

New Approaches to NMR Crystallography

Thèse N° 7523

Présentée le 13 décembre 2019

à la Faculté des sciences de base

Laboratoire de résonance magnétique

Programme doctoral en chimie et génie chimique

pour l'obtention du grade de Docteur ès Sciences

par

Federico Maria PARUZZO

Acceptée sur proposition du jury

Prof. A.-C. Corminboeuf, présidente du jury

Prof. D. L. Emsley, directeur de thèse

Prof. M. Baías, rapporteuse

Prof. L. Mafra, rapporteur

Prof. J.-Ph. Ansermet, rapporteur

2019

A Nonna Ninni

Abstract

The first full protocol for nuclear magnetic resonance (NMR) crystallography (NMRX) using chemical shifts was developed a decade ago, and it combines experimental isotropic chemical shifts with crystal structure prediction (CSP) and with the calculation of NMR parameters. In a nutshell, a set of candidate crystal structures is generated using CSP, and then the crystal structure is determined by assessing the agreement between experimental and calculated chemical shifts (usually ^1H).

While this method has proven to be very powerful for structure validation and *de novo* crystal structure determination, it has some severe limitations. The main drawbacks of this protocol, which often prevent the application of NMRX to large and/or complex molecules, are: (i) the poor resolution of ^1H solid-state NMR spectra due to the strong homonuclear dipolar coupling networks, and (ii) the computational costs of the chemical shift calculations and of the CSP protocol. This thesis focuses on the development of methods to overcome these limitations.

The resolution of ^1H solid-state NMR spectra can be improved either by using homonuclear dipolar decoupling sequences or by magic angle spinning (MAS). In the frame of homonuclear dipolar decoupling sequences, we make a broad comparison of the experimental performance of eight schemes developed in the past five decades, and then we provide simple guidelines for the optimization of these experiments. We also probe the limits on resolution using this approach by measuring transverse dephasing times under homonuclear decoupling. We find that coherence lifetimes are limited by the appearance of a coherent oscillatory behaviour that leads to a residual anisotropic splitting, and that this oscillation can be completely removed in a double spin-echo experiment.

In the context of ultra-fast MAS, we work on the development of methods to improve ^1H spectral resolution at sample spinning frequencies higher than 100.0 kHz. We first develop a method which makes use of a constant-time acquisition to remove the spectral broadening due to non-refocusable interactions. Then, assuming that part of the residual homonuclear dipolar coupling behaves like the scalar coupling in this spinning regime, we show that the anti-z-COSY pulse sequence reduces the residual line broadening of ^1H NMR spectra of powdered organic solids. These methods provide up to 50% improvement in resolution compared to conventional echo experiments.

To reduce the computational costs of chemical shift calculations, we develop a machine learning model to predict chemical shifts in molecular solids. This model predicts chemical shifts thousands of times faster than traditional DFT methods, while maintaining similar accuracy. We show that this method can be used as an alternative to GIPAW-DFT in NMR crystallography.

Finally, we propose a modification of the current NMR crystallography protocol. We first develop a method to select the conformers in the first step of the CSP protocol based on structural constraints extracted from NMR experiments. This reduces the number of conformers selected while ensuring that the correct conformer is not excluded. Then, we introduce a Bayesian framework to determine the confidence in the identification of the crystal structure, as well as a visualization approach for the similarity between candidates in terms of their chemical shifts and of their structures to critically assess the reliability of the structure determinations.

Keywords

Solid-state NMR, NMR crystallography, NMRX, homonuclear dipolar decoupling, CRAMPS, high-resolution, ^1H NMR, ultra-fast MAS, constant time, anti-z-COSY, machine learning, chemical shifts calculation, GIPAW, pharmaceutical compounds, microcrystalline solids, crystal structure prediction.

Sommario

Il primo protocollo completo per cristallografia utilizzando lo spostamento chimico della risonanza magnetica nucleare (NMR) è stato sviluppato un decennio fa, e si basa sulla combinazione della spettroscopia NMR con la crystal structure prediction (CSP) e con il calcolo computazionale di parametri NMR. In breve, un set di possibili strutture cristallografiche viene generato utilizzando CSP, e poi la struttura cristallina viene determinata paragonando gli spostamenti chimici calcolati e sperimentali (normalmente ^1H).

Nonostante questo metodo abbia dimostrato di poter essere usato per predire strutture cristallografiche *de novo*, non è privo di difetti. I maggiori problemi, che prevengono l'utilizzo di questo protocollo su strutture più grandi o complesse, sono: (i) la risoluzione dello spettro ^1H NMR allo stato solido, a causa dell'accoppiamento dipolare omonucleare, e (ii) i costi computazionali dei calcoli degli spostamenti chimici e della CSP. Questa tesi si focalizza sullo sviluppo di metodi per superare queste limitazioni.

La risoluzione degli spettri ^1H NMR allo stato solido può essere migliorata utilizzando sequenze per il disaccoppiamento dipolare omonucleare o facendo ruotare velocemente il campione attorno all'angolo magico (MAS). In questa tesi, abbiamo fatto un confronto delle performance di otto sequenze per il disaccoppiamento dipolare omonucleare sviluppate negli ultimi 50 anni, e abbiamo fornito semplici linee guida per la loro ottimizzazione. Abbiamo anche testato i limiti della risoluzione utilizzando queste sequenze attraverso la misura dei tempi di rilassamento trasversale durante disaccoppiamento. Abbiamo trovato che i tempi di vita sono limitati dalla presenza di un'oscillazione, e che questa può essere rimossa utilizzando un doppio echo.

Abbiamo lavorato sullo sviluppo di sequenze per migliorare la risoluzione a frequenze MAS più alte di 100.0 kHz. Abbiamo sviluppato la sequenza CT-MAS, che si basa su un tempo di evoluzione costante per rimuovere il contributo delle interazioni non rifocalizzabili dall'allargamento di banda. In seguito, assumendo che parte dell'accoppiamento dipolare omonucleare residuo a quelle velocità di rotazione si comporti in maniera analoga all'accoppiamento scalare dipolare, abbiamo dimostrato che l'uso della sequenza anti-z-COSY riduce la larghezza di banda in spettri protonici di solidi organici. Questi metodi migliorano la risoluzione fino al 50 % rispetto esperimenti 1D convenzionali.

Per ridurre i costi computazionali del calcolo di spostamenti chimici, abbiamo sviluppato un modello basato sul machine learning per predire spostamenti chimici. Quest modello è migliaia di volte più veloce rispetto a metodi DFT, mantenendo una simile precisione. Abbiamo dimostrato che questo può essere utilizzato come alternativa a GIPAW-DFT in cristallografia NMR.

Infine, abbiamo proposto delle modifiche del protocollo per cristallografia NMR. Abbiamo sviluppato un metodo per selezionare i conformeri nel protocollo CSP basato su vincoli strutturali estratti da esperimenti NMR. Questo riduce il numero di conformeri selezionati assicurando che il conformero corretto non sia escluso. Poi, abbiamo introdotto un nuovo metodo per determinare la confidenza nell'identificazione della struttura cristallografica, e per la visualizzazione della somiglianza delle strutture candidate in termini di spostamento chimico e di coordinate in modo da valutare in modo critico l'affidabilità delle strutture determinate.

Parole Chiave

NMR allo stato solido, cristallografia NMR, NMRX, disaccoppiamento dipolare omonucleare, CRAMPS, alta risoluzione, ^1H NMR, MAS veloce, anti-z-COSY, machine learning, calcoli spostamento chimico, GIPAW, composti farmaceutici, solidi microcristallini, crystal structure prediction.

Contents

Abstract	i
Sommario	v
List of publications	ix
Abbreviations.....	xi
Chapter 1 Introduction.....	13
1.1 Why NMR ?.....	13
1.2 NMR in a Nutshell.....	14
1.2.1 The chemical shielding.....	15
1.2.2 The dipolar coupling	16
1.3 NMR Crystallography.....	18
1.4 The Isotropic Chemical Shift Based NMR Crystallographic Approach	21
1.5 High-Resolution ^1H Solid-State NMR	22
1.6 Outline and Scope of the Thesis	23
Chapter 2 Decoupling Sequences	25
2.1 Materials and Methods	28
2.2 Comparison of Decoupling Performance	29
2.3 Pulse Sequence Optimization.....	36
2.3.1 Preparatory Steps	37
2.3.2 Optimization Steps.....	40
2.4 Limits on Resolution	44
2.5 Conclusions.....	47
2.6 Supplementary Information and Discussion	48
2.6.1 Supplementary Information Section 2.4.....	48
Chapter 3 Decoupling at Ultra-Fast MAS	51
3.1 Materials and Methods	52
3.2 The TOP-CT-MAS Experiment.....	55
3.3 The anti-z-COSY Experiment.....	58
3.4 Conclusions.....	62

3.5	Supplementary Information and Discussion	63
3.5.1	Supplementary Information Section 3.2.....	63
3.5.1	Supplementary Information Section 3.3.....	63
Chapter 4	Machine Learning Chemical Shifts.....	67
4.1	Materials and Methods	68
4.2	ShiftML	69
4.3	Conclusions.....	75
4.4	Supplementary Information and Discussion	75
Chapter 5	Rethinking NMR Crystallography	83
5.1	Materials and Methods	84
5.2	NMRX Directed by Unambiguous Prior Constraints	86
5.3	A Probabilistic Crystal Structure Selection	95
5.4	Conclusions.....	100
5.5	Supplementary Information and Discussion	101
5.5.1	Supplementary Information Section 5.2.....	101
5.5.2	Supplementary Information Section 5.3.....	111
Chapter 6	Conclusions.....	115
Appendix	117
Chapter 2.....		117
Additional Section 2.2 and 2.3 Information		117
Chapter 3.....		133
Additional Section 3.2 Information		133
Additional Section 3.3 Information		134
Chapter 4.....		135
Chapter 5.....		139
Additional Section 5.2 Information		139
Additional Section 5.3 Information		143
References		147
Acknowledgements		161
Curriculum Vitae		165

List of publications

The work shown in this thesis is based on the following publications:

1. Moutzouri, P., Paruzzo, F.M. and Emsley, L., "Homonuclear Decoupling in ^1H NMR of Solids by Remote Correlation", *article in preparation* **2019**. (**pre-print**)
2. Engel, E.A., Anelli, A., Hofstetter, A., Paruzzo, F.M., Emsley, L. and Ceriotti, M., "A Bayesian approach to NMR crystal structure determination", *Physical Chemistry Chemical Physics* **2019**, 21, 23385-23400. DOI: [10.1039/C9CP04489B](https://doi.org/10.1039/C9CP04489B) (**pre-print**)
3. Paruzzo, F. M. and Emsley, L., "High-Resolution ^1H NMR of Powdered Solids by Homonuclear Dipolar Decoupling". *Journal of Magnetic Resonance* **2019**, 309, 106598. DOI: [10.1016/j.jmr.2019.106598](https://doi.org/10.1016/j.jmr.2019.106598) (**post-print**)
4. Hofstetter, A., Balodis, M., Paruzzo, F.M., Widdifield, C.M., Stevanato, G., Pinon, A.C., Bygrave, P.J., Day, G.M. and Emsley, L., "Rapid structure determination of molecular solids using chemical shifts directed by unambiguous prior constraints". *Journal of the American Chemical Society* **2019**, 141, 42, 16624-16634. DOI: [10.1021/jacs.9b03908](https://doi.org/10.1021/jacs.9b03908) (**post-print**)
5. Paruzzo, F.M., Walder, B.J. and Emsley, L., "Line Narrowing in ^1H NMR of Powdered Organic Solids with TOP-CT-MAS Experiments at Ultra-Fast MAS". *Journal of Magnetic Resonance* **2019**, 305, 131-137. DOI: [10.1016/j.jmr.2019.06.015](https://doi.org/10.1016/j.jmr.2019.06.015) (**post-print**)
6. Paruzzo, F. M., Hofstetter, A., Musil, F., De, S., Ceriotti, M. and Emsley, L., "Chemical shifts in molecular solids by machine learning". *Nature Communications* **2018**, 9 (1), 4501. DOI: [10.1038/s41467-018-06972-x](https://doi.org/10.1038/s41467-018-06972-x) (**post-print**)
7. Paruzzo, F.M., Stevanato, G., Halse, M.E., Schlagnitweit, J., Mammoli, D., Lesage, A. and Emsley, L., "Refocused linewidths less than 10 Hz in ^1H solid-state NMR". *Journal of Magnetic Resonance* **2018**, 293, 41-46. DOI: [10.1016/j.jmr.2018.06.001](https://doi.org/10.1016/j.jmr.2018.06.001) (**post-print**)

During my PhD, I have also been working on other aspects of nuclear magnetic resonance crystallography. These results are not part of this thesis, and more details can be found in the following publications:

1. Walder, B., Prisco, N., Paruzzo, F.M., Yarava, J., Chmelka, B. and Emsley, L., "Measurement of Proton Spin Diffusivity in Hydrated Cementitious Solids". *The Journal of Physical Chemistry Letters* **2019**, 10, 17, 5064-5069. DOI: [10.1021/acs.jpcllett.9b01861](https://doi.org/10.1021/acs.jpcllett.9b01861)
2. Leclaire, J., Poisson, G., Ziarelli, F., Pepe, G., Fotiadu, F., Paruzzo, F.M., Rossini, A.J., Dumez, J.N., Elena-Herrmann, B. and Emsley, L., "Structure elucidation of a complex CO_2 -based organic framework material by NMR crystallography". *Chemical science* **2016**, 7(7), 4379-4390. DOI: [10.1039/C5SC03810C](https://doi.org/10.1039/C5SC03810C)

Abbreviations

CRAMPS	combined rotation and multi-pulse spectroscopy
CSP	crystal structure prediction
CT	constant-time
DFT	deinsity functional theory
DOR	double rotation
DUMBO	decoupling using mind-boggling optimization
EDX	electron diffraction
FPS	farthest point sampling algorithm
FSLG	frequency-switched Lee-Goldburg
GPR	gaussian process regression
HETCOR	heteronuclear correlation
HOT	high order truncation
INADEQUATE	incredible natural abundance double quantum transfer experiment
LG	Lee-Golburg
MAS	magic angle spinning
MAT	magic angle turning
ML	machine learning
NMR	nuclear magnetic resonance
NMRX	NMR crystallography
PCA	principal component analysis
PMLG	phase-modulated Lee-Goldburg
PSD	proton spin diffusion
<i>rf</i>	radiofrequency
RMSD	root mean square deviation
RMSE	root mean square error
SOAP	smooth overlap of atomic positions
XRD	X-ray diffraction

Chapter 1 Introduction

In this thesis, I will show some recent developments in the field of nuclear magnetic resonance (NMR) crystallography. These developments, which range from new sequences to improve the resolution of ^1H NMR spectra to a machine learning model to predict chemical shifts in record time, have made NMR crystallography (NMRX) faster, more reliable, and more accurate. To start, let's make a step back.

The meaning of the word "crystallography", which derives from the Greek words *krystallos* ("clear ice") and *graphe* ("description"),¹ evolved from the almost 300-years-old idea of "science that describes the geometrical shapes of minerals"² to the most modern definition of being a "branch of science devoted to the study of molecular and crystalline structure and properties, with far-reaching applications in mineralogy, chemistry, physics, mathematics, biology, metallurgy and materials science".³ The origin of this shift can be attributed to René Just Haüy, who in 1784 first realized that crystals were made by the combination of very small identical building blocks which describe the geometrical properties of the crystal (the "unit cell").⁴ However, the perhaps most relevant event that led to the creation of the "modern crystallography" is dated to 1912, when Max von Laue observed the diffraction of X-rays by a crystal.⁵ This discovery immediately attracted the attention of William Henry Bragg and William Lawrence Bragg, who developed this observation into the famous Bragg's Law ($n\lambda = 2d\sin\theta$), which describes the relationship between the X-ray wavelength (λ), the scattering angle (θ) and the interplanar spacings in the crystal lattice (d).⁶ These events, whose importance is underlined by two Nobel Prizes in Physics (von Laue in 1914 and Henry and Lawrence Bragg in 1915), marked the birth of X-ray diffraction (XRD) crystallography and, more generally, of the idea of crystallography as "the study of structures at the atomic level".

XRD was first used for the determination of the crystal structure of diamonds,⁷ and it is nowadays routinely used in many areas of chemistry, physics, biology, and material science. While single-crystal XRD is widely considered the gold standard to determine the structure of crystalline solids, it is not the only technique that provides structural information at the atomic level. Similar to X-rays, neutrons are diffracted by crystals,⁸ and can be used in an analogous way for structural determination of crystalline solids.⁹ More recently, solid-state NMR and electron diffraction (EDX) methods¹⁰⁻¹³ have shown to be powerful methods to access atomic level information.

As mentioned above, the work shown in this thesis is focused on the development of methodologies for NMRX. The following paragraphs are dedicated to a more detailed introduction of this concept.

1.1 Why NMR ?

Being a spectroscopic technique, NMR measures the intensity of absorption of a given electromagnetic radiation as a function of the frequency of irradiation. The NMR spectrum appears at relatively low frequencies in the electromagnetic spectrum, at about a few tens/hundreds of MHz, which is the region of the radio waves that corresponds to wavelengths a few m/cm long. This spectroscopic technique is based on the ability of an oscillating magnetic field to flip the magnetic orientation of nuclei immersed in a static magnetic field, which was discovered by Isidor Isaac Rabi in 1938.¹⁴ However, only after Felix Bloch and Edward Mills Purcell independently acquired the first NMR experiments on condensed matter, the interest for this technique spread.^{15,16}

NMR offers many advantages over XRD from a crystallographic point of view. The most relevant one is the fact that NMR directly probes local atomic environments, thus allowing for structural characterization without the need for long-range order. Indeed, when the sample exists only in the form of a powder, which is increasingly the case, the use of XRD is still possible, but far more challenging.¹⁷ Solid-state NMR, on the other side, has shown to be a powerful and robust method to determine crystal structures from powdered samples with an accuracy at least comparable to that of single-crystal XRD.¹⁸ Another advantage of NMR over XRD is its ability to accurately determine ^1H positions. Nuclei poor in electrons, such as hydrogen, are indeed hard to detect using XRD due to the fact that X-rays are scattered by the electron density. In contrast, ^1H is often considered the perfect nucleus for NMR because it has nearly 100% natural abundance and the highest gyromagnetic ratio (see below) of all the stable nuclei. This is a key difference between XRD and NMR, as it allows to access the rich information contained in ^1H , which is ubiquitous in chemistry, that leads to determining accurate structural information, and shines also light to interactions such as hydrogen bonds, which play a critical role in crystallography. Another difference between these two techniques is the ability of NMR to obtain selective information. Selective information means to focus only on a particular nucleus of interest at the time, as well as to observe specific interactions (see below). In both cases, this is a very precious feature of NMR as it gives access to an enormous amount of information, facilitating the analysis of complex samples. Finally, solid-state NMR has shown to be a powerful method to obtain some of the crystallographic information hardly accessible with powder XRD, such as the study of dynamics and the ability to study amorphous or disordered solids.¹⁹

On the other side, the use of NMR for crystallographic purposes is more recent and less developed than XRD. The large amount of information contained in NMR yields relatively complicated spectra, which often require the acquisition of multiple NMR experiments for a correct interpretation. This can be challenging and time-consuming, especially considering that the acquisition and interpretation of solid-state NMR data is often limited either by sensitivity or by spectral resolution. In addition, NMR alone cannot yet be used for full structure determination, and it needs to be paired with other techniques that require different areas of expertise, such as XRD or crystal structure prediction (CSP). The complexity of NMR spectra, together with the computational costs associated with the calculation of NMR parameters or to other processes involved in NMRX such as CSP, strongly limit the size and the complexity of the structures that can be determined using this method. All these disadvantages, together perhaps with the high costs associated with setting up and maintaining an NMR facility, prevent the widespread of NMRX, despite its potential.

In the following section, I will review some theoretical aspects of NMR in order to understand how this technique can be used to obtain structural information. These concepts are only briefly summarized here as they are explained with more detail in several textbooks.²⁰⁻²³

1.2 NMR in a Nutshell

In NMR we measure the net magnetization (\mathbf{M}), which is the vectorial sum of the individual magnetic moments associated with the nuclei

$$\mathbf{M} = \sum_i \boldsymbol{\mu}_i, \quad (1.1)$$

where $\boldsymbol{\mu}_i$ is the magnetic moment of the i -th nucleus, which is related to the nuclear spin (\mathbf{I}) of the nucleus by the gyromagnetic ratio (γ), an intrinsic property of each NMR active isotope

$$\boldsymbol{\mu}_i = \gamma_i \mathbf{I}_i. \quad (1.2)$$

The magnitude of the magnetic moment is given by

$$|\boldsymbol{\mu}_i| = \sqrt{I(I+1)}\gamma_i\hbar, \quad (1.3)$$

where \hbar is the reduced Planck constant ($\hbar = 1.054571817 \times 10^{-34}$ J·s), and I is the nuclear spin quantum number ($I = 0, 1/2, 1, 3/2$, etc.) which depends on the number of protons and neutrons in the nucleus.

When a nucleus is placed in an external magnetic field (B_0) whose axis defines the z-direction, its magnetic dipole moment interacts with the magnetic field (the *Zeeman interaction*), causing the loss of degeneration of the energy levels which splits into $(2I+1)$ eigenstates, with energies equal to:

$$E_m = -\mu_z B_0 = -m\hbar\gamma B_0, \quad (1.4)$$

where m is the magnetic quantum number, which varies between $-I$ and $+I$ in integer steps. This creates energy splittings which correspond to a photon of energy $\Delta E = \hbar\gamma B_0 = \hbar\omega_L$, where ω_L is known as the Larmor frequency of the nucleus (in rad·s⁻¹).

If all nuclei of a given species would resonate exactly at their Larmor frequency, NMR would not be of use in many applications (except, perhaps, elemental analysis), and it definitely would not have been useful in crystallographic applications. The real strength of NMR lies in the energy gap between eigenvalues, which is in a first order defined by the Zeeman splitting, and then shaped by weaker interactions that depend on the chemical system under analysis. For spins with $I=1/2$ such as ¹H or ¹³C, the spins subject to the work of this thesis, the Hamiltonian operator (\hat{H}) that describes the energy of the system in an external magnetic field is typically given by:

$$\hat{H} = \hat{H}_{Zeeman} + \hat{H}_\sigma + \hat{H}_D + \hat{H}_J, \quad (1.5)$$

where the corresponding interactions are \hat{H}_{Zeeman} the Zeeman interaction, \hat{H}_σ the chemical shielding, \hat{H}_D the dipolar coupling and \hat{H}_J the scalar coupling. While the Zeeman interaction, discussed above, it is not of great interest from a crystallographic point of view, the other components of **Equation 1.5** are, since they are directly related to the structure of the system under analysis.

In the following, the chemical shielding and dipolar coupling interactions are detailed as these are the interactions that are mainly used to obtain structural information with NMR. Note that, in principle, also the scalar coupling (or J-coupling), which represents the interaction between nuclear spins mediated by electrons, can be used in NMR crystallography as it contains important structural information such as dihedral angles, bond lengths and, more generally, crystal packing.²⁴⁻³⁰

1.2.1 The chemical shielding

The chemical shielding Hamiltonian describes the effect of the electron density surrounding a given nucleus on its resonance frequency. Indeed, when an atom is placed into a magnetic field, the electrons that surround the nucleus do not remain inert, but react producing a secondary field that contributes to the total magnetic field experienced by the nucleus. The electron density depends to first order on the central nucleus and to second order on the adjacent nuclei, especially the bonded ones. As a result, the nuclei of a compound have often different but typical resonance frequencies, which mostly depend on chemical connectivity. This effect has made NMR the technique of choice for the characterization of chemical compounds, and it is also very important for crystallographic purposes as it allows to obtain selective information on each atom. Even more importantly from a crystallographic point of view is the fact that *the electron density also depends, to third order, on the crystal packing surrounding the investigated nuclei*, which allows direct correlation between resonance frequency and crystal structure. The chemical shielding Hamiltonian is given by

$$\hat{H}_\sigma = -\gamma B_0 \sigma \hat{I}, \quad (1.6)$$

where σ is the chemical shielding tensor and \hat{I} is the nuclear spin operator. Here, σ is a 3x3 matrix, and it is possible to choose the axis frame so that this tensor is diagonal (called the principal axis system (PAS)). The chemical shielding tensor can be expressed in terms of its principal values σ_{ii}^{PAS} as

$$\begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{pmatrix} \xrightarrow{PAS} \begin{pmatrix} \sigma_{xx}^{PAS} & 0 & 0 \\ 0 & \sigma_{yy}^{PAS} & 0 \\ 0 & 0 & \sigma_{zz}^{PAS} \end{pmatrix} \quad (1.7)$$

with the isotropic chemical shielding (σ_{iso}) being the average of the three principal components

$$\sigma_{iso} = \frac{\sigma_{xx}^{PAS} + \sigma_{yy}^{PAS} + \sigma_{zz}^{PAS}}{3}. \quad (1.8)$$

In NMR experiments the chemical shielding tensor is not measured as absolute frequency, but relative to a reference substance and reported as *chemical shift* with respect to that reference. The chemical shift tensor (δ) and the isotropic chemical shift are thus expressed as

$$\delta_{ij} = \frac{\sigma_{ij}^{ref} - \sigma_{ij}}{1 - \sigma_{ij}^{ref}} \quad \text{and} \quad \delta_{iso} = \frac{\sigma_{iso}^{ref} - \sigma_{iso}}{1 - \sigma_{iso}^{ref}}. \quad (1.9)$$

Both components of the chemical shift, isotropic and anisotropic parts, strongly depend on the local environment around the nucleus, making full structural information accessible with NMR. The work shown in this thesis is based on an NMR crystallographic method which uses isotropic chemical shifts, which are usually easier to measure than the full anisotropic tensors, to fully determine the crystal structure of molecular solids (see **Section 1.3**).

