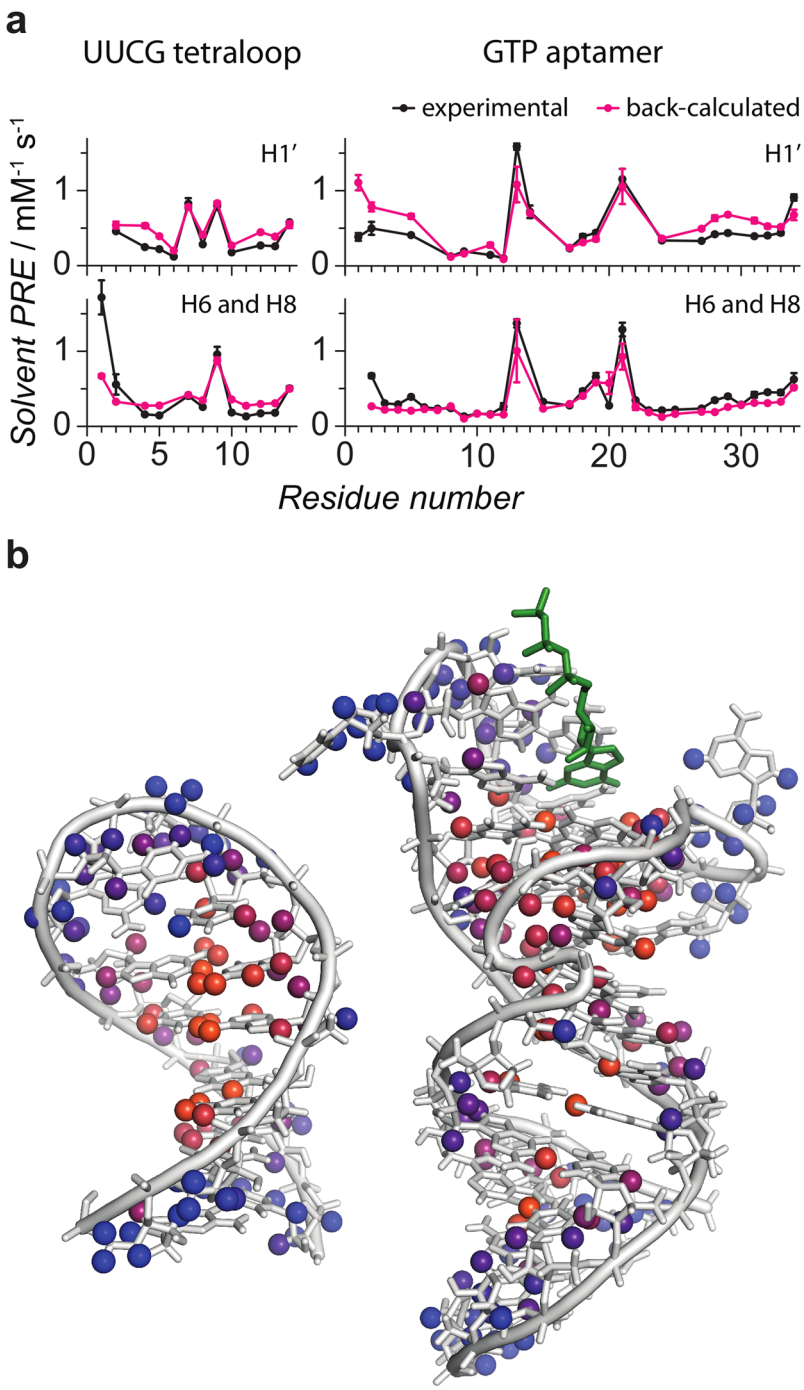


Figure 2. sPRE data correlates well with RNA structure. (**a**) Experimental sPRE data (black) are compared to predicted sPRE data based on NMR solution structures of the UUCG tetraloop (2KOC) and the GTP-bound aptamer (5LWJ). Data are shown for sugar protons (H1′) and aromatic protons (H6 and H8) as indicated. (**b**) NMR solution structures of the UUCG tetraloop[37,38] (2KOC, left) and the GTP-bound aptamer[39,40] (5LWJ, right, heavy atoms of GTP ligand shown in green) are shown. All protons for which a sPRE value was obtained are shown as spheres and colored according to the sPRE (blue corresponding to high and orange corresponding to low sPRE values).

torsion angle restraints. Strikingly, using only hydrogen bonds-derived restraints in combination with sPRE data, the structure of the UUCG tetraloop can be determined with an RMSD of 2.8 Å compared to the published NMR-based structure (Fig. 3a). Even sparse sPRE data sets, in some cases down to 25%, improve structural quality of the UUCG tetraloop (Supplementary Table 6).

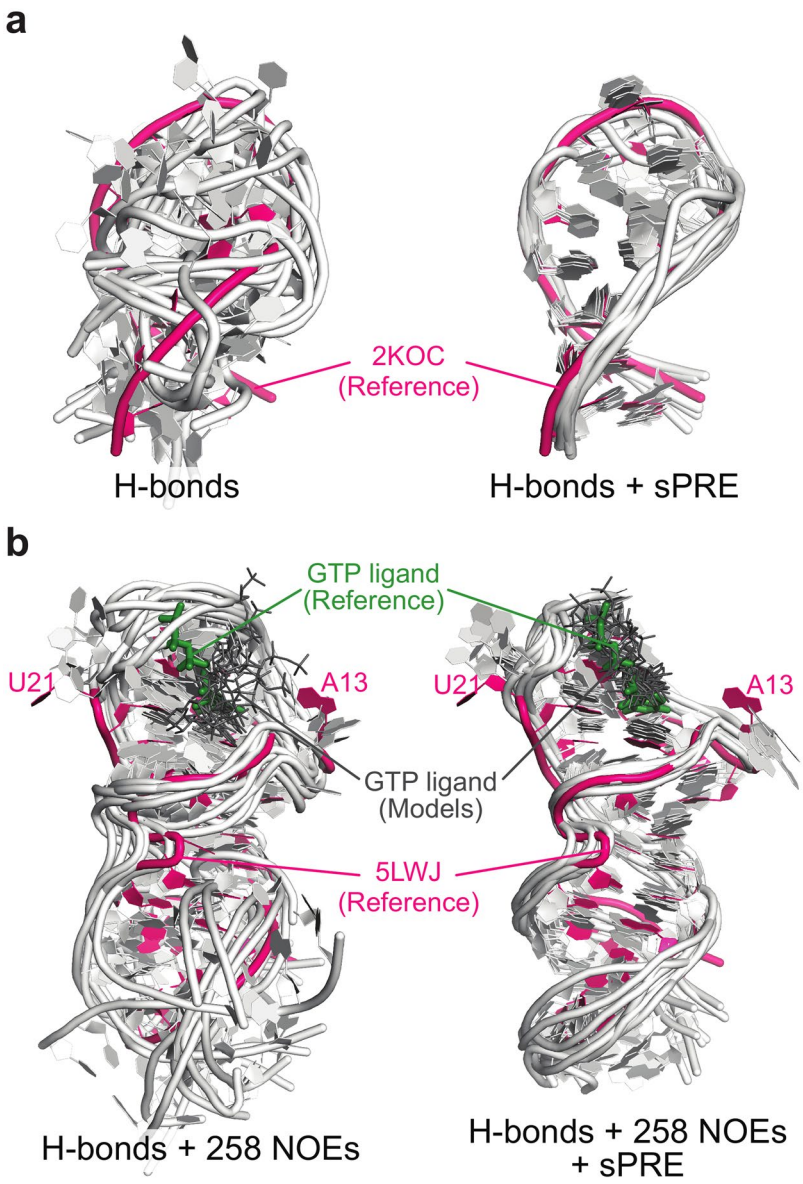


Figure 3. sPRE data improve structure determination of RNAs. Structural models of the UUCG tetraloop (**a**) and the GTP-bound GTP aptamer (**b**) were obtained without (left) and with (right) sPRE data using XplorNIH. The 10 best scored models (light gray) in terms of total energy were selected from a total of 200 models and aligned to the corresponding NMR structure (magenta). All restraints used in the computations are indicated below the respective models. In (**b**) heavy atoms of the GTP ligand are shown as sticks (Reference in green, computed models in dark gray) and the positions of the intrinsically flexible nucleotides A13 and U21[40] are indicated.

Discussion

In summary, we show that sPRE data are efficient NMR observables to probe RNA structure. sPRE data obtained with the paramagnetic compound Gd(DTPA-BMA) yield quantitative information of solvent accessibility as they correlate well with RNA structure, provide structural information for both buried and surface-exposed atoms and can be used as restraints to drive molecular dynamics-based structure determination of RNAs. Since sPRE data reflect the global fold of a RNA they are well suited to identify tertiary contacts or map interaction surfaces with other molecules[32], for example, in RNA-protein complexes. As the measurement of sPRE data does not require complete chemical shift assignments, surface accessibility restraints can be recorded for structural studies also in the case of large RNAs, in particular in combination with specific or segmental labelling, where it is difficult to obtain a comprehensive set of experimental restraints for structural analysis. Compared to proteins, several considerations have to be taken into account for sPRE-based studies of RNAs. First, different NMR pulse sequences might be needed, in particular for triple resonance experiments. Second, multiple data sets acquired for different RNA labelling schemes and different solvent conditions need to be combined. In order to compensate for minor differences between these data sets we have proposed an approach in which we normalize the sPRE data using

the sPRE of water protons in each sample. Third, lower concentrations of Gd(DTPA-BMA) are used for RNAs compared to proteins due to the higher sPRE caused by the lack of large hydrophobic cores resulting in smaller distances to the RNA surface. Last, chemical exchange of protons with water is typically more efficient in RNAs and modulates sPREs of nitrogen-bound amino and imino protons. In these cases, only the region of linear response is used for quantitative analysis.

Our results show that sPRE-derived restraints are particularly promising to define the global fold of larger RNAs and provide valuable information for structure determination that is orthogonal to other NMR-based restraints in molecular dynamics-based approaches. Recently, accuracy of chemical-shift based RNA *de novo* structure prediction using the Rosetta framework was significantly enhanced by improving the chemical shift scoring function for RNAs[52]. Combining this improved score with a sampling algorithm driven by sPRE data[33] and specific or segmental labeling could provide powerful tools for structural modeling of large RNAs based on NMR data in the future.

Methods

RNA sample preparation. A uniformly ^{13}C, ^{15}N labeled sample of the UUCG tetraloop (5′-GGCACUUCGGUGCC-3′) was purchased from Silantes (Munich, Germany). The NMR buffer contained 20 mM K_2HPO_4 and 0.4 mM EDTA and was adjusted to pH 6.4. Transcription and purification of the GTP class II aptamer in complex with GTP was been described previously[39]. The sample was measured in NMR buffer containing 25 mM KH_2PO_4, 25 mM KCl and 2 mM magnesium acetate. The buffer was adjusted to pH 6.3 and uniformly ^{13}C, ^{15}N labeled RNA was measured in the presence of a two-fold excess of unlabeled GTP. NMR assignments and a solution NMR structure of the UUCG tetraloop were obtained from literature[37,38] (PDB code 2KOC, BMRB 5705). For the GTP class II aptamer in complex with GTP, the recently reported NMR assignments and solution NMR structure were used[39,40] (PDB code 5LWJ, BMRB 25661).

NMR Experiments. NMR sPRE data were obtained by measuring proton R_1 relaxation rates as a function of the concentration of Gd(DTPA-BMA) using a saturation-recovery approach which has been described previously for protein applications[30]. Briefly, a 7.5 ms proton trim pulse followed by a gradient is used to dephase magnetization. Next, longitudinal magnetization is recovered during a recovery delay of variable length. The recovery delay is followed by a read-out spectrum, such as a ^{1}H,^{15}N or ^{1}H,^{13}C HSQC. This saturation-recovery scheme allows to use a short interscan delay (about 50 ms) which dramatically reduces experimental time compared to measuring proton R_2 rates. Pseudo-3D experiments were acquired for 4 to 7 different concentrations of the paramagnetic compound. To recover the NMR RNA sample, the paramagnetic compound was first removed by dialysis against water using a membrane with a 1 kDa molecular weight cut-off and then, the sample was lyophilized and resuspended in the corresponding NMR buffer.

sPRE data for nitrogen-bound protons of the UUCG tetraloop were recorded with a 0.8 mM uniformly ^{13}C, ^{15}N labeled sample at 283 K on an Avance III Bruker 900 MHz NMR spectrometer equipped with a cryo-TXI probe head. Proton R_1 relaxation rates were measured in the absence and presence of 0.5, 1, 1.75, 2.75 and 4.0 mM Gd(DTPA-BMA) using ^{1}H,^{15}N HSQC-based pseudo-3D experiments. sPRE data for carbon-bound protons were acquired with the same sample at 298 K and R_1 relaxation rates were measured in the absence and presence of 0.75, 1.5, 2.25, 3.25 and 4.5 mM Gd(DTPA-BMA) using ^{1}H,^{13}C HSQC-based pseudo-3D experiments. Delay times and other NMR parameters for all experiments as well as the total NMR time are presented in Supplementary Table 1. To obtain sPRE data of the GTP aptamer in complex with GTP, samples containing about 150 µM of (^{13}C, ^{15}N)-GU, (^{13}C, ^{15}N)-C or (^{13}C, ^{15}N)-A labeled RNA and 300 µM unlabeled GTP were measured at 293 K on an Avance III Bruker 900 MHz NMR spectrometer equipped with a cryo-TXI probe head and an Oxford Instruments 600 MHz NMR spectrometer equipped with a Bruker Avance III console and a cryo-TCI probe head. ^{1}H,^{13}C and ^{1}H,^{15}N HSQC-based pseudo-3D experiments were used to acquire proton R_1 relaxation rates for 4 to 7 different concentrations of Gd(DTPA-BMA). Delay times and other NMR parameters for all experiments as well as the total NMR time are presented in Supplementary Table 2. For all RNA samples, amino and imino protons were measured in buffer containing 5 to 10% D_2O whereas carbon-bound protons were acquired in buffer containing $\geq$99.990% D_2O (Sigma-Aldrich). For every titration step of Gd(DTPA-BMA), solvent water R_1 was measured and the derived sPRE of the water solvent was used to normalize the RNA sPRE values. The normalized sPRE data were then rescaled by the mean sPRE of the solvent, obtained by averaging the sPRE of water protons in all relevant experiments. This procedure allows to combine sPRE measurements of different samples or different magnetic field strengths (e.g. in different buffers or with different labeling schemes; see Supplementary Figure 6). Furthermore, by plotting the solvent water R_1 against the concentration of Gd(DTPA-BMA) errors in pipetting can be detected and compensated accordingly if necessary.

In the case of nitrogen-bound protons, plotting the proton R_1 against the concentration of Gd(DTPA-BMA) revealed a non-linear correlation with rates at low concentrations of the paramagnetic compound being lower than expected according to a linear model. Consequently, the R_1 for low concentrations (0 mM in the case of the UUCG tetraloop, as well as 0 and 0.645 mM in the case of the GTP aptamer) were omitted in the computations of the sPRE values and only the linear correlation was used (compare Supplementary Table 1 and 2 as well as Supplementary Figure 3). In all cases, at least 4 different concentrations were used to derive the sPRE.

For large RNAs, specific isotope labeling schemes are required to reduce signal overlap[9,10,53]. In a simple and straight-forward approach, only one or two nucleotide types are enriched in NMR active nuclei. While this approach reduces spectral complexity, the measurement of sPREs with different samples might introduce systematic errors, due to pipetting errors. In addition, the usage of different magnetic field strengths affects the scaling of the sPRE data[31]. To ensure that sPRE data acquired on multiple samples and field strengths are consistent, the sPRE data sets are scaled by referencing to the sPRE of the water solvent in each sample. This referencing notably improves the correlation of calculated and experimental sPRE (Supplementary Figure 6). To normalize the sPRE

data, the sPRE of water protons was acquired by measuring the R_1 rate of the water proton signal. The R_1 rate was measured using a pseudo 2D experiments with simple 1D proton read-out (the proton pulse length was set between 0.3 and 1 µs). The change of the R_1 rate of the water protons during the titration Gd(DTPA-BMA) with was used to obtain the sPRE of the water protons. The sPRE of the water solvent was then used to normalize the sPRE values of the RNA signals. The normalized sPRE values of different titration experiments (with different labeling schemes or at different field strengths) were then combined to generate a large sPRE data set for the corresponding RNA. To obtain an absolute quantity, the normalized sPRE data were then multiplied by the mean sPRE of the solvent, obtained by averaging the sPRE of water protons in all relevant experiments. This procedure allows to combine sPRE measurements of different samples or different magnetic field strengths (e.g. in different buffers or with different labeling schemes; see Supplementary Figure 6). Furthermore, by plotting the solvent water R_1 against the concentration of Gd(DTPA-BMA) errors in pipetting can be detected and compensated accordingly if necessary.

The measurement times for the collection of sPRE data can be quite long, as sPRE data are recorded as a series of proton R_1 relaxation experiments. Typical ^{1}H-R_1 rates in the absence of Gd(DTPA-BMA) are relatively slow with an average ^{1}H-R_1 rate of 0.23 s^{-1} ($\pm$0.10 s^{-1}) at 900 MHz and 0.30 s^{-1} ($\pm$0.14 s^{-1}) at 600 MHz. As a consequence, long delay times of several seconds (Supplementary Table 2) are required which in turn increase the overall experimental time. To overcome this drawback, the data sets of both RNAs were re-evaluated by determining the sPREs based on R_1 rates measured only in the presence of 0.75, 1.5, 3.25 and 4.5 mM Gd(DTPA-BMA) for the UUCG tetraloop and 0.8, 1.6, 3.2 and 4 mM for the GTP aptamer. Notably, the obtained sPRE data are fully consistent with those derived from a full set of measurements, but the optimized acquisition scheme reduces experimental time by about 40%. It should be noted, that experimental time could be further reduced by applying selective saturation schemes that employ longitudinal relaxation-enhancing techniques as demonstrated in a previous study[54].

NMR data analysis. NMR spectra were processed with the NMRpipe software package[55]. Peak positions from literature were transferred using CcpNmr Analysis[56] and peak intensities of well-resolved peaks were obtained using the nmrglue Python package[57]. Peak intensities of the pseudo-3D experiments were fitted to an exponential recovery function according to equation (1):

$$I(\tau) = I_0 - A \cdot e^{-R_1 \cdot \tau} \tag{1}$$

where τ is the recovery delay, $I(\tau)$ is the peak intensity measured for the recovery delay τ, I_0 is the maximum peak intensity, R_1 is the longitudinal proton relaxation rate and A is the amplitude of the relaxation.

To estimate the error of the peak intensities ε, the recovery delay τ_d was measured twice to obtain two intensity values $I(\tau_d, I, 1)$ and $I(\tau_d, I, 2)$, where i is the index of the peak. For every peak i, the difference of the duplicates $\Delta I_i = I(\tau_d, i, 1) - I(\tau_d, i, 2)$ was computed. The overall error of the pseudo-3D experiment $\varepsilon_{\text{pseudo-3D}}$ was then computed according to equation (2):

$$\varepsilon_{\text{pseudo}-3\text{D}} = \sqrt{\frac{1}{2N} \cdot \sum_{i}^{N} (\Delta I_i - \overline{\Delta I})^2} \tag{2}$$

where N is the number of peaks, and $\overline{\Delta I} = \sum_i^N \Delta I_i$ is the average difference of the duplicates. The subtraction of $\overline{\Delta I}$ accounts for systematic errors, and in all cases was significantly lower in magnitude than the differences of the duplicates ΔI_i.

The error of the peak intensities $\varepsilon_{\text{pseudo-3D}}$ was then used to estimate the error of the fitted proton relaxation rates ΔR_1. To this end, the experimental data was resampled using the following combined Monte Carlo-type bootstrapping approach: For every peak in a given pseudo-3D experiment, a set of delay-intensity data points, $I(\tau)$, was measured. This set is used to generate 1000 new random sets $\widetilde{I_j(\tau)}$, each having 2.5 times as many data points as the original data set $I(\tau)$ (Recovery delays measured as duplicates are only counted as one). These sets were created by randomly selecting data points from the original data set $I(\tau)$ and allowing values to be selected multiple times (Increasing the size of the new sets by a factor of 2.5 allows to generate a more diverse ensemble of sets). For every set $\widetilde{I_j(\tau)}$, the intensity values were randomly altered by adding a random number drawn from a normal distribution centered at 0 and with a standard deviation of $\varepsilon_{\text{pseudo-3D}}$. Every set $\widetilde{I_j(\tau)}$ was then fitted according to equation (1) giving rise to 1000 different R_1^j rates. The error of the proton relaxation rate ΔR_1 was computed as the standard deviation of all values R_1^j. This procedure was repeated for every peak and every pseudo-3D experiment.

To obtain the sPRE value, the proton R_1 rates and the corresponding errors ΔR_1 were collected for all measured concentrations of Gd(DTPA-BMA) c_{para} and a weighted linear regression using equation (3) was performed

$$R_1(c_{\text{para}}) = m_{\text{sPRE}} \cdot c_{\text{para}} + R_1^0 \tag{3}$$

where $R_1(c_{\text{para}})$ is the proton R_1 measured at the concentration c_{para} of the paramagnetic compound, the slope m_{sPRE} corresponds to the sPRE and R_1^0 is the fitted proton R_1 in the absence of the paramagnetic compound. The errors ΔR_1 (obtained from the resampling approach described above) were used as weights in a weighted linear regression and the error of the sPRE value Δm_{sPRE} as well as the error of the relaxation rate ΔR_1^0 were directly obtained from the weighted linear regression.

Prediction of sPRE Data. To predict sPRE data based on published NMR solution structure, a previously published grid-based approach was used[31]. Briefly, the structural model was placed in a regularly-spaced grid representing the uniformly distributed paramagnetic compound and the grid was built with a point-to-point distance of 0.1 Å and a minimum distance of 20 Å between the RNA model and the outer border of the grid. Next, grid points that overlap with the RNA model were removed assuming a molecular radius of 3.5 Å for the paramagnetic compound. To compute the sPRE for a given RNA proton $sPRE^{i}_{predicted}$, the distance-dependent paramagnetic effect[29] was numerically integrated over all remaining grid points according to equation (4):

$$sPRE^{i}_{predicted} = c \cdot \sum_{d_{i,j} < 20\ \text{Å}} \frac{1}{d_{i,j}^{6}} \tag{4}$$

where i is the index of a proton of the RNA, j is the index of the grid point, $d_{i,j}$ is the distance between the i-th proton and the j-th grid point and c is an arbitrary constant to scale the sPRE values. Here, c was chosen such that the sum of all predicted sPRE values is equal to the sum of all experimental sPRE values.