1.2.2 The dipolar coupling

The third term of **Equation 1.5** is the dipolar coupling Hamiltonian, which describes the interaction through space between the magnetic moments of two nuclei. It is possible to derive the form of this Hamiltonian from classical physics. Considering nuclei I and S at a distance r , the energy of the interaction of their magnetic moments is

$$U = \left\{ \frac{\mu_I \cdot \mu_S}{r^3} - 3 \frac{(\mu_I \cdot r)(\mu_S \cdot r)}{r^5} \right\} \frac{\mu_0}{4\pi}, \quad (1.10)$$

which leads to

$$\hat{H}_D = -\left(\frac{\mu_0}{4\pi}\right) \gamma_I \gamma_S \hbar \left(\frac{\hat{I} \cdot \hat{S}}{r^3} - 3 \frac{(\hat{I} \cdot r)(\hat{S} \cdot r)}{r^5} \right), \quad (1.11)$$

where μ_0 is the vacuum permeability constant. This equation can also be expressed using spherical polar coordinates

$$\mathbf{r} = (r \sin \theta \cos \phi, r \sin \theta \sin \phi, r \cos \theta), \quad (1.12)$$

which gives the dipolar coupling Hamiltonian as

$$\hat{H}_D = -\left(\frac{\mu_0}{4\pi}\right) \frac{\gamma_I \gamma_S \hbar}{r^3} [A + B + C + D + E + F], \quad (1.13)$$

where

$$\begin{aligned}
A &= \hat{I}_z \hat{S}_z (3 \cos^2 \theta - 1) \\
B &= -\frac{1}{4} [\hat{I}_+ \hat{S}_- + \hat{I}_- \hat{S}_+] (3 \cos^2 \theta - 1) \\
C &= -\frac{3}{2} [\hat{I}_z \hat{S}_+ + \hat{I}_+ \hat{S}_z] (\sin \theta \cos \theta e^{-i\phi}) \\
D &= -\frac{3}{2} [\hat{I}_z \hat{S}_- + \hat{I}_- \hat{S}_z] (\sin \theta \cos \theta e^{+i\phi}) \\
E &= -\frac{3}{4} [\hat{I}_+ \hat{S}_+] (\sin^2 \theta e^{-2i\phi}) \\
D &= -\frac{3}{4} [\hat{I}_- \hat{S}_-] (\sin^2 \theta e^{+2i\phi})
\end{aligned} \tag{1.14}$$

known as the dipolar alphabet. \hat{I}_+ , \hat{S}_+ , \hat{I}_- and \hat{S}_- in **Equation 1.14** are respectively the raising and lowering operators acting on spin I and S , which in terms of Cartesian operators are equal to

$$\hat{J}_{\pm} = \hat{J}_X \pm i\hat{J}_Y. \tag{1.15}$$

We can now consider this interaction in the presence of a static magnetic field. In general, the local fields arising from the interactions within the sample are orders of magnitude lower than the Zeeman interaction, and therefore at equilibrium many of them have a negligible effect on the spin states. The only components of local fields that significantly affect the spin state are: (i) components parallel or antiparallel to B_0 and (ii) components precessing at or near the Larmor frequency in the plane perpendicular to B_0 . Among the terms that compose the dipolar alphabet, only terms A and B have a non-negligible effect on the energy of the spin-system in the presence of a static magnetic field. We can thus introduce the high-field dipolar Hamiltonian as

$$\hat{H}_D = -\left(\frac{\mu_0}{4\pi}\right) \frac{\gamma_I \gamma_S \hbar}{r^3} (3 \cos^2 \theta - 1) \left[\hat{I}_z \hat{S}_z - \frac{1}{2} (\hat{I}_x \hat{S}_x + \hat{I}_y \hat{S}_y) \right]. \tag{1.16}$$

Note that the terms $\hat{I}_x \hat{S}_x$ and $\hat{I}_y \hat{S}_y$ represent components of the nuclear spin in the plane perpendicular to B_0 . As mentioned above, these components have a non-negligible effect on the spin state only if they precess at or near the Larmor frequency. We can thus distinguish two cases, depending on the species of the spin I and S . If I and S are the same nuclide (homonuclear dipolar coupling) then this component affects the spin state, otherwise (heteronuclear dipolar coupling) this term is negligible. We can then describe the homonuclear and heteronuclear dipolar coupling Hamiltonians respectively as

$$\hat{H}_D^{homo} = -\left(\frac{\mu_0}{4\pi}\right) \frac{\gamma_I \gamma_S \hbar}{r^3} (3 \cos^2 \theta - 1) \left[\hat{I}_z \hat{S}_z - \frac{1}{2} (\hat{I}_x \hat{S}_x + \hat{I}_y \hat{S}_y) \right] \tag{1.17}$$

and

$$\hat{H}_D^{hetero} = -\left(\frac{\mu_0}{4\pi}\right) \frac{\gamma_I \gamma_S \hbar}{r^3} (3 \cos^2 \theta - 1) [\hat{I}_z \hat{S}_z] \tag{1.18}$$

This difference has a huge impact on the effect of these couplings on the spectra and on the strategies that need to be implemented in order to remove or scale them. This, however, does not influence the structural information contained in this interaction. From a crystallographic point of view, the direct dipolar couplings are interesting thanks to their dependence on the cube of the distance between nuclei ($\propto \frac{1}{r^3}$), which implies that *the distances between all nuclei are encoded in this coupling*, and therefore accessible with NMR. In **Chapter 5** I will show an example of how this information can be extracted from correlation experiments and used to drive the structural determination of molecular solids, and in **Chapter 2** and **Chapter 3** I will present the work we have done to average the homonuclear dipolar couplings and improve the resolution of ^1H NMR spectra.

1.3 NMR Crystallography

If we consider the meaning of NMR crystallography in its broader sense, as the use of NMR experiments in order to obtain structural and crystallographic information, then most of the solid-state NMR works published so far would fit in this category. This is due to the fact that the amount of structural information that can be extracted using NMR is massive, and this power was recognized already in the early days of NMR.³¹ The potential of solid-state NMR in this field has been perfectly described by Robin Harris, who wrote:³²

“A few points about NMR and its potential applications to crystallography are as follows:

- *NMR responds to the short-range environment of relevant atoms and is not directly influenced by long-range order.*
- *It can therefore be applied to amorphous materials as well as crystalline ones, though with broader lines for the former to encompass the variations in nuclear environments.*
- *It can be readily used to determine the chemical nature of a solid compound, including crystallographically important information such as conformation and tautomeric form.*
- *Chemical shifts give information about intermolecular interactions.*
- *Inter- and intra-molecular hydrogen-bond linkages can be identified.*
- *Information on crystallographic asymmetric units is especially readily available, usually merely by counting lines.*
- *Polymorphs are usually easily distinguished.*
- *Phase transitions can be monitored.*
- *Crystallographic disorder is detectable, and distinctions between spatial and temporal disorder can be made.*
- *Motions such as internal rotation and ring inversion can be detected and their rates obtained, even in cases of mutual exchange (e.g. 180° ring flips of phenyl groups).*
- *More general information about molecular-level mobility can be obtained by measurements of relaxation times.*
- *Measurement of dipolar coupling constants yields through-space inter-atomic (i.e. internuclear) distances, though these will be modulated by local mobility.*
- *NMR data can be used as restraints in carrying out full structure determination from powder diffraction data.*
- *Heterogeneous materials can be studied and selective spectra for particular domains obtained by the use of special pulse sequences.*
- *Intensities can be made quantitative (e.g. for polymorph ratios or crystallinity proportions).*
- *NMR is a multinuclear technique, each relevant isotope and element having its own specific frequency range.*
- *NMR experiments can be tailored to produce particular results by suitable choice of pulse sequences.”*

which explains the success of solid-state NMR as a structural investigator, subject of many reviews,^{19,33-49} in several fields of chemistry, life and material science. In the following, I will focus only on the application of solid-state NMR for crystallographic purposes in organic chemistry (in particular, for small organic compounds), although it has been successfully widely applied also in other fields such as bio-⁵⁰⁻⁶⁴ and inorganic chemistry.⁶⁵⁻⁸⁵

Today, the term “NMR crystallography” refers to the approaches that use NMR data in order to determine, refine, select or validate crystal structures.⁴⁸ The structural interest for solid-state NMR was initially developed due to the fact that it provides structural information which is complementary to the ones obtained with XRD. In particular, it was first used to determine the number of inequivalent molecules in the asymmetric unit,⁸⁶⁻⁸⁹ distinguish between polymorphs,⁸⁷⁻⁸⁹ as well as to obtain information difficult to access with XRD such as the characterization of hydrogen bonds^{90,91} or the nature of disorder in crystals.⁹²⁻⁹⁴ This information was mostly used

to validate or refine the crystal structures obtained using single-crystal XRD,^{87-91,93} and in some cases also to aid the refinement of diffraction data.⁹²

The use of NMR in combination with XRD, however, became more prominent only when it was necessary to determine the crystal structures from powdered samples. In these cases, indeed, XRD could still be used for structure determination,^{17,95} but it is far more challenging compared to single-crystal XRD.⁴⁹ The key for the success of powder XRD is strongly correlated to the quality of the starting guess structure, which should be as close as possible to the correct solution, or to the ability to suggest structural constraints.⁹⁶ It was soon realized that this problem could have been solved by combining the powder XRD protocol for structural determination with structural information obtained from solid-state NMR,⁹⁷⁻¹⁰⁰ leading to faster and more accurate structure determination compared to powder XRD alone.⁴⁹

An alternative NMR crystallographic approach consists in using NMR data to perform the crystal structure determination, instead of using it to assist the XRD protocol. Pioneering works in this context were done by Taulelle¹⁰¹ and Elena *et al.*,¹⁰² who showed that NMR can provide information on space groups and on unit cell parameters respectively, setting up the stage for crystallographic approaches entirely driven by NMR. These first examples rapidly led to the first complete protocol for crystal structure determination of small molecular compounds by NMR, introduced in 2006 by Elena and co-workers.¹⁰³ This method involves the combination of molecular modeling and experimental proton spin diffusion (PSD) data, which is correlated with internuclear distances (and therefore the structure). The development of a model to back-calculate PSD buildup curves from structures, allowed, the determination of the crystal structure of β -AspAla starting from a randomly generated ensemble of structures. This structure determined on a powder sample has a root mean square deviation (RMSD) of 0.33 Å to the structure determined with single-crystal XRD.

In the same period, the idea of using chemical shifts for the determination of crystal structures started to develop. This idea was made possible by the introduction of density functional theory (DFT) methods that allow to accurately reproduce the exquisite sensitivity of NMR chemical shifts to the atomic positions.¹⁰⁴⁻¹⁰⁷ This concept was first demonstrated by Yates *et al.*, who explored the agreement between experimental chemical shifts and the chemical shift calculated for the known structure of flurbiprofen.¹⁰⁸ In the following years many works focused on the agreement between experimental and calculated chemical shifts, mainly to refine or validate crystal structures previously determined with XRD or to explore the opportunity to automatically assign experimental chemical shifts.^{100,109-112} In 2006, Harper and Grant have showed that based on the agreement between experimental and calculated chemical shifts the number of candidates selected by CSP as possible experimental structures can be significantly reduced (in their case, the use of anisotropic chemical shift in particular excluded more than 80 % of the CSP's structures).¹¹³ Pickard *et al.* also showed that the agreement between experimental and calculated chemical shifts can be used to evaluate the quality of model structures.¹¹⁴ Starting from the structure of Elena *et al.*,¹⁰³ they determined the crystal structure of β -AspAla with a heavy-atoms RMSD of 0.12 Å, and they found that the RMSD between calculated and experimental ¹H and ¹³C chemical shifts was considerably smaller than that for the initial structure (which has higher positional RMSD). This protocol was then further developed by Salager *et al.* in 2009, with the determination of the crystal structure of thymol.¹¹⁵ In a first stage, a set of model structures was created using the combination of PSD build up curves and molecular modeling previously introduced.¹⁰³ Then, the structures obtained were refined using a DFT geometry optimization, and finally the crystal structure was determined by comparing the experimentally measured and calculated ¹H and ¹³C isotropic chemical shifts. The crystal structure of thymol determined in this way has an all-atoms RMSD of 0.32 Å with respect to the structure determined using single-crystal XRD.

All these progress yielded the development of the first complete protocol for *ab-initio* isotropic chemical shift based NMRX of powdered solids, introduced by Salager *et al.* in 2010.¹¹⁶ This protocol is based on the combination of CSP, DFT chemical shift calculations and the measurements of isotropic chemical shifts. First, a set of candidate crystal structures was created *ab initio* using the CSP protocol introduced by Day *et al.*,¹¹⁷⁻¹¹⁹ starting from the

chemical formula and without any structural hypothesis. Then, they selected the most stable structures within the ensemble of structures generated with CSP, and calculated the chemical shifts using GIPAW DFT.^{104,105} The crystal structure was then determined by finding the structure with the highest agreement between the calculated ^1H and ^{13}C chemical shifts and the experimentally measured ones (not assigned), considering the uncertainty of calculated chemical shifts which was estimated to be 0.33 ± 0.16 ppm for ^1H and 1.9 ± 0.4 ppm for ^{13}C . This protocol resulted in the determination of the crystal structure of thymol with an all-atoms RMSD of 0.29 \AA with respect to the structure determined using single-crystal XRD. Interestingly, the authors showed that this protocol does not necessarily identify the CSP's lowest energy structure as the experimentally present one. The result obtained was also validated using the "more traditional" NMRX protocol based on PSD build up curves.

This approach was then extended by Baías *et al.* to the determination of the crystal structure of four compounds of pharmaceutical interest (cocaine, flutamide, flufenamic acid, and theophylline).¹²⁰ This work is key in the development of the isotropic chemical shift based NMRX approach, not only because it tests the *general applicability of the method to molecular organic solids*, but also because it clarifies many aspects of this approach, shaping the method as it is commonly used today. In particular, it highlights the importance of assigning the chemical shifts (at least partially) to succeed in the crystal structure determination, and it reveals that ^1H chemical shifts appear to provide a much stronger discrimination than the ^{13}C ones, as they appear to be more sensitive to structural changes. Finally, the importance of this work lies also in the fact that it probes the limits of the method, showing that it could not successfully be used in the structural determination of theophylline due to the low amount of inequivalent ^1H chemical shifts. The crystal structures of cocaine, flutamide and flufenamic acid determined using this protocol have all-atom RMSDs with respect to the structure determined using single-crystal XRD of only 0.07 , 0.10 and 0.12 \AA respectively.

The progress made with these last works allowed to reach another milestone in NMRX: the first *de novo* crystal structure determined using isotropic chemical shift based NMRX. In 2013, Baías *et al.* used the isotropic chemical shift based NMRX approach to determine the crystal structure of the form 4 of the drug 4-[4-(2-adamantylcarbamoyl)-5-tert-butylpyrazol-1-yl]benzoic acid (also known as AZD8329).¹²¹ This work clearly shows the power of NMR crystallography by (i) unambiguously determining the crystal structure of the drug, despite the predicted structure being far higher than the lowest energy structure from CSP, and (ii) giving information on a double H-bond motif that connects translationally related molecules along the crystallographic c-axis (not possible with XRD). It should be also noticed that AZD8329 was the largest molecule tackled by this NMRX protocol at the time of the publication of this work, with a molecular weight of 422 g/mol !

Recently, it was shown that this NMRX approach, which has been successfully used for many structural studies up to date,¹²²⁻¹²⁷ has an accuracy at least comparable with single-crystal XRD.¹⁸ The positional uncertainties in crystal structures determined by isotropic chemical shift based NMRX were quantified with a method based on experimental ^1H chemical shifts, molecular dynamics simulations, GIPAW DFT calculations, and machine learning methods. The authors found that the average positional uncertainty in five structures studied with NMRX (cocaine, flutamide, flufenamic acid, AZD8329, and the K salt of penicillin G) yields an RMSD of 0.17 \AA , which corresponds to a gain in positional accuracy of around a factor 2 compared to XRD structure determination.¹⁸

While this section was mostly focused on isotropic chemical shift based NMRX, which is the topic of this thesis, it must be mentioned that this is not the only method currently used in NMR crystallography of molecular solids. Examples of the crystallographic strategies recently developed include, but are not limited to, the use of dipolar couplings,¹²⁸⁻¹³⁰ and chemical shift anisotropy.¹³¹⁻¹³⁵ Finally, another strategy for structure determination that has been widely used over the past few years is based on the combination of solid-state NMR, powder XRD and first-principles calculations, also known as "SMARTER crystallography".¹³⁶⁻¹⁴⁸ This combined approach is particularly robust, as it provides with the determination of a structure that is assessed both against XRD and solid-state NMR data, and it proved to be another powerful method for *de novo* structure determination using NMR.^{149,150} All of the work described in this thesis can of course be integrated into the SMARTER approaches.

1.4 The Isotropic Chemical Shift Based NMR Crystallographic Approach

As mentioned above, the isotropic chemical shift based NMRX approach^{116,120} is based on the combination of CSP, chemical shift calculations (which are traditionally performed using DFT) and experimental NMR isotropic chemical shifts (typically ^1H). In a nutshell, the idea behind this method is to create a (comprehensive) set of candidate crystal structures using CSP, and then to determine the correct crystal structure from within this set by comparing the experimental chemical shifts with the chemical shifts calculated for each structure. The success of this method in determining the crystal structure of molecular solids is given by three main factors: the ability of CSP algorithms to predict the observed polymorph starting from minimal information (such as the chemical formula only),¹¹⁹ the sensitivity of chemical shifts to changes in local atomic environments (^1H in particular),¹²⁰ and the introduction of DFT methods able to accurately calculate chemical shifts in solids.¹⁰⁴⁻¹⁰⁷

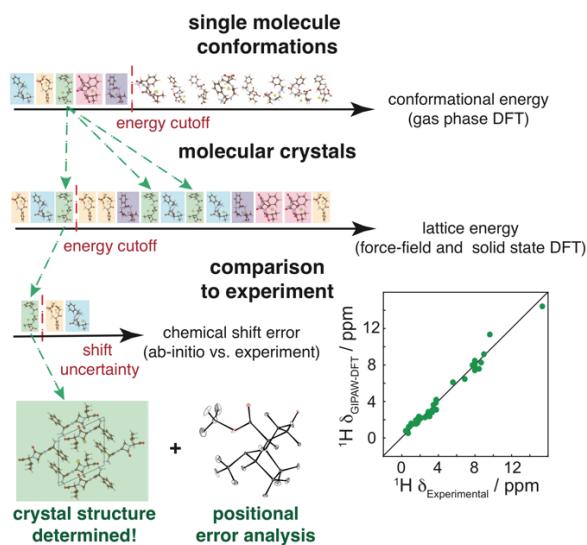


Figure 1.1. Schematic representation of the current isotropic chemical shift based NMRX protocol. The coloured boxes are intended as a guide to the eye, as to which conformer results in which crystal structures. The green crystal structure is the “correct” crystal structure.

The CSP workflow for molecular solids is schematically shown in **Figure 1.1**. It can be divided into three main steps: (i) the exploration of the conformational preferences of the target molecule, (ii) the generation of plausible crystal-packing arrangements and (iii) the ranking of the likelihood of the resulting crystal structures.¹¹⁹

In the first step, the conformational flexibility of the target molecule is explored in the gas phase. A screening of the torsional degrees of freedom generates a comprehensive ensemble of energetically stable single-molecule conformers. This ensemble is then sorted according to the calculated conformational energies and the lowest energy conformers are selected to proceed to the next step, based on empirical cut-off energies. Although flexible molecules often do not assume their lowest energy molecular conformation in their observed crystal structures,¹⁵¹ the assumption here is that low energy crystal structures, including the correct (present) polymorph, will generally result from low energy molecular conformers.

Next, each selected conformation is subjected to a crystal structure search, during which trial structures are generated by varying the degrees of freedom of the unit cell, the number of molecules per asymmetric unit and the position and orientation of molecules in the unit cell. This process can lead to hundreds or thousands of possible crystal structures from each single molecular conformer.

In the final step, the energy of each structure is minimized, typically using atom-atom force fields and DFT.¹¹⁹ This ensemble is ranked by calculated lattice energy and again only the structures below a given cut-off energy are retained. These structures are the candidate crystal structures. Their geometries are further optimized, usually

using periodic boundary DFT calculations, and then their chemical shifts are calculated, typically using GIPAW DFT.^{104,105}

Meanwhile, the ^1H experimental chemical shifts are acquired and assigned. This requires the acquisition of a set of 2D correlation experiments such as ^1H - ^{13}C HETCOR,^{152,153} ^{13}C - ^{13}C INADEQUATE,^{154,155} and/or ^1H - ^1H Double Quantum Single Quantum experiments.¹⁵⁶ The crystal structure is then determined by *finding the candidate that shows the best agreement between the calculated and the experimental chemical shifts*. For this structure, positional errors are then calculated using a molecular dynamics approach.¹⁸

While this method has shown impressive results both for crystal structure validation and for *de novo* crystal structure determination (see **Section 1.3**), there are also severe limitations. From the computational point of view, limitations include the cost of the chemical shift calculations and of the CSP protocol. In particular, for GIPAW DFT calculations the cubic scaling of the computational cost with system size prevents its application to larger and more complex crystals. If one wanted to use more accurate *ab initio* calculations, the expense is currently prohibitive. The computational cost also limits the system size in the CSP protocol. However, in this step the expensive computational cost is cut with selections of a reduced number of candidate structures at each CSP stage. Note, that the computational cost rises sharply when moving from the energy calculations of a single molecule to lattice energy calculations to GIPAW DFT chemical shift calculations, thus requiring the use of severe cuts to make this process affordable. This selection raises a second problem that can lead to failure of the crystal structure determination: the exclusion of the correct structure from the set of candidate structures. This exclusion is particularly likely to happen at the first CSP stage (the conformer selection), where the molecular conformer present in the crystal structure might have high relative energy in the gas phase.

The main limitation from the experimental side is the extraction of the ^1H chemical shift due to the poor resolution of ^1H solid-state NMR spectra (see **Section 1.5**).

The development of novel strategies is necessary to overcome all these drawbacks, and make ^1H isotropic chemical shift based NMRX a more powerful and reliable method.

1.5 High-Resolution ^1H Solid-State NMR

^1H is often considered a perfect nucleus for nuclear magnetic resonance (NMR): it is ubiquitous in chemistry, it has the highest gyromagnetic ratio of all the stable nuclei, and nearly 100% natural isotopic abundance. However, this ensemble of qualities is not sufficient to guarantee success in solid-state NMR. Indeed, the use of ^1H is very limited in solids, where spectra of other spin $\frac{1}{2}$ nuclei such as carbon-13, phosphorus-31 or silicon-29 are more common. This is not because of differences in the information content, but primarily for technical reasons, since the strong ^1H homonuclear dipolar coupling network leads to linewidths in solids orders of magnitude larger compared to those in solution (where dipolar couplings are largely averaged out by rapid molecular tumbling). This broadening masks underlying chemical shifts, significantly reducing the utility of solid-state ^1H NMR spectra.

Figure 1.2a shows the ^1H spectrum of powdered alanine acquired at 3.0 kHz magic angle spinning (MAS) rate. At this spinning regime, the broadening observed in ^1H spectra is of about the same order of magnitude than for a static sample. In this spectrum, the full width at half-maximum height (in the following referred to as “linewidth”) is about 35 kHz, which corresponds to 70 ppm on a 500 MHz spectrometer. This is far larger than the typical ^1H chemical shift range for organic solids, which is about 10/15 ppm.

^1H spectral resolution in solids can be improved dramatically by MAS (**Figure 1.2b**).^{157,158} Due to the strength of the ^1H homonuclear dipolar coupling, and due to the fact that it is a homogeneous interaction in the nomenclature of Maricq and Waugh,¹⁵⁹ low to moderate spinning regimes (<30.0 kHz MAS rates) lead, in general, to ^1H NMR spectra with line narrowing by a factor ≈ 40 -50, but this is not sufficient to alleviate the resolution problem.¹⁶⁰⁻¹⁶³

At this spinning regime, further decoupling can be achieved with the so-called combined rotation and multi-pulse spectroscopy (CRAMPS) approaches,¹⁶⁴ which involve the simultaneous application of MAS and pulse sequences

designed to remove homonuclear dipolar couplings while retaining isotropic chemical shifts.¹⁶⁵⁻¹⁷⁴ This approach yields the best resolution available today for spinning rates up to ~ 65.0 kHz,¹⁷⁵⁻¹⁷⁹ and has enabled a number of methods to characterise structure and dynamics in powdered solids, among which NMRX. A typical ^1H CRAMPS spectrum of a molecular solid, shown in **Figure 1.2c**, has linewidths on the order of 0.3/0.4 ppm.

The recent introduction of probes capable of spinning samples up to 100/150 kHz^{53,180-183} made the acquisition of reasonably high-resolution solid-state ^1H NMR spectra possible with MAS alone, as shown in **Figure 1.2b** for the ^1H 111.0 kHz MAS spectrum of alanine. This approach yields linewidths comparable to those obtained using state-of-the-art CRAMPS methods at lower MAS frequencies,^{172,173,177,184} of about 0.3 ppm for powdered molecular solids.

With these methods it is possible to reduce the linewidths of the proton peaks by orders of magnitude, which is enough to distinguish the isotropic chemical shifts in many organic solids. However, despite the good resolution, *the residual line broadening (which is estimated to be roughly 1% of the original broadening for static organic solids) still limits the use of ^1H NMR in the solid-state*, especially for complex systems. As long as we do not reach sample spinning rates fast enough to provide linewidth comparable to solution-state NMR (which has been estimated to be around 250.0 kHz for proteins),¹⁸⁵ the development of novel strategies to further improve the resolution is necessary to extend the capabilities of NMR applications such as NMRX.

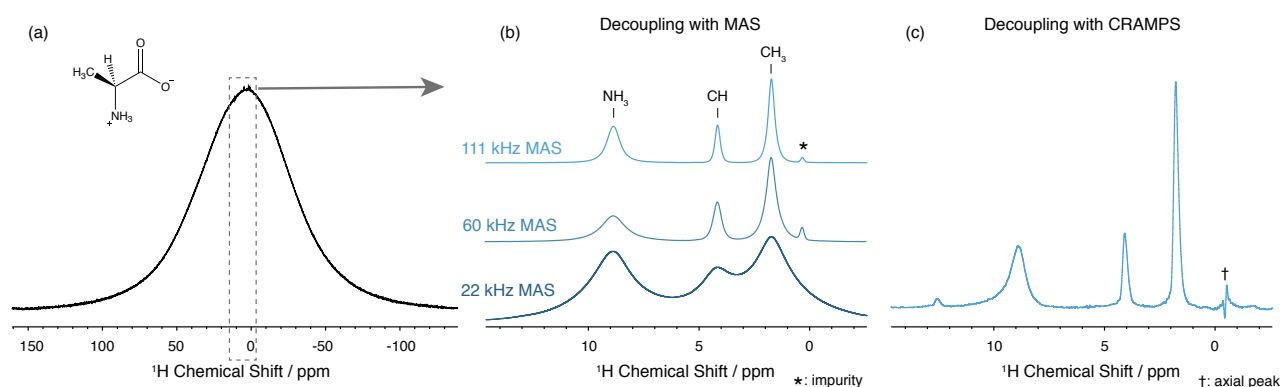


Figure 1.2. ^1H solid-state NMR spectra of powdered alanine. (a) ^1H MAS spectrum acquired with a sample spinning rate of 3.0 kHz (quasi-static regime). In the top left corner of the panel is shown the chemical formula of alanine. The dotted grey box shows the region of the spectrum which is expanded in (b) and (c). (b) Example of homonuclear dipolar decoupling using MAS only. ^1H MAS spectra acquired with sample spinning rates of 22.0, 60.0 and 111.0 kHz. (c) Example of homonuclear dipolar decoupling combining MAS and homonuclear dipolar decoupling sequences (CRAMPS approach). ^1H CRAMPS spectra acquired decoupling with eDUMBO-1₂₂ at 22.0 kHz MAS. Note that while the comparison of these spectra gives an idea of the line narrowing achieved in the different conditions, the resolution are not directly comparable as the spectrum in (a), the 22.0 kHz MAS spectrum in (b) and the spectrum in (c) were acquired at 11.7 T with a 3.2 mm rotor, while the spectra at 60.0 and 111.0 kHz MAS in (b) were acquired at 21.1 T with a 0.7 mm rotor.

1.6 Outline and Scope of the Thesis

In this chapter, I have presented an overview of the role of solid-state NMR in crystallography. I have started by highlighting the power of NMR as structural investigator, and I have briefly introduced some theoretical concepts of this technique, focusing in particular on two NMR observables rich in structural information: the chemical shielding and the dipolar coupling. Then, I have summarized some of the progress that was done by NMR in the crystallographic field over the past years, highlighting particularly the ones that resulted in the development of the isotropic chemical shift based NMRX approach. Finally, I have detailed the protocol followed in this NMRX method, highlighting some of its computational and experimental limitations. During my Ph.D. studies, I have worked mostly on the development of new approaches to overcome the limitations of the isotropic chemical shift based NMRX approach as it was presented in this chapter (in the following I will refer to isotropic chemical shift based NMRX as “NMRX” or “NMR crystallography”). **Chapter 2** and **Chapter 3** are focused on the acquisition of high-resolution ^1H spectra, while **Chapter 4** and **Chapter 5** are mostly about the development of strategies that aim to improve the computational aspects of NMRX.

Chapter 2 focuses on the use of homonuclear dipolar decoupling sequences for the acquisition of high-resolution ^1H spectra of powdered molecular solids. We first investigate the experimental performance of eight homonuclear dipolar decoupling schemes in a low-moderate MAS regime, through the acquisition of 1D CRAMPS spectra and T_2' measurements. We present these results together with step by step guidelines for the optimization of these pulse sequences. We then probe the limits on resolution by measuring transverse dephasing time under the application of homonuclear dipolar decoupling. We find that coherence lifetimes are limited by the appearance of a coherent oscillatory behaviour that leads to a residual anisotropic splitting. We show that the symmetry of the interaction is such that it responds to π pulses in different ways, and that the oscillation can be completely removed in a double spin-echo experiment.

Chapter 3 discusses the development of pulse sequences in the ultra-fast MAS regime ($\nu_R > 100$ kHz). We first show that, for β -AspAla, refocusable and non-refocusable interactions contribute roughly equally to the residual line broadening under 111.0 kHz MAS at high-fields (21.14 T). We suggest that the removal of the non-refocusable part will produce a significant increase in spectral resolution, and we demonstrate an experiment for the indirect acquisition of constant-time experiments at ultra-fast MAS (CT-MAS) which verifies this hypothesis. The combination of this experiment with the two-dimensional one pulse (TOP) transformation reduces the experimental time to a fraction of the original cost while retaining the narrowing effects. Then, we propose a different approach to improve the resolution in this spinning regime, which makes use of the anti-z-COSY pulse sequence, used for broadband homonuclear J decoupling in solution-state NMR. We show that at 100.0 kHz MAS this experiment leads to significant reduction of the residual line broadening of ^1H NMR spectra of powdered organic solids.

Chapter 4 presents an alternative to GIPAW DFT for the calculation of chemical shifts of organic molecular solids. We develop a machine learning method based on local environments to accurately predict chemical shifts of molecular solids to within DFT accuracy. We also demonstrate that the trained model can be used instead of GIPAW DFT calculations in the NMRX approach, showing the results we obtained on the structures of cocaine and AZD8329.

Chapter 5 shows modifications of the traditional NMRX protocol to further extend the applicability of the method. We first propose a CSP method that includes unambiguous experimental constraints during conformer selection, which we extract based on the analysis of absent cross-peaks in solid-state NMR correlation experiments. We then show that this method is able to correctly determine the crystal structure of ampicillin, which would have failed using the traditional approach because it adopts a high energy conformer in its crystal structure. Finally, we propose a Bayesian framework for determining the confidence in the identification of the experimental crystal structure. We introduce a visualization approach for the similarity between candidates in terms of their chemical shifts and their structures to critically assess the reliability of the structure determinations, and we show that in this framework it is consistently possible to use ^{13}C shifts to improve the accuracy of the structure determination.

Chapter 6 summarizes the achieved results and presents a general outlook on future developments for NMRX

Chapter 2 Decoupling Sequences

This chapter has been adapted with permission from:

Paruzzo, F. M. and Emsley, L., "High-Resolution ^1H NMR of Powdered Solids by Homonuclear Dipolar Decoupling". *Journal of Magnetic Resonance* **2019**, 309, 106598. (post-print)

and

Paruzzo, F.M., Stevanato, G., Halse, M.E., Schlagnitweit, J., Mammoli, D., Lesage, A. and Emsley, L., "Refocused linewidths less than 10 Hz in ^1H solid-state NMR". *Journal of Magnetic Resonance* **2018**, 293, 41-46. (post-print)

As mentioned in **Chapter 1.5**, a powerful way to increase the resolution of ^1H spectra is to use homonuclear dipolar decoupling sequences. Just over 50 years ago, the pioneering works of Lee and Goldburg^{165,186} and Waugh and co-workers¹⁶⁶ revolutionized the world of solid-state NMR with the introduction of the Lee-Golburg (LG) and WAHUHA (WHH-4) pulse sequences respectively. Following the principles of coherent averaging,¹⁸⁷ these pulse sequences can remove the broadening effects of homonuclear dipolar couplings in solid samples, while leaving the chemical shift interactions (at least partially) intact. The dramatic line narrowing effect was first demonstrated for WHH-4 with the ^{19}F spectrum of a single-crystal of CaF_2 as reproduced from the original article in **Figure 2.1**. Despite their technical differences, both LG and WHH-4 work in a very similar way: they generate a rotation of the spin operators in the spin space around an effective field tilted of 54.74° with respect to the main static magnetic field. This rotation is achieved with a continuous off-resonance irradiation in the case of LG, and with four $\pi/2$ pulses in the case of WHH-4. The result is that the homonuclear dipolar coupling is averaged to zero (to first order), in a manner similar to that achieved by MAS,^{157,158} while the chemical shift interaction is only scaled by a factor of $1/\sqrt{3}$. Another way to do this was demonstrated with the idea of a "magic-echo" in which a burst of radiofrequency irradiation produces a (partially) time reversed dipolar interaction, and which then leads to the formation of a dipolar echo at some time later.¹⁸⁸

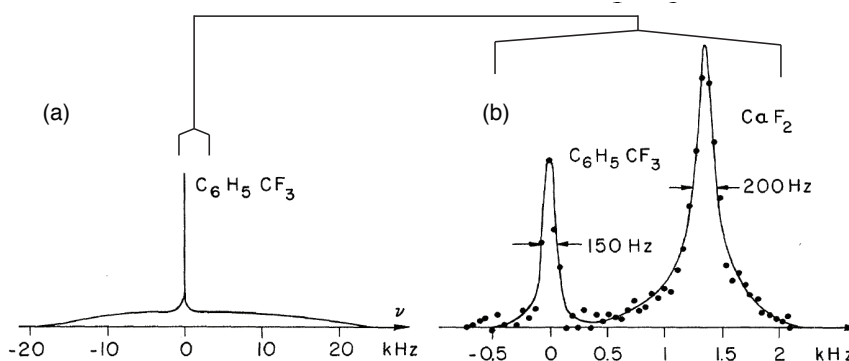


Figure 2.1. First example of the performances of homonuclear dipolar decoupling sequences. (a) ^{19}F spectrum of a single-crystal of CaF_2 wet with liquid $\text{C}_6\text{H}_5\text{CF}_3$. (b) ^{19}F spectrum of the same compound obtained decoupling the ^{19}F - ^{19}F dipolar couplings using WHH-4. Figure adapted with permission from Ref. 166.

While the introduction of LG and WHH-4 was a conceptual landmark, there were two key roadblocks to obtaining truly high-resolution ^1H NMR spectra of solids. First, in static samples, the anisotropy of the chemical shift remains. In certain cases, this can be useful, but in general one would like to acquire spectra from powders with only isotropic shifts. A solution to this is to combine MAS and multi-pulse spectroscopy, in the so-called CRAMPS approach, introduced by Gerstein *et al.* in 1977.¹⁶⁴ As long as the decoupling sequence and the sample rotation do not interfere, CRAMPS leads to a double averaging process that yields high-resolution isotropic ^1H spectra with significantly better resolution than MAS alone (at least so far up to MAS rates of 65.0 kHz).¹⁷⁵⁻¹⁷⁹

The second roadblock is due to the fact that both LG and WHH-4 only lead to the removal of the dipolar coupling under ideal conditions to first order. Even if this means that residual broadening might be only about 1% of the initial width (a remarkable achievement), this would lead to residual linewidths in the early work on the order of ~ 500 Hz (e.g. 1 ppm at 500 MHz), which still severely limits the applicability to complex chemistry. Thus, from these very first developments, the acquisition of ^1H spectra in solids with better resolution has been a challenge faced by many, and which still remains partially unresolved (so to speak).

Using concepts of coherent averaging the WHH-4 and LG blocks were rapidly combined with concepts such as cyclic decoupling and reflection symmetry^{167,168} to develop decoupling sequences with better decoupling efficiency. WHH-4 led to MREV-8 in the 1970s,^{167,188} while the combination of LG with symmetry properties led to flip-flop Lee-Goldburg.¹⁸⁹

The introduction of the CRAMPS experiment then led the development of schemes such as BR-24, BR52¹⁶⁸ and BLEW12¹⁹⁰ using WHH-4 as the basic building block, while LG was developed into frequency-switched Lee-Goldburg (FSLG),¹⁶⁹ which was later implemented as phase-modulated Lee-Goldburg (PMLG),¹⁷⁰ and LG4.^{173,184} In addition to these, many other schemes were developed aiming at better decoupling. For example, extensive analysis of the terms composing the dipolar Hamiltonian with average Hamiltonian theory¹⁸⁷ and their symmetry properties led to the so-called high order truncation (HOT) multiple pulse techniques, such as MSHOT-3.^{191,192} Another elegant example of symmetry considerations applied in the development of decoupling schemes was introduced by Levitt with the development of the rotor synchronized RN_n^V sequences.¹⁹³

A different approach was used for the development of another class of decoupling sequences: the so-called decoupling using mind-boggling optimization (DUMBO) family. These were sequences developed using numerical optimization of continuous phase-modulated radiofrequency (*rf*) irradiation schemes based on either numerical simulations (DUMBO-1)¹⁷¹ or direct experimental optimization (eDUMBO-1_{12.5} and eDUMBO-1₂₂).¹⁷²

In this chapter, we focus on the performance and the optimization of these homonuclear dipolar decoupling sequences. We will start by providing a comparison of CRAMPS experiments using eight homonuclear dipolar decoupling sequences: LG, WHH-4, MREV-8, BR-24, PMLG, DUMBO-1, eDUMBO-1₂₂ and LG4. (We note that FSLG and PMLG are formally equivalent. Here we denote the sequence as PMLG in the following). We assess the experimental properties and performances of these schemes through the acquisition of 1D spectra and T_2' measurements¹⁹⁴ on three different samples (β -AspAla, alanine, and glycine (**Figure 2.2**)) and at three different MAS rates (3.0, 12.5 and 22.0 kHz MAS). To our knowledge, this is the broadest experimental comparison of homonuclear dipolar decoupling sequences provided so far using modern NMR spectrometers.^{172,173,175,184,190,191,195-207}

Then, we provide step by step guidelines for their optimization which in principle can be extended to any other homonuclear dipolar decoupling schemes aiming to facilitate the implementation of these experiments for all NMR users. Indeed, we note that despite being powerful tools, these experiments are still comparatively rarely used in routinely NMR analysis, perhaps because they have acquired a reputation of being technically very demanding. With the technology of NMR consoles and *rf* amplifiers of the 70s-90s, implementing homonuclear dipolar decoupling was an art in itself, requiring careful optimisation along the whole chain of spectrometer components.^{196,208-211} However, with the introduction of modern digital NMR spectrometers at the beginning of

this century and the development of reliable high-performance routine MAS probes there are no longer any significant technical challenges associated with implementing these pulse sequences.

Finally, we probe the limits on resolution of these schemes by measuring transverse dephasing times (T_2')¹⁹⁴ under homonuclear decoupling. We find that coherence lifetimes are limited by the appearance of a coherent oscillatory behaviour that leads to a residual anisotropic splitting. We characterise this behaviour in series of organic solids, where it appears to be general, and we show that the symmetry of the interaction is such that it responds to π pulses in different ways and that can be completely removed in a double spin-echo experiment.

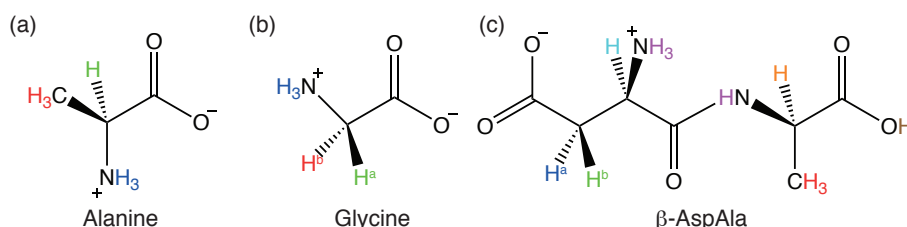


Figure 2.2. Chemical formulae of alanine (a), glycine (b) and β -AspAla (c). Hydrogen atoms are colored in order to correspond to the peaks in the spectra that follow.

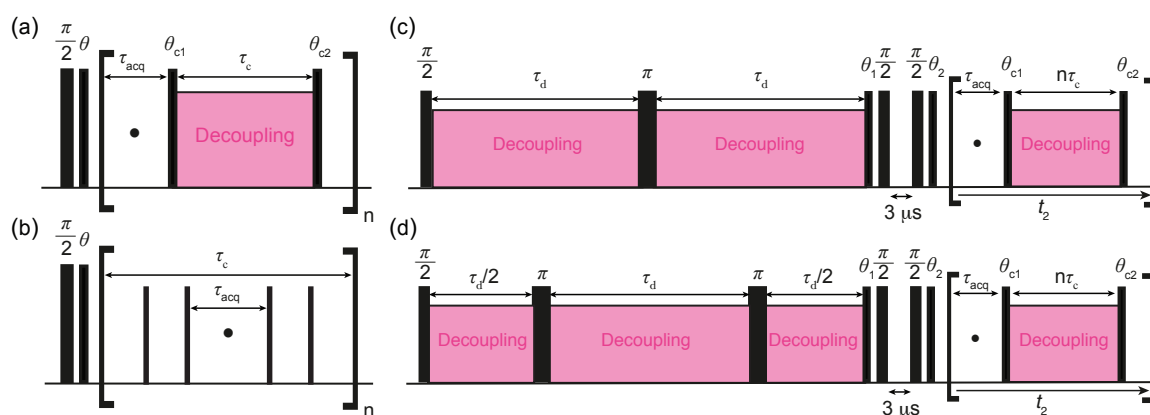


Figure 2.3. Pulse sequences used for the acquisition of: (a-b) 1D 1H CRAMPS experiments with windowed (a) and windowless (b) decoupling schemes (definitions of windowed and windowless schemes is given in **Paragraph 2.1**), and (c-d) T_2' dephasing curves with single echo (c) and double echo (d).

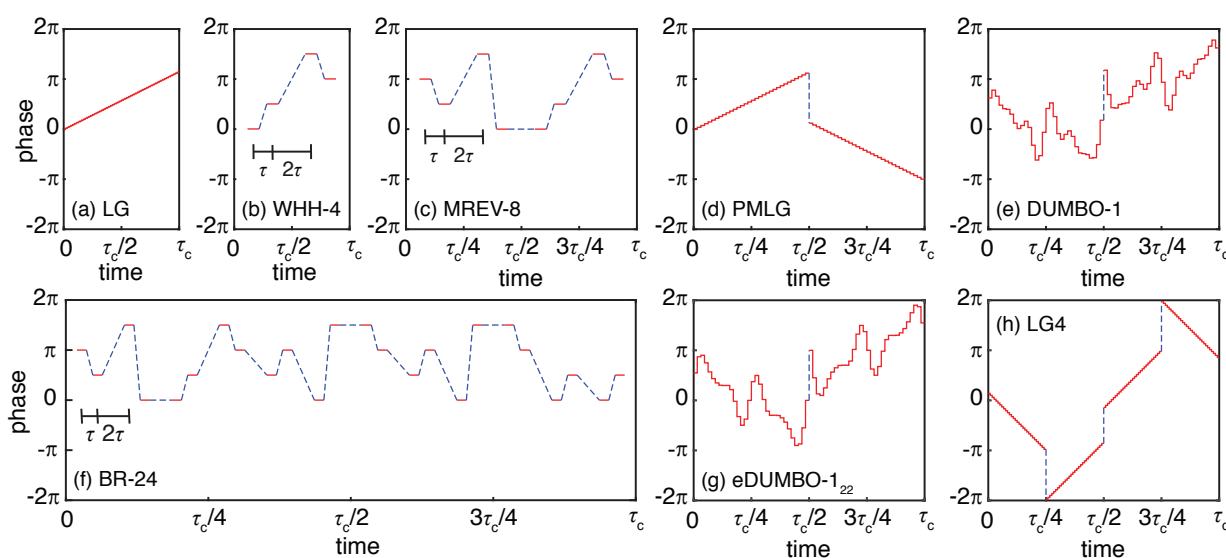


Figure 2.4. (a-h) Phase modulation waveforms for LG (a), WHH-4 (b), MREV-8 (c), PMLG (d), DUMBO-1 (e), BR-24 (f), eDUMBO-1₂₂ (g), and LG4 (h) decoupling sequences. The width of each plot is proportional to the typical length of the corresponding decoupling cycle.

2.1 Materials and Methods

Experiments of **Section 2.2** and **2.3** were performed on a Ascend 500 wide-bore Avance III HD NMR spectrometer, equipped with H/X/Y 3.2 mm CPMAS probe. Experiments of **Section 2.4** were performed on Bruker Ascend 400 and Ascend 500 wide-bore Avance III NMR spectrometers, equipped with H/X/Y 3.2mm and 1.3mm probes. When the 3.2mm probe was used, samples were restricted to the central third of a rotor with inner diameter of 2.2 mm to maximize rf homogeneity. This is done by inserting a 12 μ l Teflon spacer above the sample and below the drive cap, while in the lower part of the rotor the sample is already restricted by the thick bottom wall of the reduced volume 3.2 mm rotor.

1D ^1H CRAMPS experiments were acquired using the sequences shown in **Figure 2.3a-b**. After the initial $\pi/2$ pulse, a θ pulse (the so-called pre-pulse) rotates the magnetization into the tilted transverse plane that will be generated by the effective field of the homonuclear decoupling sequence^{184,208,212} and then the FID is sampled in acquisition windows inserted within the homonuclear dipolar decoupling period. We divide the decoupling schemes into two categories: windowed schemes, which alternate pulses and delays, and windowless schemes, which are continuous pulses. Windowed schemes (WHH-4, MREV-8 and BR-24 - **Figure 2.4b,c,f**) contain delays which allow the insertion of an acquisition window within the decoupling cycle (**Figure 2.3b**),^{166-168,188} while windowless schemes (LG, PMLG, DUMBO-1, eDUMBO-1₂₂, and LG4 - **Figure 2.4a,d,e,g,h**) necessitate the addition of an extra delay for the signal to be sampled in the absence of rf irradiation (**Figure 2.3a**).^{198,213,214} Note that all the sequences implemented here are implemented in the original versions and not in the z-rotation versions.^{215,216} As discussed in **Section 2.3** below, the z-rotation implementation produces a significant reduction of the scaling factor, and therefore resolution. As a consequence we recommend using sequences that are not supercycled, and where the optimization process described below provides spectra which are equally clean. Also note that the LG, FSLG/PMLG, LG4 and the DUMBO type sequences were all implemented in a phase-modulated fashion.

Transverse dephasing times (T_2')¹⁹⁴ shown in **Section 2.2** and **2.3** were measured using the sequence shown in **Figure 2.3d**, those shown in **Section 2.4** were measured using the sequences shown in **Figure 2.3c** and **d**. **Figure 2.3c** shows the single echo pulse sequence introduced by Elena and co-workers.¹⁷² It is a pseudo-2D experiment with the application of a spin echo sequence ($\tau - \pi - \tau$) and with a homonuclear decoupling sequence applied during the τ_d periods. In **Figure 2.3d** we show a double spin-echo sequence. It consists of two consecutive spin echo blocks, with homonuclear decoupling applied during τ_d periods. In both sequences, the phase of the homonuclear decoupling sequence is carefully chosen in order to align the plane of the effective field during decoupling to be perpendicular to the magnetization after the first $\pi/2$ pulse.²¹² At the end of the 2τ evolution a pulse of angle θ_1 is applied, which flips the magnetization from the tilted transverse plane to the (x,y) plane in the rotating frame, and is then followed by a z-filter, and acquisition of the high resolution ^1H spectra using CRAMPS techniques. Note that the sequence used for acquisition of the signal in t_2 has no bearing on the dephasing behaviour studied here occurring during the τ periods. We chose to use BR-24 at 3.0 kHz MAS and eDUMBO-1₂₂ at 12.5 and 22.0 kHz MAS. The dephasing curves were then obtained by plotting the area under each resonance as a function of the dephasing time ($2\tau_d$). Finally, T_2' values were extracted by fitting the dephasing curves to an exponential decay $S(2\tau_d) = a \cdot \exp(-2\tau_d/T_2')$. The homonuclear dipolar decoupling scheme applied during direct acquisition (t_2) and the corresponding parameters were kept unchanged between T_2' measurements performed at the same spinning rate.