Structure Determination Benchmark. To demonstrate the potential of sPRE data for the structure determination of RNAs, structural models of the UUCG tetraloop as well as the GTP class II aptamer in complex with GTP were computed using the XplorNIH 2.40 framework[49,50]. The used protocol is based on the gb1_rdc example that is included in the XplorNIH package and was adjusted to fold the respective RNA during an annealing procedure starting from an extended structure. The initial temperature was set to 3500 and the simulation at this high temperature was stopped after reaching 1000 ps or 10,000 steps. The temperature was then reduced to a final temperature of 25 in steps of 12.5 degrees and simulations were run for at least 4 ps or 2000 steps at every temperature step. The annealing procedure was repeated to obtain 200 models. The protocol made use of the recently published torsion potential RNA-ff1[51] and the sPRE data was included in the structure determination using the nbTargetPot module[27] of the framework. nbTargetPot is the scaling factor (or weight) of the sPRE potential. The value of the scaling factor mainly depends on the scaling factor of the other energy terms. The weight of the sPRE potential should not be proportional to the size of the RNA. For the mentioned benchmark, the weight of the sPRE potential was determined by testing different weights and the weight that produced the best models was used for all runs. To this end, one weight was determined for every RNA and the obtained weight was used for all computations of the corresponding RNA. The lower value of the UUCG tetraloop can be explained with the fact that the Xplor runs of the UUCG tetraloop included more energy functions than the ones of the GTP aptamer (in the case of the UUCG tetraloop, torsion angles and RDC, both obtained from PDB entry 2KOC, were used). The sPRE data sets of the UUCG tetraloop and GTP aptamer were filtered to remove all data points with an experimental error above 10% and all data points of nitrogen-bound protons. The nbTargetPot energy was initialized using the slope and intercept parameter as returned by the calibrate function of the nbTargetPotTools module in combination with the NMR structure of the corresponding RNA. The weight of the nbTargetPot energy was set to 1000 in the case of the UUCG tetraloop and 3000 for the GTP aptamer. For more details such as weight factors of all potentials, please refer to the python code of the protocols which are shown at the end of this document for both RNAs.

The annealing protocol was then used to benchmark the benefit of the sPRE data for structure prediction. To this end, structural models of the UUCG tetraloop and the GTP-bound aptamer were computed in the absence and presence of the sPRE potential and in combination with experimental restraints derived from hydrogen bond information and experimental NOEs (PDB entries 2KOC and 5LWJ)[37,38,40]. To simulate several NOE assignments, random experimental NOE subsets were created by randomly drawing a certain percentage of restraints from the full NOE set (compare Supplementary Tables 3 and 4). To account for the effect of the random selection process, at least 3 different random sets with the same number of NOE restrains were created and used to benchmark the impact of the sPRE data. It should be noted, that the randomly created NOE subsets not only account for different levels of assignments, but also simulate different qualities of NOE data as random subsets with the same number of NOE restraints perform differently in driving the structure determination to the correct fold. For every given set of restraints, 200 structure models were computed with and without the sPRE data. To quantify the impact of the sPRE data, the 20 best models of each run (10%, scored according to the total energy) were selected and the RMSD to the published NMR structures was computed. Computation of RMSD was performed on all carbon, nitrogen and phosphorus atoms in the case of the UUCG tetraloop. For the GTP-bound aptamer, the RMSD was computed using all carbon, nitrogen and phosphorus atoms of the GTP ligand and all non-terminal nucleotides except the intrinsically flexible nucleotides A13 and U21[40].

A second benchmark was performed to address the performance of the sPRE potential in the absence of any other experimental restraint as well as in combination with other experimental NMR restraints, such hydrogen bonds, NOEs, RDCs and torsion angles. Experimental restraints for the UUCG tetraloop were obtained from PDB entry 2KOC[37,38]. Two hundred structure models of the UUCG tetraloop were computed with and without the sPRE data for every restraint set (compare Supplementary Tables 5) and the 20 best scored models according to the total energy were selected. The average RMSD to the published NMR structure was computed for the top 20 models and used to quantify the impact of the sPRE data (computation of RMSD was performed on all carbon, nitrogen and phosphorus atoms).

Data Availability. The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request. All scripts used for data analysis and back-calculation of sPRE data are openly available (http://mbbc.medunigraz.at/forschung/forschungseinheiten-und-gruppen/forschungsgruppe-tobias-madl/software/).

References

1. Morris, K. V. & Mattick, J. S. The rise of regulatory RNA. *Nat Rev Genet* **15**, 423–437, doi:10.1038/nrg3722 (2014).
2. Furtig, B., Richter, C., Wohnert, J. & Schwalbe, H. NMR spectroscopy of RNA. *Chembiochem* **4**, 936–962, doi:10.1002/cbic.200300700 (2003).
3. Latham, M. P., Brown, D. J., McCallum, S. A. & Pardi, A. NMR methods for studying the structure and dynamics of RNA. *Chembiochem* **6**, 1492–1505, doi:10.1002/cbic.200500123 (2005).
4. Bothe, J. R. *et al*. Characterizing RNA dynamics at atomic resolution using solution-state NMR spectroscopy. *Nat Methods* **8**, 919–931, doi:10.1038/nmeth.1735 (2011).
5. Dominguez, C., Schubert, M., Duss, O., Ravindranathan, S. & Allain, F. H. Structure determination and dynamics of protein-RNA complexes by NMR spectroscopy. *Prog Nucl Magn Reson Spectrosc* **58**, 1–61, doi:10.1016/j.pnmrs.2010.10.001 (2011).
6. Wijmenga, S. S. & van Buuren, B. N. M. The use of NMR methods for conformational studies of nucleic acids. *Progress in Nuclear Magnetic Resonance Spectroscopy* **32**, 287–387, doi:10.1016/s0079-6565(97)00023-x (1998).
7. Wu, H., Finger, L. D. & Feigon, J. Structure determination of protein/RNA complexes by NMR. *Methods Enzymol* **394**, 525–545, doi:10.1016/S0076-6879(05)94022-6 (2005).
8. Varani, G., Aboul-ela, F. & Allain, F. H. NMR investigation of RNA structure. *Prog Nucl Magn Reson Spectrosc* **29**, 51–127, doi:10.1016/0079-6565(96)01028-X (1996).
9. Duss, O., Lukavsky, P. J. & Allain, F. H. Isotope labeling and segmental labeling of larger RNAs for NMR structural studies. *Adv Exp Med Biol* **992**, 121–144, doi:10.1007/978-94-007-4954-2_7 (2012).
10. Liu, Y. *et al*. Synthesis and applications of RNAs with position-selective labelling and mosaic composition. *Nature* **522**, 368–372, doi:10.1038/nature14352 (2015).
11. Lu, K., Miyazaki, Y. & Summers, M. F. Isotope labeling strategies for NMR studies of RNA. *J Biomol NMR* **46**, 113–125, doi:10.1007/s10858-009-9375-2 (2010).
12. Milligan, J. F., Groebe, D. R., Witherell, G. W. & Uhlenbeck, O. C. Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates. *Nucleic Acids Res* **15**, 8783–8798 (1987).
13. Milligan, J. F. & Uhlenbeck, O. C. Synthesis of small RNAs using T7 RNA polymerase. *Methods Enzymol* **180**, 51–62 (1989).
14. Hennig, M., Williamson, J. R., Brodsky, A. S. & Battiste, J. L. Recent advances in RNA structure determination by NMR. *Curr Protoc Nucleic Acid Chem* **Chapter 7**, Unit 7 7, doi:10.1002/0471142700.nc0707s02 (2001).
15. Wunderlich, C. H. *et al*. Synthesis of (6-(13)C)pyrimidine nucleotides as spin-labels for RNA dynamics. *J Am Chem Soc* **134**, 7558–7569, doi:10.1021/ja302148g (2012).
16. Johnson, J. E. Jr., Julien, K. R. & Hoogstraten, C. G. Alternate-site isotopic labeling of ribonucleotides for NMR studies of ribose conformational dynamics in RNA. *J Biomol NMR* **35**, 261–274, doi:10.1007/s10858-006-9041-x (2006).
17. Alvarado, L. J. *et al*. Regio-selective chemical-enzymatic synthesis of pyrimidine nucleotides facilitates RNA structure and dynamics studies. *Chembiochem* **15**, 1573–1577, doi:10.1002/cbic.201402130 (2014).
18. Longhini, A. P. *et al*. Chemo-enzymatic synthesis of site-specific isotopically labeled nucleotides for use in NMR resonance assignment, dynamics and structural characterizations. *Nucleic Acids Res* **44**, e52, doi:10.1093/nar/gkv1333 (2016).
19. Wenter, P., Reymond, L., Auweter, S. D., Allain, F. H. & Pitsch, S. Short, synthetic and selectively 13C-labeled RNA sequences for the NMR structure determination of protein-RNA complexes. *Nucleic Acids Res* **34**, e79, doi:10.1093/nar/gkl427 (2006).
20. Tzakos, A. G., Easton, L. E. & Lukavsky, P. J. Complementary segmental labeling of large RNAs: economic preparation and simplified NMR spectra for measurement of more RDCs. *J Am Chem Soc* **128**, 13344–13345, doi:10.1021/ja064807o (2006).
21. Xu, J., Lapham, J. & Crothers, D. M. Determining RNA solution structure by segmental isotopic labeling and NMR: application to Caenorhabditis elegans spliced leader RNA 1. *Proc Natl Acad Sci USA* **93**, 44–48 (1996).
22. Nelissen, F. H. *et al*. Multiple segmental and selective isotope labeling of large RNA for NMR structural studies. *Nucleic Acids Res* **36**, e89, doi:10.1093/nar/gkn397 (2008).
23. Clore, G. M. & Iwahara, J. Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes. *Chem Rev* **109**, 4108–4139, doi:10.1021/cr900033p (2009).
24. Lipsitz, R. S. & Tjandra, N. Residual dipolar couplings in NMR structure analysis. *Annu Rev Biophys Biomol Struct* **33**, 387–413, doi:10.1146/annurev.biophys.33.110502.140306 (2004).
25. Dallmann, A. & Sattler, M. Detection of hydrogen bonds in dynamic regions of RNA by NMR spectroscopy. *Curr Protoc Nucleic Acid Chem* **59**, 7 22 21–19, doi:10.1002/0471142700.nc0722s59 (2014).
26. Grzesiek, S., Cordier, F., Jaravine, V. & Barfield, M. Insights into biomolecular hydrogen bonds from hydrogen bond scalar couplings. *Progress in Nuclear Magnetic Resonance Spectroscopy* **45**, 275–300, doi:10.1016/j.pnmrs.2004.08.001 (2004).
27. Wang, Y., Schwieters, C. D. & Tjandra, N. Parameterization of solvent-protein interaction and its use on NMR protein structure determination. *J Magn Reson* **221**, 76–84, doi:10.1016/j.jmr.2012.05.020 (2012).
28. Modig, K., Liepinsh, E., Otting, G. & Halle, B. Dynamics of protein and peptide hydration. *Journal of the American Chemical Society* **126**, 102–114, doi:10.1021/ja038325d (2004).
29. Hocking, H. G., Zangger, K. & Madl, T. Studying the structure and dynamics of biomolecules by using soluble paramagnetic probes. *Chemphyschem* **14**, 3082–3094, doi:10.1002/cphc.201300219 (2013).
30. Madl, T., Bermel, W. & Zangger, K. Use of relaxation enhancements in a paramagnetic environment for the structure determination of proteins using NMR spectroscopy. *Angew Chem Int Ed Engl* **48**, 8259–8262, doi:10.1002/anie.200902561 (2009).
31. Pintacuda, G. & Otting, G. Identification of protein surfaces by NMR measurements with a pramagnetic Gd(III) chelate. *J Am Chem Soc* **124**, 372–373 (2002).
32. Madl, T., Guttler, T., Gorlich, D. & Sattler, M. Structural analysis of large protein complexes using solvent paramagnetic relaxation enhancements. *Angew Chem Int Ed Engl* **50**, 3993–3997, doi:10.1002/anie.201007168 (2011).
33. Hartlmuller, C., Gobl, C. & Madl, T. Prediction of Protein Structure Using Surface Accessibility Data. *Angew Chem Int Ed Engl* **55**, 11970–11974, doi:10.1002/anie.201604788 (2016).
34. Venditti, V., Niccolai, N. & Butcher, S. E. Measuring the dynamic surface accessibility of RNA with the small paramagnetic molecule TEMPOL. *Nucleic Acids Res* **36**, e20, doi:10.1093/nar/gkm1062 (2008).
35. Bernini, A. *et al*. NMR studies on the surface accessibility of the archaeal protein Sso7d by using TEMPOL and Gd(III)(DTPA-BMA) as paramagnetic probes. *Biophys Chem* **137**, 71–75, doi:10.1016/j.bpc.2008.07.003 (2008).
36. Caravan, P., Ellison, J. J., McMurry, T. J. & Lauffer, R. B. Gadolinium(III) Chelates as MRI Contrast Agents: Structure, Dynamics, and Applications. *Chem Rev* **99**, 2293–2352 (1999).
37. Furtig, B., Richter, C., Bermel, W. & Schwalbe, H. New NMR experiments for RNA nucleobase resonance assignment and chemical shift analysis of an RNA UUCG tetraloop. *J Biomol NMR* **28**, 69–79, doi:10.1023/B:JNMR.0000012863.63522.1f (2004).
38. Nozinovic, S., Furtig, B., Jonker, H. R., Richter, C. & Schwalbe, H. High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. *Nucleic Acids Res* **38**, 683–694, doi:10.1093/nar/gkp956 (2010).
39. Wolter, A. C. *et al*. NMR resonance assignments for the class II GTP binding RNA aptamer in complex with GTP. *Biomol NMR Assign* **10**, 101–105, doi:10.1007/s12104-015-9646-7 (2016).
40. Wolter, A. C. *et al*. A Stably Protonated Adenine Nucleotide with a Highly Shifted pKa Value Stabilizes the Tertiary Structure of a GTP-Binding RNA Aptamer. *Angewandte Chemie*. doi:10.1002/anie.201609184 (2016).

41. Matsumura, K. & Komiyama, M. Enormously fast RNA hydrolysis by lanthanide(III) ions under physiological conditions: eminent candidates for novel tools of biotechnology. *J Biochem* **122**, 387–394 (1997).
42. Deng, N. J. & Cieplak, P. Free energy profile of RNA hairpins: a molecular dynamics simulation study. *Biophys J* **98**, 627–636, doi:10.1016/j.bpj.2009.10.040 (2010).
43. Giambasu, G. M., York, D. M. & Case, D. A. Structural fidelity and NMR relaxation analysis in a prototype RNA hairpin. *RNA* **21**, 963–974, doi:10.1261/rna.047357.114 (2015).
44. Villa, A., Widjajakusuma, E. & Stock, G. Molecular dynamics simulation of the structure, dynamics, and thermostability of the RNA hairpins uCACGg and cUUCGg. *J Phys Chem B* **112**, 134–142, doi:10.1021/jp0764337 (2008).
45. Borkar, A. N., Vallurupalli, P., Camilloni, C., Kay, L. E. & Vendruscolo, M. Simultaneous NMR characterisation of multiple minima in the free energy landscape of an RNA UUCG tetraloop. *Phys Chem Chem Phys* **19**, 2797–2804, doi:10.1039/c6cp08313g (2017).
46. Vallurupalli, P. & Kay, L. E. A suite of 2H NMR spin relaxation experiments for the measurement of RNA dynamics. *J Am Chem Soc* **127**, 6893–6901, doi:10.1021/ja0427799 (2005).
47. Williams, D. J. & Hall, K. B. Experimental and computational studies of the G[UUCG]C RNA tetraloop. *J Mol Biol* **297**, 1045–1061, doi:10.1006/jmbi.2000.3623 (2000).
48. Akke, M., Fiala, R., Jiang, F., Patel, D. & Palmer, A. G. 3rd. Base dynamics in a UUCG tetraloop RNA hairpin characterized by 15N spin relaxation: correlations with structure and stability. *RNA* **3**, 702–709 (1997).
49. Schwieters, C. D., Kuszewski, J. J. & Clore, G. M. Using Xplor-NIH for NMR molecular structure determination. *Progress in nuclear magnetic resonance spectroscopy* **48**, 47–62 (2006).
50. Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. The Xplor-NIH NMR molecular structure determination package. *Journal of magnetic resonance* **160**, 65–73 (2003).
51. Bermejo, G. A., Clore, G. M. & Schwieters, C. D. Improving NMR Structures of RNA. *Structure* **24**, 806–815, doi:10.1016/j.str.2016.03.007 (2016).
52. Sripakdeevong, P. *et al*. Structure determination of noncanonical RNA motifs guided by (1)H NMR chemical shifts. *Nat Methods* **11**, 413–416, doi:10.1038/nmeth.2876 (2014).
53. Dallmann, A. *et al*. Site-Specific Isotope-Labeling of Inosine Phosphoramidites and NMR Analysis of an Inosine-Containing RNA Duplex. *Chemistry*. doi:10.1002/chem.201602784 (2016).
54. Farjon, J. *et al*. Longitudinal-relaxation-enhanced NMR experiments for the study of nucleic acids in solution. *Journal of the American Chemical Society* **131**, 8571–8577, doi:10.1021/ja901633y (2009).
55. Delaglio, F. *et al*. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6**, 277–293 (1995).
56. Vranken, W. F. *et al*. The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* **59**, 687–696, doi:10.1002/prot.20449 (2005).
57. Helmus, J. J. & Jaroniec, C. P. Nmrglue: an open source Python package for the analysis of multidimensional NMR data. *J Biomol NMR* **55**, 355–367, doi:10.1007/s10858-013-9718-x (2013).

Acknowledgements

This work was supported by the Integrative Metabolism Research Center Graz (to T.M.), the Austrian infrastructure programme 2016/2017 (to T.M.), BioTechMed/Graz (to T.M.), the Omics Center Graz (to T.M.), the Bavarian Ministry of Sciences, Research and the Arts (Bavarian Molecular Biosystems Research Network, to T.M.), the President's International Fellowship Initiative of CAS (No. 2015VBB045, to T.M.), the National Natural Science Foundation of China (No. 31450110423, to T.M.), the Austrian Science Fund (FWF: P28854 and W1226-B18 to T.M.) as well as the Deutsche Forschungsgemeinschaft (DFG) with the grants Wo901/1-1 (to J.W.), SFB1035 (to M.S.), GRK1721 (to M.S.) and MA5703/1-1 (to T.M.).

Author Contributions

C.H. and T.M. designed the studies, established and tested the NMR experiments, and performed the prediction of sPRE data. J.G., A.W., J.W. and M.S. prepared the isotope labelled RNA samples. C.H. and T.M., C.H. and J.G. performed the NMR titrations. C.H. processed and fitted the NMR data, performed the structure prediction computations and analyzed the obtained models. C.H. and T.M. wrote the manuscript and all authors have given critical feedback and approved the final version of the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-05821-z

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2017

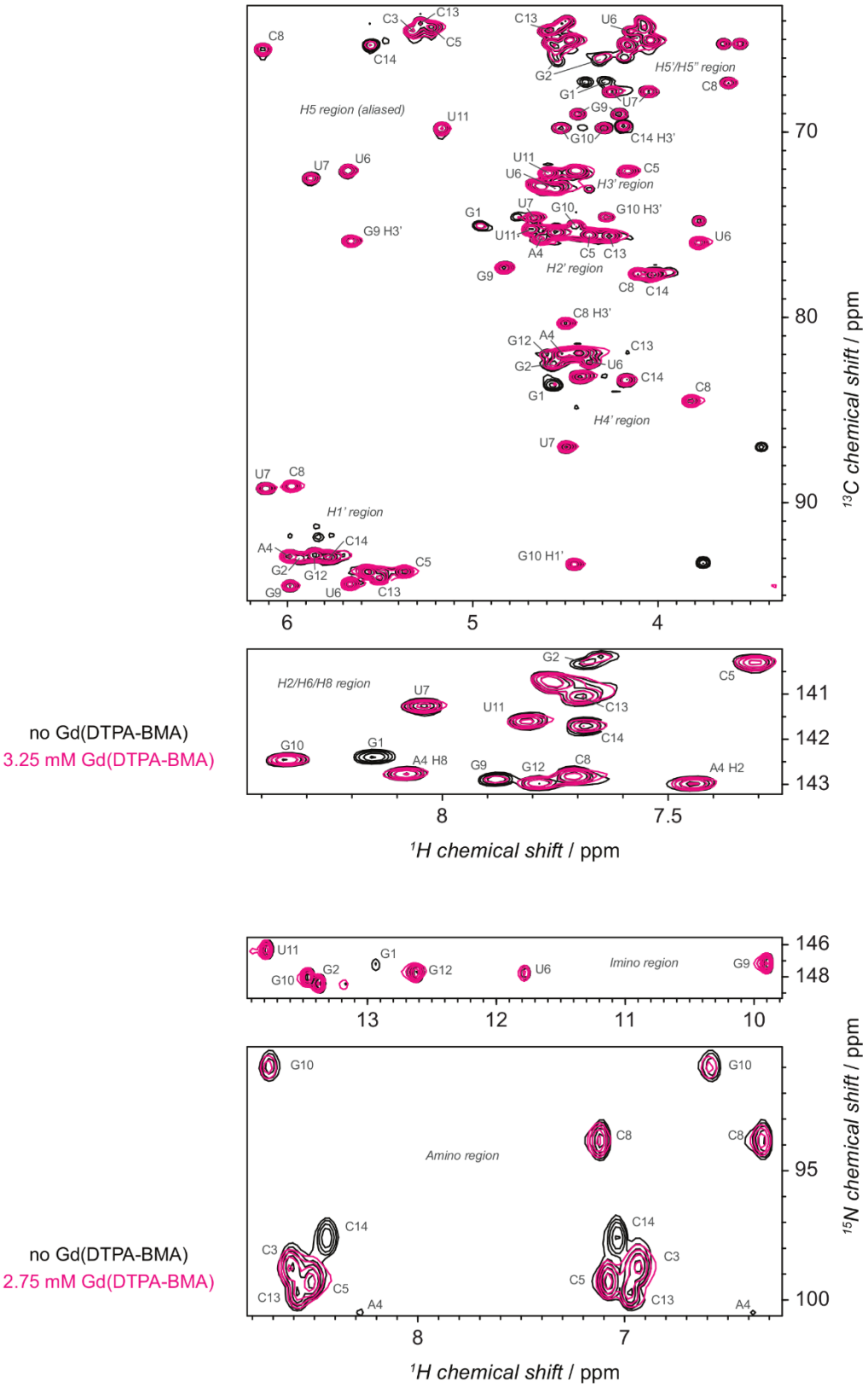
Supplementary Information

RNA structure refinement using NMR solvent accessibility data

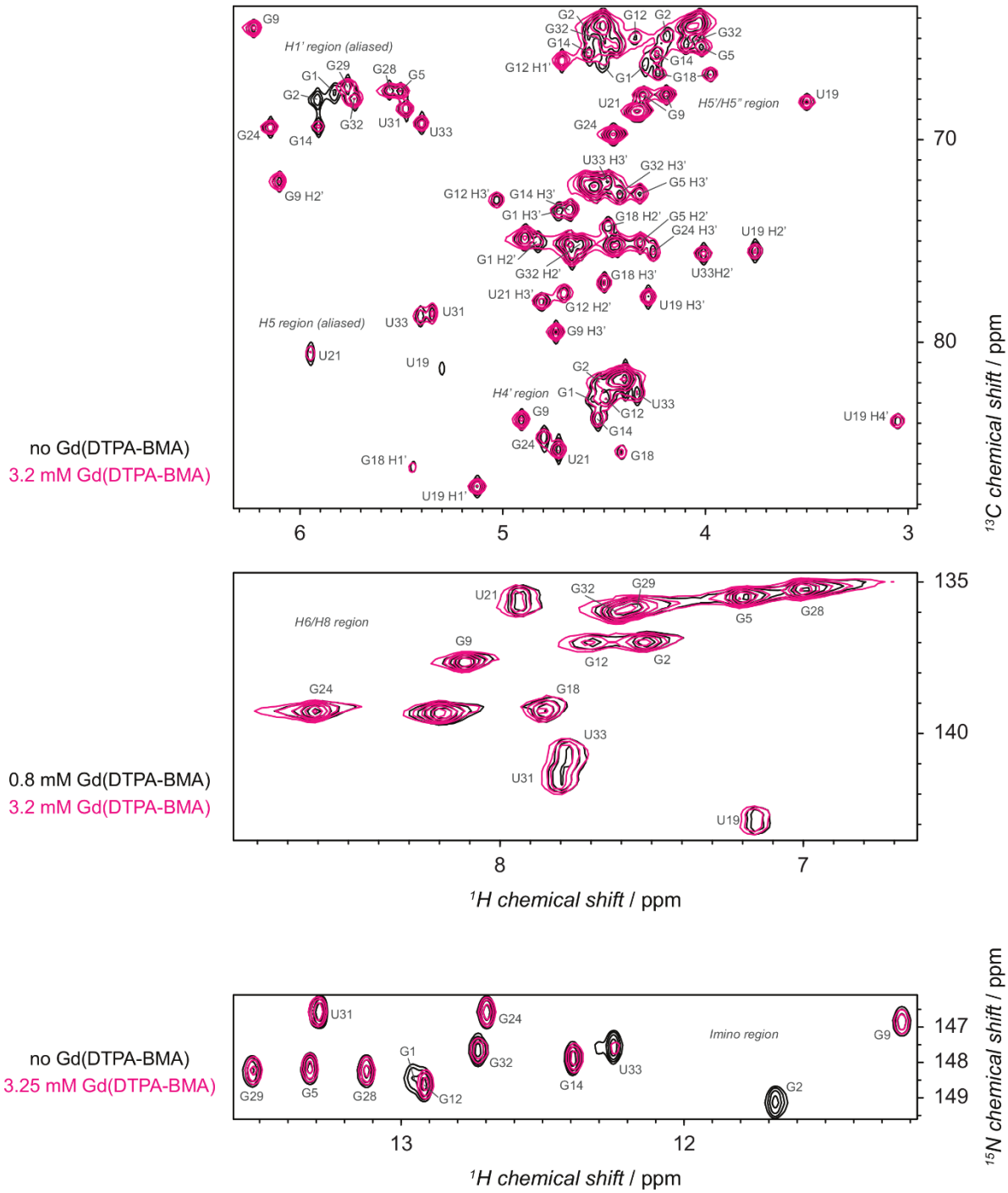
Christoph Hartlmüller[1,2] *, Johannes C. Günther[1,2] *, Antje C. Wolter[3], Jens Wöhnert[3], Michael Sattler[1,2] & Tobias Madl[1,2,4]

[1]Center for Integrated Protein Science Munich, Department Chemie, Technical University of Munich, Lichtenbergstr. 4, 85748 Garching, Germany. [2]Institute of Structural Biology, Helmholtz Zentrum München, Ingolstadter Landstr. 1, 85764 Neuherberg, Germany. [3]Institut für Molekulare Biowissenschaften and Zentrum für Biomolekulare Magnetische Resonanz (BMRZ), Goethe-Universität Frankfurt, Max-von-Laue Str. 9, 60438 Frankfurt/M, Germany. [4]Institute of Molecular Biology and Biochemistry, Center of Molecular Medicine, Medical University of Graz, Har-rachgasse 21, 8010 Graz, Austria. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to T.M. (email: tobias.madl@medunigraz.at).

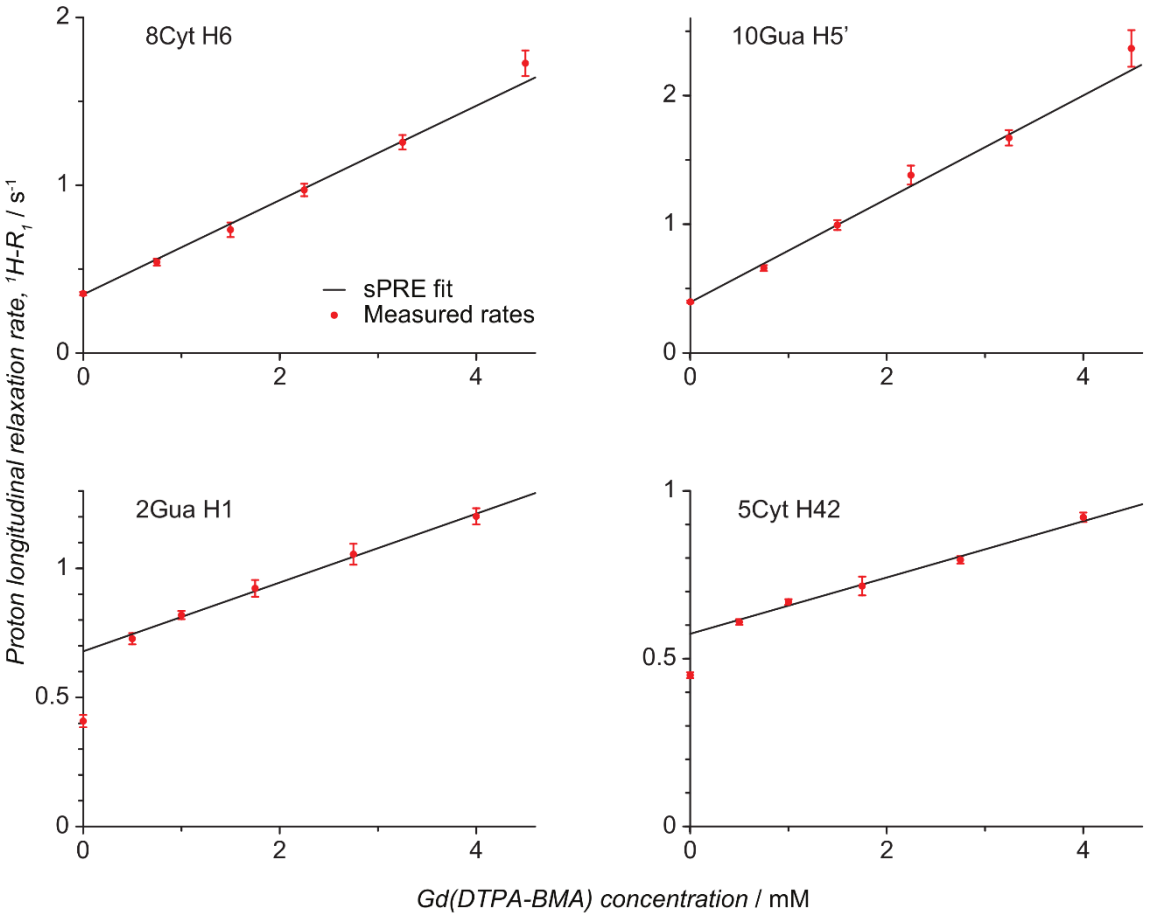
Supplementary Figures



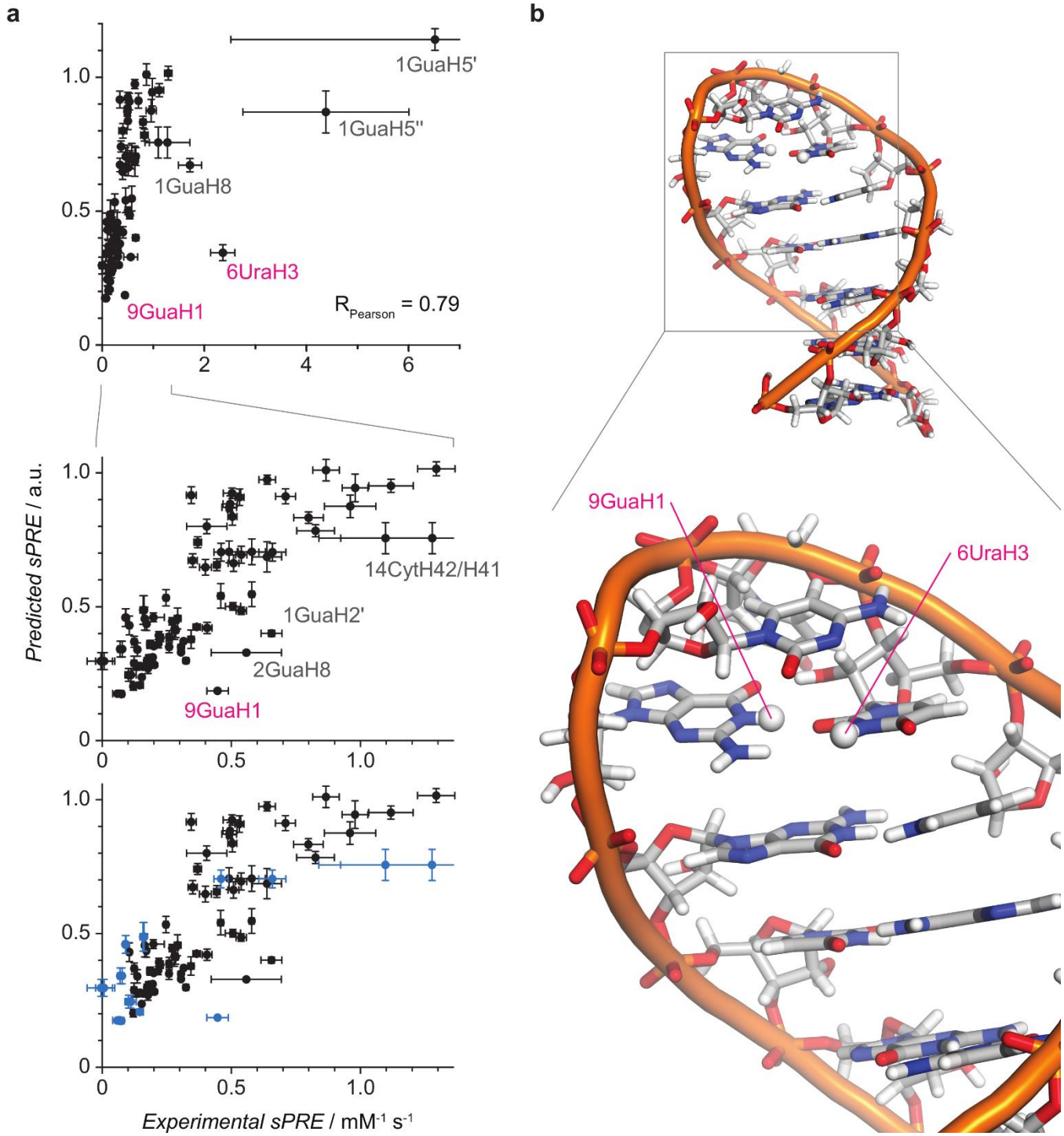
Supplementary Figure 1 | NMR spectra of the UUCG tetraloop in the absence (black) and presence of paramagnetic compound (magenta). The shown regions of the spectra are (from top to bottom) the sugar (including H5), the bases (H6 and H8), the imino and the amino region. Note that some resonances appear at the aliased frequency. Assignments were obtained from BMRB entry 5705[1].



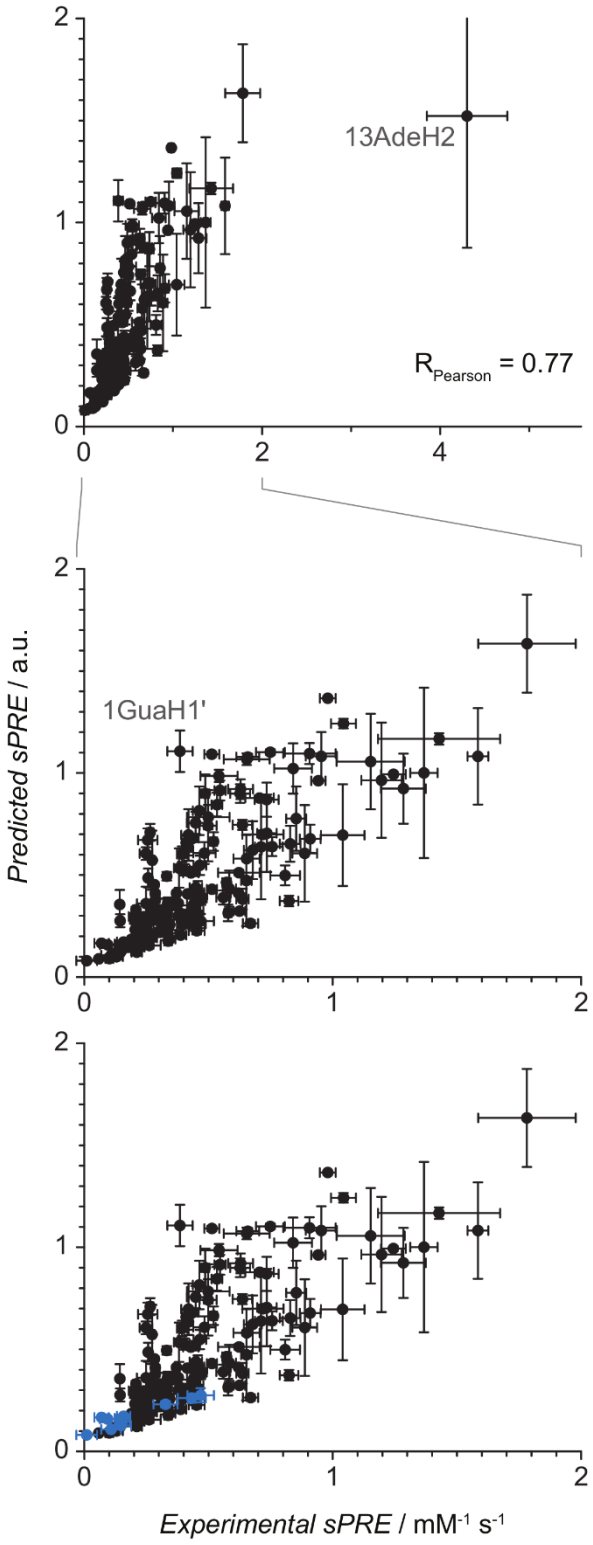
Supplementary Figure 2 | NMR spectra of a (^{13}C, ^{15}N)-GU labeled sampled of the GTP aptamer in the absence (black) and presence of paramagnetic compound (magenta). The shown regions of the spectra are (from top to bottom) the sugar (including H5), the bases (H6 and H8) and the imino region. Note that some resonances appear at the aliased frequency. Assignments were obtained from BMRB entry 25661[2].



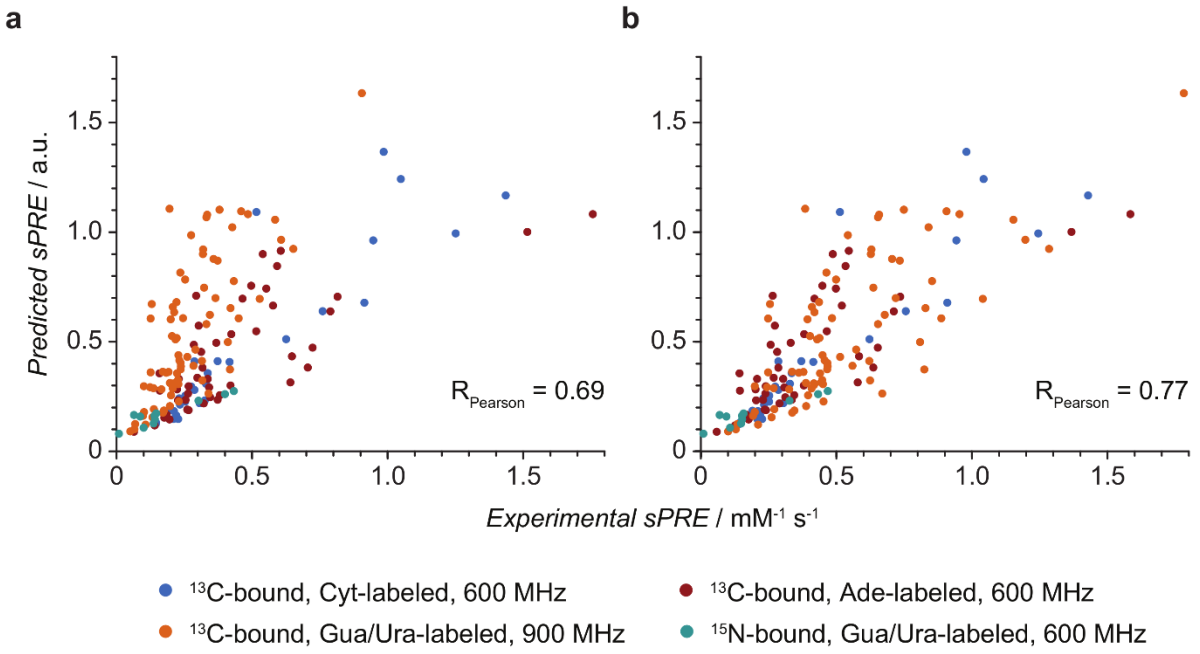
Supplementary Figure 3 | Increase of proton R_1 as function of the concentration of Gd(DTPA-BMA). Rates are plotted as red dots for carbon-bound (8Cyt H6 and 10Gua H5') as well as for nitrogen-bound (2Gua H1 and 5Cyt H42) protons of the UUCG tetraloop. The parameters obtained from fitting the linear sPRE model are shown as a black line. For the nitrogen-bound protons, R_1 data points in the absence of paramagnetic compound were not used in the weighted linear regression to fit the sPRE model (compare Supplementary Table 1).



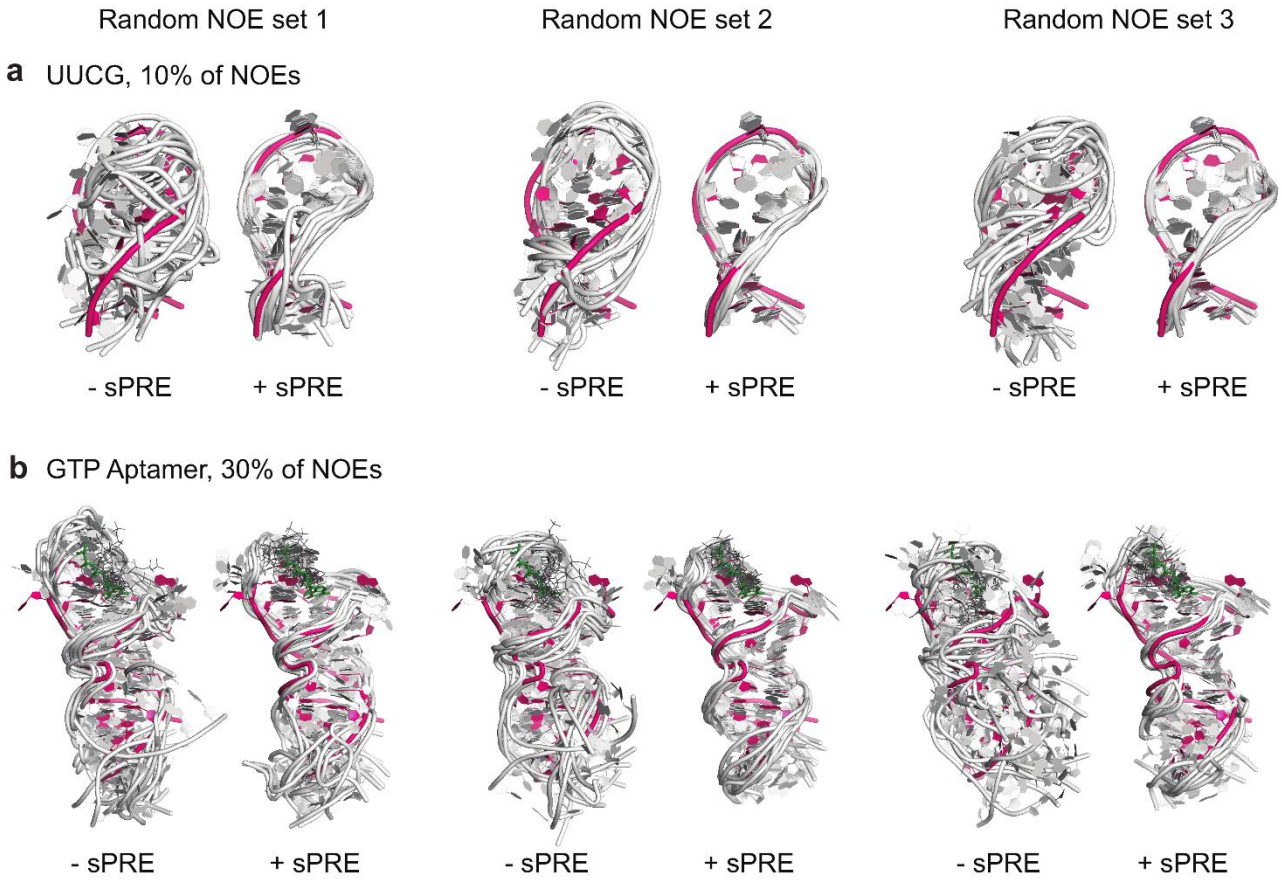
Supplementary Figure 4 | sPRE data of the UUCG tetraloop correlate well with the corresponding structure. (**a**) Correlation of experimental and predicted sPRE data for the UUCG tetraloop is shown. The middle panel is an expansion of the top panel. Outliers located at terminal nucleotides are labeled in grey and those located in the loop region are labeled in magenta. In the bottom plot, the carbon-bound protons are shown in black and the nitrogen-bound protons are drawn in blue. The Pearson correlation coefficient was calculated using all data points for which the errors of the experimental and predicted sPRE value is below 10 %. (**b**) The complete structural model of the UUCG tetraloop (PDB code 2KOC) is shown on the top and a close-up of the loop region is shown on the bottom. The positions of the outliers are indicated and both protons are shown as larger spheres.



Supplementary Figure 5 | sPRE data of the GTP-bound GTP class II aptamer correlate well with the corresponding structure. The correlation of experimental and predicted sPRE data for the ligand-bound GTP aptamer is shown and outliers are labeled in grey. In the bottom plot, the carbon-bound protons are shown in black and the nitrogen-bound protons are drawn in blue. The Pearson correlation coefficient was calculated using all data points for which the errors of the experimental and predicted sPRE value is below 10 %.



Supplementary Figure 6 | Effect of normalization using the sPRE of the water solvent. Scatter plots show the measured and predicted sPRE data of the GTP-bound GTP aptamer before (**a**) and after (**b**) normalization of the experimental sPRE data. The data was obtained in several experiments (using 3 different labeling schemes, H_2O and D_2O-based buffer and two different field strengths) as indicated by the different colors. For each of the experiments, the sPRE of water solvent was measured and used to normalize the sPRE. The normalized data sets were then rescaled by the average sPRE of water solvent in all experiments. The Pearson correlation coefficients were calculated using all data points for which the errors of the experimental and predicted sPRE value is below 10 %. sPRE data for carbon- and nitrogen-bound protons were acquired in D_2O and H_2O buffer, respectively.