Experiments were carried out using rotor spinning rates (ν_r) of 3.0, 12.5 and 22.0 kHz. The parameters of each homonuclear dipolar decoupling scheme were carefully optimized using the procedure described in **Section 2.3**. This optimization was performed independently at each spinning rate for each sample. All chemical shifts were referenced to pure liquid tetramethylsilane (TMS) at 0 ppm. Scaling factors were determined experimentally by comparing the peak positions in the decoupled spectra with those obtained with a single pulse experiment under MAS, except for adamantane spectra where we used the theoretical scaling factors since the two ^1H resonances are not resolved. The comparison was made using the CH_3 and NH_3^+ resonances for alanine (1.13 and 8.54 ppm),

the CH₃ and OH resonances for β -AspAla (0.86 and 12.85 ppm) and the CH₂ peak at lowest chemical shifts and NH₃⁺ resonances for glycine (2.3 and 8.0 ppm).

The processing of the spectra and the calculation of the scaling factors were done using the Bruker program TopSpin 3.5. Extraction of peak positions and experimental ¹H linewidths were done in MATLAB using home-written scripts which make use of the peakfit function.²¹⁷ The post-processing procedures for the T₂' measurements (extraction of 1D spectra, evaluation of the area under each peak, fitting and extraction of T₂' value) were also done in MATLAB.

2.2 Comparison of Decoupling Performance

Figure 2.5, Figure 2.6 and Figure 2.7 show the 1D ¹H MAS (a) and ¹H CRAMPS (b-f) spectra of alanine, glycine and β -AspAla respectively, obtained with WHH-4, MREV-8, BR-24, LG, PMLG, DUMBO-1, eDUMBO-1₂₂ and LG4 at 3.0, 12.5 and 22.0 kHz MAS. Note that the results obtained on the PMLG sequence reflect FSLG performance, as these two sequences are equivalent, except for the way they are implemented at the spectrometer.

We immediately see that all the CRAMPS spectra yield a remarkable improvement in resolution as compared to the MAS only spectra. This has also been shown to be the case at 65.0 kHz MAS for DUMBO type^{176,177} and PMLG type^{175,178,179} schemes if the pulse sequences are adapted slightly. (Note that at 65.0 kHz MAS is sufficiently efficient that the chemical shift scaling factors can be higher than the limit of 1/√3 imposed if the pulse sequence averages the dipolar couplings to zero to first order).^{175-179,218} The improvement in resolution comes at the expense of the sensitivity, which here was reduced by a factor between 4 and 22 compared to the spectra acquired at 22.0 kHz MAS.

We also immediately see that the original LG sequence, perhaps not unexpectedly, yields significantly worse results than any of the other sequences across the board.

All three compounds show the same trends. First we see that among the three “first-generation” multiple-pulse sequences (WHH-4, MREV-8 and BR-24) the decoupling efficiency in the quasi-static regime (3.0 kHz MAS) improves dramatically as we move down the series, from the least (WHH-4) to the most (BR-24) compensated sequence. However, the performance of all three sequences degrades as the spinning speed is increased. This degradation is most marked for BR-24, since the cycle time of BR-24 is significantly longer than for WHH-4 or MREV-8 (**Figure 2.4**) leading to the prediction that it is naturally more susceptible to the time-dependence introduced by faster MAS.

However, we note that BR-24 performs as well as *any* of the other sequences in the 3.0 kHz regime, including the “second-generation” phase-modulated schemes. When we look at the four phase-modulated schemes PMLG, DUMBO-1, eDUMBO-1₂₂, and LG4 we first see that again they all give very significantly better performance than the original LG scheme. Within this group of four sequences, we see that PMLG performs slightly less well than the other three sequences. This is expected since it has been shown that both the DUMBO sequences and (by design) LG4 can be seen as iterated versions of PMLG.^{173,184,204}

Indeed, the LG4 cycle is composed of two successive PMLG rotations around different angles in the xy plane. Thus, the relation between PMLG and LG4 is similar (but not the same) as the relation between WHH-4 and MREV-8. Comparing DUMBO-1, eDUMBO-1₂₂ and LG4 we see that DUMBO-1 tends to yield the best performance of the three at slow spinning (3.0 kHz), which is not surprising since it was developed in the static limit, while eDUMBO-1₂₂ tends to yield the best performance at faster rates (12.5 and 22.0 kHz), and that LG4 tends to yield good performance across the range of spinning speeds.

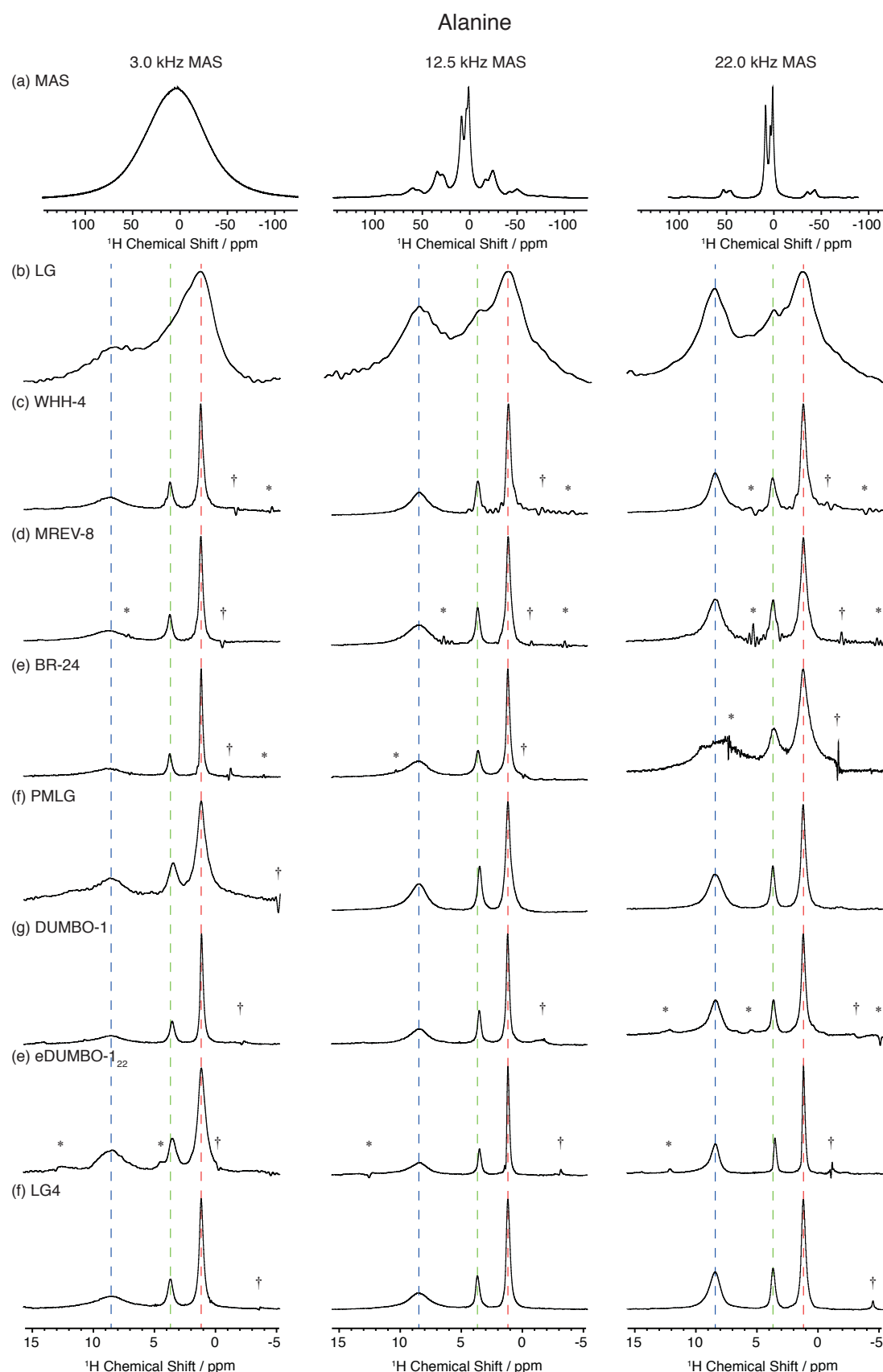


Figure 2.5. 1D ^1H MAS spectra (a) and 1D ^1H CRAMPS spectra (b-i) of powdered alanine. Spectra were obtained decoupling with LG (b), WHH-4 (c), MREV-8 (d), BR-24 (e), PMLG (f), DUMBO-1 (g), eDUMBO-1₂₂ (h) and LG4 (i). Spectra in the left, central and right columns were acquired at 3.0, 12.5 and 22.0 kHz MAS rate respectively. Full details on the experimental parameters are given in **Appendix**. Asterisks denote artefacts in the spectra and crosses denote axial peaks. The assignment is done using the color-code of **Figure 2.2**. Spectra are shown after correction for the experimentally measured chemical shift scaling factor, as detailed in **Section 2.1**.

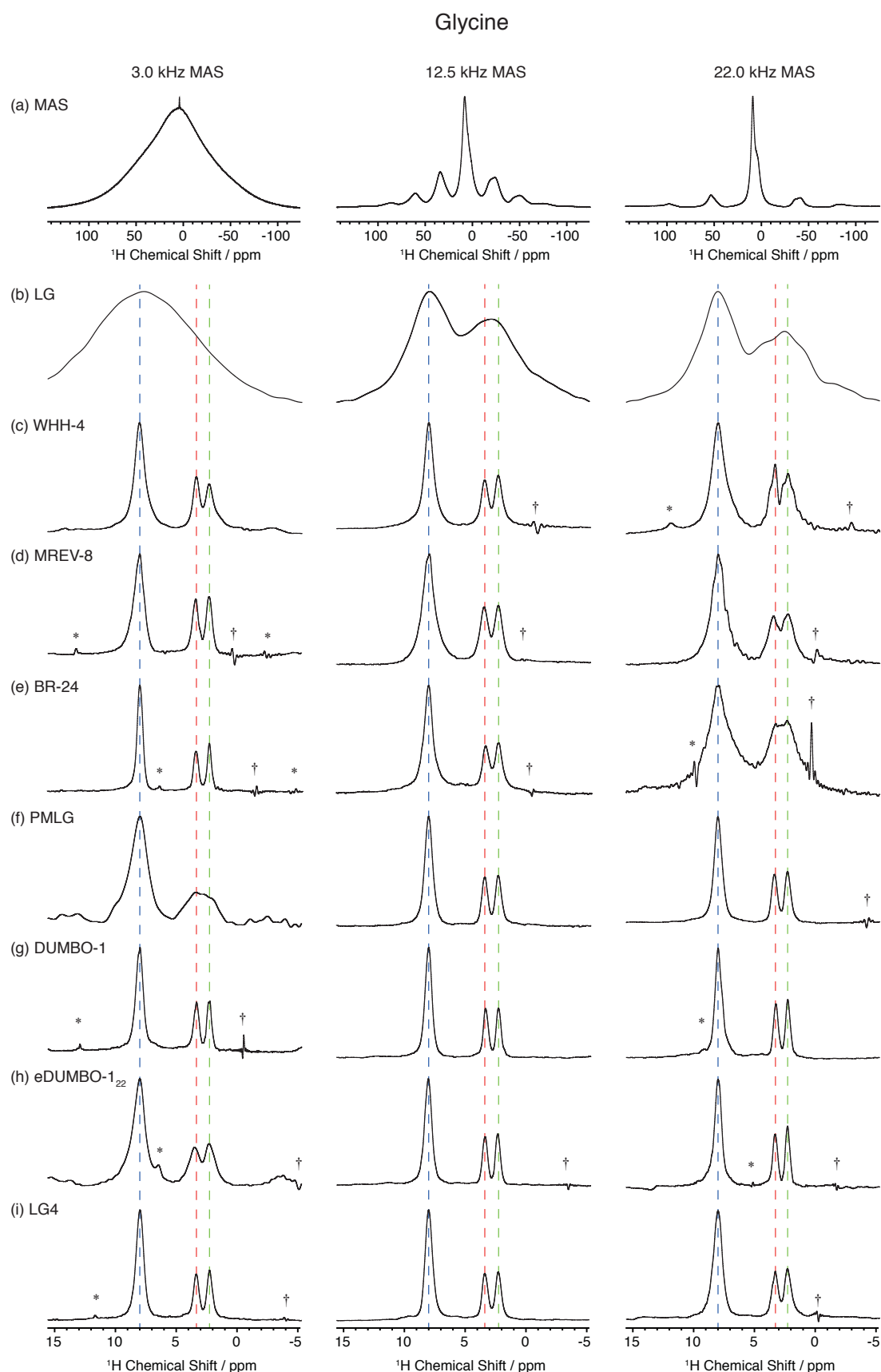


Figure 2.6. 1D ^1H MAS spectra (a) and 1D ^1H CRAMPS spectra (b-i) of powdered glycine. Spectra were obtained decoupling with LG (b), WHH-4 (c), MREV-8 (d), BR-24 (e), PMLG (f), DUMBO-1 (g), eDUMBO-1₂₂ (h) and LG4 (i). Spectra in the left, central and right columns were acquired at 3.0, 12.5 and 22.0 kHz MAS rate respectively. Full details on the experimental parameters are given in **Appendix**. Asterisks denote artefacts in the spectra and crosses denote axial peaks. The assignment is done using the color-code of **Figure 2.2**. Spectra are shown after correction for the experimentally measured chemical shift scaling factor, as detailed in **Section 2.1**.

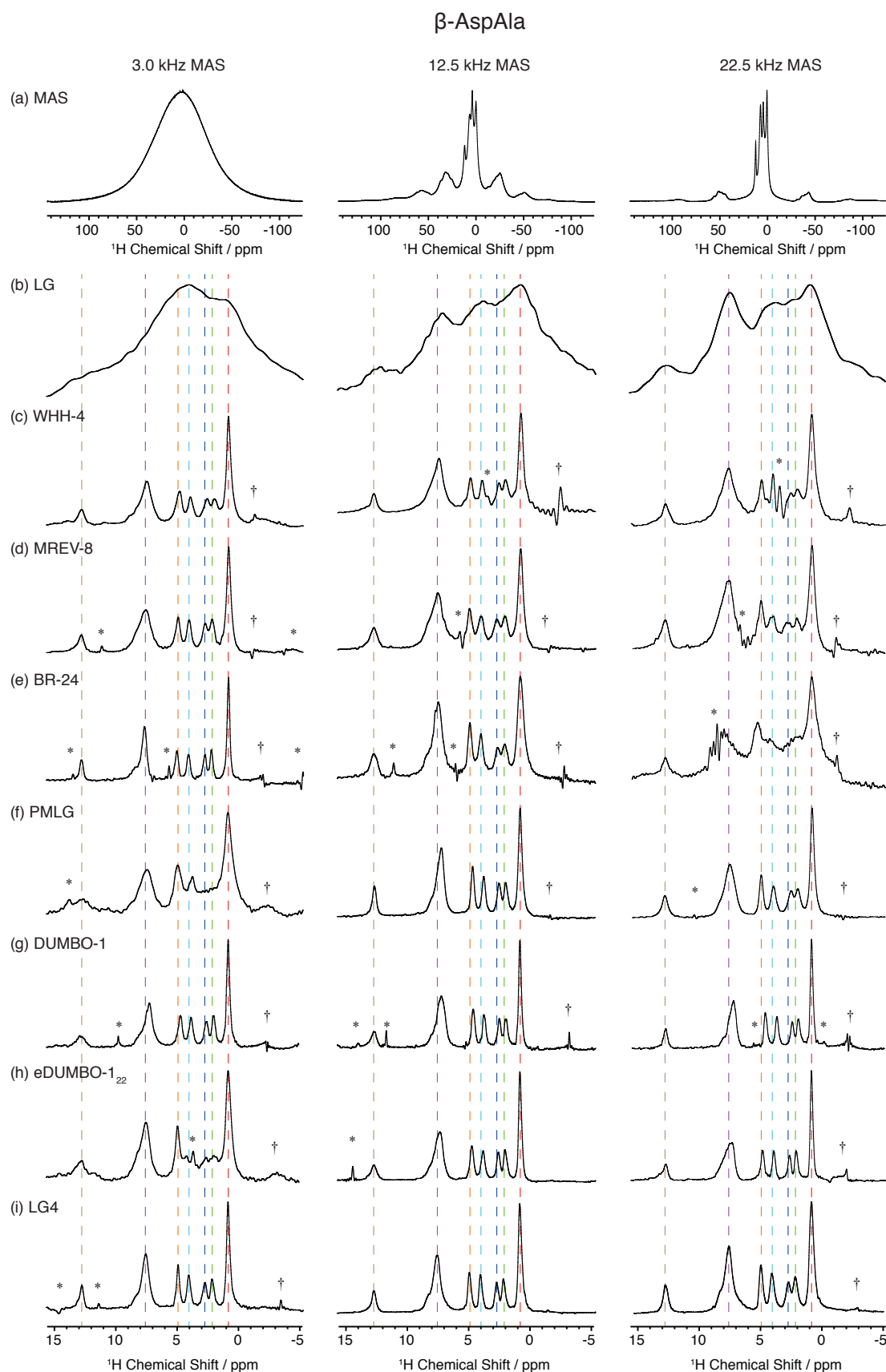


Figure 2.7. 1D ^1H MAS spectra (a) and 1D ^1H CRAMPS spectra (b-i) of powdered β -AspAla. Spectra were obtained decoupling with LG (b), WHH-4 (c), MREV-8 (d), BR-24 (e), PMLG (f), DUMBO-1 (g), eDUMBO-1₂₂ (h) and LG4 (i). Spectra in the left, central and right columns were acquired at 3.0, 12.5 and 22.0 kHz MAS rate respectively. Full details on the experimental parameters are given in **Appendix**. Asterisks denote artefacts in the spectra and crosses denote axial peaks. The assignment is done using the color-code of **Figure 2.2**. Spectra are shown after correction for the experimentally measured chemical shift scaling factor, as detailed in **Section 2.1**.

For alanine (**Figure 2.5**) this results in the best linewidths at 3.0 kHz being (from BR-24) 115, 180 and 1530 Hz for CH_3 , CH and NH_3 respectively. At 12.5 kHz we obtain 104, 150 and 878 Hz with eDUMBO-1₂₂. At 22.0 kHz we observe the best results again with eDUMBO-1₂₂, with linewidths of 113, 117 and 399 Hz. Note that the significant broadening of the NH_3^+ resonance is attributed to interference between the decoupling sequence and the molecular dynamics of the NH_3^+ moiety (rotation around the C-N bond).

For glycine (**Figure 2.6**) we obtain the best results at 3.0 kHz with BR-24 and DUMBO-1. The former yields linewidths of 176, 210 and 241 Hz for CH_{2a} , CH_{2b} and NH_3 respectively, while the latter yields 172, 211 and 315 Hz. At 12.5 kHz we obtain 189, 191 and 282 Hz with DUMBO-1, and at 22.0 kHz 180, 204 and 325 Hz with eDUMBO-1₂₂.

For β -AspAla (**Figure 2.7**) the best results at 3.0 kHz MAS are again obtained with BR-24, which provides all linewidths narrower than 220 Hz (with the CH_3 linewidth of 125 Hz being the narrowest). At 12.5 kHz we observe the narrowest linewidths using DUMBO-1 and eDUMBO-1₂₂, which give very similar results and yield 119 Hz for the CH_3 linewidth. At 22.0 kHz MAS we observe again the best results using eDUMBO-1₂₂, which yields a linewidth of 109 Hz for the CH_3 group.

In order to further understand the residual broadening under homonuclear decoupling, **Figure 2.8**, **Figure 2.9** and **Figure 2.10** show the ^1H T_2' dephasing curves¹⁹⁴ measured using a double spin echo sequence (**Figure 2.3d**) for alanine, glycine and β -AspAla. The T_2' values extracted by fitting these curves are given in **Appendix**. Transverse dephasing times are important both to further probe the mechanisms of decoupling, but also to measure the coherence lifetimes that are directly limiting for the efficiency of many correlation experiments such as J-based HSQC type experiments.²¹⁹

Analogously to the decoupled spectra in the previous figures, we see that the shortest dephasing times under all conditions and across all three samples are obtained with the basic LG (for the first generation) and PMLG (for the second generation) sequences. For the “first-generation” sequences we also see a “textbook” increase in T_2' as we go from LG to WHH-4 to MREV-8 to BR-24 (see **Figure 2.10a**) at the two lower spinning speeds. Interestingly, at 22.0 kHz, all four sequences yield essentially the same T_2' value. The T_2' value consistently decreases for all these sequences as we move to faster spinning speeds, in line with expectations.

For alanine, at 3.0 kHz, the longest measured T_2' for the CH_3 group is 11.2 ms obtained decoupling with BR-24. Note that this is a factor ten longer than the shortest measured value of ~ 1.4 ms with both eDUMBO-1₂₂ and PMLG.

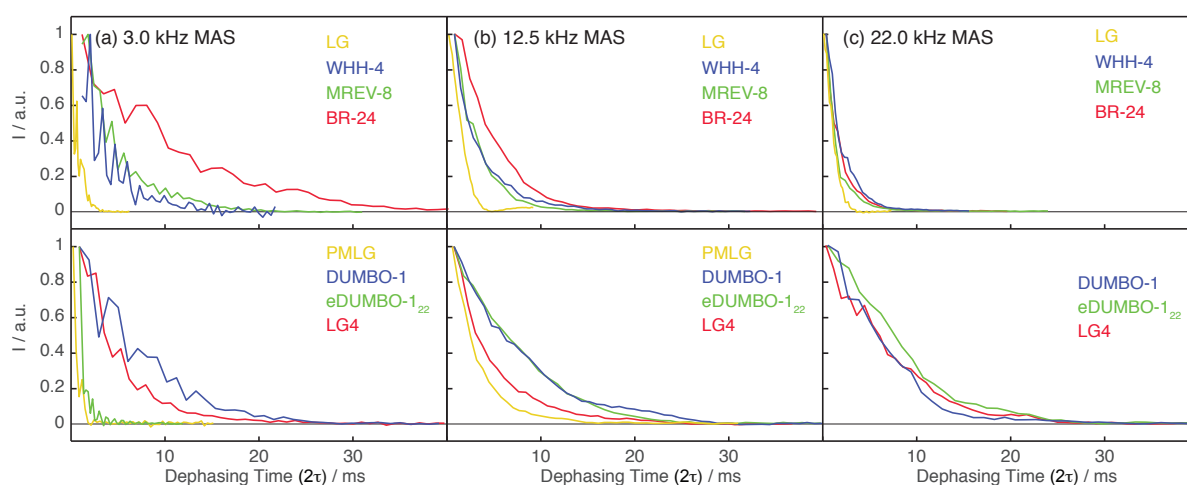


Figure 2.8. Dephasing curves for CH_3 resonance of alanine recorded using the pulse sequence of **Figure 2.3d**. The spectra were acquired using a MAS rate of 3.0 (a), 12.5 (b), and 22.0 (c) kHz. For decoupling we used: (top) LG (yellow), WHH-4 (blue), MREV-8 (green) and BR-24 (red), (bottom) PMLG (yellow), DUMBO-1 (blue), eDUMBO-1₂₂ (green) and LG4 (red). Full details on the experimental parameters are given in **Appendix**.

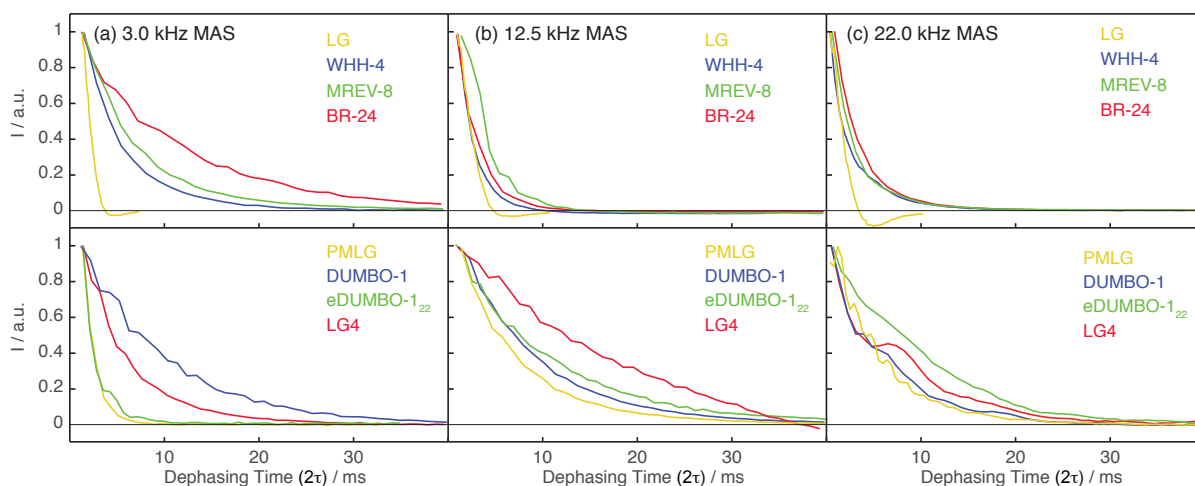


Figure 2.9. Dephasing curves for CH_2 resonances of glycine recorded using the pulse sequence of **Figure 2.3d**. The spectra were acquired using a MAS rate of 3.0 (a), 12.5 (b), and 22.0 (c) kHz. For decoupling we used: (top) LG (yellow), WHH-4 (blue), MREV-8 (green) and BR-24 (red), (bottom) PMLG (yellow), DUMBO-1 (blue), eDUMBO-1₂₂ (green) and LG4 (red). Full details on the experimental parameters are given in **Appendix**.