Supplementary Figure 7 | sPRE data improve structure determination of RNAs with sparse data sets. Structural models of the UUCG tetraloop (a) and the GTP-bound GTP aptamer (b) were obtained without (-sPRE) and with (+sPRE) sPRE data using XplorNIH (see also supplementary tables 3 and 4). The 10 best scored models (light gray) in terms of total energy were selected from a total of 200 models and aligned to the corresponding NMR structure (magenta). All restraints used in the computations are indicated. In (b) heavy atoms of the GTP ligand are shown as sticks (Reference in green, computed models in dark gray).

Supplementary Tables

Supplementary Table 1 | Experimental details for the acquisition of sPRE data for the UUCG tetraloop. All experiments were acquired on a 900 MHz spectrometer equipped with a cryo probe head.

^{15}N amino region

Conc. of Gd(DTPA-BMA) [mM]	Number of delays	Longest delay [s]	Number of scans	Number of increments	Approx. NMR time
0 [a]	12	6	8	20	1:38 h
0.5	12	6	16	20	3:16 h
1	14	6	16	20	3:40 h
1.75	12	4	16	20	2:11 h
2.75	12	3.5	20	20	2:24 h
4	12	3	28	20	2:54 h

[a] Not used for fitting the linear sPRE model

^{15}N imino region

Conc. of Gd(DTPA-BMA) [mM]	Number of delays	Longest delay [s]	Number of scans	Number of increments	Approx. NMR time
0 [a]	12	6	16	6	0:59 h
0.5	12	6	20	6	1:13 h
1	14	6	20	6	1:22 h
1.75	12	4	20	6	0:49 h
2.75	12	3.5	24	6	0:52 h
4	12	3	36	6	1:07 h

[a] Not used for fitting the linear sPRE model

^{13}C sugar region

Conc. of Gd(DTPA-BMA) [mM]	Number of delays	Longest delay [s]	Number of scans	Number of increments	Approx. NMR time
0	9	11	8	81	9:17 h
0.75	9	8	8	80	6:41 h
1.5	9	5.5	8	80	4:37 h
2.25	11	5	8	80	3:55 h
3.25	11	4	8	80	3:08 h
4.5	11	4.5	8	80	3:31 h

^{13}C base region

Conc. of Gd(DTPA-BMA) [mM]	Number of delays	Longest delay [s]	Number of scans	Number of increments	Approx. NMR time
0	9	10	16	19	3:57 h
0.75	9	8	16	19	3:10 h
1.5	9	5.5	16	19	2:11 h
2.25	11	5	16	19	1:51 h
3.25	11	5	16	19	1:51 h
4.5	11	4.5	16	19	1:40 h

Supplementary Table 2 | Experimental details for the acquisition of sPRE data for the GTP aptamer.

^{15}N imino region, GU-labeled sample, 600 MHz spectrometer with cryo probe head

Conc. of Gd(DTPA-BMA) [mM]	Number of delays	Longest delay [s]	Number of scans	Number of increments	Approx. NMR time
0 [a]	22	7	48	6	5:43 h
0.645 [a]	22	7	56	6	6:40 h
1.29	22	6	56	6	5:43 h
1.94	22	4	72	6	4:56 h
2.58	22	3	72	6	3:43 h
3.23	22	3	100	6	5:09 h

[a] Not used for fitting the linear sPRE model

^{13}C sugar region, GU-labeled sample, 900 MHz spectrometer with cryo probe head

Conc. of Gd(DTPA-BMA) [mM]	Number of delays	Longest delay [s]	Number of scans	Number of increments	Approx. NMR time
0	11	10	40	50	24:20 h
0.8	11	8	40	50	19:28 h
1.6	11	6	42	50	15:23 h
2.4	11	5.5	44	50	14:47 h
3.2	11	5	44	50	13:28 h
4	11	4.5	42	50	11:35 h

^{13}C base region, GU-labeled sample, 900 MHz spectrometer with cryo probe head

Conc. of Gd(DTPA-BMA) [mM]	Number of delays	Longest delay [s]	Number of scans	Number of increments	Approx. NMR time
0	11	10	40	20	9:43 h
0.8	11	8	38	20	7:23 h
1.6	11	6	38	20	5:33 h
2.4	11	5	38	20	4:38 h
3.2	11	5	38	20	4:38 h
4	11	4.5	34	20	3:44 h

^{13}C sugar region, A-labeled sample, 600 MHz spectrometer with cryo probe head

Conc. of Gd(DTPA-BMA) [mM]	Number of delays	Longest delay [s]	Number of scans	Number of increments	Approx. NMR time
0	11	8.5	30	40	16:23 h
0.8	11	6	36	40	13:55 h
1.6	11	4.75	50	40	15:21 h
2.4	11	3.5	66	40	15:00 h
3.2	17	3	92	40	25:07 h
4	11	2.5	80	40	13:06 h

^{13}C base region, A-labeled sample, 600 MHz spectrometer with cryo probe head

Conc. of Gd(DTPA-BMA) [mM]	Number of delays	Longest delay [s]	Number of scans	Number of increments	Approx. NMR time
0	11	8.5	38	15	7:47 h
0.8	11	6	38	15	5:30 h
1.6	11	6	56	15	6:09 h
2.4	17	6	56	15	8:23 h
3.2	17	6	60	15	8:59 h
4	17	4	50	15	5:01 h

^{13}C sugar region, C-labeled sample, 600 MHz spectrometer with cryo probe head

Conc. of Gd(DTPA-BMA) [mM]	Number of delays	Longest delay [s]	Number of scans	Number of increments	Approx. NMR time
0	11	10	56	25	22:27 h
0.8	11	7.5	50	32	19:18 h
1.6	11	4.5	58	32	13:30 h
2.4	11	3.5	66	32	12:00 h
3.2	11	2.5	74	32	13:31 h

4	11	2.5	74	32	7:23 h
4.8	11	2.5	64	32	6:23 h

^{13}C base region, C-labeled sample, 600 MHz spectrometer with cryo probe head

Conc. of Gd(DTPA-BMA) [mM]	Number of delays	Longest delay [s]	Number of scans	Number of increments	Approx. NMR time
0	11	10	56	8	7:11 h
0.8	11	7.5	64	8	6:10 h
1.6	11	4.5	70	8	4:04 h
2.4	11	3.5	80	8	3:38 h
3.2	11	2	80	8	2:06 h
4	11	2	80	8	1:36 h
4.8	11	2	72	8	1:27

Supplementary Table 3 | sPRE data improves structure determination of the UUCG tetraloop. Structural models of the UUCG tetraloop were computed using hydrogen bonds-derived restraints in combination with different sets of experimental NOEs obtained from PDB entry 2KOC. The size of the NOE restraints set was varied and for every size, 3 randomly created sets were generated. For every restraint set, 200 models were computed in the absence and presence of the sPRE potential and the best 20 models based on the total energy were selected. RMSD values are computed by comparing all carbon, nitrogen and phosphorus atoms to the published NMR structure and the average RMSDs of the 20 best scored models are shown in the table.

Number of NOE restraints (percentage of all unambig. NOEs)	sPRE data used	Average RMSD of best scored models [Å]		
		Random set 1	Random set 2	Random set 3
25 (10%)	no	7.32	4.75	5.80
25 (10%)	yes	2.10	1.35	1.77
50 (20%)	no	3.49	6.57	2.03
50 (20%)	yes	1.39	1.49	1.41
75 (30%)	no	1.90	2.64	4.52
75 (30%)	yes	1.58	1.39	1.60
100 (40%)	no	1.94	1.88	1.54
100 (40%)	yes	1.30	1.30	1.46
125 (50%)	no	1.56	1.31	1.28
125 (50%)	yes	1.23	1.39	1.13
188 (75%)	no	1.44	1.27	1.05
188 (75%)	yes	1.05	1.28	1.04
251 (100%)	no	1.13		
251 (100%)	yes	1.05		

Supplementary Table 4 | sPRE data improves structure determination of the GTP class II aptamer in complex with GTP. Structural models of the ligand-bound aptamer were computed using experimental hydrogen bonds-derived restraints in combination with different sets of experimental NOEs. The size of the NOE restraints set was varied and for every size, at least 3 randomly created sets were generated. For every restraint set, 200 models were computed in the absence and presence of the sPRE potential and the best 20 models based on the total energy were selected. RMSD values are computed by comparing all carbon, nitrogen and phosphorus atoms of the GTP ligand and all non-terminal nucleotides (except the flexible A13 and U21) to the published NMR structure. The average RMSDs of the 20 best scored models are shown in the table.

Number of NOE restraints (percentage of all NOEs)	sPRE data used	Average RMSD of best scored models [Å]				
		Random set 1	Random set 2	Random set 3	Random set 4	Random set 5
43 (5%)	no	11.31	11.44	10.77		
43 (5%)	yes	8.61	6.98	7.79		
86 (10%)	no	10.24	10.13	10.71	11.30	9.92
86 (10%)	yes	6.44	5.45	5.21	5.69	5.96
172 (20%)	no	7.90	9.06	9.64		
172 (20%)	yes	5.27	5.35	6.11		
258 (30%)	no	6.54	7.80	9.56		
258 (30%)	yes	5.96	3.51	4.01		
345 (40%)	no	4.49	4.56	5.13		
345 (40%)	yes	3.44	2.96	4.71		
431 (50%)	no	3.55	4.55	3.77		
431 (50%)	yes	2.44	3.46	2.61		
647 (75%)	no	3.39	3.40	2.84		
647 (75%)	yes	2.33	2.78	2.81		
863 (100%)	no	2.30				
863 (100%)	yes	2.18				

Supplementary Table 6 | sPRE data is an orthogonal restraint that can be combined with classical NMR restraints. Structural models of the UUCG tetraloop were computed in the absence and presence of the sPRE potential and in combination with different experimental NMR restraints obtained from PDB entry 2KOC. For every restraint set, 200 models were computed and the best 20 models based on the total energy were selected. RMSD values are computed by comparing all carbon, nitrogen and phosphorus atoms to the published NMR structure and the average RMSDs of the 20 best scored models are shown in the table.

Restraints used	sPRE data used	Average RMSD of best scored models [Å]
Force field only	no sPRE	12.49
	with sPRE data	8.69
Hydrogen bonds	no sPRE	7.77
	with sPRE data	2.83
Hydrogen bonds + RDCs	no sPRE	6.90
	with sPRE data	2.62
Hydrogen bonds + Torsion angles	no sPRE	1.75
	with sPRE data	1.28
Hydrogen bonds + all NOEs (ambig. and unambig.)	no sPRE	1.06
	with sPRE data	1.05
Hydrogen bonds + Torsion angles + RDCs + all NOEs (ambig. and unambig.)	no sPRE	0.76
	with sPRE data	0.79

Supplementary Table 6 | Sparse sPRE data sets improve structure determination of the UUCG tetraloop. Structural models of the UUCG tetraloop were computed using hydrogen bonds-derived restraints in combination with different sets of synthetic sPREs obtained from PDB entry 2KOC using the back-calculation as described in methods. The size of the sPRE restraints set was varied and for every size, 5 randomly created sets were generated. For every restraint set, 200 models were computed in the presence of the sPRE potential and the best 20 models based on the total energy were selected. RMSD values are computed by comparing all carbon, nitrogen and phosphorus atoms to the published NMR structure and the average RMSDs of the 20 best scored models are shown in the table. Note that the result obtained for 97 experimental sPREs (Supplementary Table 6, RMSD = 2.83 Å) and 101 synthetic sPREs (this table, 2.75 Å) is comparable indicating that our approach performs equally well with experimental and synthetic sPRE data.

Percentage of sPRE restraints [%]	Number of sPRE restraints	Average RMSD of best scored models [Å]				
		Random set 1	Random set 2	Random set 3	Random set 4	Random set 5
6.25	6	7.50	7.87	7.03	7.03	7.55
12.5	13	6.46	7.64	6.64	7.46	6.19
25	25	7.93	4.93	4.30	7.54	6.68
50	50	3.58	3.41	3.92	3.99	3.09
75	76	3.38	2.87	3.34	3.07	3.14
100	101	2.75				

XplorNIH Protocols

UUCG tetraloop

```
# protocol is based on gb1_rdc example

xplor.requireVersion("2.24")

xplor.parseArguments()

import os
import sys, traceback
import protocol
output_folder="output"
outFilename = os.path.join(output_folder, "SCRIPT_STRUCTURE.sa")

numberOfStructures=200
protocol.initRandomSeed()
command = xplor.command

# load RNA topology
protocol.topology['nucleic'] = "nucleic-3.1_GTP_AP7.top"
protocol.parameters['nucleic'] = "nucleic-3.1_GTP_AP7.par"

from psfGen import seqToPSF
xplor.command('''
 topology
    @TOPPAR:nucleic-3.1_GTP_AP7.top
 end
 parameter
    @TOPPAR:nucleic-3.1_GTP_AP7.par
 end

 segment                                  {*Generate protein              *}
   name="    "                            {*This name has to match the    *}
                                          {*four characters in columns 73 *}
                                          {*through 76 in the coordinate  *}
                                          {*file, in XPLOR this name is   *}
                                          {*name is referred to as SEGId. *}
   chain
     @TOPPAR:toph11.nuc                   {*Read peptide bond file        *}
     sequence ''' +
# sequence file for RNA
open("UUCG.seq").read() +
     ''' end           {*interpret sequence file to  *}
   end                                    {*obtain the sequence           *}
 end
''')


protocol.genExtendedStructure()

from potList import PotList
potList = PotList()

from simulationTools import MultRamp, StaticRamp, InitialParams

rampedParams=[]
highTempParams=[]

from posDiffPotTools import create_PosDiffPot
refRMSD = create_PosDiffPot("refRMSD",
                            "(name P* or name N* or name C*) and not (resi 1 and name P*)",
                            pdbFile='2KOC__model1.pdb' )


noe=PotList('noe')
potList.append(noe)
from noePotTools import create_NOEPot
for (name,scale,file) in [('noes',1,"noe_2koc.tbl"),
                          ('hbonds',1,"hbond_2koc.tbl")
                          ]:
    pot = create_NOEPot(name,file)
    pot.setScale(scale)
    noe.append(pot)
rampedParams.append( MultRamp(2,30, "noe.setScale( VALUE )") )


# create Potential for sPRE
import nbTargetPotTools
spre_pot = nbTargetPotTools.create_NBTargetPot("spre", restraints=open("UUCG_sPRE.tbl").read() , selection='all', restraintFormat="xplor")
spre_pot.setCutoffDist(20)
spre_pot.setAveType("center")     # sum or center
spre_pot.setPotType("correlation")    # rmsd or correlation
spre_pot.setAveExp(6)          # 6 is the default anyway
spre_pot.setInvPow(2)          # 2 is the default

# set intercept and slope obtained from calibrate function
spre_pot.setIntercept(-0.257261)
spre_pot.setSlope(2.25292)
potList.append(spre_pot)
rampedParams.append( MultRamp(1000,1000, "spre_pot.setScale( VALUE )") )
```

```
from varTensorTools import create_VarTensor
media={}
for (medium,Da,Rh) in [ ('medium1',   7.5, 0.5) ]:
    oTensor = create_VarTensor(medium)
    oTensor.setDa(Da)
    oTensor.setRh(Rh)
    oTensor.setFreedom("varyDa, varyRh")
    media[medium] = oTensor
    pass

from rdcPotTools import create_RDCPot, scale_toCH
rdcs = PotList('rdc')
for (medium,expt,file,weight) in     [
     ('medium1','CH','rdc-CH_2koc.tbl' ,1)
     ]:
    rdc = create_RDCPot("%s_%s"%(medium,expt),file,media[medium])

    rdc.setScale(weight)
    if expt != 'CH':
        scale_toCH(rdc)
    rdcs.append(rdc)
    pass
potList.append(rdcs)
rampedParams.append( MultRamp(0.01,1.0, "rdcs.setScale( VALUE )") )


from xplorPot import XplorPot
xplor.command("@plane_2koc.tbl")
potList.append(XplorPot("plan",xplor.simulation))


dihedralRestraintFilename="torsion_2koc.tbl"
protocol.initDihedrals(dihedralRestraintFilename)
potList.append( XplorPot('CDIH') )
highTempParams.append( StaticRamp("potList['CDIH'].setScale(10)") )
rampedParams.append( StaticRamp("potList['CDIH'].setScale(200)") )
potList['CDIH'].setThreshold( 5 )


# use new torsion potential RNA-ff1
#   Bermejo, G. A.; Clore, G. M.; Schwieters, C. D. Structure 2016, 24, 806
import torsionDBPotTools
torsiondb = torsionDBPotTools.create_TorsionDBPot(name='torsiondb',database='rna09_v0.dat')
potList.append(torsiondb)
rampedParams.append(MultRamp(0.1, 1, "torsiondb.setScale(VALUE)"))


potList.append( XplorPot('VDW') )
rampedParams.append( StaticRamp("protocol.initNBond()") )
rampedParams.append( MultRamp(1.0,0.9,
                              "command('param nbonds repel VALUE end end')") )
rampedParams.append( MultRamp(.004,4,
                              "command('param nbonds rcon VALUE end end')") )
highTempParams.append( StaticRamp("""protocol.initNBond(cutnb=100,
                                                        rcon=0.004,
                                                        tolerance=45,
                                                        repel=1.2,
                                                        selStr="name C1'")""") )
potList.append( XplorPot("BOND") )
potList.append( XplorPot("ANGL") )
potList['ANGL'].setThreshold( 5 )
rampedParams.append( MultRamp(0.4,1,"potList['ANGL'].setScale(VALUE)") )
potList.append( XplorPot("IMPR") )
potList['IMPR'].setThreshold( 5 )
rampedParams.append( MultRamp(0.1,1,"potList['IMPR'].setScale(VALUE)") )

protocol.massSetup()

from ivm import IVM
dyn = IVM()

protocol.torsionTopology(dyn)

minc = IVM()
protocol.initMinimize(minc)

protocol.cartesianTopology(minc)

from simulationTools import AnnealIVM
init_t  = 3500.
cool = AnnealIVM(initTemp =init_t,
                 finalTemp=25,
                 tempStep =12.5,
                 ivm=dyn,
                 rampedParams = rampedParams)

scale_anneal_time = 10

def calcOneStructure(loopInfo):
    """ this function calculates a single structure, performs analysis on the
    structure, and then writes out a pdb file, with remarks.
    """
    from monteCarlo import randomizeTorsions
```

```
    randomizeTorsions(dyn)
    try:
        protocol.fixupCovalentGeom(maxIters=100, useVDW=1)
    except protocol.CovalentViolation:
        pass


    protocol.writePDB(loopInfo.filename()+".init")

    InitialParams( rampedParams )

    InitialParams( highTempParams )

    protocol.initDynamics(dyn,
                          potList=potList,
                          bathTemp=init_t,
                          initVelocities=1,
                          finalTime=100*scale_anneal_time,
                          numSteps=1000*scale_anneal_time,
                          printInterval=100)

    dyn.setETolerance( init_t/100 )
    dyn.run()

    InitialParams( rampedParams )

    protocol.initDynamics(dyn,
                          potList=potList,
                          numSteps=200*scale_anneal_time,
                          finalTime=.4*scale_anneal_time,
                          printInterval=100)

    cool.run()

    protocol.initMinimize(dyn, printInterval=50)
    dyn.run()

    protocol.initMinimize(minc,
                          potList=potList,
                          dEPred=10)
    minc.run()


from simulationTools import StructureLoop, FinalParams
StructureLoop(numStructures=numberOfStructures,
              doWriteStructures=True,
              pdbTemplate=outFilename,
              structLoopAction=calcOneStructure,
              genViolationStats=True,
              averageTopFraction=0.5,
              averageSortPots=[potList['BOND'],potList['ANGL'],potList['IMPR'],noe,rdcs],
              averageContext=FinalParams(rampedParams),
              averageCrossTerms=refRMSD,
              averageFilename="SCRIPT_ave.pdb",
              averagePotList=potList).run()
```

GTP-bound aptamer

```
# protocol is based on gb1_rdc example

xplor.requireVersion("2.24")

xplor.parseArguments()

import os
import sys, traceback
import protocol
output_folder="output"
outFilename = os.path.join(output_folder, "SCRIPT_STRUCTURE.sa")

numberOfStructures=200
protocol.initRandomSeed()   #set random seed - by time
command = xplor.command

# load RNA topology
protocol.topology['nucleic'] = "nucleic-3.1_GTP_AP7.top"
protocol.parameters['nucleic'] = "nucleic-3.1_GTP_AP7.par"

from psfGen import seqToPSF
xplor.command('''
 topology
    @TOPPAR:nucleic-3.1_GTP_AP7.top
 end
 parameter
    @TOPPAR:nucleic-3.1_GTP_AP7.par
 end

 segment                                  {*Generate protein              *}
   name="    "                            {*This name has to match the    *}
                                          {*four characters in columns 73 *}
                                          {*through 76 in the coordinate  *}
                                          {*file, in XPLOR this name is   *}
                                          {*name is referred to as SEGId. *}
   chain
     @TOPPAR:toph11.nuc                   {*Read peptide bond file        *}
     sequence ''' +
# sequence file for RNA
open("GTP_aptamer.seq").read() +
     ''' end           {*interpret sequence file to  *}
   end                                    {*obtain the sequence           *}
 end
segment
   name="    "                            {*This name has to match the    *}
                                          {*four characters in columns 73 *}
                                          {*through 76 in the coordinate  *}
                                          {*file, in XPLOR this name is   *}
                                          {*name is referred to as SEGId. *}

   number=99                              {*Residue number                *}


   chain
      sequence GTP end
   end
end
''')



protocol.genExtendedStructure()

from potList import PotList
potList = PotList()

from simulationTools import MultRamp, StaticRamp, InitialParams

rampedParams=[]
highTempParams=[]

from posDiffPotTools import create_PosDiffPot
refRMSD = create_PosDiffPot("refRMSD",
                            "(name P* or name N* or name C*) and not (resi 1 and name P*)",
                            pdbFile='Reference_aptamer__model1.pdb' )


noe=PotList('noe')
potList.append(noe)
from noePotTools import create_NOEPot
for (name,scale,file) in [('noes',1,"noe_list_xplor.tbl"),
                          ('hbonds',1,"hbond_list_xplor.tbl"),
                          ('lowerlimits',1,"noe_lol_list_xplor.tbl")
                          ]:
    pot = create_NOEPot(name,file)
    pot.setScale(scale)
    noe.append(pot)
rampedParams.append( MultRamp(2,30, "noe.setScale( VALUE )") )


# create Potential for sPRE
import nbTargetPotTools
```

```
spre_pot = nbTargetPotTools.create_NBTargetPot("spre", restraints=open("GTPaptamer_sPRE.tbl").read() , selection='all',
restraintFormat="xplor")
spre_pot.setCutoffDist(20)
spre_pot.setAveType("center")     # sum or center
spre_pot.setPotType("correlation")    # rmsd or correlation
spre_pot.setAveExp(6)          # 6 is the default anyway
spre_pot.setInvPow(2)          # 2 is the default

# set intercept and slope obtained from calibrate function
spre_pot.setIntercept(-0.157217)
spre_pot.setSlope(3.25629)
potList.append(spre_pot)
rampedParams.append( MultRamp(3000,3000, "spre_pot.setScale( VALUE )") )



from xplorPot import XplorPot
xplor.command("@plane.inp")
potList.append(XplorPot("plan",xplor.simulation))


# use new torsion potential RNA-ff1
#   Bermejo, G. A.; Clore, G. M.; Schwieters, C. D. Structure 2016, 24, 806
import torsionDBPotTools
torsiondb = torsionDBPotTools.create_TorsionDBPot(name='torsiondb',database='rna09_v0.dat')
potList.append(torsiondb)
rampedParams.append(MultRamp(0.3, 0.3, "torsiondb.setScale(VALUE)"))


potList.append( XplorPot('VDW') )
rampedParams.append( StaticRamp("protocol.initNBond()") )
rampedParams.append( MultRamp(1.0,0.9,
                              "command('param nbonds repel VALUE end end')") )
rampedParams.append( MultRamp(.004,4,
                              "command('param nbonds rcon VALUE end end')") )
# nonbonded interaction only between CA atoms
highTempParams.append( StaticRamp("""protocol.initNBond(cutnb=100,
                                                        rcon=0.004,
                                                        tolerance=45,
                                                        repel=1.2,
                                                        selStr="name C1'")""") )


potList.append( XplorPot("BOND") )
potList.append( XplorPot("ANGL") )
potList['ANGL'].setThreshold( 5 )
rampedParams.append( MultRamp(0.4,1,"potList['ANGL'].setScale(VALUE)") )
potList.append( XplorPot("IMPR") )
potList['IMPR'].setThreshold( 5 )
rampedParams.append( MultRamp(0.1,1,"potList['IMPR'].setScale(VALUE)") )

protocol.massSetup()

from ivm import IVM
dyn = IVM()

protocol.torsionTopology(dyn)

minc = IVM()
protocol.initMinimize(minc)

protocol.cartesianTopology(minc)

from simulationTools import AnnealIVM
init_t  = 3500.
cool = AnnealIVM(initTemp =init_t,
                 finalTemp=25,
                 tempStep =12.5,
                 ivm=dyn,
                 rampedParams = rampedParams)

scale_anneal_time = 10

def calcOneStructure(loopInfo):
    """ this function calculates a single structure, performs analysis on the
    structure, and then writes out a pdb file, with remarks.
    """
    from monteCarlo import randomizeTorsions
    randomizeTorsions(dyn)
    try:
        protocol.fixupCovalentGeom(maxIters=100, useVDW=1)
    except protocol.CovalentViolation:
        pass


    protocol.writePDB(loopInfo.filename()+".init")

    InitialParams( rampedParams )

    InitialParams( highTempParams )

    protocol.initDynamics(dyn,
                          potList=potList,
                          bathTemp=init_t,
                          initVelocities=1,
```

```
                            finalTime=100*scale_anneal_time,
                            numSteps=1000*scale_anneal_time,
                            printInterval=100)

    dyn.setETolerance( init_t/100 )
    dyn.run()

    InitialParams( rampedParams )

    protocol.initDynamics(dyn,
                            potList=potList,
                            numSteps=200*scale_anneal_time,
                            finalTime=.4*scale_anneal_time,
                            printInterval=100)

    cool.run()

    protocol.initMinimize(dyn, printInterval=50)
    dyn.run()

    protocol.initMinimize(minc,
                            potList=potList,
                            dEPred=10)
    minc.run()


from simulationTools import StructureLoop, FinalParams
StructureLoop(numStructures=numberOfStructures,
              doWriteStructures=True,
              pdbTemplate=outFilename,
              structLoopAction=calcOneStructure,
              genViolationStats=True,
              averageTopFraction=0.5,
              averageSortPots=[potList['BOND'],potList['ANGL'],potList['IMPR'],noe],
              averageContext=FinalParams(rampedParams),
              averageCrossTerms=refRMSD,
              averageFilename="SCRIPT_ave.pdb",
              averagePotList=potList).run()
```

References

1 Furtig, B., Richter, C., Bermel, W. & Schwalbe, H. New NMR experiments for RNA nucleobase resonance assignment and chemical shift analysis of an RNA UUCG tetraloop. *J Biomol NMR* **28**, 69-79, doi:10.1023/B:JNMR.0000012863.63522.1f (2004).

2 Wolter, A. C. *et al.* NMR resonance assignments for the class II GTP binding RNA aptamer in complex with GTP. *Biomol NMR Assign* **10**, 101-105, doi:10.1007/s12104-015-9646-7 (2016).

Journal of Biomolecular NMR (2019) 73:305–317
https://doi.org/10.1007/s10858-019-00248-2

ARTICLE

NMR characterization of solvent accessibility and transient structure in intrinsically disordered proteins

Christoph Hartlmüller[1] · Emil Spreitzer[2] · Christoph Göbl[3] · Fabio Falsone[4] · Tobias Madl[2,5]

Received: 3 February 2019 / Accepted: 11 April 2019 / Published online: 11 July 2019
© The Author(s) 2019

Abstract

In order to understand the conformational behavior of intrinsically disordered proteins (IDPs) and their biological interaction networks, the detection of residual structure and long-range interactions is required. However, the large number of degrees of conformational freedom of disordered proteins require the integration of extensive sets of experimental data, which are difficult to obtain. Here, we provide a straightforward approach for the detection of residual structure and long-range interactions in IDPs under near-native conditions using solvent paramagnetic relaxation enhancement (sPRE). Our data indicate that for the general case of an unfolded chain, with a local flexibility described by the overwhelming majority of available combinations, sPREs of non-exchangeable protons can be accurately predicted through an ensemble-based fragment approach. We show for the disordered protein α-synuclein and disordered regions of the proteins FOXO4 and p53 that deviation from random coil behavior can be interpreted in terms of intrinsic propensity to populate local structure in interaction sites of these proteins and to adopt transient long-range structure. The presented modification-free approach promises to be applicable to study conformational dynamics of IDPs and other dynamic biomolecules in an integrative approach.

Keywords Solvent paramagnetic relaxation enhancement · Intrinsically disordered proteins · Residual structure · p53 · FOXO4 · α-Synuclein

Introduction

The well-established structure–function paradigm has been challenged by the discovery of intrinsically disordered proteins (IDPs) (Dyson and Wright 2005). It is suggested that about 40% of all proteins have disordered regions of 40 or more residues, with many proteins existing solely in the unfolded state (Tompa 2012; Romero et al. 1998). Although they lack stable secondary or tertiary structure elements, this large class of proteins plays a crucial role in various cellular processes (Theillet et al. 2014; Wright and Dyson 2015; van der Lee et al. 2014; Uversky et al. 2014). Disorder serves a biological role, where conformational heterogeneity granted by disordered regions enables proteins to exert diverse functions in response to various stimuli. Unlike structured proteins, which are essential for catalysis and transport, disordered proteins are crucial for regulation and signaling. Due to their intrinsic flexibility they can act as network hubs interacting with a wide range of biomolecules forming dynamic regulatory networks (Dyson and Wright 2005; Tompa 2012; Babu et al. 2011; Flock et al. 2014; Wright and Dyson 1999; Uversky 2011; Habchi et al.

Christoph Hartlmüller and Emil Spreitzer are the shared first authors.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10858-019-00248-2) contains supplementary material, which is available to authorized users.

✉ Tobias Madl
tobias.madl@medunigraz.at

1 Center for Integrated Protein Science Munich (CIPSM) at the Department of Chemistry, Technische Universität München, Lichtenbergstrasse 4, 87548 Garching, Germany

2 Gottfried Schatz Research Center for Cell Signaling, Metabolism and Aging, Institute of Molecular Biology & Biochemistry, Medical University of Graz, Neue Stiftingtalstrasse 6, 8010 Graz, Austria

3 The Campbell Family Institute for Breast Cancer Research at Princess Margaret Cancer Centre, 610 University Avenue, Toronto, ON M5G 2M9, Canada

4 Institute of Pharmaceutical Sciences, University of Graz, Schubertstrasse 1, 8010 Graz, Austria

5 BioTechMed-Graz, Graz, Austria

2014). Given the plethora of potential interaction partners, it is not surprising that the interaction of IDPs with binding partners are often tightly regulated via and intricate 'code' of post-translational modifications, including phosphorylation, methylation, acetylation, and various others (Wright and Dyson 2015; Bah and Forman-Kay 2016). These proteins, and distortions in their interaction networks, for example by mutations and aberrant post-translational modifications (PTMs), are closely linked to a range of human diseases, including cancers, neurodegeneration, cardiovascular disorders and diabetes, they are currently considered difficult to study (Dyson and Wright 2005; Tompa 2012; Babu et al. 2011; Habchi et al. 2014; Metallo 2010; Uversky et al. 2008; Dyson and Wright 2004). Complications arise from the following factors: these proteins lack well-defined stable structure, they exist in a dynamic equilibrium of distinct conformational states, and the number of experimental techniques and observables renders IDP conformational characterization underdetermined (Mittag and Forman-Kay 2007; Eliezer 2009). Thus, an integration of new sets of experimental and analytical techniques are required to characterize the conformational behavior of IDPs.

Although IDPs are highly dynamic, they often contain transiently-folded regions, such as transiently populated secondary or tertiary structure, transient long-range interactions or transient aggregation (Marsh et al. 2007; Shortle and Ackerman 2001; Bernado et al. 2005; Mukrasch et al. 2007; Wells et al. 2008). These transiently-structured regions are of particular interest to study the biological function of IDPs as they can report on biologically-relevant interactions and encode biological function. Examples are aggregation, liquid–liquid phase separation, binding to folded co-factors, or modifying enzymes (Yuwen et al. 2018; Brady et al. 2017; Choy et al. 2012; Maji et al. 2009; Putker et al. 2013).

NMR spectroscopy is exceptionally well-suited to study IDPs, and in particular to detect transiently folded regions (Meier et al. 2008; Wright and Dyson 2009; Jensen et al. 2009). Several NMR observables provide atomic resolution, and ensemble-averaged information reporting on the conformational energy landscape sampled by each amino acid, including chemical shifts, residual dipolar couplings (RDCs), and paramagnetic relaxation enhancement (PRE) (Dyson and Wright 2004; Eliezer 2009; Marsh et al. 2007; Shortle and Ackerman 2001; Meier et al. 2008; Gobl et al. 2014; Gillespie and Shortle 1997; Clore et al. 2007; Huang et al. 2014; Ozenne et al. 2012; Clore and Iwahara 2009; Otting 2010; Hass and Ubbink 2014; Gobl et al. 2016). RDCs, and PREs, either alone or in combination have been used successfully in recent years to characterize the conformations and long-range interactions of IDPs (Bernado et al. 2005; Ozenne et al. 2012; Dedmon et al. 2005; Bertoncini et al. 2005; Parigi et al. 2014; Rezaei-Ghaleh et al. 2018). However, both techniques rely on a modification of the IDP of interest, either by external alignment media in case of RDCs or the covalent incorporation of paramagnetic tags in the case of PREs.

We and others have proposed applications of soluble paramagnetic agents to obtain structural information by NMR without any modifications of the molecules of interest (Gobl et al. 2014; Guttler et al. 2010; Hartlmuller et al. 2016; Hocking et al. 2013; Madl et al. 2009, 2011; Respondek et al. 2007; Zangger et al. 2009; Pintacuda and Otting 2002; Bernini et al. 2009; Wang et al. 2012; Sun et al. 2011; Gong et al. 2017; Gu et al. 2014; Hartlmuller et al. 2017). The addition of soluble paramagnetic compounds leads to a concentration-dependent and therefore tunable increase of relaxation rates, the so-called paramagnetic relaxation enhancement (here denoted as solvent PRE, sPRE; also known as co-solute PRE, Fig. 1a). This effect depends on the distance of the spins of interest (e.g. ^{1}H, ^{13}C) to the biomolecular surface. The nuclei on the surface are affected the strongest by the sPRE effect, and this approach has been shown to correlate well with biomolecular structure in the case of proteins and RNA (Madl et al. 2009; Pintacuda and Otting 2002; Bernini et al. 2009; Hartlmuller et al. 2017). sPREs have gained popularity for structural studies of biomolecules, including in the structure determination of proteins (Madl et al. 2009; Wang et al. 2012), docking of protein complexes (Madl et al. 2011), and qualitative detection of dynamics (Hocking et al. 2013; Sun et al. 2011; Gong et al. 2017; Gu et al. 2014).

The most commonly used paramagnetic agent for measuring sPRE data is the inert complex Gd(DTPA-BMA) (gadolinium diethylenetriaminepenta-acetic acid bismethylamide, commercially available as 'Omniscan'), that is known to not specifically interact on protein surfaces (Guttler et al. 2010; Madl et al. 2009, 2011; Pintacuda and Otting 2002; Wang et al. 2012; Respondek et al. 2007; Zangger et al. 2009; Göbl et al. 2010). Previously, we and others could show that sPRE data provide in-depth structural and dynamic data for IDP analysis (Madl et al. 2009; Sun et al. 2011; Gong et al. 2017; Emmanouilidis et al. 2017; Johansson et al. 2014). For example, sPRE data helped to characterize α-helical propensity in a previously postulated flexible region in the folded 42 kDa maltodextrin binding protein (Madl et al. 2009), and dynamic ligand binding to the human "survival of motor neuron" protein (Emmanouilidis et al. 2017). While writing this manuscript, and based on sPRE data for exchangeable amide protons, the Tjandra lab has shown that sPREs can detect native-like structure in denatured ubiquitin (Kooshapur et al. 2018).

Here, we present an integrative ensemble approach to predict the sPREs of IDPs. This ensemble representation is used to calculate conformationally averaged sPREs, which fit remarkably well to the experimentally-measured sPREs. We show for the disordered protein α-synuclein,

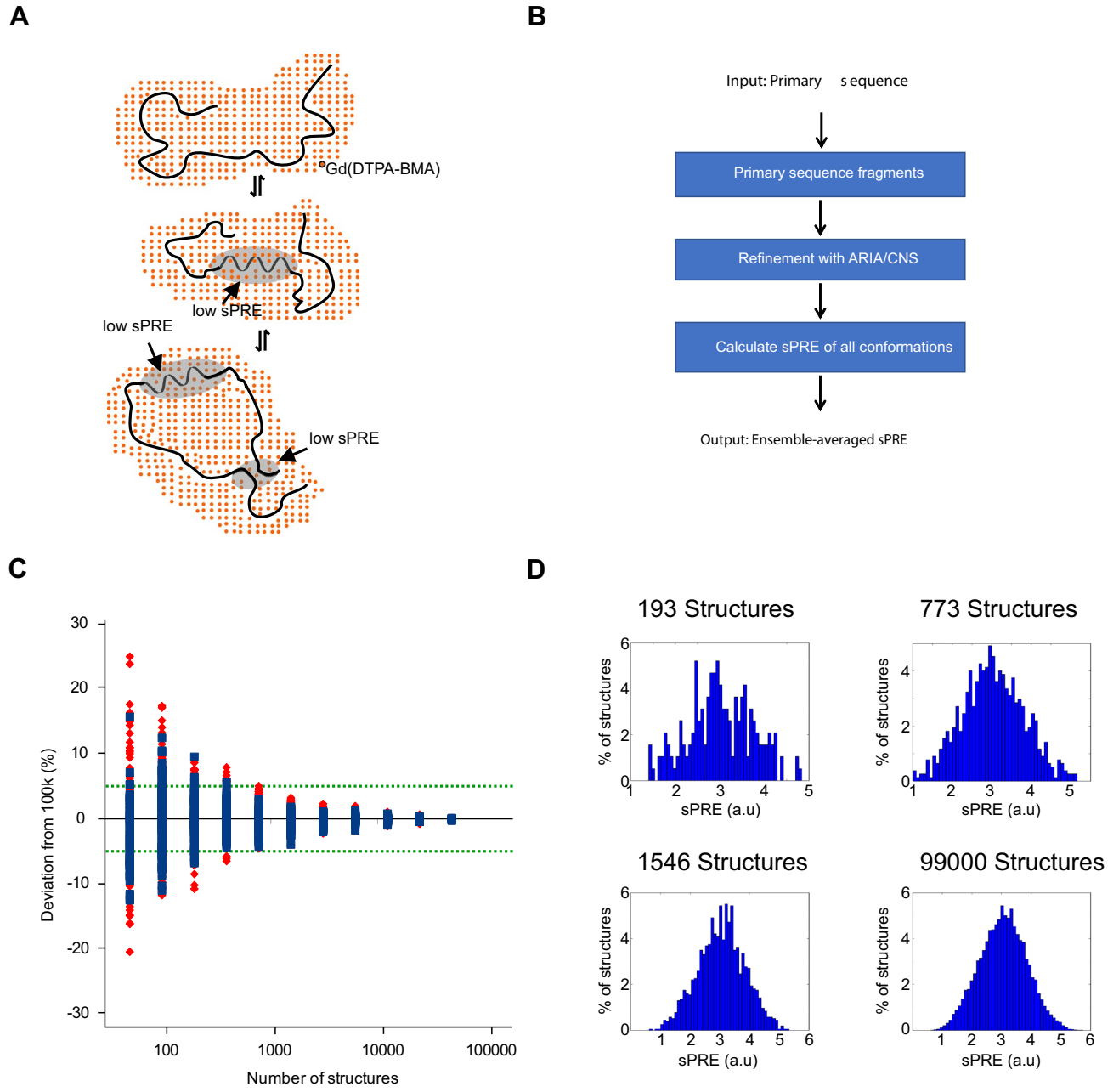


Fig. 1 Principle and workflow for solvent PRE. **a** Transient secondary structures of IDPs are characteristic for protein–protein interaction sites and are therefore crucial for various cellular functions. NMR sPRE data provide quantitative and residue specific information on the solvent accessibility as the effect of paramagnetic probes such as Gd(DTPA-BMA) is distance dependent, which can be used to detect secondary structures within otherwise unfolded regions and long-range contacts within a protein. **b** Prediction of sPRE is based on an ensemble approach of a library of peptides. Each peptide has a length of 5 residues, and is flanked by triple-Ala on both termini (e.g. AAAXXXXXAAA, where XXXXX is a 5-mer fragment of the target primary sequence). Following water refinement using ARIA/CNS, sPRE values of all conformations are calculated and the average solvent PRE value of the ensemble is returned. **c** Predicted C^{α} sPRE (blue) and standard deviation (red) of AAAVVAVVAAA ensembles consisting of 99,000 down to 48 structural conformations. The green-dotted line indicates 5% deviation from the ensemble with 99,000 conformations. **d** Histograms of different ensemble sizes showing the distribution of predicted sPRE values

and disordered regions of the proteins FOXO4 and p53 that deviation from random coil behavior can indicate intrinsic propensity to populate transient local structures and long-range interactions. In summary, this method provides a unique modification-free approach for studying IDPs, that is compatible with a wide range of NMR pulse sequences and biomolecules.

Materials and methods

Protein expression and purification

For expression of human FOXO4^{TAD} (residues 198–505), p53^{TAD} (residues 1–94), pETM11-His$_6$-ZZ cDNA and including an N-terminal TEV protease cleavage site coding for the respective proteins were transformed into *E. coli* BL21-DE3. To obtain $^{13}C/^{15}N$ isotope labeled protein, cells were grown for 1 day at 37 °C in minimal medium (100 mM KH_2PO_4, 50 mM K_2HPO_4, 60 mM Na_2HPO_4, 14 mM K_2SO_4, 5 mM $MgCl_2$; pH 7.2 adjusted with HCl and NaOH with 0.1 dilution of trace element solution (41 mM $CaCl_2$, 22 mM $FeSO_4$, 6 mM $MnCl_2$, 3 mM $CoCl_1$, 2 mM $ZnSO_4$, 0.1 mM $CuCl_2$, 0.2 mM $(NH_4)_6Mo_7O_{17}$, 24 mM EDTA) supplemented with 2 g of $^{13}C_6H_{12}O_6$ (Cambridge isotope) and 1 g of $^{15}NH_4Cl$ (Sigma). At an OD (600 nm) of 0.8, cells were induced with 0.5 mM isopropyl-β-D-thiogalactopyranosid (IPTG) for 16 h at 20 °C. Cell pellets were harvested and sonicated in denaturing buffer containing 50 mM Tris–HCl pH 7.5, 150 mM NaCl, 20 mM imidazole, 2 mM tris(2-carboxyethyl)phosphine (TCEP), 20% glycerol and 6 M urea. His$_6$-ZZ proteins were purified using Ni–NTA agarose (QIAGEN) and eluted in 50 mM Tris–HCl pH 7.5, 150 mM NaCl, 200 mM imidazole, 2 mM TCEP and subjected to TEV protease cleavage. Untagged proteins were then isolated by performing a second affinity purification step using Ni–NTA beads for removal of TEV and uncleaved substrate. A final exclusion chromatography purification step was performed in the buffer of interest on a gel filtration column (Superdex peptide (10/300) for p53 and Superdex 75 (16/600) for FOXO4, GE Healthcare).

α-Synuclein was expressed and purified as described (Falsone et al. 2011). Briefly, pRSETB vector containing the human AS gene was transformed into BL21 (DE3) Star Cells. $^{13}C/^{15}N$-labeled α-synuclein was expressed in minimal medium (6.8 g/l Na_2HPO_3, 4 g/l KH_2PO_4, 0.5 g/l NaCl, 1.5 g/l $(15NH_4)_2SO_2$, 4 g/l ^{13}C glucose, 1 µg/l biotin, 1 µg/l thiamin, 100 µg/ml ampicillin, and 1 ml 1000 × microsalts). Cells were grown to an OD (600 nm) of 0.7. Protein was expressed by addition of 1 mM IPTG for 4 h. After harvesting cells were resuspended in 20 mM Tris–HCl, 50 mM NaCl, pH 7.4, supplemented with a Complete® protease inhibitor mix (Roche, Basel, Switzerland). Protein purification was then achieved using a Resource Q column and gel filtration using a Superdex 75 gel filtration column (GE Healthcare, Uppsala, Sweden).

Generation of conformational ensembles

Conformational ensembles were generated using the ARIA/CNS software-package, comprising of 1500 random backbone conformations of all possible 5-mer peptides of the protein of interest, and flanked by triple-alanine. Every backbone conformation served as starting structure in a full-atom water refinement using ARIA (Bardiaux et al. 2012). For every refined structure the solvent PRE is calculated and the averaged solvent PRE of the central residue is stored in the database. To predict sPRE data, a previously published grid-based approach was used (Hartlmuller et al. 2016; Pintacuda and Otting 2002). Briefly, the structural model was placed in a regularly-spaced grid representing the uniformly distributed paramagnetic compound and the grid was built with a point-to-point distance of 0.5 Å and a minimum distance of 10 Å between the protein model and the outer border of the grid. Next, grid points that overlap with the protein model were removed assuming a molecular radius of 3.5 Å for the paramagnetic compound. To compute the sPRE for a given protein proton $sPRE^i_{predicted}$, the distance-dependent paramagnetic effect (Hartlmuller et al. 2016; Hocking et al. 2013; Pintacuda and Otting 2002) was numerically integrated over all remaining grid points according to Eq. (1):

$$sPRE^i_{predicted} = c \cdot \sum_{d_{i,j}<10\text{Å}} \frac{1}{d^6_{i,j}} \quad (1)$$

where i is the index of a proton of the protein, j is the index of the grid point, $d_{i,j}$ is the distance between the ith proton and the jth grid point and c is an arbitrary constant to scale the sPRE values (1000). Theoretical sPRE values were normalized by calculating the linear fit of experimental and predicted sPRE followed by shifting and scaling of the theoretical sPRE. To predict the solvent PRE of the entire IDP sequence, each peptide with the five matching amino acids is searched and the corresponding solvent PRE values are combined. sPRE data of the two N- and C-terminal residues were not predicted in this setup. All scripts and sample runs can be obtained downloaded from the homepage of the authors (https://mbbc.medunigraz.at/forschung/forschungseinheiten-und-gruppen/forschungsgruppe-tobias-madl/software/).