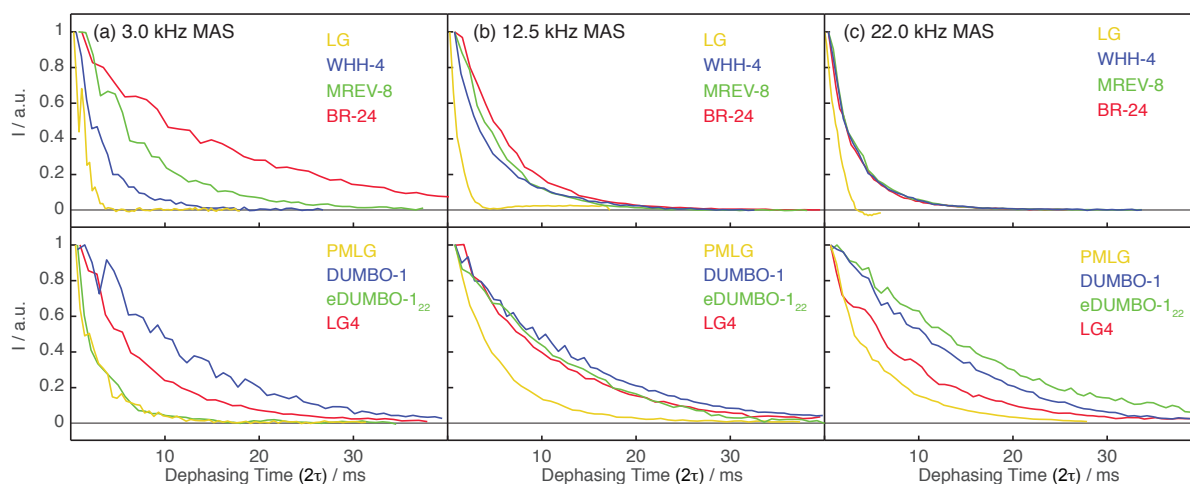


Figure 2.10. Dephasing curves for CH_3 resonance of β -AspAla recorded using the pulse sequence of **Figure 2.3d**. The spectra were acquired using a MAS rate of 3.0 (a), 12.5 (b), and 22.0 (c) kHz. For decoupling we used: (top) LG (yellow), WHH-4 (blue), MREV-8 (green) and BR-24 (red), (bottom) PMLG (yellow), DUMBO-1 (blue), eDUMBO-1₂₂ (green) and LG4 (red). Full details on the experimental parameters are given in **Appendix**.

Interestingly, LG4 at 12.5 kHz MAS yields a T_2' of 13.4 ms, which is the longest we have observed for the CH_3 protons of alanine. This, however, is not directly reflected in the spectral resolution as this resonance is more than 60 Hz broader than in the 12.5 kHz MAS eDUMBO-1₂₂ spectrum. At 22.0 kHz, the longest T_2' values are from eDUMBO-1₂₂ and LG4 (9.7 and 8.4 ms respectively), while the shortest are from the windowed schemes, which all have T_2' values around 2.5 ms.

Similar T_2' results are obtained for glycine and β -AspAla with the same trends. The longest T_2' measured here is 15.3 ms for BR-24 at 3.0 kHz (corresponding to a refocused linewidth of 21 Hz), 13.4 ms for LG4 at 12.5 kHz (24 Hz), and 14.9 ms with eDUMBO-1₂₂ at 22.0 kHz (21 Hz).

By combining the analysis of the spectral resolution and the T_2' values from the three compounds under analysis, we conclude that the best performing sequences in our hands in terms of decoupling performance are BR-24 and DUMBO-1 at 3.0 kHz MAS, and eDUMBO-1₂₂ at 12.5 and 22.0 kHz MAS, which give both the best spectral resolution and the longest T_2' . The better performances of eDUMBO-1₂₂ over the other sequences might be likely attributed to the way this sequence was developed. Indeed, being optimized directly on experimental results, the sequence

might compensate for effects due to pulse imperfections (such as phase transients) and the insertion of the acquisition window. On the other side, we have systematically obtained worse results using eDUMBO-1₂₂ and PMLG at 3.0 kHz MAS, and MREV-8 and BR-24 at 12.5 kHz and, especially, at 22.0 kHz MAS, where they led to very poor resolution and short T_2' . LG4 never yields the absolute best performance, but it is close to optimum across the whole range of spinning speeds and samples. While in general these results are in line with expectations, the excellent performance of BR-24 at low spinning speeds on modern consoles, in direct comparison with the second-generation sequences considered here, was unexpected by us.

As mentioned above, we note some inconsistencies between the spectral resolution and the measured T_2' values. Indeed, some of the sequences we tested, such as PMLG, provide fairly good spectral resolution but comparatively short T_2' , while others, such as LG4, led to poorer resolution than sequences yielding shorter T_2' .

These inconsistencies might be due to the fact that for windowless schemes we have omitted the acquisition windows when performing the indirect T_2' measurements, where these windows do directly have an effect on the resolution of 1D CRAMPS spectra. Also, when comparing T_2' values and spectral resolution we tend to assume that the residual broadening is mainly caused by residual homonuclear dipolar coupling (which is measured by the T_2'). However, this is not necessarily the case, especially in the low-to-moderate regime of MAS. Indeed, it has been shown that also at the highest spinning rate available today (~111 kHz MAS) the interactions refocused by a π pulse (chemical shift distribution) contribute significantly to the line broadening (see **Chapter 3** and Ref. 163). If some of the sequences average also part of the residual interactions that are refocused by the π pulse, then the correlation between T_2' and spectral resolution would not be straightforward.

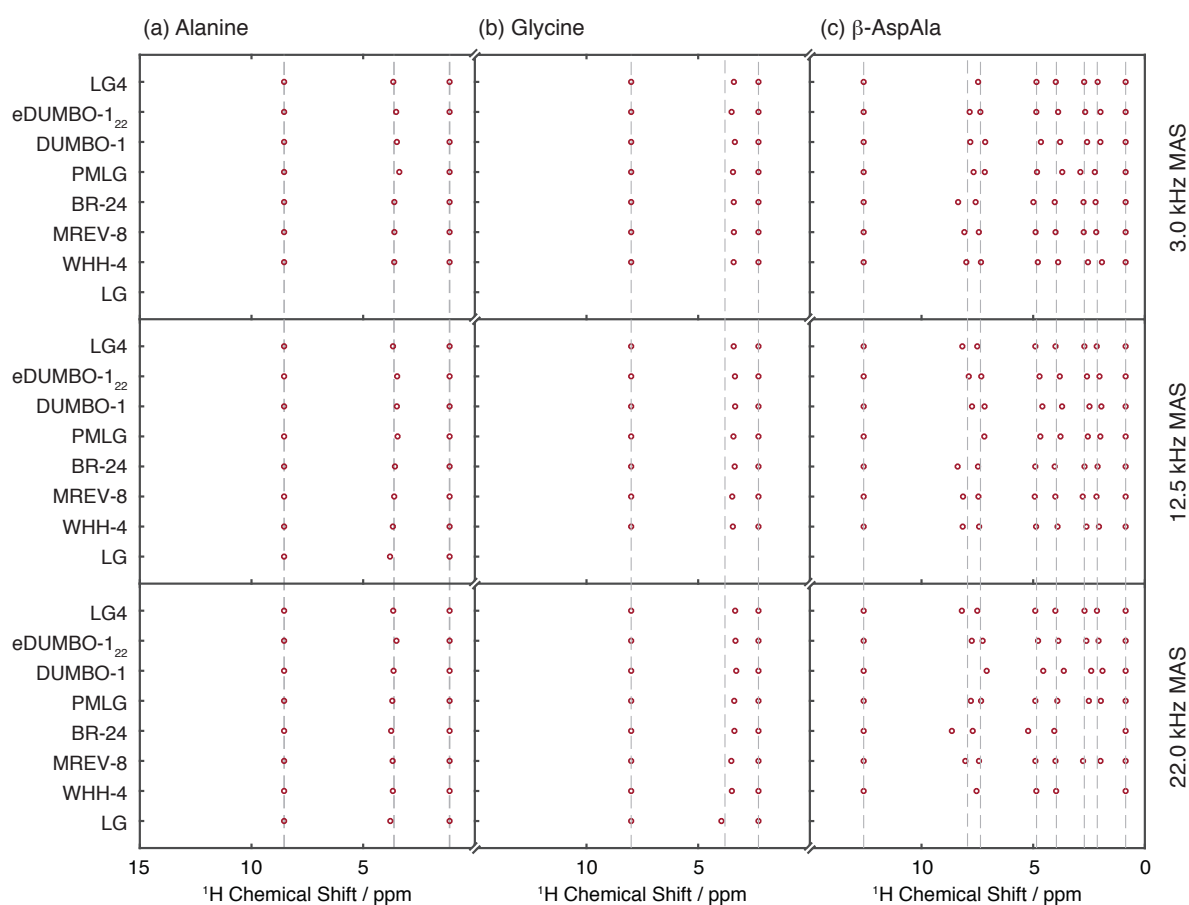


Figure 2.11. Isotropic chemical shifts observed in the ^1H CRAMPS spectra of powdered alanine (a), powdered glycine (b) and β -AspAla (c). The spectra were acquired using a MAS rate of 3.0 (top row), 12.5 (middle row), and 22.0 (bottom row) kHz. The observed isotropic chemical shifts are shown as red dots, while the expected isotropic chemical shifts (obtained with MAS only) are shown as grey dashed lines. The missing points are chemical shifts that could not be resolved due to the low spectral resolution. All observed chemical shift values are given in **Appendix**.

Another important feature of homonuclear dipolar decoupling sequences is the chemical shift scaling factor. Overall, the experimental scaling factors measured here are in agreement with the expected values (details in **Appendix**). More interesting than the absolute value of the average scaling factor is, instead, the homogeneity of the scaling factor across the spectra. When applying homonuclear dipolar decoupling sequences, indeed, the scaling factor is rarely perfectly homogeneous among the whole spectra.^{172,184,216,220-222} As a result, the isotropic chemical shifts of the resonances we observe with CRAMPS differ slightly from the isotropic chemical shifts we obtain using MAS. **Figure 2.11** shows the isotropic chemical shifts values extracted from the ^1H CRAMPS spectra (red dots) compared to the expected values, obtained by decoupling with MAS only (grey dotted lines). The overall RMSE between all observed and expected positions is 0.15 ppm. This is certainly not negligible, and might also be underestimated due to the fact that two out of the three compounds under analysis contain only three resonances. Indeed, if we consider only β -AspAla the RMSE goes up to 0.17 ppm.

At 3.0 kHz MAS, the least good agreement between observed and expected shifts is given by PMLG, with an average RMSE for the three compounds of 0.12 ppm, while the best agreement is provided by LG4 and MREV-8, with average RMSEs of 0.02 ppm and 0.03 ppm respectively. At 12.5 kHz, the poorest agreement is given by DUMBO-1 (average RMSE of 0.12 ppm) while the best match between expected and observed shifts is obtained with LG4, with an average RMSE of 0.05 ppm. At 22.0 kHz MAS, the most accurate decoupling sequence is PMLG with an average RMSE of 0.06 ppm, while the worst results are obtained with BR-24 and DUMBO-1 (average RMSEs of 0.18 and 0.13 ppm respectively).

If we consider all spectra across the three spinning rates, LG4 clearly emerges as the most robust, with the most homogeneous scaling factor (the average RMSE is 0.05 ppm, and it never exceeded 0.12 ppm in any of the spectra) while the most inhomogeneous scaling factors are from DUMBO-1 and BR-24 (average RMSEs of 0.12 and 0.11 ppm).

2.3 Pulse Sequence Optimization

Figure 2.12 shows the procedure we used to set up the 1D ^1H CRAMPS experiments shown in **Section 2.2** and **Section 2.4**. We divide the setup of CRAMPS experiments into two main phases: the preparatory steps (in red in **Figure 2.12**), which consist in a set of guidelines to consider before starting the optimization of the pulse scheme, and the optimization steps (in yellow in **Figure 2.12**), where the experimental parameters are tuned in order to provide the best possible decoupling. Note that this second phase it is not exactly the same for all pulse schemes, as some of them may have extra parameters to consider. We will group the optimization step into two categories: windowed (WHH-4, MREV-8, and BR-24) and windowless (LG, PMLG, DUMBO-1, eDUMBO-1₂₂, and LG4) schemes. Despite some small technical differences, the guidelines for the optimization are general and can be extended to other homonuclear dipolar decoupling schemes not investigated here.

Note that there are a few other examples of guidelines to set up homonuclear dipolar decoupling experiments in the literature.^{184,196,212,223,224} Despite being more focused on specific decoupling schemes, we strongly encourage the reader to consult them to complement the material reported here. In particular, Ref. 223 offers a comprehensive study of experimental aspects for the optimization of LG-based decoupling schemes, and compared to our work it provides additional details on the effects of several experimental parameters on the scaling factor, and studies the dependence of the spectral resolution from relevant spectrometer default timings, such as the transmitter blanking time (*BLKTR*) and the phase presetting time (*PHASPR*).

The overall quality of a CRAMPS spectrum depends on three factors: the spectral resolution, the intensity of the axial peak, and the presence of artefacts.²¹² Ideally, the resolution should be as high as possible, while the axial peak and the artefacts should be absent. To generalize, the spectral resolution is mostly affected by the transmitter offset, the cycle time (τ_c), the length of the acquisition window (τ_{acq} - for windowless sequences only) and the decoupling power level. The intensity of the axial peak is mostly affected by the overall phase angle and the phases and lengths of the pre-pulse (the pulse θ after the initial $\pi/2$) and compensation pulses θ_{c1} and θ_{c2} .

(when present). Note that this separation between parameters that influence resolution and the axial peak is indicative, as there can be significant cross talk. Finally, the intensity of the artefacts can be adjusted by changing the length of the acquisition window, and their position by moving the transmitter offset. However, to obtain spectra completely free of artefacts is not trivial, and it is often achieved in conditions which do not maximize the spectral resolution. As a consequence, it is usually better to adjust the transmitter position so that the artefacts are far from the peaks of interest.

Note that the intensity of the axial peak and the number of artefacts can be lowered (often suppressed) also through the implementation of supercycled sequences.^{215,216} Supercycling here means to apply a phase increment of $2\pi/n$ (where n is usually equal to 2) between consecutive decoupling elements. This leads to the generation of an effective field along the z -axis, which can produce cleaner spectra, with fewer (and smaller) artefacts and lower intensity of the axial peak without further optimization. However, this also produces a significant reduction of the scaling factor, and therefore resolution, which is a major drawback of such an implementation. As a consequence, we recommend using sequences that are not supercycled, and where the optimization process described here provides spectra which are equally clean.

The following step by step protocol for the optimization of CRAMPS experiments makes references to the pulse sequences (**Appendix**) and “au” programs (given at <https://doi.org/10.1016/j.jmr.2019.106598>). These pulse and “au” programs were written for a Bruker Avance III HD console with Topspin 3.5, and might require modifications if used on a different system.

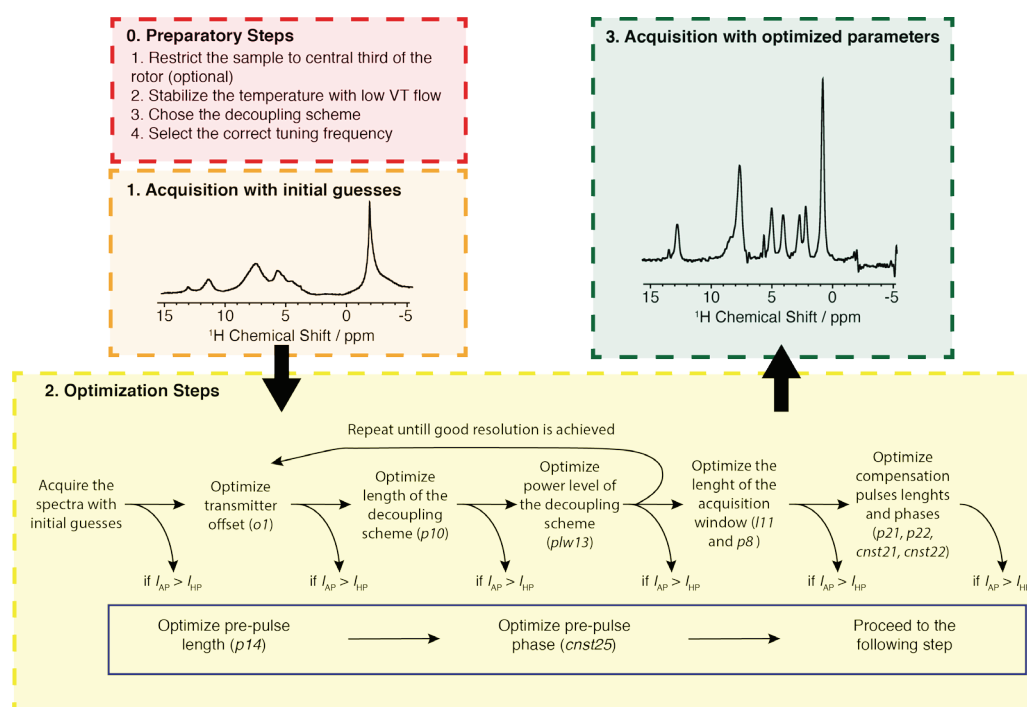


Figure 2.12. Schematic representation of the procedure to setup CRAMPS experiments described in **Section 2.3**. I_{AP} indicates the intensity of the axial peak, I_{HD} indicates the intensity of the highest peak in the spectra (which is not the axial peak).

2.3.1 Preparatory Steps

Good practice for the acquisition of CRAMPS spectra starts when packing the sample. At this stage while this may depend on individual probes, we systematically restrict the sample to the central third of the rotor with the thought that this will reduce sensitivity to rf inhomogeneity. Restricting the sample will not always make a large difference,²²³ but it is never detrimental in our hands. Using inserts with spherical samples has been suggested in the past,^{174,225} but in our hands this has always proven to be very difficult to implement and has not led to observably better performance than simple restriction to the central third, which is what we thus recommend here.

Figure 2.13a shows an example of the gain in resolution obtained by confining the sample, in the case of a spectrum of β -AspAla acquired using eDUMBO-1₂₂. Here, the linewidth of the methyl resonance (at 0.86ppm) decreases from 186 Hz (blue spectrum) to 160 Hz (red spectrum) when adding the Teflon insert, which corresponds to about a 14% gain in resolution.

Once the sample is packed and inserted into the probe, it is important to always actively control the sample temperature with the variable temperature (VT) gas. Indeed, changes in temperature can lead to small changes in probe tuning, which will change decoupling performance. Additionally, the application of high power *rf* pulses leads to additional heating, which can also lead to extra line broadening contributions. The application of active temperature control limits these effects. However, note that using overly high VT flows sometimes leads to micro vibrations of the rotor (visible on the probe tuning wobble curve) which can also affect the quality of the spectra by increasing the quantity and intensity of the artefacts. (All our experiments were done using a VT flow of 600 l/h).

The next step is the selection of the decoupling scheme, in the light of the guidelines given in **Section 2.2** depending on the MAS rate desired. Note that at higher spinning rates than 22.0 kHz, sequences such as PMLG, DUMBO-1, and eDUMBO-1₂₂ have been shown to still provide good resolution (and presumably LG4),¹⁷⁵⁻¹⁷⁷ but the best decoupling performances at 65.0 kHz MAS were so far obtained with sequences specifically developed to work in the very-fast spinning regime such as eDUMBO-PLUS-1¹⁷⁷ or TIMES/TIMES₀.^{178,179}

The next parameter to consider is the probe tuning frequency (ν_t). **Figure 2.13b** shows three spectra of β -AspAla acquired by detuning the probe from 500.43 MHz (the transmitter offset) to 500.46 MHz. We note the strong effect of the tuning frequency on the peak positions, which are significantly shifted to higher frequencies when ν_t increases. This frequency shift has been explained by Vega in terms of phase transients.²²¹ Note that changes of ν_t also affect the spectral resolution. In the spectra of **Figure 2.13b** the linewidth of the methyl resonance goes from 202 Hz in the spectra acquired using ν_t of 500.43 and 500.46 MHz, to 182 Hz using ν_t of 500.45 MHz; about a 10% gain in resolution. To begin, the probe should be tuned at the frequency at which the probe gives the minimum reflected power (as shown by the probe tuning wobble curve). The tuning can then be adjusted toward the end of the optimization to find the ν_t value that gives the highest resolution.

The spectrometer pulse program to use will be different depending on the choice of the decoupling scheme. Windowless decoupling schemes (LG, PMLG, DUMBO-1, eDUMBO-1₂₂ and LG4 – **Figure 2.4a,d,e,g,h**) require the use of a pulse program corresponding to **Figure 2.3a**, where an acquisition window is added between decoupling pulses in order to periodically sample the FID and pulse are implemented directly using quadrature phases. The second-generation decoupling schemes are all implemented as continuous phase-modulated pulses (as is LG). The parameters to select when generating a shaped pulse are: (i) the number of phase steps, (ii) the number of decoupling cycles to repeat between each acquisition window and (iii) the offset. The number of phase steps depends on the pulse resolution of the console in use. The Avance III HD system has a resolution of 0.1 μ s, so a pulse of 10 μ s can be described using up to 100 steps. We found that the number of phase steps used to describe the sequence does not have a significant influence on the resolution, unless too few. Here we used 64 steps to describe each cycle of DUMBO-1, eDUMBO-1₂₂, LG, and PMLG (despite the latter is usually implemented using only 6, 10 or 18 steps each cycle – respectively dubbed as wPMLG3, wPMLG5, and wPMLG9)^{213,226} and 120 steps to describe LG4. The number of decoupling cycles between each acquisition window should be optimized. We found that the best resolution was achieved with 2 cycles for DUMBO-1 and eDUMBO-1₂₂, 3 cycles for LG4 and 1 cycle for LG and PMLG. In the following, we will use the term “decoupling element” to refer to the group of decoupling cycles applied between two consecutive acquisition windows. The last parameter to choose during the generation of the shaped pulse is the starting phase, which changes the direction of the effective field in the xy plane by incrementing the phase of the whole waveform. In the next section, we will describe in detail how to perform the optimization of this phase.

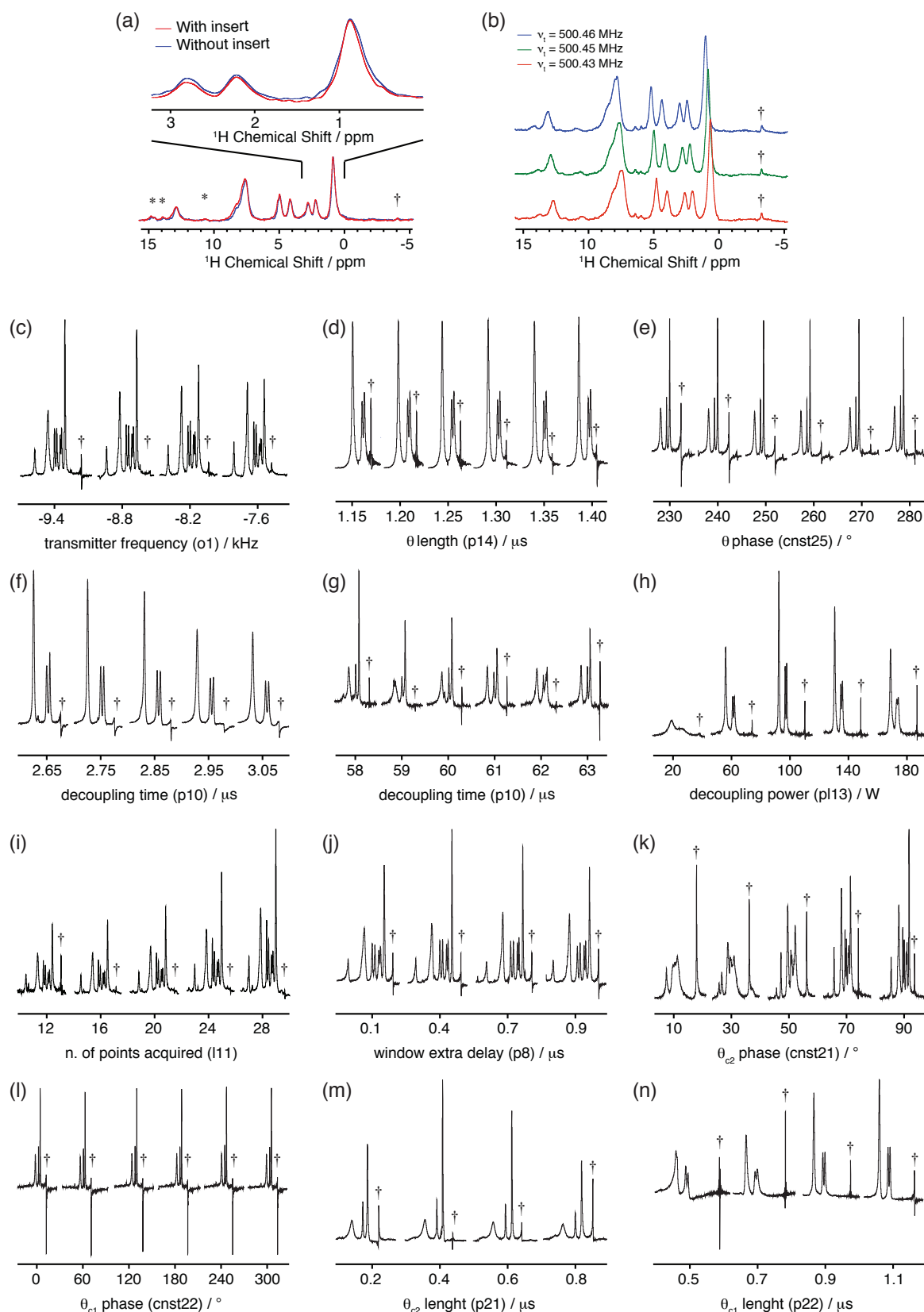


Figure 2.13. Optimization of ^1H CRAMPS experiments. (a) Spectra acquired with (red) and without (blue) a Teflon insert. (b) Spectra acquired with different tuning frequencies (ν_t). (c-p) Examples of decoupling performance as a function of (c) transmitter offset, (d) pre-pulse length, (e) pre-pulse phase, (f-g) length of the decoupling element for windowed (f) and windowless (g) schemes, (h) power level of the decoupling element, (i) number of FID points acquired during each acquisition window, (j) extra delay in the acquisition window, (k-l) phase of the compensation pulses θ_{c2} (k) and θ_{c1} (l), (m-n) length of the compensation pulses θ_{c2} (m) and θ_{c1} (n). Spectra (a-c, i-k) were acquired using β -AspAla, (d, f, h, and n) using glycine and (e, g, j, and m) using alanine. The decoupling scheme was eDUMBO-1₂₂ in (a, b, g, j, and m), LG4 in (c, i, and k), WHH-4 in (d), BR-24 in (f) and PMLG in (e, h, and l). MAS rates were 22.0 kHz in (a, b, e, g, j, k, and l), 12.5 kHz in (c, d, h, m, and n) and 3.0 kHz in (f and i).

Windowed schemes (WHH-4, MREV-8, and BR-24 – **Figure 2.4b,c,f**) require the pulse scheme of **Figure 2.3b**, where the FID is sampled in acquisition windows placed within the decoupling cycle. Note that the presence of an intrinsic acquisition window within the decoupling scheme is an advantage for homonuclear decoupling, since, adding a delay for acquisition may degrade performance. We use one acquisition window per cycle, centered in the first 2τ delay. Note that to maximize the signal to noise ratio it is possible to use multiple acquisition windows per decoupling cycle if a proper post-acquisition FID reconstruction is performed.¹⁹⁹ The overall phase again plays an important role also for these sequences and needs to be adjusted to obtain the correct orientation of the effective field. Since these schemes do not require the generation of a shaped pulse, this is done using the constant *cnst10* in the pulse program, which adds to the phase of all the $\pi/2$ pulses. Further details on the optimization of this phase can be found in the next section.

2.3.2 Optimization Steps

The optimization then starts with the acquisition of a spectrum using guess parameters (for example, which were previously optimized for another system, or alternatively using the set of initial parameters given in **Appendix**) as a starting point. Note that before acquiring the first spectrum it is important for the safety of the probe to adjust the acquisition time. Due to the application of high power *rf* pulses between acquisition windows, the acquisition time should be kept as short as possible for the sample under analysis. In our experiments, we never exceed 35 ms acquisition times, and we typically use about 20 ms. This limitation is not only related to the duty cycle of the probe itself, but also to the fact that the longer the application of high power *rf* pulses, the higher the chance to significantly heat up the sample, introducing extra line broadening effects.

Transmitter Frequency

After the acquisition of the initial spectrum, the optimization of the pulse program parameters should begin with a rough adjustment of the transmitter offset (*o1* in the pulse program, in steps of 500 Hz). As shown in **Figure 2.13c**, the position of the transmitter offset affects both the spectral resolution and the intensity of the axial peak. The aim of this first step is to avoid overlap with any of the resonances of interest, as the transmitter frequency and the zone nearby will appear distorted in the spectrum. In terms of resolution, best results are usually obtained when the transmitter is positioned on the right hand side of the spectrum, as close as possible to the peaks of interest but far enough to avoid distortion of the zone of interest.^{212,221} While optimizing the transmitter frequency, it is also important to ensure that the positions of the artefacts do not overlap with the peaks. The position of the transmitter offset will then be refined in the following steps.

Suppression of the Axial Peak

During the optimization of the transmitter offset, as well as of other parameters that influence the intensity of the axial peak, the axial peak might likely become so intense to overwhelm the spectra itself. In such a case, the adjustment of the pre-pulse length and the overall phase angle became necessary to continue with the optimization. One way to experimentally determine the values of these two parameters was originally introduced by Lesage and co-workers.²¹² Despite being a very precise method to determine both pre-pulse length and offset angle, able in most cases to completely suppress the axial peak, this method has also some disadvantages. In particular, it (i) requires the use of a different pulse program than the one used for the direct acquisition of ^1H CRAMPS spectra, (ii) requires the generation of multiple phase-modulated waveforms when windowless schemes are used and (iii) assumes continuous decoupling without considering the effect of the acquisition window. As a result, the process is time consuming and used as a last resort. The direct optimization of the pre-pulse phase (*cnst25* in the pulse program) is a faster way to suppress the axial peak in 1D ^1H spectra, and usually provides satisfactory results. This optimization can be run directly in the 1D spectra, in concert with the optimization of the pre-pulse length (*p14* in the pulse program) to quickly remove the axial peak (**Figure 2.13d-e**). In the initial stages of optimization, where there is no need of completely suppress the axial peak, the phase can be optimized in steps of $20/30^\circ$ (from 0° to 360°) and the length in steps of $0.2/0.25\ \mu\text{s}$ (from 0 to $1.5\ \mu\text{s}$). The step sizes should then be

decreased later on in the optimization, when approaching the final result. These two parameters should be refined whenever the intensity of the axial peak is higher than the most intense peak in the spectra.

Length and Power Level of Decoupling Element

The following two parameters to optimize are the length and the power level of the decoupling element, which both play a central role in the maximization of the spectral resolution. The optimization protocol is slightly different for windowed and windowless schemes.

Windowed Schemes. In windowed schemes, the length of the decoupling cycle is defined by the delay τ shown in **Figure 2.4b,c,f** (*p10* in the pulse programs). The minimum value of τ is limited by the length of the acquisition window, which is positioned within a 2τ delay in the decoupling cycle, and by the length of the $\pi/2$ pulses, which must be subtracted from the τ and 2τ delays to maintain the symmetry typical of windowed sequences. A good strategy to find the optimal length of the decoupling cycle is to start setting up τ as small as possible, and then to increment it of 2 to 3 μs in steps of 0.05 to 0.1 μs . For the windowed decoupling schemes we test in this work, we found that shorter decoupling cycles provide better resolution, as shown in **Figure 2.13f**. We usually set τ to 0.05 or 0.1 μs larger than the minimum possible value, as the spectral resolution remains almost unchanged, but the intensity of the artefacts decreases. For optimal resolution, the power level for these sequences (*p13* in the pulse program) should be set as high as the duty cycle of the probe allows, due to the fact that the length of the $\pi/2$ pulses limits the length of the decoupling cycle. Besides the effect on the cycle length, we do not observe any significant direct influence of the power level on the spectral resolution (keeping τ constant) with the *rf* amplitudes used here. Note that the calibration of the $\pi/2$ pulse is very important to maximize the performance of these sequences, as miscalibrated pulses provide poorer resolution and lower scaling factors. Once both the length and the power level of the decoupling cycle are set, the transmitter offset should be tuned (in steps of 100/200 Hz) to obtain optimal resolution. For windowed sequences, we found that the transmitter position is relatively independent of the length of the decoupling cycle.

Windowless Schemes. Windowless schemes require more careful optimization of length and power level of the decoupling element. The spectral resolution is strongly influenced by the interplay of these two parameters and the transmitter offset, and a mis-setting of one of these values can dramatically decrease the quality of the spectra. From our experience, the most efficient way to find optimal values for these three parameters is through an iterative optimization of each of them while keeping the other two fixed.

We usually start by adjusting the length of the decoupling element (*p10* in the pulse program) while keeping the power level constant (**Figure 2.13g**). At the beginning of the optimization this parameter can be screened from the smallest possible pulse length (which depends on the number of steps used to describe the shaped pulse) up to a few tens of μs , in large steps of about 5 μs (depending on the decoupling scheme used and the number of decoupling cycles repeated, the length of the decoupling element might also exceed 100 μs). The range of decoupling lengths that provide better resolution in this first screening should be further investigated using smaller steps (down to 1-2 μs). Note that the optimal decoupling time strongly depends on the rotor period (τ_r). In particular, strong recoupling conditions which lead to very poor resolution are observed when the decoupling cycle time and the rotor period are synchronized ($\tau_c/\tau_r = n/2$, where n is an integer).^{212,226,227} While the conditions where recoupling occurs are clear and can be calculated, there conversely is not a known relation of cycle time and rotor period that guarantee good resolution. As a guideline, in our experiments we often observed optimal resolution with τ_c/τ_r around 0.2, 0.7 and 1.2 for windowless homonuclear dipolar decoupling schemes. These are however purely empirical observations and do not guarantee optimal resolution.

Once the length of the decoupling element is set to a value that provides good spectral resolution, the decoupling power level (*p13* in the pulse program) is adjusted. An example of the optimization of this parameter is shown in **Figure 2.13h**. Initially, we suggest screening the decoupling power from low power levels (few Watts) up to the

maximum value allowed by the duty cycle in steps of 10/20 W. Then, the region of power levels that provides best decoupling should be further investigated using smaller step sizes (around 5 W).

Following the optimization of length and power level of the decoupling element, the transmitter frequency should be re-adjusted with smaller step size as compared to the initial optimization (around 100/200 Hz). To find optimal resolution, the optimization of these three parameters should be iteratively repeated decreasing gradually both the screening window and the step size (down to about 0.25 μ s, 2 W and 100 Hz for the length and power level of the decoupling element and transmitter frequency respectively). Note that to reach a satisfying spectral resolution might require several iterative loops, especially when the starting point is far from the optimal conditions. Note also that the combination of decoupling length and power level affects the scaling factor of the sequence, which should be taken into account while finding the optimal values.

Acquisition Window

The next step is the optimization of the acquisition window. On an Avance III spectrometer, the acquisition window is composed of an initial receiver deadtime of 1.6 μ s, a delay of variable length to turn the receiver on and sample n complex points, 0.1 μ s to turn off the receiver, and an extra variable delay. The two parameters to tune are the number of complex points to acquire during each acquisition window (*l11* in the pulse program) and (for windowless sequences only) the delay that follows the receiver closure (*p8* in the pulse program).

Windowed Schemes. In windowed schemes the acquisition window is inserted within the decoupling cycle, limiting its minimum duration. It plays a critical role in spectral resolution, which we found to be remarkably better with shorter cycle lengths (as mentioned above). As a consequence, the acquisition window should be as short as possible. We obtained optimal resolution acquiring 3 complex points per decoupling cycle (*l11*=6). Note that currently for windowed sequences the delay after receiver closure is not a tunable parameter. It has a fixed value of 1.4 μ s, and it is used to centre the acquisition window in the 2τ delay of the decoupling scheme (this delay is encoded in the pulse program and is not under the control of the user).

Windowless Schemes. In windowless sequences the acquisition window is positioned between two decoupling elements, giving a higher degree of freedom for its optimization compared to windowed schemes. However also for these sequences the acquisition windows should be kept short due to the absence of homonuclear dipolar decoupling, which leads to broad spectra when long windows are used. **Figure 2.13i** shows an example of the optimization of the number of complex points acquired after each decoupling cycle. Increasing this value increases the spectral sensitivity, and at the same time affects slightly both the spectral resolution and the scaling factor. Since for organic solids the ^1H sensitivity is usually not a limiting problem, we selected the correct values based solely on the resolution, which is given by the combination of linewidth and scaling factor. For most schemes, we found optimal resolution using *l11* equal to 16 (except for LG4 where the optimal value was 26, probably due to the longer decoupling element). **Figure 2.13j** shows an example of the optimization of the delay after the closure of the receiver. To find the optimal value, we suggest screening this parameter from 0 to 1 μ s in steps of 0.1 μ s. In general, we found the best values for this parameter to be often lower than 0.5 μ s, and that values higher than 1 μ s provide systematically poorer resolution.

Compensation Pulses

This step has to be performed for *windowless sequences only*, which have two compensation pulses (θ_{c1} and θ_{c2}) at the beginning and at the end of each decoupling element. These pulses were originally developed for PMLG and LG4 to remove the chemical shift scaling factor distortion that occurs near the transmitter frequency in 1D CRAMPS experiments.¹⁸⁴ We found that phenomenologically the addition of these two pulses improves the spectral resolution also when using the other windowed decoupling schemes tested in this work.

In their original implementation, these pulses were of the same length, always smaller than 0.5 μ s. In contrast, we found that the best results are obtained when both the lengths (as well as the phases) of these two pulses are

independently optimized. Typically, we start their optimization by fixing the pulse lengths to a relatively short value, about $0.3\ \mu\text{s}$, and by screening over 360° the phase of the θ_{c2} pulse in steps of 5° (this can be done by changing the constant *cnst21* in the pulse program). The value of the θ_{c2} phase affects significantly both the spectral resolution and the intensity of the axial peak, as shown in **Figure 2.13k**. However, the choice of the optimal value for this parameter should be based solely on the spectral resolution, as the axial peak can be suppressed in a later stage by modifying the pre-pulse parameters. The optimization of the phase of the pulse θ_{c1} can be performed in a similar way (changing the constant *cnst22* in the pulse program), but from our experience we found that it has a much smoother effect on the spectral resolution compared to the θ_{c2} phase (**Figure 2.13l**). Note that screening these phases over 360° is not necessary if, initially, the axial peak is adjusted by changing the phase of the decoupling element instead of the pre-pulse. In this case, the correct values for the parameters *cnst21* and *cnst22* should be close to 0, and the optimization can be run in a smaller range (from -30° to $+30^\circ$ in steps of 5°).

Once the θ_{c1} and θ_{c2} phases are set, we focus on their lengths, whose effects on the quality of the spectra are shown in **Figure 2.13m** and **Figure 2.13n** respectively for θ_{c2} and θ_{c1} . This optimization should be performed in an iterative way, changing one pulse length at the time while keeping all other parameters constant. We start by screening the length of θ_{c2} (*p21* in the pulse program) from 0 to $1.2\ \mu\text{s}$ in steps of $0.1\ \mu\text{s}$, followed by the same optimization for the θ_{c1} length (*p22* in the pulse program). Then, this procedure should be iteratively repeated decreasing the step size (down to $0.025\ \mu\text{s}$) until optimal resolution is achieved.

Final Adjustments

At this stage, the optimization of both windowed and windowless decoupling schemes is nearly finished, and the quality of the spectra should be comparable to the ones shown in the section above. The last step is the fine tuning of the pre-pulse length and phase, in order to completely suppress the axial peak. Being the last step, these two parameters should be screened using small step sizes ($0.025\ \mu\text{s}$ and 2° for length and phase of the pre-pulse respectively). The tuning of these two parameters should lead to a high-resolution ^1H spectrum with a minimized axial peak.

Scaling Factor Correction & Referencing

The axis of the acquired spectrum then needs to be corrected for the chemical shift scaling factor. This can be measured by comparing the distance between two peaks in the ^1H CRAMPS spectrum with the distance between the same peaks in the MAS spectrum. The correction of the spectrum can then be done at the spectrometer by dividing the acquisition spectral width by the scaling factor (using the command “s swh” in TopSpin), and Fourier transforming again. The chemical shift values extracted from the MAS spectrum should be used to calibrate on an absolute scale the chemical shift axis of the CRAMPS spectrum.

When it is not possible to resolve at least two peaks in the MAS spectrum, it is necessary to use an external reference for the calculation of the chemical shift scaling factor and for the calibration of the chemical shift axis on an absolute scale. In this case, the optimization of the CRAMPS parameters should be performed on the compound under analysis, and then the CRAMPS spectrum of the reference compound should be acquired under the exact same conditions. Note that the optimization of the CRAMPS parameters on the sample of interest is preferred, however in some cases (such as complex samples, samples with low sensitivity or long relaxation times) it is easier to carry out also the optimization on an external reference (we suggest to use one of the three standard compounds used in **Section 2.2**). Each time an external reference is used, it is important to take into account the influence of the probe tuning frequency on the observed chemical shifts, mentioned in **Section 2.3.1** and shown in **Figure 2.13b**. To avoid inconsistent referencing, it is indeed very important to use the same value of ν_t to acquire the CRAMPS spectra of the compound under analysis and of the reference.

2.4 Limits on Resolution

Figure 2.14a shows the measured spin-echo dephasing curve (in red) for the methyl resonance of β -AspAla recorded under MREV-8 decoupling with 12.5 kHz sample spinning. The same resonance in the corresponding 1D spectra has a linewidth (Δ) of 182 Hz (which yields $T_2^* = 1/\pi\Delta = 1.7$ ms). As shown in **Figure 2.14a** in blue, the measured spin-echo dephasing curve best fits to an exponential decay ($S(2\tau) = a \cdot \exp(-2\tau/T_2')$) yielding $T_2' = 5.95$ ms, which in turn corresponds to a residual linewidth due to coherent non-refocusable interactions (the “homogeneous” interactions accordingly to the definition of Maricq and Waugh)¹⁵⁹ of 54 Hz, as shown in **Figure 2.14b**. The homogeneous interactions (not corrected here for the scaling factor of the sequence, which conceptually should only be applied to the refocusable terms) are thus estimated to be about half of the inhomogeneous contributions (the refocusable ones). This is broadly in line with previous observations.²²⁸

When using the highest performance methods available today, we expect to see a lengthening of the coherence lifetimes, accompanying the reduction in observable spectral linewidths. However, we discovered unexpectedly that at longer coherence lifetimes, *the dephasing curves deviate significantly from simple exponential decays*. As a representative example, **Figure 2.15** shows the dephasing curves for four of the eight resonances of β -AspAla acquired using eDUMBO-1₂₂ at a spinning rate of 12.5 kHz. In all cases the dephasing curves have significantly longer T_2' than for MREV-8, as expected, but they deviate from the exponential case with the appearance of a strongly oscillating component. The Fourier transforms of these curves seen in **Figure 2.15** show that these residual coherent homogeneous interactions present during decoupling lead to a splitting of up to 80 Hz. This behaviour is observed for several decoupling sequences, such as eDUMBO-1₂₂, LG4 and BR-24, over a range of MAS rates between 3.0 and 22.0 kHz. This oscillating contribution is the primary limiting factor for the lifetime of the decay. Removing this residual splitting would lead to line narrowing in the spectra. We note that the 80 Hz splitting here is roughly 0.3% of the typical ^1H - ^1H dipolar coupling between geminal protons. Identifying the origin of the residual splitting from a theoretical standpoint is thus a very challenging problem that will be addressed in the future. However, we have been able to characterise the transformation properties of the interaction from an experimental standpoint.

Figure 2.16 shows the difference between the dephasing curves acquired with the basic single π pulse Hahn echo sequence,²²⁹ and a series of different experiments in which we use two or three π pulses to form an echo with different pulse phases and timings. This is motivated by the idea that while the residual that we observe in **Figure 2.15** is not refocused by a single π pulse, there are many examples in the literature of how permutation in general might lead to better averaging of residuals,^{167,168,174,191-193,230-234} and for π pulses in particular the archetypal example is the Meiboom-Gill modification of the Carr-Purcell echo train experiment.²³⁵

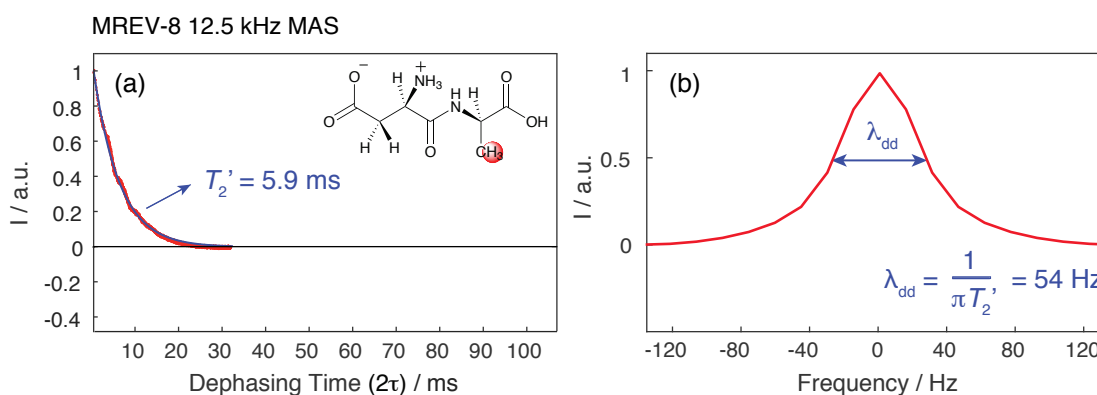


Figure 2.14. (a) Dephasing curve for the CH_3 resonance of powdered β -AspAla recorded using the pulse sequence of **Figure 2.3c**. The spectra were acquired with a MAS rate of 12.5 kHz. During 2τ the MREV-8 sequence was used for decoupling, with 156.25 kHz ^1H rf power (90° pulse length of $1.6\mu\text{s}$) and length of a decoupling cycle (τ_c) of $31.8\mu\text{s}$. A total of 512 2τ values were used, each one acquired with 32 scans. Experimental points are shown in red, and the blue line corresponds to the best fit to an exponential decay, leading to a T_2' of 5.9 ms. (b) Fourier transform of the experimental dephasing curve shown in (a). The full linewidth at half-height is measured to be 54 Hz.

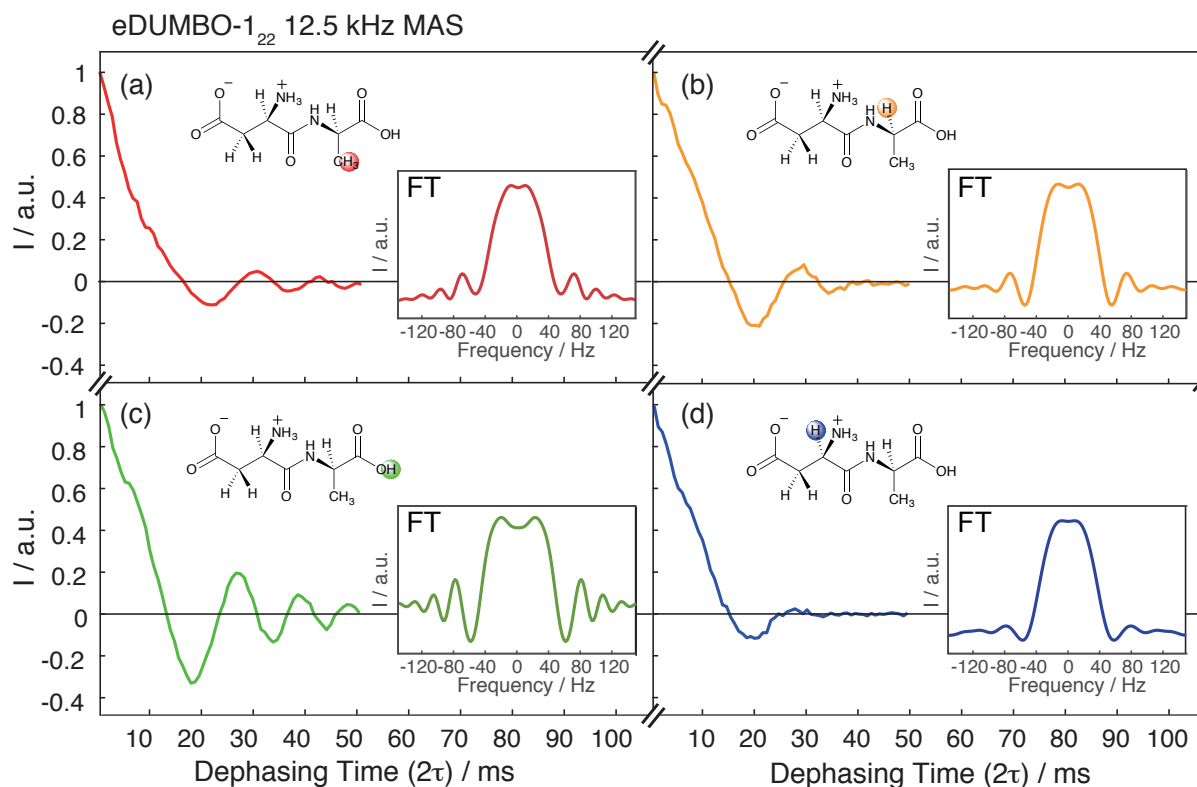


Figure 2.15. Dephasing curves for CH_3 (a) $\text{CH}_{(\text{Ala})}$ (b) OH (c) and $\text{CH}_{(\text{Asp})}$ (d) resonances of β -AspAla recorded using the pulse sequence of **Figure 2.3c**. The spectra were acquired using a MAS rate of 12.5 kHz. During 2τ a 64 phase-step eDUMBO- 1_{22} sequence was used for decoupling. The proton rf power during decoupling was 130 kHz with a time of 24.5 μs for each cycle. A total of 64 t_1 points with 32 scans were collected.