NMR experiments

The setup of sPRE measurements using NMR spectroscopy was performed as described previously (Hartlmuller et al. 2016, 2017). To obtain sPRE data, a saturation-based approach was used. The ^1H-R_1 relaxation rates were determined by a saturation-recovery scheme followed by a

read-out experiment such as a ^{1}H, ^{15}N HSQC, ^{1}H, ^{13}C HSQC or a 3D CBCA(CO)NH experiment. The read-out experiments were combined with the saturation-recovery scheme in a Pseudo-3D (HSQCs) or Pseudo-4D [CBCA(CO)NH] experiment, with the recovery time as an additional dimension. The CBCA(CO)NH was recorded using non-uniform sampling. Alternatively, ^{1}H-R_2 relaxation rates can be as described (Clore and Iwahara 2009).

A 7.5 ms ^{1}H trim pulse followed by a gradient was applied for proton saturation. During the recovery, ranging from several milliseconds up to several seconds, z-magnetization is built up. The individual recovery delays are applied in an interleaved manner, with short and long delays occurring in alternating fashion. For every ^{1}H-R_1 measurement 10 delay times were recorded and for error estimation, at 1 delay time was recorded as a duplicate.

Measurement of ^{1}H-R_1 rates were repeated for increasing concentrations of the relaxation-enhancing Gd(DTPA-BMA)/Omniscan (GE Healthcare, Vienna, Austria) and the solvent PRE was obtained as the average change of the proton R_1 rate per concentration of the paramagnetic agent. After each addition of Gd(DTPA-BMA), the recovery delays were shortened such that for the longest delay all NMR signals were sufficiently recovered. The interscan delay was set to 50 ms, as the saturation-recovery scheme does not rely on an equilibrium z-magnetization at the start of each scan. All NMR samples contained 10% $^{2}H_2O$. Spectra were processed using NMRPipe and analyzed with the NMRView and CcpNmr Analysis software packages (Johnson 2004; Delaglio et al. 1995; Skinner et al. 2016).

Measurement of sPRE data used in this study

Assignment of p53^{TAD} was achieved using HNCACB, CBCA(CO)NH and HCAN spectra and analyzed using ccpNMR (Skinner et al. 2016). sPRE data of 300 µM samples of uniformly ^{13}C/^{15}N labeled p53^{TAD} was measured on a 600 MHz Bruker Avance Neo NMR spectrometer equipped with a TXI probehead at 298 K in a buffer containing 50 mM sodium phosphate buffer, 0.04% sodium azide, pH 7.5. ^{1}H-R_1 rates of $^{1}H^{N}$, H^{α} and H^{β} were determined using ^{1}H,^{13}C HSQC and ^{1}H, ^{15}N HSQC as read-out spectra (4/4 scans, 200/128 complex points in F2). For assignment of α-synuclein, previously reported chemical shifts were obtained from the BMRB (accession code 6968) and the assignment was confirmed using HNCACB and CBCA(CO)NH spectra. ^{1}H-R_1 rates of aliphatic protons and amide protons of a 100 µM sample (50 mM bis(2-hydroxyethyl)amino-tris(hydroxymethyl) methane (bis–Tris), 20 mM NaCl, 3 mM sodium azide, pH 6.8) were determined using ^{1}H, ^{13}C HSQC and ^{1}H, ^{15}N HSQC read-out spectra, respectively, at 282 K in the presence of 0, 1, 2, 3, 4 and 5 mM Gd(DTPA-BMA). For assignment of FOXO4^{TAD} HNCACB, CBCA(CO)NH and HCAN spectra were recorded and assigned using ccpNMR (Skinner et al. 2016). Measurements of ^{13}C, ^{15}N labeled FOXO4^{TAD} at 390 µM in 20 mM sodium phosphate buffer, pH 6.8, 50 mM NaCl, 1 mM DTT were performed on a 600 MHz magnet (Oxford Instruments) equipped with an AV III console and cryo TCI probe head (Bruker Biospin). Pseudo-4D CBCA(CO)NH spectra served as read-out for ^{1}H-R_1 rates and were recorded on a 250 µM sample on a 900 MHz Bruker Avance III spectrometer equipped with a TCI cryoprobe using non-uniform sampling (4 scans, 168/104 complex points in F1 (^{13}C)/F2 (^{15}N) sampled with 13.7% resulting a total number of 600 complex points). Spectra were processed using hmsIST/NMRPipe (Hyberts et al. 2014).

Analysis of NMR data

sPRE data of the model proteins was analyzed as described previously. Briefly, peak intensities were extracted using nmrglue python package and fitted to a mono-exponential build up curve the SciPy python package and Eq. (2).

$$I(t) = -A \cdot e^{-R_1 * t} + C \tag{2}$$

where $I(t)$ is the peak intensity of the saturation-recovery experiment, t is the recovery delay, A is the amplitude of the z-magnetization build-up, C is the plateau of the curve and R_1 is the longitudinal relaxation rate. Duplicate recovery delays were used to determine the error for the fitted rates R_1.

$$\epsilon_{\text{exp}} = \sqrt{\frac{1}{2N} \cdot \sum_{i=1}^{N} \delta_i} \tag{3}$$

where N is the number of peaks in the spectrum, i is the index of the peak, and δ_i is the difference of the duplicates for the ith peak. The error of the rates R_1 was then obtained using a Monte Carlo-type resampling strategy. The solvent PRE is obtained by performing a weighted linear regression using the equation

$$R_1(c) = sPRE \cdot c + R_1^0 \tag{4}$$

where c is the concentration of Gd(DTPA-BMA), $R_1(c)$ is the fitted R_1 rate at the present of Gd(DTPA-BMA) with a concentration c, R_1^0 is the R_1 in the absence of Gd(DTPA-BMA) and sPRE is the slope and the desired sPRE value. For the weighted linear regression, the previously determined errors ΔR_1 for R_1 was used, and the error on the concentration c was neglected.

Results and discussion

To detect transient structural elements in IDPs, an efficient back-calculation of sPREs of IDPs is essential. Whereas back-calculation of sPREs is relatively straightforward for folded rigid structures and can be carried out efficiently using a grid-based approach by integration of the solvent environment (Hartlmuller et al. 2016, 2017), this approach fails in the case of highly conformationally heterogeneous IDPs. In our approach, sPREs are best represented as an average sPRE of an ensemble. NMR observables and nuclear spin relaxation phenomena, including sPREs, directly sense chemical exchange through the distinct magnetic environments that nuclear spins experience while undergoing those exchange processes. The effects of the dynamic exchange on the NMR signals can be described by the McConnell Equations (Mcconnell 1958) In the case of a two-site exchange process, and assuming that the exchange rate is faster than the difference in the sPREs observed in both states, the observed sPRE is a linear, population-weighted average of the sPRE observed in both states, as seen for covalent paramagnetic labels (Clore and Iwahara 2009). Moreover, the correlation time for relaxation is assumed to be faster than the exchange time among different conformations within the IDP (Jensen et al. 2014; Iwahara and Clore 2010). The effective correlation time for longitudinal relaxation depends on the rotational correlation time of the biomolecule, the electron relaxation time and the lifetime of the rotationally correlated complex of the biomolecule and the paramagnetic agent (Madl et al. 2009; Eletsky et al. 2003). For ubiquitin, the effective correlation time for longitudinal relaxation was found to be in the sub-ns time scale (Pintacuda and Otting 2002), whereas that conformational exchange in IDPs typically appears at slower timescales (Jensen et al. 2014).

Calculating the average of sPREs over an ensemble of protein conformations presents serious practical difficulties that affect both the accuracy and the portability of the calculation. For RDCs it has been shown that convergence to the average requires an unmanageably large number of structures (e.g. 100,000 models for a protein with 100 amino acids), and that the convergence strictly depends on the length of the protein (Bernado et al. 2005; Nodet et al. 2009). To simplify the back-calculation of sPREs we use a strategy proposed for RDCs by the Forman-Kay and Blackledge groups (Marsh et al. 2008; Huang et al. 2013).

To back-calculate the sPRE from a given primary sequence of an IDP we generated fragments of five amino acids of the sequence of interest and flanked them with triple-alanine sequences at the N- and C-termini to simulate the presence of upstream/downstream amino acids (Fig. 1b). An ensemble of structures for these sequences is then generated using ARIA/CNS including water refinement (Bardiaux et al. 2012). To predict the solvent PRE of the entire IDP, the peptide with the five matching residues is searched and the corresponding solvent PREs averaged for the entire conformational ensemble are returned. This approach is highly parallelizable and dramatically reduces the computational effort compared to simulating the conformations of the full-length IDP.

To determine the number of conformers necessary to converge the back-calculated sPRE of the defined 11-mers, we generated an ensemble of 100,000 structures for a 11-mer AAAVVAVVAAA peptide using ARIA/CNS (Bardiaux et al. 2012) and back-calculated the sPRE for subsets with decreasing number of structures. We find that 1500 conformers are sufficient to reproduce the sPRE with a deviation compared to the maximum ensemble below 5% (Fig. 1c, d).

Back-calculation of the sPRE by fast grid-based integration has some advantages compared to alternative approaches relying on surface accessibility (Kooshapur et al. 2018). First, sPREs can be obtained for atoms without any surface accessibility in grid-based integration approaches as they still take into account the distance-dependent paramagnetic effect. This is expected to provide more accurate predictions for regions with a high degree of bulky side chains or transient folding.

To validate our computational approach, we recorded several sets of experimental ^{1}H-sPREs for the disordered regions of the human proteins FOXO4, p53, and α-synuclein. Similar to many other transcription factors, p53 and FOXO4 are largely disordered outside their DNA binding domains.

In order to demonstrate that surface accessibility data can be obtained for a challenging IDP, we recorded sPRE data for the 307 residue transactivation domain of FOXO4. The FOXO4 transcription factor is a member of the forkhead box O family of proteins that share a highly conserved DNA-binding motif, the forkhead box domain (FH). The FH domain is surrounded by large N- and C-terminal intrinsically disordered regions which are essential for the regulation of FOXO function (Weigel and Jackle 1990). FOXOs control a plethora of cellular functions, such as cell growth, survival, metabolism and oxidative stress, by regulating the expression of hundreds of target genes (Burgering and Medema 2003; Hornsveld et al. 2018). Expression and activity of FOXOs are tightly controlled by PTMs such as phosphorylation, acetylation, methylation and ubiquitination, and these modifications impact on FOXO stability, sub-cellular localization and transcriptional activity (Essers et al. 2004; de Keizer et al. 2010; van den Berg et al. 2013). Because of their anti-proliferative and pro-apoptotic functions, FOXOs have been considered as bona fide tumor suppressors. However, FOXOs can also support tumor development and progression by maintaining cellular homeostasis, facilitating metastasis and inducing therapeutic resistance (Hornsveld

et al. 2018). Thus, targeting FOXO activity might hold promise in cancer therapy.

The C-terminal FOXO4 transactivation domain has been suggested to be largely disordered and to be the binding site for many cofactors. Because it also harbors most of the post-translational modifications (Putker et al. 2013; Burgering and Medema 2003; Hornsveld et al. 2018; Bourgeois and Madl 2018), we set off to study this biologically important domain using our sPRE approach. ^{1}H,^{15}N and ^{1}H, ^{13}C HSQC NMR spectra of FOXO4^{TAD} are of high quality and showed no detectable ^{1}H, ^{13}C, or ^{15}N chemical shift changes between the spectra recorded in the absence or presence of Gd(DTPA-BMA) (Fig. 2a). sPRE data of FOXO4 had to be recorded in pseudo-4D saturation-recovery CBCA(CO)NH spectra due to the severe signal overlap observed in the 2D HSQC spectra. It should be noted that any kind of NMR experiment could be combined in principle with a sPRE saturation recovery measurement block in order to obtain ^{1}H- or ^{13}C sPRE data. The sPRE data of FOXO4^{TAD} yield differential solvent accessibilities in a residue-specific manner (Fig. 2b, c). H$^{\alpha}$ atoms located in regions rich in bulky residues are showing lower sPREs and H$^{\alpha}$ atoms located in more exposed glycine-rich regions display higher sPREs. H$^{\beta}$ sPRE data was obtained for a limited number of residues and shows overall elevated sPREs due to the higher degree of exposure and a reasonable agreement of predicted and experimental data (Supporting Fig. 1). A comparison of the predicted sPRE data with a bioinformatics bulkiness prediction shows that some features are reproduced by the bioinformatics prediction (Supporting Fig. 2A). However, the experimental sPRE is better described by our approach. Strikingly, the predicted sPRE pattern reproduces the

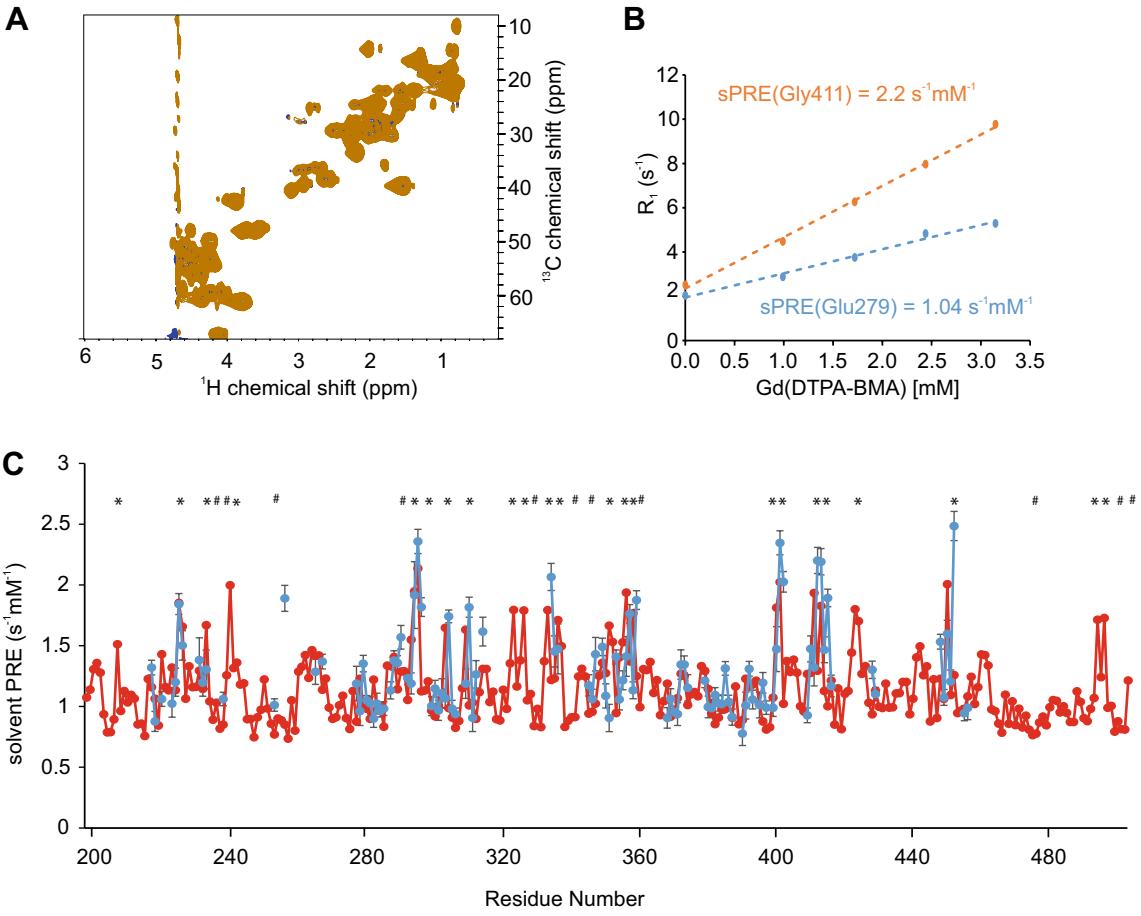


Fig. 2 Comparison of predicted and measured solvent PRE of FOXO4^{TAD}. **a** Overlay of ^{1}H,^{13}C HSQC spectra, with full recovery time of a 390 µM ^{13}C,^{15}N labeled FOXO4^{TAD} sample in the absence (blue) and presence of 3.25 mM Gd(DTPA-BMA) (orange). **b** ^{1}H-R$_1$ rates of two selected residues of FOXO4^{TAD} at different Gd(DTPA-BMA) concentrations. **c** Predicted (red) and experimentally-determined (blue) solvent PRE values using CBCA(CO)NH as readout spectrum, of assigned H$^{\alpha}$ peaks of FOXO4^{TAD}. Experimental sPRE values are calculated by fitting the data with a linear regression equation. Predicted sPRE values are based on the previously described ensemble approach. Residues with bulky side chains (Phe, Trp, Tyr) are labeled with #, and exposed glycine residues are labeled with * (see Supporting Fig. 2A for a bulkiness profile). Errors of the measured ^{1}H-R$_1$ rates were calculated using a Monte Carlo-type resampling strategy and are shown in the diagram as error bars

experimental sPRE pattern exceptionally well, indicating that the FOXO4^{TAD} is largely disordered and does not adopt any stable or transient tertiary structure in the regions for which sPRE data could be obtained.

In order to demonstrate that surface accessibility data can be obtained for a IDP with potential formation of transient local secondary structure we recorded sPRE data for the 94-residue transactivation domain of p53. p53 is a homotetrameric transcription factor composed of an N-terminal trans-activation domain, a proline-rich domain, a central DNA-binding domain followed by a tetramerization domain and the C-terminal negative regulatory domain. p53 is involved in the regulation of more than 500 target genes and thereby controls a broad range of cellular processes, including apoptosis, metabolic adaptation, DNA repair, cell cycle arrest, and senescence (Vousden and Prives 2009). The disordered N-terminal p53 transactivation domain (p53^{TAD}) is a key interaction motif for regulatory protein–protein interactions (Fernandez-Fernandez and Sot 2011): it possesses two binding motifs with α-helical propensity, named p53^{TAD1} (residues 17–29) and p53^{TAD2} (residues 40–57). These two motifs act independently or in combination in order to allow p53 to bind to several proteins regulating either p53 stability or transcriptional activity (Shan et al. 2012; Jenkins et al. 2009; Rowell et al. 2012). Because of its pro-apoptotic function, p53 is recognized as tumor suppressor, and is found mutated in more than half of all human cancers affecting a wide variety of tissues (Olivier et al. 2010). Within this biological and disease context the N-terminal p53-TAD plays a key role: it mediates the interaction with folded co-factors, and comprises most of the regulatory phosphorylation sites.

^{1}H, ^{15}N and ^{1}H, ^{13}C HSQC NMR spectra recorded of p53^{TAD} are of high quality and showed no detectable ^{1}H, ^{13}C, or ^{15}N chemical shift changes between the spectra recorded in the absence or presence of Gd(DTPA-BMA) (Fig. 3a, Supporting Fig. 3A). The sPRE data of p53^{TAD} display differential solvent accessibilities in a residue-specific manner: due to different excluded volumes for the paramagnetic agent H$^{\alpha}$ atoms located in regions rich in bulky residues show lower sPREs and H$^{\alpha}$ atoms located in more exposed regions show higher sPREs (Fig. 3b, c, Supporting Fig. 2B).

sPRE data of structured proteins are often recorded for amide protons. However, chemical exchange of the amide proton with fast-relaxing water solvent protons might lead to an increase of the experimental sPRE, as has been observed for the disordered linker regions in folded proteins and in RNA (Hartlmuller et al. 2017; Gobl et al. 2017). For imino and amino protons of the UUCG tetraloop RNA and a GTP class II aptamer, for example, the increase of ^{1}H-R$_1$ rates is larger at small concentrations of the paramagnetic compound, and becomes linear at higher concentrations. Thus, we decided to focus here on experimental and back-calculated sPRE data of H$^{\alpha}$ protons. Nevertheless, ^{1}H^{N}-sPREs are shown for comparison in the supporting information (Supporting Fig. 4A).

Comparison of the back-calculated and experimental p53^{TAD}-sPREs shows that several regions within p53^{TAD} yield lower sPREs than predicted, indicating that p53^{TAD} populates residual local structure or shows long-range tertiary interactions. In line with this, ^{15}N NMR relaxation data and ^{13}C secondary chemical shift data display reduced flexibility of p53^{TAD} and transient α-helical structure (Supporting Fig. 5). This is in line with previous studies which found that the p53^{TAD1} domain adopts a transiently populated α-helical structure formed by residues Phe19-Leu26 and that the p53^{TAD2} domain adopts a transiently populated turn-like structure formed by residues Met40-Met44 and Asp48-Trp53 (Lee et al. 2000). Given that p53^{TAD} has been reported to interact with several co-factors, our data indicate that sPRE data can indeed provide important insight into the residual structure of this key interaction motif (Bourgeois and Madl 2018; Raj and Attardi 2017).

In order to address the question of whether sPREs can be used to detect transient long-range interactions in disordered proteins we recorded ^{1}H sPRE data for the 141-residue IDP α-synuclein using ^{1}H, ^{13}C and ^{1}H, ^{15}N, HSQC-based saturation recovery experiments at increasing concentrations of Gd(DTPA-BMA). α-Synuclein controls the assembly of presynaptic vesicles in neurons and is required for the release of the neurotransmitter dopamine (Burre et al. 2010). The aggregation of α-synuclein into intracellular amyloid inclusions coincides with the death of dopaminergic neurons, and therefore constitutes a pathologic signature of synucleinopathies such as Parkinson's disease, dementia with Lewy bodies, and multiple system atrophy (Alafuzoff and Hartikainen 2017). Formation of transient long-range interactions has been proposed to protect α-synuclein from aggregation.