The different experiments will be denoted using a combination of the symbols E_α and π_α , where E and π are used to describe the building block used to construct the sequence, and α represents the phase of the π pulse inside the corresponding building block. E_α represents a Hahn echo building block composed of τ - π_α - τ . The phase α is defined with respect to the phase of the first $\pi/2$ pulse, with x being the same phase as the initial pulse, y a phase shift of 90° , \bar{x} of 180° and \bar{y} of 270° . So, for example, the symbol E_y can be used to represent the sequence composed by $\pi/2_x$ - τ - π_y - τ , or $\pi/2_y$ - τ - π_x - τ . In the same way, it is possible to define the double spin-echo experiment ($\pi/2_x$ - τ - π_y - 2τ - π_y - τ) as E_yE_y , ($\pi/2_x$ - τ - π_y - 2τ - π_x - τ) as E_yE_x and so on. The symbol π_α is a building block composed of only a single π pulse of phase α , with α defined as described above. For example, a sequence composed of two echoes with a π pulse in the middle ($\pi/2_x$ - τ - π_y - τ - π_y - τ - π_y - τ) is described as $E_y\pi_yE_y$.

From the data shown in **Figure 2.16** we first notice that the oscillating component previously introduced in the dephasing curves of **Figure 2.15** is present also for other samples, at different spinning rates, and for decoupling with different homonuclear dipolar decoupling sequences. **Figure 2.16** also shows how the application of a different number of spin echo blocks changes the behaviour of the dephasing curves dramatically. In all cases the double spin-echo E_yE_y sequence (in green) removes the oscillation that leads to negative intensities when the curve is acquired with the E_y sequence. In contrast, if a π pulse is inserted between the two spin echoes (experiment $E_y\pi_yE_y$, in light blue in **Figure 2.16**) the oscillation is re-introduced, but with a frequency which is about half of the original one. The data for triple and quadruple echoes, $E_yE_yE_y$ in orange and $E_yE_yE_yE_y$ in blue, seem to indicate that they also largely remove the oscillating component. Other combinations of echoes and π pulses were tested and are shown in **Section 2.6**.

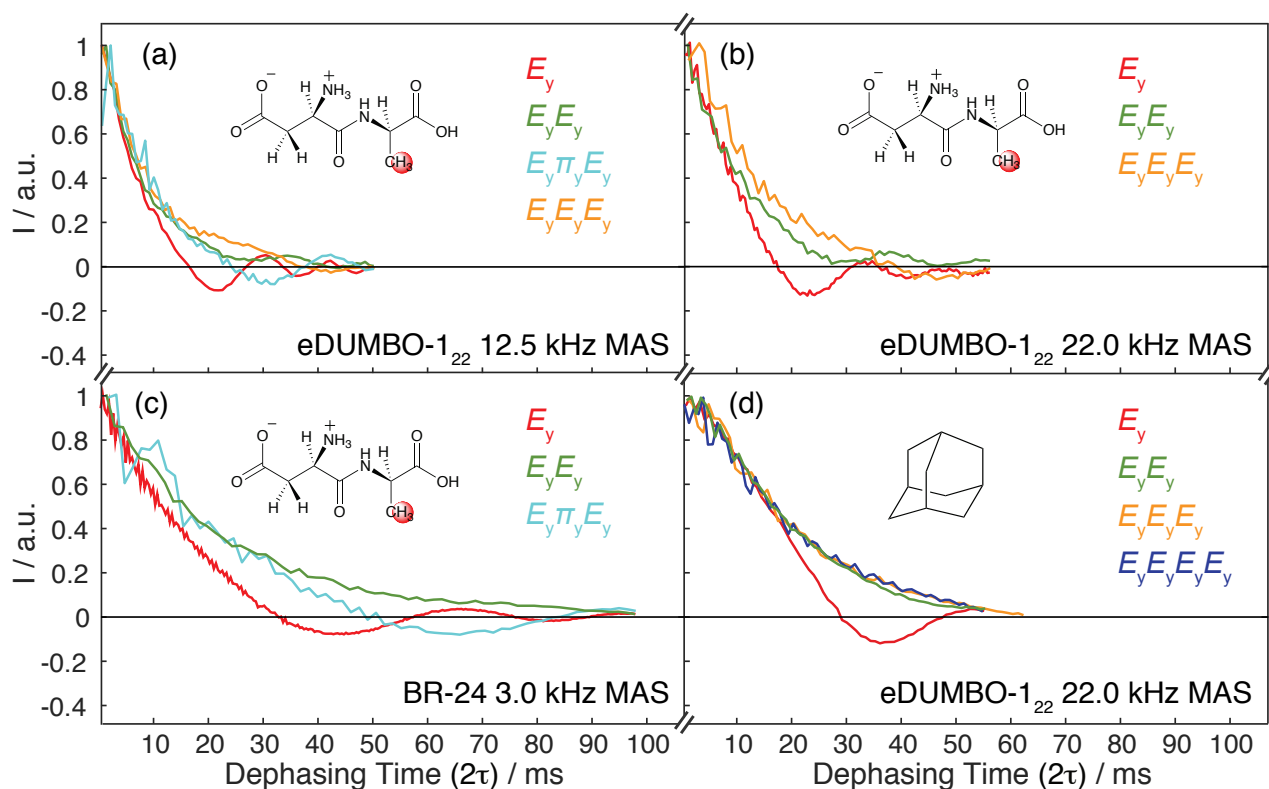


Figure 2.16. Behaviour of dephasing curves for different numbers and timings of π pulses. Results from different experiments are shown in different colours, corresponding to the labels shown and with the nomenclature explained in the text. (a), (b) and (c) are the dephasing curves of the CH_3 resonance of β -AspAla recorded at 500 MHz. (a) and (b) were acquired with eDUMBO-1₂₂ at 12.5 and 22.0 kHz MAS respectively. (c) was acquired with BR-24 at 3.0 kHz MAS. During decoupling ^1H rf power was set to 130 kHz in (a), 125 kHz in (b) and 156.25 kHz in (c). Decoupling cycles length was 24.5 μs in (a), 27.4 μs in (b) and 95.4 μs in (c). (d) shows the dephasing curves of the CH and CH_2 resonances of adamantane acquired at 400 MHz. Here the spectrum was decoupled with eDUMBO-1₂₂ using a MAS rate of 22.0 kHz. ^1H rf power during decoupling was 125 kHz with a time of 27 μs for each cycle.

The use of multiple echo sequences leads to longer apparent decays, which correspond to narrower homogeneous broadening. In **Figure 2.17** we show results that demonstrate the better performance of the double spin-echo E_yE_y sequence if compared to the E_y version. From the Fourier transforms of the dephasing curves it is possible to see the remarkable narrowing in the linewidth obtained with the application of the E_yE_y experiment. In the case of eDUMBO-1₂₂ at 22.0 kHz MAS, the linewidth of the CH_3 resonance of alanine obtained in a 1D experiment, Δ^* , is 112 Hz. Of this residual broadening, the contribution of coherent non-refocusable interactions is 70 Hz as revealed from the $E_y T_2'$ experiment. With the E_yE_y experiment these contributions can be reduced to about half of their original values, leading to a residual broadening of 40 Hz (**Figure 2.17a**). **Figure 2.17b** shows the dephasing curves of the methyl resonance of the β -AspAla, acquired decoupling with BR-24 at 3.0 kHz MAS. In a 1D experiment this resonance has $\Delta^* = 166$ Hz, of which 38 Hz are due to coherent interactions. The application of a double spin-echo leads to a T_2' of about 22.7 ms, which corresponds approximately to a broadening of 14 Hz. The last two curves of **Figure 2.17** were obtained for the CH_2 and CH resonances of adamantane using eDUMBO-1₂₂ at 22.0 kHz MAS using BR-24 at 3.0 kHz MAS. Due to their small difference in isotropic chemical shift, the two resonances cannot be resolved and therefore the dephasing curves are the sum of both. In the 1D CRAMPS spectra we obtained linewidths (Δ^*) of 68 and 86 Hz by decoupling with eDUMBO-1₂₂ and BR-24 respectively. The $E_y T_2'$ experiments reveal that the contribution of coherent homogeneous interactions to the total linewidths corresponds to 36 Hz out of 68 Hz when decoupling with eDUMBO-1₂₂ and 40 Hz out of 86 Hz when using BR-24. In both cases the application of the E_yE_y experiment is able to substantially reduce the amount of non-refocusable interactions. In the first case, the line broadening is reduced to 21 Hz, about half of the starting value. This gain in lifetime is even greater for the second curve, for BR-24, with a residual non-refocusable broadening of only 7 Hz ($T_2' = 45.2$ ms) recorded with the E_yE_y experiment, which corresponds to an increase of more than a factor four as compared to the natural resolution. We note that the difference in magnetic fields used should have little effect on the

refocused linewidths. The difference in performance here should then be related to the intrinsic performance of the two sequences under these conditions.

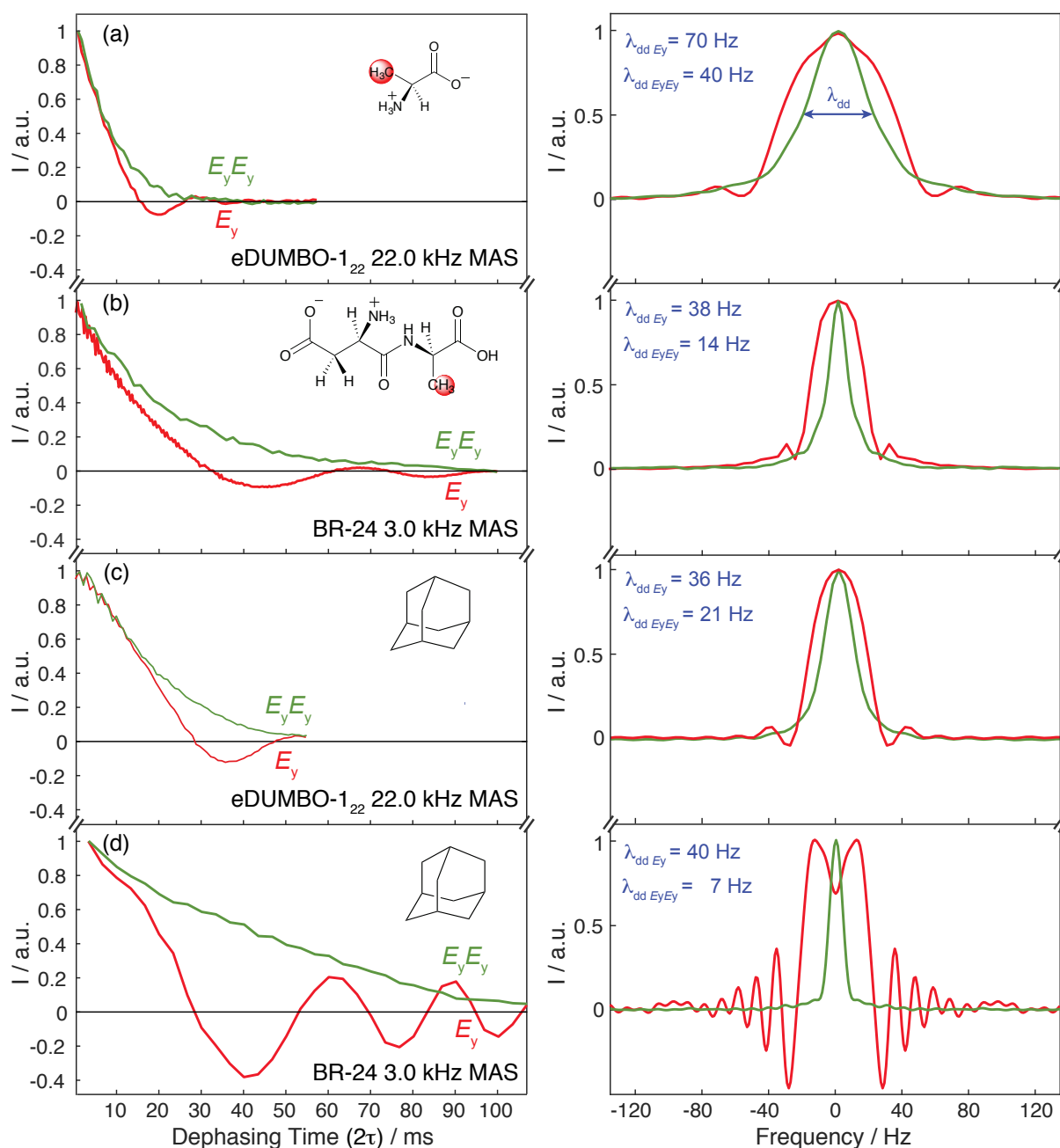


Figure 2.17. The left column shows dephasing curves for the CH₃ resonance of Ala (a), β-AspAla (b) and the CH and CH₂ resonances of adamantane (c-d). Dephasing curves were acquired with the pulse sequences in **Figure 2.3c** (red) and **Figure 2.3d** (green). MAS rate was set to 22.0 kHz in (a) and (c) and 3.0 kHz in (b) and (d). For decoupling in (a) and (c) we used eDUMBO-1₂₂, with ¹H rf power of 125 kHz and decoupling cycles lengths (τ_c) of 27.4 μs in (a) and 27 μs in (c). (b) and (d) were acquired decoupling with BR-24, using 156.25 kHz ¹H rf power (90° pulse length of 1.6 μs) and τ_c of 95.4 μs in (b) and 125 kHz ¹H rf power (90° pulse length of 2 μs) and τ_c of 104.4 μs in (d). A total of 64 τ points with 32 scans were collected for both single and double echo experiments in (a), (c) and (d). In (b) 256 τ points were collected for the single echo and 64 for the double echo, in both cases with 32 scans. The right column shows the Fourier transforms of the corresponding dephasing curves. The curves in (a), (b) and (d) were acquired at 500 MHz, the one in (c) at 400 MHz.

2.5 Conclusions

In this chapter, we have focused on the acquisition of high-resolution ¹H spectra using homonuclear dipolar decoupling sequences. First, we compare the experimental performances of ¹H CRAMPS experiments using eight homonuclear dipolar decoupling sequences (LG, WHH-4, MREV-8, BR-24, PMLG/FSLG, DUMBO-1, eDUMBO-1₂₂,

and LG4) on three different powdered samples (alanine, glycine, and β -AspAla) and at three different MAS rates (3.0, 12.5, and 22.0 kHz). We compare these sequences by analysing spectral resolution, T_2' values and homogeneity of the scaling factors. We show that the best spectral resolution and the longest T_2' values are given by BR-24 and DUMBO-1 at 3.0 kHz MAS, and eDUMBO-1₂₂ at 12.5 and 22.0 kHz MAS. LG4 never yields the absolute best performance, but it is close to optimum across the whole range of spinning speeds and samples and is the sequence that gives the most homogeneous scaling factor (with an average RMSE between chemical shifts observed with LG4 CRAMPS and with MAS experiments of 0.05 ppm). Then, we provide a detailed set of guidelines to optimise the performances of ^1H CRAMPS experiments with the aim of making these experiments accessible to all NMR users.

Finally, we have explored the current limits of ^1H resolution in NMR of organic solids at natural abundance. To better understand the contributions to the residual line broadening, we perform a set of T_2' measurements with multiple echo configurations. The measurements are performed on different samples of powdered organic solids (β -AspAla, alanine, and adamantane), decoupling with different schemes over a range of MAS rates between 3.0 and 22.0 kHz, and at two different ^1H frequencies (400 and 500 MHz). We discover that the coherence lifetime, and consequently the spectral resolution, is limited by the presence of an oscillation when the experiments are done using homonuclear dipolar decoupling sequences. We observe this oscillation for the sequences which provide the best resolution in 1D ^1H spectra, such as eDUMBO-1₂₂ and BR-24. This oscillation produces a residual anisotropic splitting of up to 80 Hz, a significant contribution to the total observed linewidths which are on average about 100-150 Hz for organic solids. The study of the behaviour of this oscillation under the influence of a higher number of π pulses reveals that it can be removed by the application of a double spin-echo ($E_y E_y$) experiment. As a consequence, the coherence lifetime is extended, and line broadening due to non-refocusable interactions is significantly reduced. In summary, we were able to characterize a limiting factor for ^1H resolution in solid-state NMR, and we were able to remove it by using a double spin-echo sequence (**Figure 2.3d**), reducing the contribution of non-refocusable interactions to as little as 7 Hz.

2.6 Supplementary Information and Discussion

2.6.1 Supplementary Information Section 2.4

Single and double spin-echo experiments with LG4 decoupling

Figure 2.18 shows further examples of the application of single and double spin-echo experiments. These dephasing curves were obtained decoupling with LG4. All of them are relative to the CH resonance of the alanine residue of β -AspAla ($\text{CH}_{(\text{Ala})}$). LG4 decoupling was applied using ^1H RF power ($\omega_{\text{RF}}/2\pi$) of 110 kHz (a), 130 kHz (b), 150 kHz (c) and 170 kHz (d) while rotor spinning frequency was 12.5 kHz in each case. Experiments were acquired at 500 MHz, using a H/X 1.3 mm probe.

T_2' Behavior Under Different Echo Sequences

Figure 2.19 shows the effect on the dephasing curve of the number of π pulses during the 2τ period and their relative phases. The curves acquired with the same phase for all π pulses and the ones which have a phase shift of 180° between π pulses behave similarly. While a different behavior was observed when the phase of the second π pulse was shifted of 90° with respect to the others. All curves were acquired on a 400 MHz spectrometer, with a rotor spinning frequency 22.0 kHz MAS, and decoupling with eDUMBO-1₂₂.

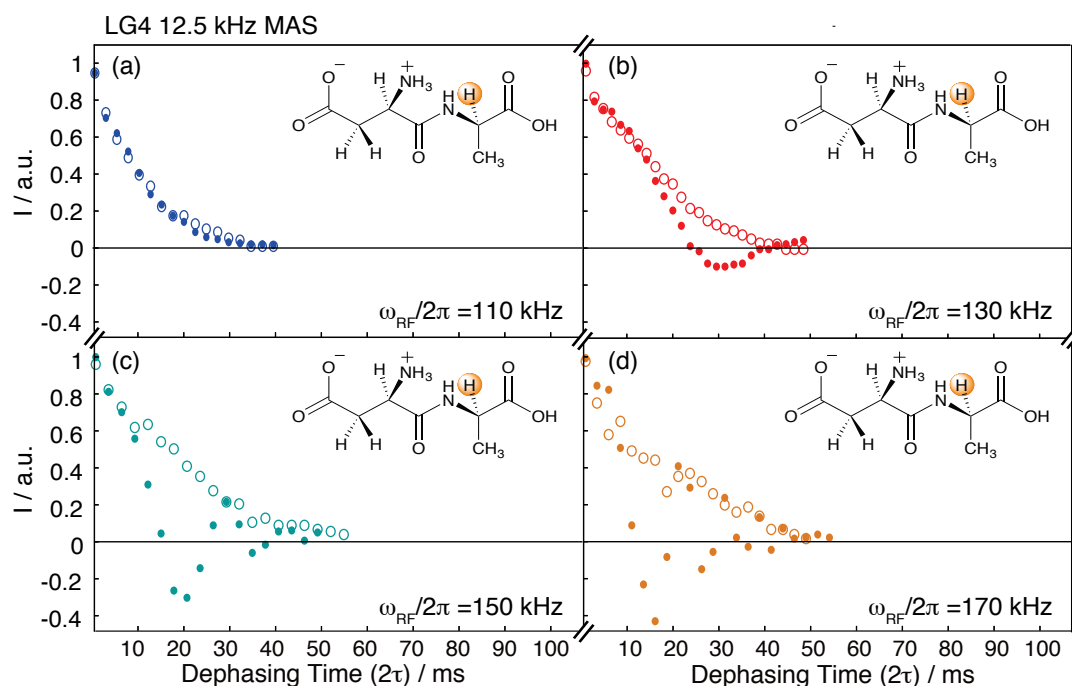


Figure 2.18. Examples of the application of single and double spin-echo experiments where LG4 decoupling is applied during τ periods. Open circles represent data obtained using the E_y sequence, filled circles data obtained using the $E_y E_y$ experiment.

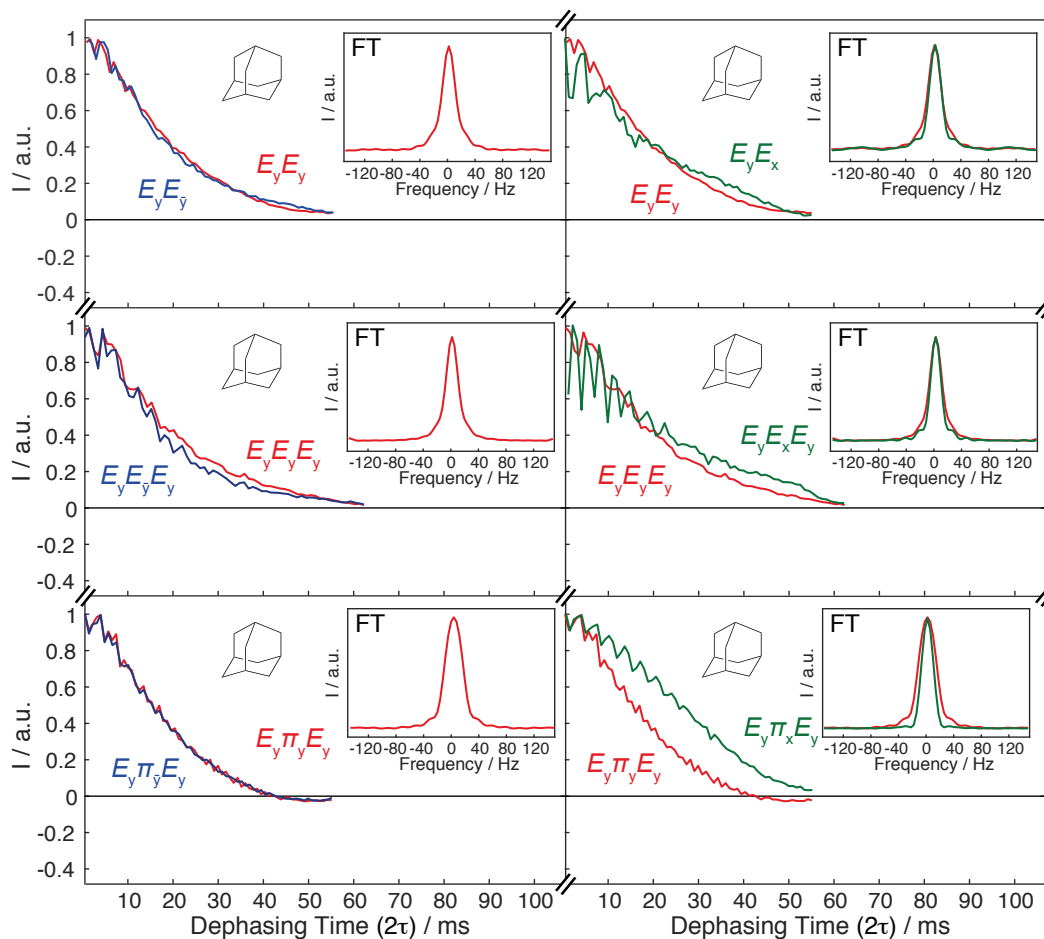


Figure 2.19. Dephasing curves of adamantane acquired using two or three π pulses combined together to form echoes with different pulse phases. The sequence used for the acquisition of the curves is reported in each plot, following the nomenclature introduced in Section 2.4.