^{1}H, ^{15}N and ^{1}H, ^{13}C HSQC NMR spectra of α-synuclein are of high quality and showed no detectable ^{1}H, ^{13}C, or ^{15}N chemical shift changes between the spectra recorded in the absence or presence of 5 mM Gd(DTPA-BMA) (Fig. 4a). The sPRE data of α-synuclein display variable solvent accessibilities in a residue-specific manner (Fig. 4b), with Hα atoms located in regions rich in bulky residues showing lower sPREs and H$^{\alpha}$ atoms located in more exposed regions showing higher sPREs (see also Supporting Fig. 2C for a comparison with the bioinformatics bulkiness profile and Supporting Fig. 4B for the ^{1}H^{N} sPRE data). Thus, the sPRE value provides local structural information about the disordered ensemble. Strikingly, we observed decreased sPREs, and therefore lower surface accessibility, in several regions, such as between residues 15–20, 26–30, 52–57, 74–79, 87–92, 102–110, and 112–121, respectively (Fig. 4c). Comparison of these regions with recently–published ensemble modeling using extensive sets of RDC and PRE data (Salmon et al. 2010) shows that the previously–observed

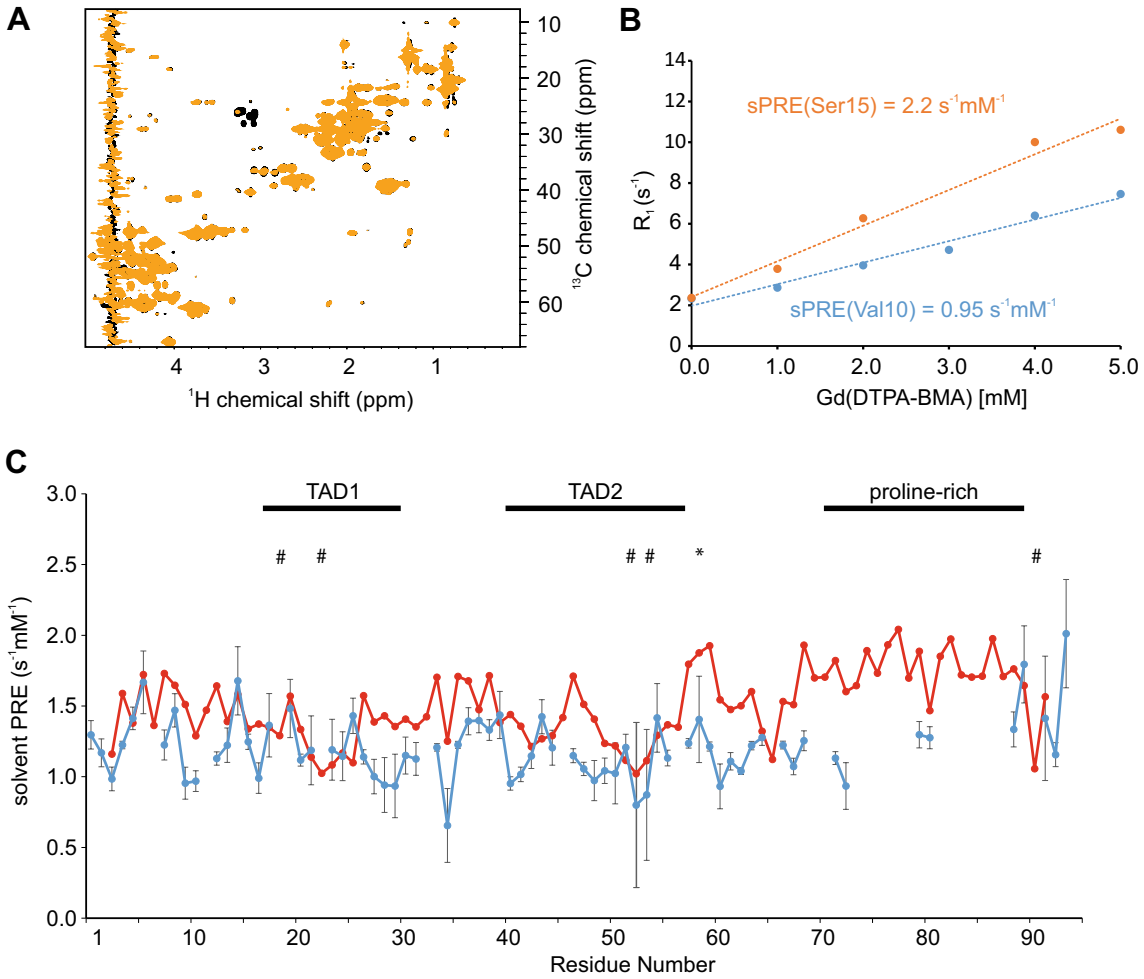


Fig. 3 Comparison of predicted and measured solvent PRE of p53^{TAD}. **a** Overlay of ^{1}H, ^{13}C HSQC read-out spectra, with full recovery time of a 300 µM ^{13}C, ^{15}N labeled p53^{TAD} in absence (black) and presence of 5 mM Gd(DTPA-BMA) (orange). **b** Gd(DTPA-BMA)-concentration-dependent R_1 rates of two selected residues. **c** Diagram showing predicted (red) and measured (blue) solvent PRE values of each H^{α} atom of p53^{TAD}. Experimental sPRE values are calculated by fitting the data with a linear regression equation. Predicted sPRE values are based on the previously described ensemble approach. Regions binding to co-factors (TAD1, TAD2) and the proline rich region are labeled. Residues with bulky side chains (Phe, Trp, Tyr) are labeled with #, and exposed glycine residues are labeled with * (see Supporting Fig. 2B for a bulkiness profile). Errors of the measured ^{1}H-R_1 rates were calculated using a Monte Carlo-type resampling strategy and are shown in the diagram as error bars

transient intra-molecular long-range contacts involving mainly the regions 1–40, 70–90, and 120–140 within α-synuclein are reproduced by the sPRE data. Thus, sPRE data are highly sensitive to low populations of residual structure in disordered proteins.

Conclusions

In order to understand the conformational behavior of IDPs and their biological interaction networks, the detection of residual structure and long-range interactions is required. The large number of degrees of conformational freedom of IDPs require extensive sets of experimental data. Here, we provide a straightforward approach for the detection of residual structure and long-range interactions in IDPs and show that sPRE data contribute important and easily-accessible restraints for the investigation of IDPs. Our data indicate that for the general case of an unfolded chain with a local flexibility described by the overwhelming majority of available combinations, sPREs can be accurately predicted through our approach. It can be envisaged that a database of all potential combinations of the 20 amino acids within the central 5-mer peptide can be generated in the future. Nevertheless, generation of sPRE datasets for the entire 3.2 million possible combinations is beyond the current computing capabilities.

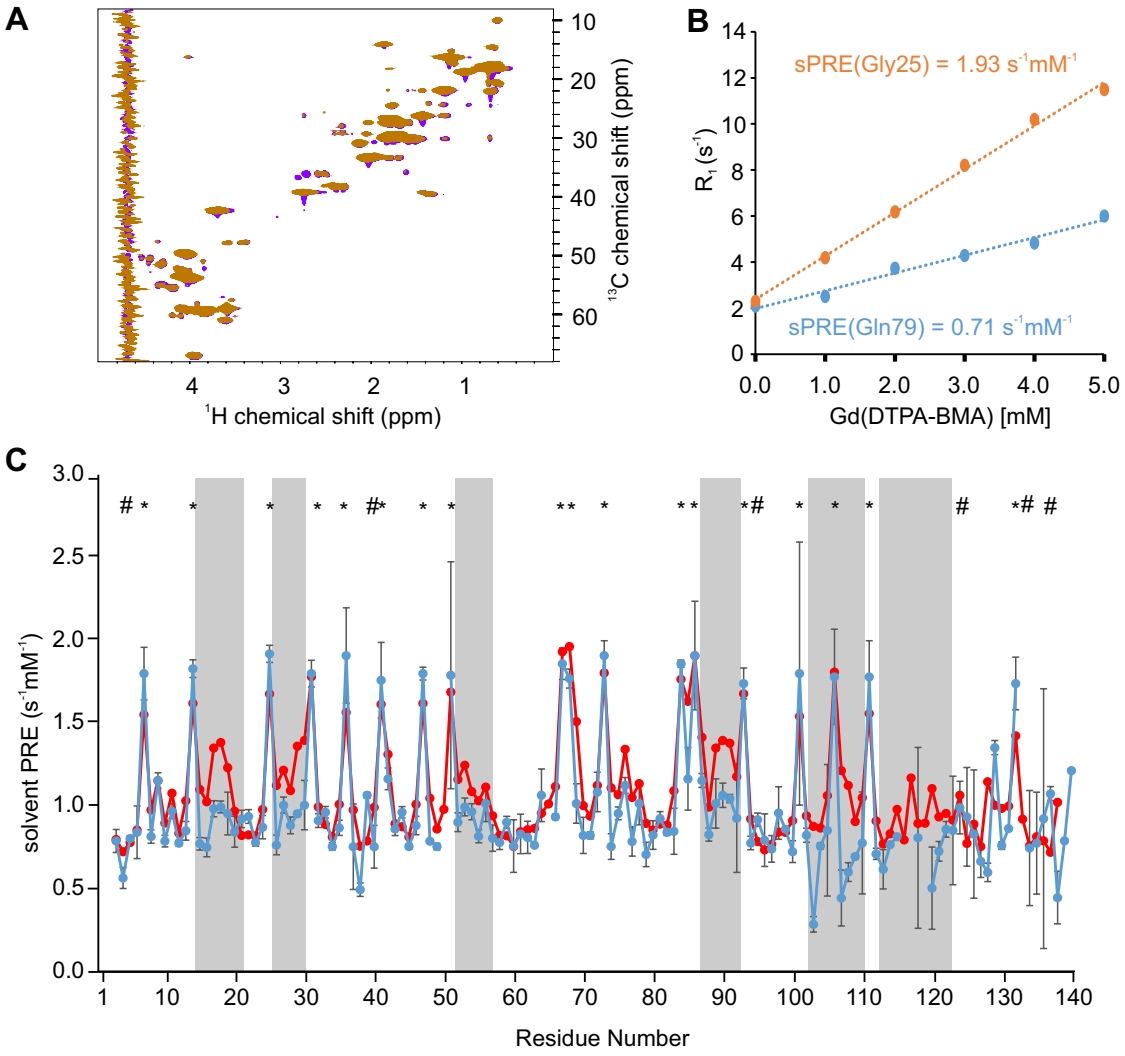


Fig. 4 Comparison of predicted and measured solvent PRE of α-synuclein. **a** Overlay of ^{1}H, ^{13}C HSQC Read-out spectra, with full recovery time of 100 µM ^{13}C,^{15}N labeled α-synuclein in absence (violet) and presence of 5 mM Gd(DTPA-BMA) (orange). **b** Linear fit of relaxation rate ^{1}H-R$_1$ and Gd(DTPA-BMA) concentration of two selected residues of α-synuclein. **c** Predicted (red) and experimentally determined (blue) sPRE values from ^{1}H,^{13}C HSQC read-out spectra. Regions of strong variations between predicted and measured sPRE values are highlighted by grey boxes. Experimental sPRE values are calculated by fitting the data with a linear regression equation. Predicted sPRE values are based on the previously described ensemble approach. Residues with bulky side chains (Phe, Trp, Tyr) are labeled with #, and exposed glycine residues are labeled with * (see Supporting Fig. 2C for a bulkiness profile). Errors of the measured ^{1}H-R$_1$ rates were calculated using a Monte Carlo-type resampling strategy and are shown in the diagram as error bars

Our approach promises to be a straightforward screening tool to exclude potential specific interactions of the soluble paramagnetic agent with IDPs and to guide positioning of covalent paramagnetic spin labels which are often used to detect long-range interactions within IDPs (Gobl et al. 2014; Clore and Iwahara 2009; Otting 2010; Jensen et al. 2014). Paramagnetic spin labels are preferable placed close to, but not within regions involved in transient interactions in order to avoid potential interference of the spin label with weak and dynamic interactions.

In summary, we used three highly disease-relevant biological model systems for determining the solvent accessibility information provided by sPREs. This information can be easily determined experimentally and agrees well with the sPREs predicted for non-exchangeable protons using our grid-based approach. Our method proves to be highly sensitive to low populations of residual structure and long-range contacts in disordered proteins. This approach can be easily combined with ensemble-based calculations such as implemented in flexible-meccano/ASTEROIDS (Mukrasch et al. 2007; Nodet et al. 2009), Xplor-NIH (Kooshapur et al. 2018), or other programs (Estana et al. 2019) to interpret residual structure of IDPs quantitatively and in combination with complementary restraints obtained from RDCs and

PREs. In particular for IDP ensemble calculations relying on sPRE data it is essential to exclude specific interactions of the paramagnetic agent with the IDP of interest which would lead to an enhanced experimental sPRE compared to the predicted sPRE.

Acknowledgements Open access funding provided by Austrian Science Fund (FWF). This research was supported by the Austrian Science Foundation (P28854, I3792, DK-MCD W1226 to TM), the President's International Fellowship Initiative of CAS (No. 2015VBB045, to TM), the National Natural Science Foundation of China (No. 31450110423, to TM), the Austrian Research Promotion Agency (FFG: 864690, 870454), the Integrative Metabolism Research Center Graz, the Austrian infrastructure program 2016/2017, the Styrian government (Zukunftsfonds) and BioTechMed/Graz. E.S. was trained within frame of the PhD program Molecular Medicine. We thank Dr. Vanessa Morris for carefully reading the manuscript.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Alafuzoff I, Hartikainen P (2017) Alpha-synucleinopathies. Handb Clin Neurol 145:339–353

Babu MM, van der Lee R, de Groot NS, Gsponer J (2011) Intrinsically disordered proteins: regulation and disease. Curr Opin Struct Biol 21:432–440

Bah A, Forman-Kay JD (2016) Modulation of intrinsically disordered protein function by post-translational modifications. J Biol Chem 291:6696–6705

Bardiaux B, Malliavin T, Nilges M (2012) ARIA for solution and solid-state NMR. Methods Mol Biol 831:453–483

Bernado P, Bertoncini CW, Griesinger C, Zweckstetter M, Blackledge M (2005) Defining long-range order and local disorder in native alpha-synuclein using residual dipolar couplings. J Am Chem Soc 127:17968–17969

Bernini A, Venditti V, Spiga O, Niccolai N (2009) Probing protein surface accessibility with solvent and paramagnetic molecules. Prog Nucl Magn Reson Spectrosc 54:278–289

Bertoncini CW et al (2005) Release of long-range tertiary interactions potentiates aggregation of natively unstructured alpha-synuclein. Proc Natl Acad Sci USA 102:1430–1435

Bourgeois B, Madl T (2018) Regulation of cellular senescence via the FOXO4-p53 axis. FEBS Lett 592:2083–2097

Brady JP et al (2017) Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. Proc Natl Acad Sci USA 114:E8194–E8203

Burgering BM, Medema RH (2003) Decisions on life and death: FOXO Forkhead transcription factors are in command when PKB/Akt is off duty. J Leukoc Biol 73:689–701

Burre J et al (2010) Alpha-synuclein promotes SNARE-complex assembly in vivo and in vitro. Science 329:1663–1667

Choy MS, Page R, Peti W (2012) Regulation of protein phosphatase 1 by intrinsically disordered proteins. Biochem Soc Trans 40:969–974

Clore GM, Iwahara J (2009) Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes. Chem Rev 109:4108–4139

Clore GM, Tang C, Iwahara J (2007) Elucidating transient macromolecular interactions using paramagnetic relaxation enhancement. Curr Opin Struct Biol 17:603–616

de Keizer PL et al (2010) Activation of forkhead box O transcription factors by oncogenic BRAF promotes p21cip1-dependent senescence. Cancer Res 70:8526–8536

Dedmon MM, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, Dobson CM (2005) Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations. J Am Chem Soc 127:476–477

Delaglio F et al (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 6:277–293

Dyson HJ, Wright PE (2004) Unfolded proteins and protein folding studied by NMR. Chem Rev 104:3607–3622

Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6:197–208

Eletsky A, Moreira O, Kovacs H, Pervushin K (2003) A novel strategy for the assignment of side-chain resonances in completely deuterated large proteins using 13C spectroscopy. J Biomol NMR 26:167–179

Eliezer D (2009) Biophysical characterization of intrinsically disordered proteins. Curr Opin Struct Biol 19:23–30

Emmanouilidis L et al (2017) Allosteric modulation of peroxisomal membrane protein recognition by farnesylation of the peroxisomal import receptor PEX19. Nat Commun 8:14635

Essers MA et al (2004) FOXO transcription factor activation by oxidative stress mediated by the small GTPase Ral and JNK. EMBO J 23:4802–4812

Estana A et al (2019) Realistic ensemble models of intrinsically disordered proteins using a structure-encoding coil database. Structure 27:381–391e2

Falsone SF et al (2011) The neurotransmitter serotonin interrupts alpha-synuclein amyloid maturation. Biochim Biophys Acta 1814:553–561

Fernandez-Fernandez MR, Sot B (2011) The relevance of protein-protein interactions for p53 function: the CPE contribution. Protein Eng Des Sel 24:41–51

Flock T, Weatheritt RJ, Latysheva NS, Babu MM (2014) Controlling entropy to tune the functions of intrinsically disordered regions. Curr Opin Struct Biol 26:62–72

Gillespie JR, Shortle D (1997) Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. J Mol Biol 268:158–169

Gobl C, Madl T, Simon B, Sattler M (2014) NMR approaches for structural analysis of multidomain proteins and complexes in solution. Prog Nucl Magn Reson Spectrosc 80:26–63

Gobl C et al (2016) Increasing the chemical-shift dispersion of unstructured proteins with a covalent lanthanide shift reagent. Angew Chem Int Ed Engl 55:14847–14851

Gobl C et al (2017) Flexible IgE epitope-containing domains of Phl p 5 cause high allergenic activity. J Allergy Clin Immunol 140:1187–1191

Göbl C, Kosol S, Stockner T, Rückert HM, Zangger K (2010) Solution structure and membrane binding of the toxin fst of the par addiction module. Biochemistry 49:6567–6575

Gong Z, Gu XH, Guo DC, Wang J, Tang C (2017) Protein structural ensembles visualized by solvent paramagnetic relaxation enhancement. Angew Chem Int Ed Engl 56:1002–1006

Gu XH, Gong Z, Guo DC, Zhang WP, Tang C (2014) A decadentate Gd(III)-coordinating paramagnetic cosolvent for protein relaxation enhancement measurement. J Biomol NMR 58:149–154

Guttler T et al (2010) NES consensus redefined by structures of PKI-type and Rev-type nuclear export signals bound to CRM1. Nat Struct Mol Biol 17:1367–1376
Habchi J, Tompa P, Longhi S, Uversky VN (2014) Introducing protein intrinsic disorder. Chem Rev 114:6561–6588
Hartlmuller C, Gobl C, Madl T (2016) Prediction of protein structure using surface accessibility data. Angew Chem Int Ed Engl 55:11970–11974
Hartlmuller C et al (2017) RNA structure refinement using NMR solvent accessibility data. Sci Rep 7:5393
Hass MA, Ubbink M (2014) Structure determination of protein-protein complexes with long-range anisotropic paramagnetic NMR restraints. Curr Opin Struct Biol 24:45–53
Hocking HG, Zangger K, Madl T (2013) Studying the structure and dynamics of biomolecules by using soluble paramagnetic probes. ChemPhysChem 14:3082–3094
Hornsveld M, Dansen TB, Derksen PW, Burgering BMT (2018) Re-evaluating the role of FOXOs in cancer. Semin Cancer Biol 50:90–100
Huang JR, Ozenne V, Jensen MR, Blackledge M (2013) Direct prediction of NMR residual dipolar couplings from the primary sequence of unfolded proteins. Angew Chem Int Ed Engl 52:687–690
Huang JR et al (2014) Transient electrostatic interactions dominate the conformational equilibrium sampled by multidomain splicing factor U2AF65: a combined NMR and SAXS study. J Am Chem Soc 136:7068–7076
Hyberts SG, Arthanari H, Robson SA, Wagner G (2014) Perspectives in magnetic resonance: NMR in the post-FFT era. J Magn Reson 241:60–73
Iwahara J, Clore GM (2010) Structure-independent analysis of the breadth of the positional distribution of disordered groups in macromolecules from order parameters for long, variable-length vectors using NMR paramagnetic relaxation enhancement. J Am Chem Soc 132:13346–13356
Jenkins LM et al (2009) Two distinct motifs within the p53 transactivation domain bind to the Taz2 domain of p300 and are differentially affected by phosphorylation. Biochemistry 48:1244–1255
Jensen MR et al (2009) Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. Structure 17:1169–1185
Jensen MR, Zweckstetter M, Huang JR, Blackledge M (2014) Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. Chem Rev 114:6632–6660
Johansson H et al (2014) Specific and nonspecific interactions in ultraweak protein-protein associations revealed by solvent paramagnetic relaxation enhancements. J Am Chem Soc 136:10277–10286
Johnson BA (2004) Using NMRView to visualize and analyze the NMR spectra of macromolecules. Methods Mol Biol 278:313–352
Kooshapur H, Schwieters CD, Tjandra N (2018) Conformational ensemble of disordered proteins probed by solvent paramagnetic relaxation enhancement (sPRE). Angew Chem Int Ed Engl 57:13519–13522
Lee H et al (2000) Local structural elements in the mostly unstructured transcriptional activation domain of human p53. J Biol Chem 275:29426–29432
Madl T, Bermel W, Zangger K (2009a) Use of relaxation enhancements in a paramagnetic environment for the structure determination of proteins using NMR spectroscopy. Angew Chem Int Ed Engl 48:8259–8262
Madl T, Bermel W, Zangger K (2009b) Use of relaxation enhancements in a paramagnetic environment for the structure determination of proteins using NMR spectroscopy. Angew Chem Int Ed Engl 48:8259–8262
Madl T, Guttler T, Gorlich D, Sattler M (2011) Structural analysis of large protein complexes using solvent paramagnetic relaxation enhancements. Angew Chem Int Ed Engl 50:3993–3997
Maji SK et al (2009) Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. Science 325:328–332
Marsh JA et al (2007) Improved structural characterizations of the drkN SH3 domain unfolded state suggest a compact ensemble with native-like and non-native structure. J Mol Biol 367:1494–1510
Marsh JA, Baker JM, Tollinger M, Forman-Kay JD (2008) Calculation of residual dipolar couplings from disordered state ensembles using local alignment. J Am Chem Soc 130:7804–7805
Mcconnell HM (1958) Reaction rates by nuclear magnetic resonance. J Chem Phys 28:430–431
Meier S, Blackledge M, Grzesiek S (2008) Conformational distributions of unfolded polypeptides from novel NMR techniques. J Chem Phys 128:052204
Metallo SJ (2010) Intrinsically disordered proteins are potential drug targets. Curr Opin Chem Biol 14:481–488
Mittag T, Forman-Kay JD (2007) Atomic-level characterization of disordered protein ensembles. Curr Opin Struct Biol 17:3–14
Mukrasch MD et al (2007) Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation. J Am Chem Soc 129:5235–5243
Nodet G et al (2009) Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. J Am Chem Soc 131:17908–17918
Olivier M, Hollstein M, Hainaut P (2010) TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harb Perspect Biol 2:a001008
Otting G (2010) Protein NMR using paramagnetic ions. Annu Rev Biophys 39:387–405
Ozenne V et al (2012) Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics 28:1463–1470
Parigi G et al (2014) Long-range correlated dynamics in intrinsically disordered proteins. J Am Chem Soc 136:16201–16209
Pintacuda G, Otting G (2002a) Identification of protein surfaces by NMR measurements with a paramagnetic Gd(III) chelate. J Am Chem Soc 124:372–373
Pintacuda G, Otting G (2002b) Identification of protein surfaces by NMR measurements with a pramagnetic Gd(III) chelate. J Am Chem Soc 124:372–373
Putker M et al (2013) Redox-dependent control of FOXO/DAF-16 by transportin-1. Mol Cell 49:730–742
Raj N, Attardi LD (2017) The transactivation domains of the p53 protein. Cold Spring Harb Perspect Med 7:a026047
Respondek M, Madl T, Gobl C, Golser R, Zangger K (2007a) Mapping the orientation of helices in micelle-bound peptides by paramagnetic relaxation waves. J Am Chem Soc 129:5228–5234
Respondek M, Madl T, Gobl C, Golser R, Zangger K (2007b) Mapping the orientation of helices in micelle-bound peptides by paramagnetic relaxation waves. J Am Chem Soc 129:5228–5234
Rezaei-Ghaleh N et al (2018) Local and global dynamics in intrinsically disordered synuclein. Angew Chem Int Ed Engl 57:15262–15266
Romero P et al (1998) Thousands of proteins likely to have long disordered regions. Pac Symp Biocomput 3:437–448
Rowell JP, Simpson KL, Stott K, Watson M, Thomas JO (2012) HMGB1-facilitated p53 DNA binding occurs via HMG-Box/p53 transactivation domain interaction, regulated by the acidic tail. Structure 20:2014–2024

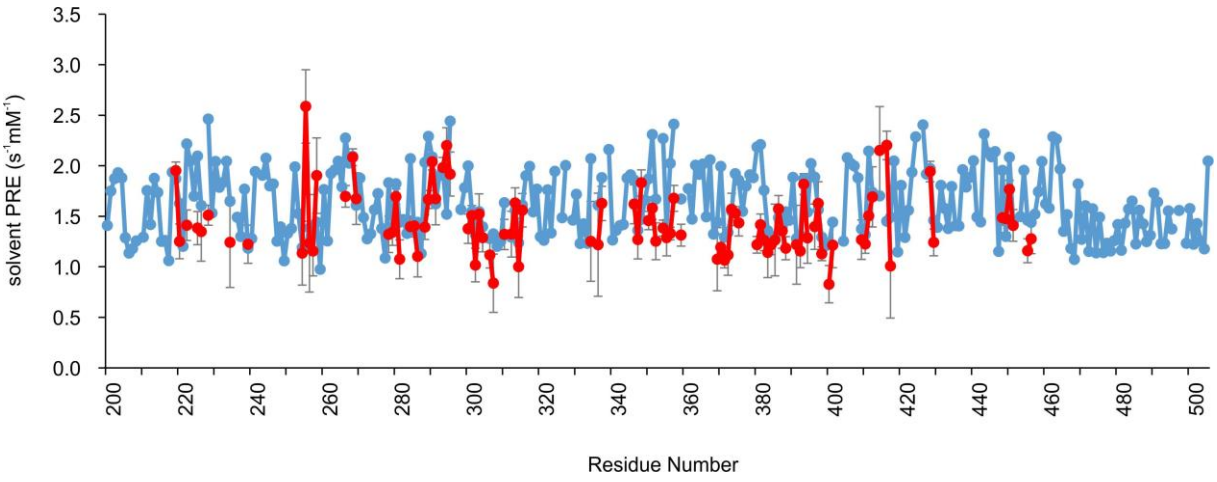
Salmon L et al (2010) NMR characterization of long-range order in intrinsically disordered proteins. J Am Chem Soc 132:8407–8418
Shan B, Li DW, Bruschweiler-Li L, Bruschweiler R (2012) Competitive binding between dynamic p53 transactivation subdomains to human MDM2 protein: implications for regulating the p53. MDM2/MDMX interaction. J Biol Chem 287:30376–30384
Shortle D, Ackerman MS (2001) Persistence of native-like topology in a denatured protein in 8 M urea. Science 293:487–489
Skinner SP et al (2016) CcpNmr AnalysisAssign: a flexible platform for integrated NMR analysis. J Biomol NMR 66:111–124
Sun Y, Friedman JI, Stivers JT (2011) Cosolute paramagnetic relaxation enhancements detect transient conformations of human uracil DNA glycosylase (hUNG). Biochemistry 50:10724–10731
Theillet FX et al (2014) Physicochemical properties of cells and their effects on intrinsically disordered proteins (IDPs). Chem Rev 114:6661–6714
Tompa P (2012) Intrinsically disordered proteins: a 10-year recap. Trends Biochem Sci 37:509–516
Uversky VN (2011) Intrinsically disordered proteins from A to Z. Int J Biochem Cell Biol 43:1090–1103
Uversky VN, Oldfield CJ, Dunker AK (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys 37:215–246
Uversky VN et al (2014) Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. Chem Rev 114:6844–6879
van den Berg MC et al (2013) The small GTPase RALA controls c-Jun N-terminal kinase-mediated FOXO activation by regulation of a JIP1 scaffold complex. J Biol Chem 288:21729–21741
van der Lee R et al (2014) Classification of intrinsically disordered regions and proteins. Chem Rev 114:6589–6631
Vousden KH, Prives C (2009) Blinded by the Light: The Growing Complexity of p53. Cell 137:413–431
Wang Y, Schwieters CD, Tjandra N (2012a) Parameterization of solvent-protein interaction and its use on NMR protein structure determination. J Magn Reson 221:76–84
Wang Y, Schwieters CD, Tjandra N (2012b) Parameterization of solvent–protein interaction and its use on NMR protein structure determination. J Magn Reson 221:76–84
Weigel D, Jackle H (1990) The fork head domain: a novel DNA binding motif of eukaryotic transcription factors? Cell 63:455–456
Wells M et al (2008) Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. Proc Natl Acad Sci USA 105:5762–5767
Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. J Mol Biol 293:321–331
Wright PE, Dyson HJ (2009) Linking folding and binding. Curr Opin Struct Biol 19:31–38
Wright PE, Dyson HJ (2015) Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol 16:18–29
Yuwen T et al (2018) Measuring solvent hydrogen exchange rates by multifrequency excitation (15)N CEST: application to protein phase separation. J Phys Chem B 122:11206–11217
Zangger K et al (2009a) Positioning of micelle-bound peptides by paramagnetic relaxation enhancements. J Phys Chem B 113:4400–4406
Zangger K et al (2009b) Positioning of micelle-bound peptides by paramagnetic relaxation enhancements. J Phys Chem B 113:4400–4406

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

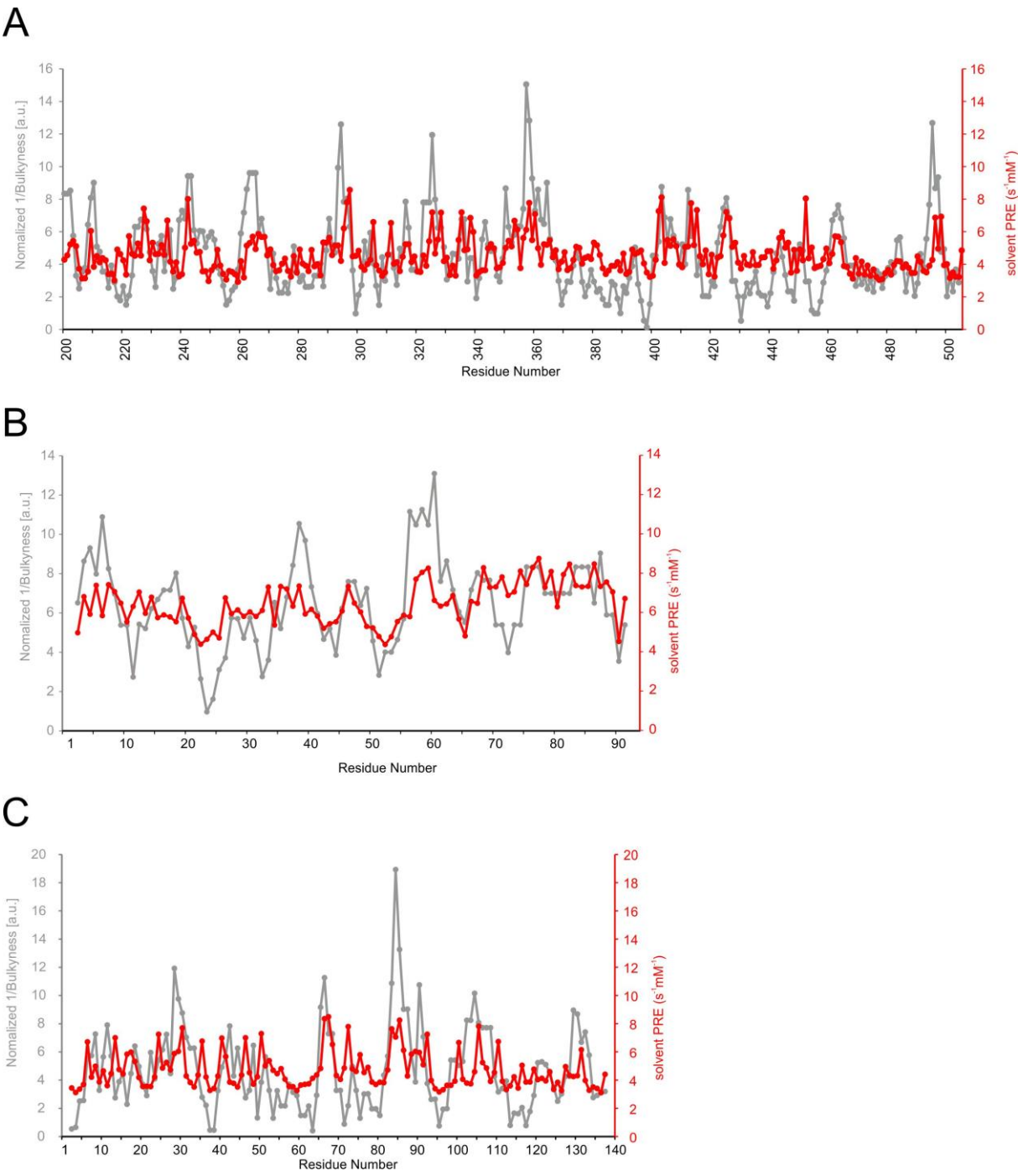
NMR Characterization of Solvent Accessibility and Transient Structure in Intrinsically Disordered Proteins

Christoph Hartlmüller[1,#], Emil Spreitzer[2,#], Christoph Göbl[3], Fabio Falsone[4], Tobias Madl[1,5,*]

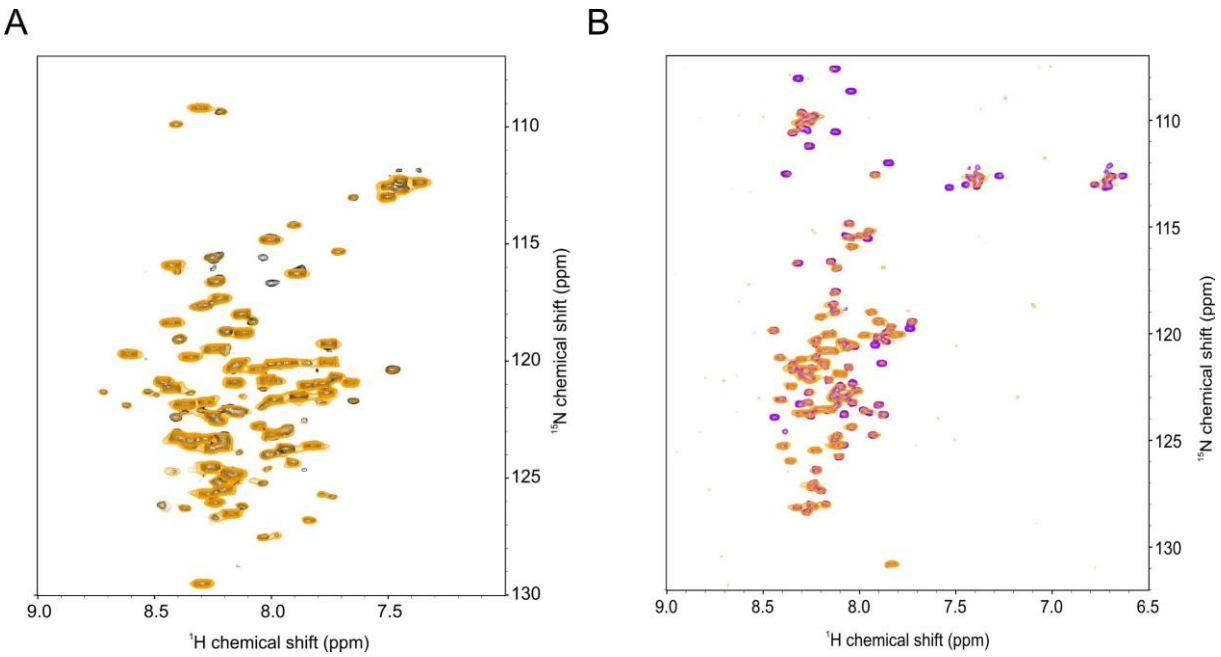
Supporting Information



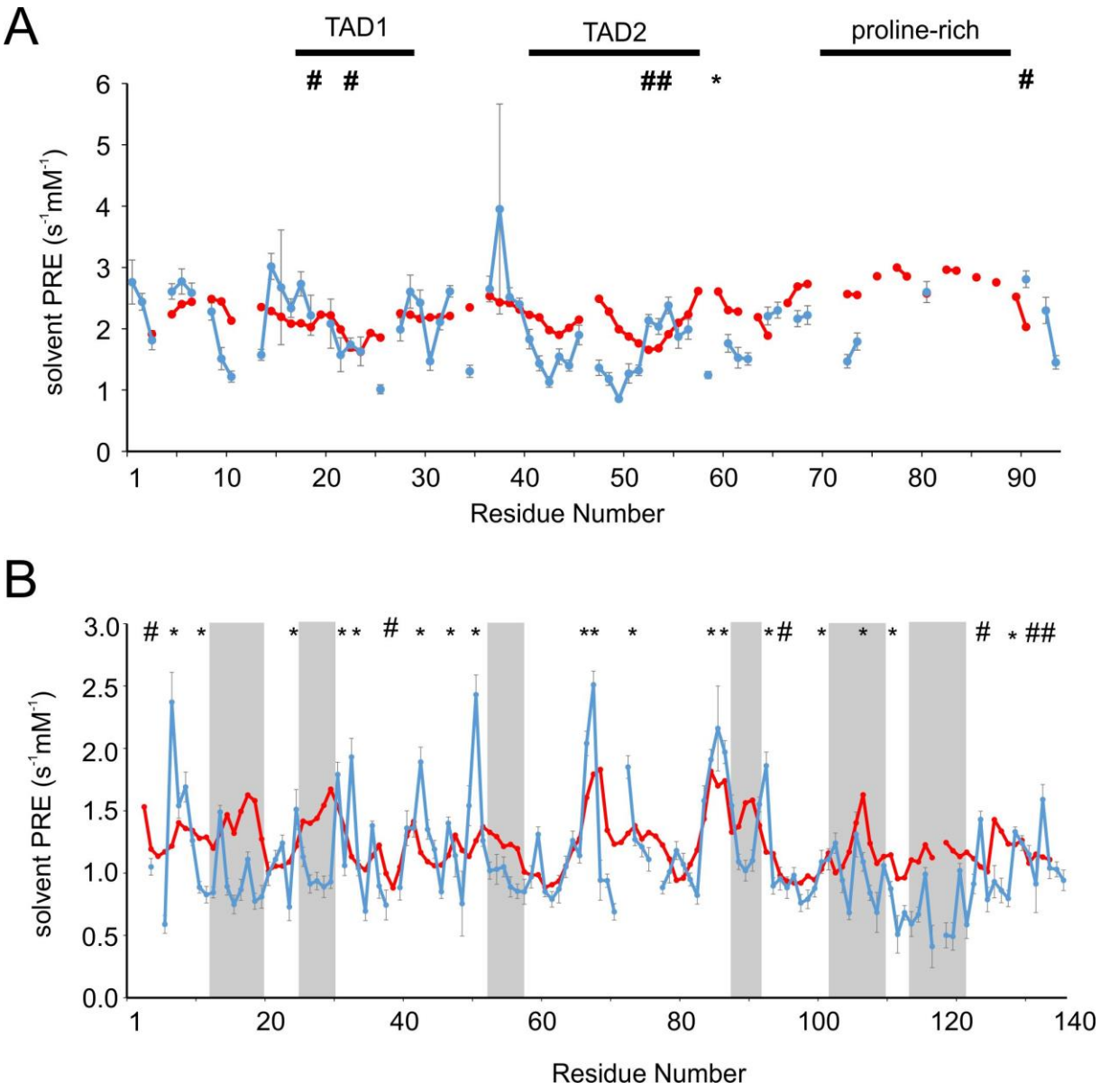
Supporting Figure 1: Experimentally-determined (red) and predicted (blue) solvent PRE values using CBCA(CO)NH as readout spectrum, of assigned H^{β} peaks of $FOXO4^{TAD}$. Experimental sPRE values are calculated by fitting the data with a linear regression equation. Errors of the measured $^{1}H\text{-}R_1$ rates were calculated using a Monte Carlo-type resampling strategy and are shown in the diagram as error bars.



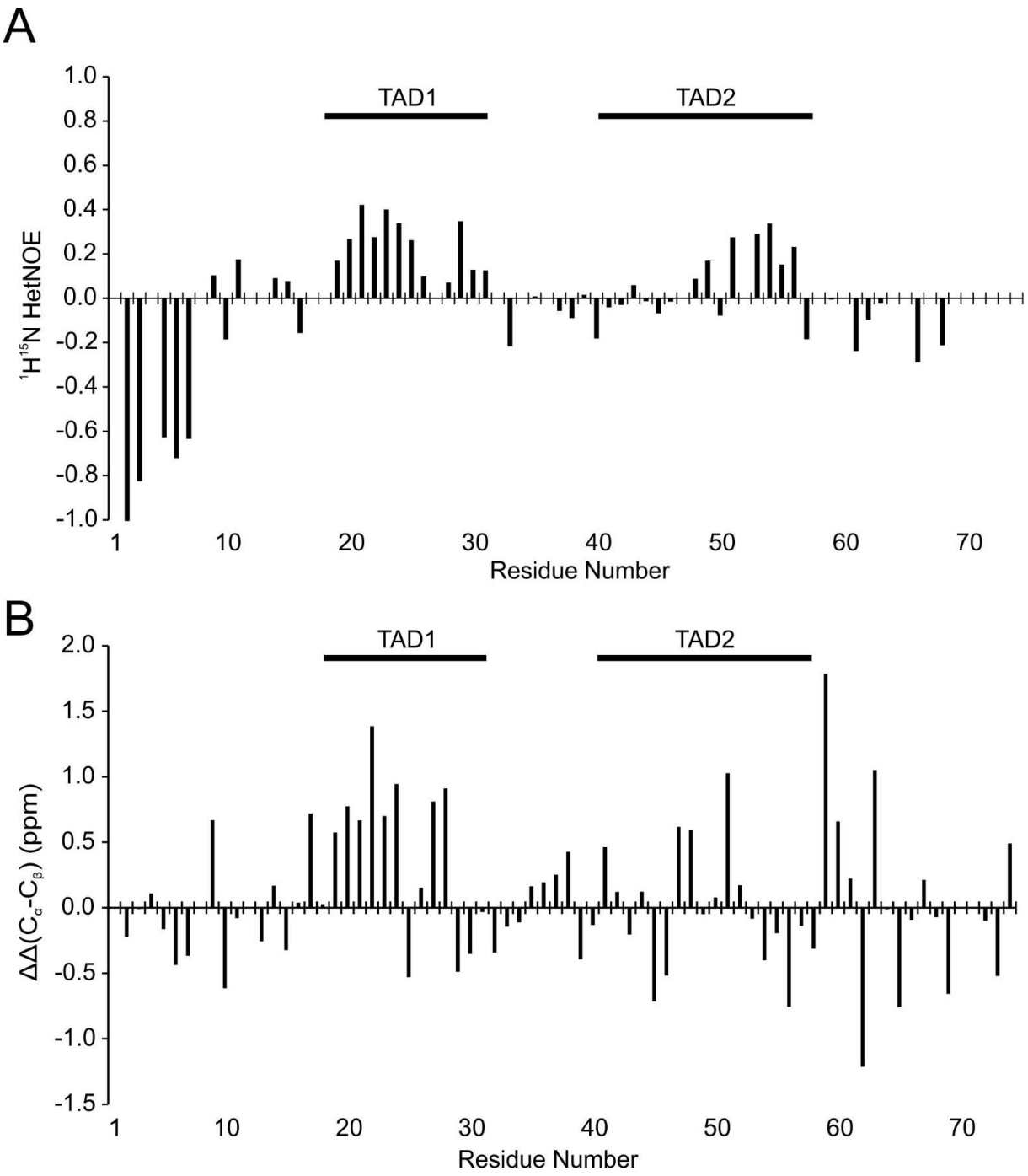
Supporting Figure 2: Comparison of experimental sPRE values of Hα protons (red) and bulkiness (grey) of FOXO4 **(A)**, $p53^{TAD}$ **(B)** and α-synuclein **(C)**. Bulkyness was calculated using the ProtScale web server (https://web.expasy.org/protscale/, 01.04.2019) using a window size of five residues and standard parameters. To compare bulkiness and sPRE the bulkiness was inverted, a linear fit calculated and the bulkiness shifted and scaled to the predicted sPRE data.



Supporting Figure 3: (A) Overlay of ^{1}H,^{15}N HSQC spectra, with full recovery time of a 300 µM ^{13}C,^{15}N labeled p53^{TAD} sample in the absence (black) and presence of 3.25 mM Gd(DTPA-BMA) (orange). **(B)** Overlay of ^{1}H, ^{15}N HSQC spectra, with full recovery time of 100 µM ^{13}C,^{15}N labeled α-synuclein in absence (violet) and presence of 5 mM Gd(DTPA-BMA) (orange).



Supporting Figure 4: Comparison of predicted and measured $^{1}H^{N}$ solvent PREs of $p53^{TAD}$ (A) and α–synuclein (B). Predicted (red) and experimentally determined (blue) sPRE values from ^{1}H,^{15}N HSQC read-out spectra are shown. Regions binding to co-factors (TAD1, TAD2) and the proline rich region are labeled ($p53^{TAD}$). Regions of strong variations between predicted and measured sPRE values are highlighted by grey boxes and reproduce the shielding observed for the $^{1}H^{\alpha}$ sPRE data (α-synuclein). Experimental sPRE values are calculated by fitting the data with a linear regression equation. Predicted sPRE values are based on the previously described ensemble approach. Residues with bulky side chains (Phe, Trp, Tyr) are labeled with #, and exposed glycine residues are labeled with * (see Supporting Figure 2 for bulkiness profiles). Errors of the measured ^{1}H-R_1 rates were calculated using a Monte Carlo-type resampling strategy and are shown in the diagram as error bars.



Supporting Figure 5: p53^{TAD} {^{1}H}-^{15}N heteronuclear NOE (A) and secondary chemical shifts (B). {^{1}H}-^{15}N heteronuclear NOEs were recorded on an Avance Neo 600 MHz spectrometer using the same conditions as described in materials and methods. Secondary chemical shifts were calculated from $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts obtained from backbone resonance assignment.