Chapter 3 Decoupling at Ultra-Fast MAS

This chapter has been adapted with permission from:

Paruzzo, F.M., Walder, B.J. and Emsley, L., "Line Narrowing in ^1H NMR of Powdered Organic Solids with TOP-CT-MAS Experiments at Ultra-Fast MAS". *Journal of Magnetic Resonance* **2019**, 305, 131-137. (*post-print*)

and

Moutzouri, P., Paruzzo, F.M. and Emsley, L., "Homonuclear Decoupling in ^1H NMR of Solids by Remote Correlation", article in preparation **2019**. (*pre-print*)

Thanks to the introduction of probes capable of spinning samples up to 100/150 kHz,^{53,180-183} the use of MAS alone has become a powerful alternative to CPMAS experiments for the acquisition of high-resolution ^1H spectra. Key advantages over the use of homonuclear dipolar decoupling sequences are: (i) higher sensitivity, due to the fact that the FID acquisition is not interrupted by the application of *rf* pulses, (ii) easier experimental set up, as the optimization of the decoupling sequence is not required (see **Chapter 2.3**), (iii) broader spectral width due to the absence of chemical shift scaling factors, and (iv) lower duty cycle, which results in lower chances of damaging instrumentation, lower heating of the sample (which yields line broadening, see **Chapter 2**), and gives the opportunity to acquire longer FIDs.

As discussed already in **Chapter 1.5** and shown in **Figure 1.2**, the resolution obtained at 111.0 kHz MAS is comparable to that obtained using state-of-the-art CPMAS methods at lower MAS frequencies,^{172,173,177,184} and the residual line broadening is still the main bottleneck to more a widespread use of ^1H NMR in the solid-state.

In this chapter, we focus on the development of methodologies to improve the resolution of ^1H spectra in the ultra-fast MAS regime (>100.0 kHz MAS). So far, little to no advantage has been reported using CPMAS approaches at MAS rates higher than 65.0 kHz, and given the disadvantages mentioned above, the development of alternative strategies to further increase the resolution should be explored. To this end, here we develop two new approaches that yield line narrowing in ^1H spectra of molecular solids acquired at MAS rates > 100.0 kHz.

The first involves the use of constant-time acquisition in the indirect dimension of a 2D experiment to remove residual non-refocusable broadening from the spectrum, based on the idea proposed by Lesage and co-workers.²²⁸ We show that this method yields up to 40% enhancement in resolution compared to the one-pulse acquisition experiment at the same MAS rate. We also show that this method can be combined with the TOP processing, which makes use of affine transformations on the 2D dataset to swap the spectral windows of the direct and indirectly acquired dimensions, allowing the indirectly detected signals to inherit the large spectral window of the directly acquired dimension.²³⁶⁻²³⁸ The combination of these two methods (the TOP-CT-MAS experiment) yields peaks up to 30% narrower than in the one-pulse experiment, acquired in less than 10 minutes.

The second method makes use of the anti-z-COSY experiment,^{239,240} which is used in solution-state NMR for decoupling homonuclear scalar couplings. We show that this experiment can also be used to reduce the residual line broadening of ^1H NMR spectra of powdered organic solids. Results obtained with the anti-z-COSY sequence

at 100.0 kHz MAS on three different samples (thymol, β -AspAla, and strychnine) show an improvement in resolution up to a factor of two compared to conventional echo spectra acquired at the same spinning rate.

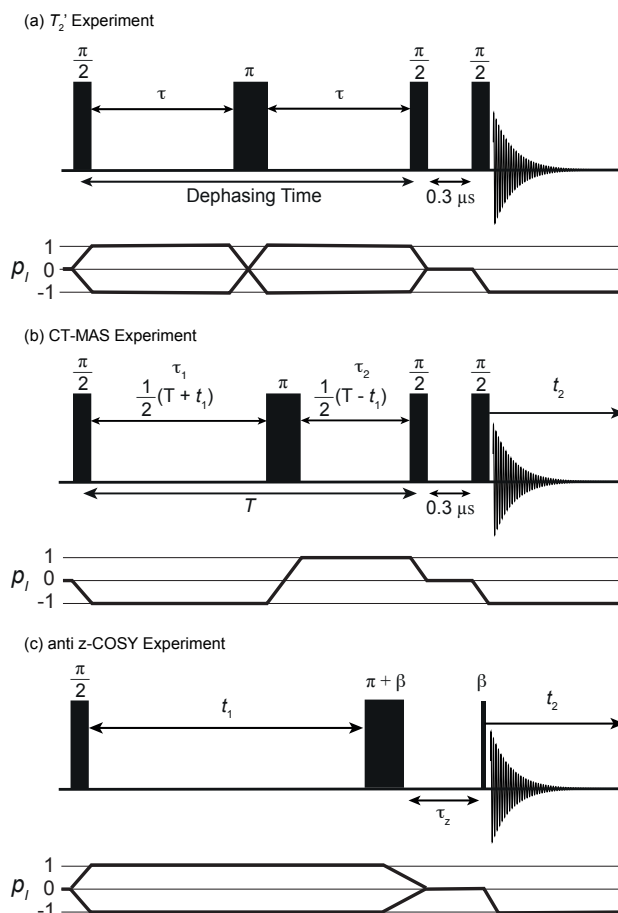


Figure 3.1. Pulse sequences for: (a) measurement of the transverse dephasing time (T_2'), (b) CT-MAS experiments and (c) anti-z-COSY experiment. In (c) β indicates a small flip angle pulse. The symmetry pathways shown are the ones that were selected by phase cycling.

3.1 Materials and Methods

NMR

Experiments of **Section 3.2** and **Section 3.3** were performed respectively on a Bruker 900 US² wide-bore Avance Neo and on a Bruker 800 US Plus Avance Neo narrow-bore NMR spectrometers. Both spectrometers were equipped with H/C/N 0.7mm CPMAS probes. All spectra in **Section 3.2** were recorded using a rotor spinning rate of 111.0 kHz, and in **Section 3.3** using a rotor spinning rate of 100.0 kHz.

The transverse dephasing times (T_2') were measured using the Hahn echo sequence,²²⁹ shown in **Figure 3.1a**. The measurement was done through a pseudo-2D experiment, based on the acquisition of a series of 1D MAS ^1H spectra where the time τ of decoupling was systematically increased in a multiple of the rotor period. The dephasing curve was then obtained by plotting the area under each resonance as a function of the dephasing time (2τ). T_2' values were then extracted by fitting the dephasing curve to an exponential decay $S(2\tau) = a \cdot \exp(-2\tau / T_2')$.

Figure 3.1b shows the pulse sequence used for the acquisition of constant-time MAS (CT-MAS) experiments. The initial $\pi/2$ pulse is followed by a τ_1 – π – τ_2 block (t_1 dimension), then a z-filter is applied before direct ^1H detection in t_2 . During the experiment, the total time T of the block τ_1 – π – τ_2 is kept constant. The evolution of the magnetization during t_1 is induced by concomitantly increasing τ_1 in a multiple of the rotor period, while τ_2 is decreased. During t_1 evolution, the π pulse is incrementally moved from the center (where $t_1 = 0$) to the end of

the T period. Analogously to the CT-CRAMPS experiment,²²⁸ this experiment makes use of hypercomplex acquisition in the manner of States, Haberkorn, and Ruben²⁴¹ to acquire the negative t_1 quadrant. The result is a two-dimensional dataset with the ordinary MAS ^1H spectrum in ω_2 and the ^1H spectrum free of non-refocusable broadening (the homogeneous interactions) in ω_1 .²²⁸ All CT-MAS experiments were acquired with 256 points in the indirect dimension. The indirect sampling intervals (Δt_1) were 0.036 ms, 0.054 ms, 0.072 ms, and 0.144 ms respectively for the experiments with constant time intervals of $T = 4.6$ ms, 6.9 ms, 9.2 ms, and 18.5 ms. Note that the increment used for $T = 18.5$ ms corresponds to a 7 kHz window, which leads to aliasing of the ^1H spectrum. The 1D spectrum shown was reconstructed from the 2D dataset by merging the 1D spectra obtained summing the signals in the columns between 0.25 and 2.38 ppm, 2.39 and 9.87 ppm, 9.88 and 13.80 ppm.

Formally, the TOP-CT-MAS experiments are implemented with a pulse sequence identical to the sequence used in the CT-MAS experiment, shown in **Figure 3.1b**. Each TOP-CT-MAS experiment was run with an indirect sampling interval of $\Delta t_1 = 1.081$ ms. For the experiments at constant time intervals of $T = 4.9$ ms, 9.8 ms, and 17.8 ms, there were 8, 16, and 32 points in the indirect dimension, respectively. The application of the TOP transformation is facilitated by phase cycling for the symmetry pathway shown in **Figure 3.1b** and running the π pulse from the beginning to the end of the constant time interval. This results in direct, phase-sensitive acquisition of the negative t_1 signal quadrant to permit absorption mode lineshapes without resorting to hypercomplex acquisition,^{241,242} as chosen for the conventional CT-MAS implementation.

Figure 3.1c shows the anti-z-COSY pulse sequence.^{239,240} The anti-z-COSY spectra of thymol were acquired with 256 points in the indirect dimension for thymol, 448 for β -AspAla and 180 for strychnine. All spectra were acquired with an indirect sampling interval of 0.040 ms ($4\tau_r$). In the spectra with flip angle β of 90° , 45° , 20° , 15° each point in the indirect dimension was acquired using 8 scans, while 16, 32 and 64 scans were used for β values of 8° , 5° , and 3° respectively (with the only exception being the $\beta = 8^\circ$ experiment of strychnine which was acquired with 32 scans). The ZQF-COSY and z-COSY experiments shown in **Section 3.3** were acquired with 256 points in the indirect dimension and indirect sampling interval of 0.040 ms ($4\tau_r$).

^1H chemical shifts were referenced to the CH_3 resonance observed for β -AspAla at 0.86 ppm with respect to the signal for neat tetramethylsilane (TMS).¹⁷² All linewidths reported are full width at half maximum linewidths. The processing of the spectra was done either using the Bruker program TopSpin 3.5 for 1D and 2D CT-MAS experiments or using RMN 1.8.4 for 2D TOP-CT-MAS.²⁴³ The affine transformations for the processing of anti-z-COSY experiments were done using the “au” program `ajp_shear`, written by Andrew Pell.²⁴⁰ The 1D spectra were then extracted using an integral projection onto F2 of the F1 range spanned by the lineshape of the signals and the presence of cross peaks. The post-processing procedures for the T_2' measurements (extraction of 1D spectra, evaluation of the area under each peak, fitting and extraction of T_2' value) as well as the extraction of the experimental ^1H linewidths for all spectra were done in MATLAB using home-written scripts.

TOP-CT-MAS Processing

In the conventional CT-MAS experiment, the indirectly acquired spectrum is acquired point-by-point, which leads to very long experiment times for the acquisition of high-resolution constant-time spectra. Strategies utilizing spectral aliasing,^{244,245} non-uniform sampling,²⁴⁶ or covariance processing²⁴⁷ are means of enhancing the resolution and information of such spectra in a time effective manner and are being actively developed.

In a unique alternative approach, Davis *et al.* demonstrated that the spectral windows of the direct and indirectly acquired dimensions in the 2D PASS experiment for sideband separation in rotating solids can be effectively swapped by the action of an affine transformation on the 2D dataset.²³⁶ This allowed the indirectly detected signals to inherit the large spectral window of the directly acquired dimension. This method of transferring spectral windows by affine transformation of uniformly sampled data was referred to as TOP processing and was later shown to generalize to other sideband separation experiments.^{237,238}

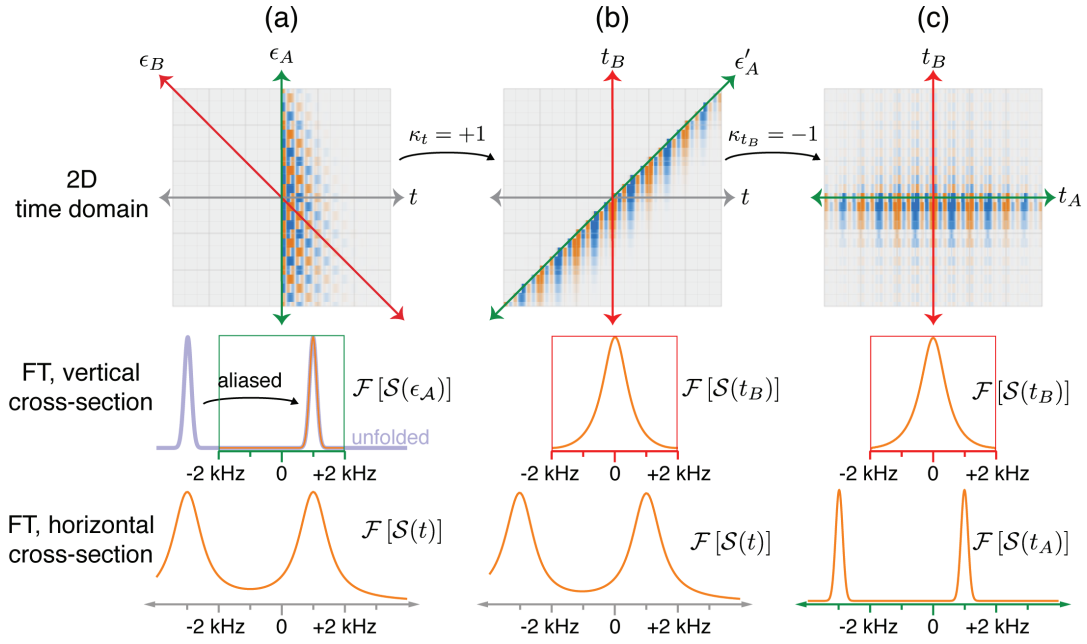


Figure 3.2. The TOP double shear transformation of a simulated signal comprised of two peaks. (a) Representation of the undersampled CT-MAS dataset. The constant-time shift evolution (interaction “A”) evolves along the vertical axis designated by the time coordinate ϵ_A (green) while the decay from non-refocusable components (interaction “B”) evolves along the line $t = -\epsilon_A$, designated by the time coordinate ϵ_B (red). Fourier transform of the signal along ϵ_A yields an aliased constant-time shift spectrum where the peak at -3 kHz from the full spectrum (overlaid in blue) is aliased into the window, directly onto the peak at +1 kHz. The directly acquired t coordinate (horizontal, grey), having a much higher bandwidth, contains the unaliased but broadened spectrum. (b) After the first active shearing transformation parallel to t (shear factor $\kappa_t = +1$), the signal along ϵ_A is shifted onto the vertical axis, along the coordinate designated t_B , while the ϵ_A coordinate now runs parallel to the line $t = +t_B$. Fourier transform of the signal along t_B yields the lineshape due to non-refocusable components, whereas the spectrum along t is unchanged. (c) The second active shearing transformation, parallel to t_B (shear factor $\kappa_{t_B} = -1$), shifts the signal along the ϵ'_A axis onto the horizontal axis, along the coordinate designated t_A . Fourier transform of the signal along t_A presents the unaliased constant-time shift spectrum. Note the coordinate grid itself is unaffected by the active transformations.

Figure 3.2 gives a visual outline of the TOP processing algorithm applied to simulated CT-MAS data. The two-dimensional dataset as acquired separates the constant-time signal evolution, which proceeds along the vertical ϵ_A axis (which corresponds to t_1 in the pulse sequence of **Figure 3.1b**), from the decay due to non-refocusable signal components, which evolves purely along the diagonal axis $\underline{t} = -\epsilon_A$. This coordinate is designated ϵ_B . Along the horizontal t axis (which corresponds to t_2 in the pulse sequence of **Figure 3.1b**), the directly detected MAS spectrum evolves with both refocusable and non-refocusable contributions to its linewidth. This structure, where separated interactions evolve along nonorthogonal dimensions and coevolve in the directly acquired signal, is like that of the 2D PASS experiment, where use of the TOP transformation to effectively swap the sampling intervals of the direct and indirectly acquired dimensions was first described.²³⁶ Indeed, we can achieve the same effect in the CT-MAS experiment using the affine transformation:

$$\begin{bmatrix} t_B \\ t_A \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}}_{\mathbf{K}_{t_B}} \underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}}_{\mathbf{K}_t} \begin{bmatrix} \epsilon_A \\ t \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \epsilon_A \\ t \end{bmatrix}. \quad (3.1)$$

The strategy is to actively transform the data such that the two interaction dimensions, ϵ_A and ϵ_B , lie along the horizontal and vertical dimensions of the coordinate grid. The new respective coordinates, t_A and t_B , inherit the sampling properties of the grid axis onto which they are transformed. The first shear, represented by the matrix \mathbf{K}_t , maps the ϵ_B coordinate onto the vertical axis of the grid. Because only signal amplitude changes along this coordinate, the spectrum is centered about zero with a linewidth that reflects the non-refocusable linewidth of the MAS spectrum, suitable for the narrow spectral window of the vertical dimension. This is evident in the spectrum corresponding to the vertical cross-section of the 2D dataset after the first shear shown in **Figure 3.2b**. After the second shear, represented by the matrix \mathbf{K}_{t_B} , the original ϵ_A axis is mapped onto the horizontal axis of the grid. In practice, the sampling interval of this dimension can be set very small, and the originally undersampled

spectrum is unaliased as the correspondingly large bandwidth of the horizontal axis is inherited, as can be seen in **Figure 3.2c**.

In most useful circumstances, applying the TOP transformation matrix $\mathbf{K}_{t_B}\mathbf{K}_t$ directly to the data vectors will map most of them off the grid nodes. Therefore, interpolation back onto the sampling grid is required. This is conveniently achieved by applying the shearing transformations as their operational equivalents in the mixed time-frequency domain according to the Fourier Affine Theorem.

Note that not all linearly sampled 2D signals can have their sampling intervals effectively swapped by TOP processing. Qualifying conditions are discussed in **Section 3.2**.

3.2 The TOP-CT-MAS Experiment

Lineshape Contributions

Figure 3.3c shows the dephasing curves of the eight ^1H resonances of β -AspAla, acquired using the sequence of **Figure 3.1a**. As described above, by fitting these curves to an exponential decay it is possible to obtain the corresponding T_2' values, which by definition corresponds to the residual broadening induced by non-refocusable interactions, such as homonuclear dipolar coupling.¹⁹⁴ From these curves we can estimate that *non-refocusable interactions contribute up to about a third of the measured line broadening* observed in a one-pulse acquire experiment of the type shown in **Figure 3.3b**, as it is shown in **Figure 3.3d**. This is in line with previous observations under both CRAMPS and fast MAS.^{163,173,228} Removing this contribution would in principle lead to a significant improvement in spectral resolution, without affecting chemical shift information (which would be refocused by the application of the π pulse).

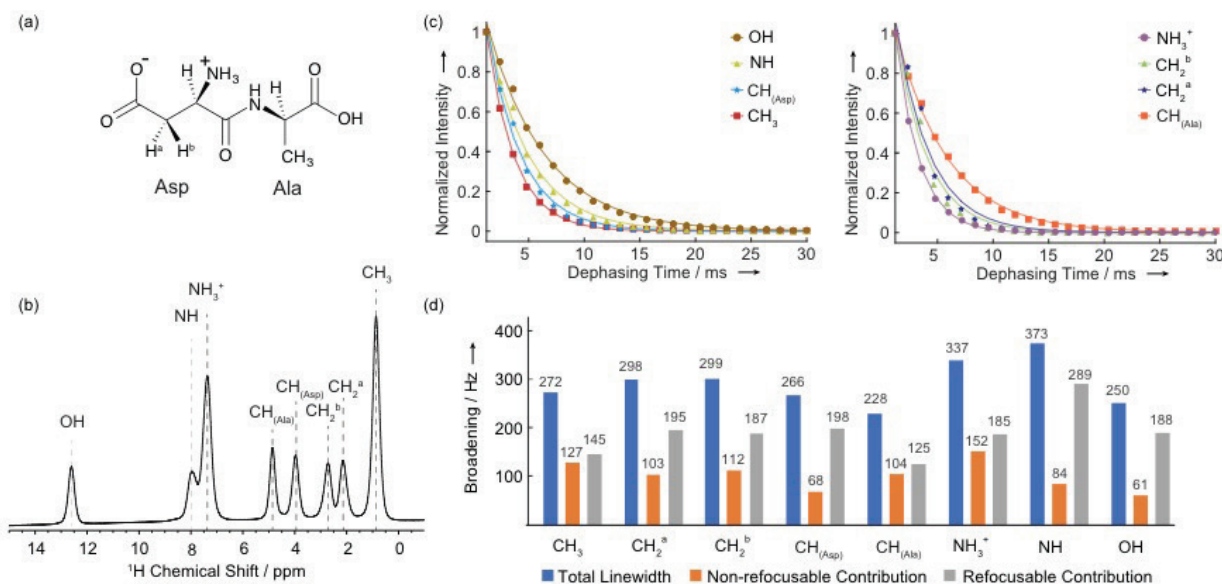


Figure 3.3. (a) Chemical formula of β -AspAla. (b) 1D ^1H spectra of powdered β -AspAla acquired at 111 kHz MAS rate using a one-pulse acquisition sequence. (c) Dephasing curves for the ^1H resonances of powdered β -AspAla recorded using the pulse sequence of **Figure 3.1a**. Each curve is normalized with respect to the corresponding highest intensity point. (d) Histogram of the contributions to the line broadening of the resonances of powdered β -AspAla at 111 kHz MAS. The total linewidth (in blue) is the linewidth measured from the experimental spectrum shown in (b), the contributions of non-refocusable interactions (in orange) were calculated from the T_2' values ($\Delta\nu_{\text{homo}} = 1/\pi T_2'$) and the contributions of the refocusable interactions (in grey) were obtained from the difference between the total and non-refocusable broadenings.

CT-MAS

The constant-time (CT) principle²⁴⁸⁻²⁵³ has already proven to be an effective way to remove broadening due to non-refocusable interactions from ^1H spectra in the solid-state. This idea was introduced by Lesage *et al.* with the CT-CRAMPS experiment.²²⁸ They have shown that constant time acquisition in the indirect dimension of a 2D dataset leads to pure refocusable spectra, which have 2 to 3 times narrower resonances compared to CRAMPS experiments at 12.5 kHz MAS. In this experiment, the application of high power homonuclear decoupling was necessary to enter the weak coupling regime where the CT principle is likely to be most valid.

Today, with the line narrowing achieved using the highest MAS rates available, we can make use of the CT principle with MAS alone.²⁵⁴ We can then obtain pure refocusable spectra by using the CT-MAS experiment, shown in **Figure 3.1b**. **Figure 3.4a** shows the 111.0 kHz CT-MAS spectra of β -AspAla acquired using constant time periods (T) ranging from 4.6 to 18.5ms.

When compared with the one-pulse ^1H spectra recorded at 111.0 kHz MAS (**Figure 3.3b**), all spectra show a significant improvement in spectral resolution. Line narrowing is already achieved at $T = 4.6$ ms, despite truncation of the signal due to the short CT period. In this spectrum the experimental linewidths are on average 17% narrower than the ones obtained with the one-pulse experiment. The CH_2 resonances give the highest gain in resolution with 30% narrower linewidths (from about 300 to 210 Hz). On the other hand, nearly no line narrowing effect was observed for the OH resonance, probably due to the presence of dynamics or exchange.

The spectral resolution can be further improved by increasing the T value. This however comes at the expense of sensitivity, which decreases when increasing T at a rate that depends on the T_2' . The dependence from the T_2' affects also the signal intensities in the CT-MAS spectra, which vary from those obtained with simple MAS experiments. Indeed, the resonances with longer T_2' will decay at a slower rate than the other resonances, resulting in apparent higher intensities. At $T = 6.8$ ms the resonances are on average 23% sharper than in the one-pulse experiment, with each resonance being at least 8% narrower than the corresponding peak in the one-pulse spectrum. $T = 6.8$ ms is also short enough to allow the observation of all the ^1H resonances in our sample. Further increase of T still provides additional line narrowing effects, but decreases the sensitivity down to a point where some of the resonances are no longer visible. At the longest CT period we used ($T = 18.5$ ms) only three peaks remain: the CH_3 , $\text{CH}_{(\text{Ala})}$ and the OH. Of these, the CH_3 shows a remarkable improvement in resolution, with a linewidth of 171 Hz, about 37% narrower than the 272 Hz observed in the one-pulse experiment. The improvement for the other two resonances is less: 16% for the $\text{CH}_{(\text{Ala})}$ and only 3% for the OH. Note that in the spectra there are also two additional peaks at 1.6 and 4.5 ppm (which appear already at $T=9.2$ ms). These might be due to impurities with long T_2' which are not observed in the one-pulse experiment. The full set of measured linewidths for these experiments is given in **Appendix**.

At this point it is important to underline that in this experiment the line narrowing effect is not only due to the CT effect. As shown previously for the CT-CRAMPS experiment,²²⁸ part of the reduction in linewidth is due to the presence of the long echo which delays the signal acquisition. During this delay the signal of protons in the more strongly coupled average orientations will dephase preferentially, leaving relatively stronger signals from protons in more weakly coupled average orientations, leading to line narrowing.

This is supported by the fact that in these spectra we observe significant line narrowing both in ω_1 and in ω_2 (where there is no CT effect). In particular, we observed that at 111.0 kHz MAS rate the narrowing due to the CT effect contributes only 35% of the total line narrowing. This is in contrast with what has been observed for CT-CRAMPS experiments where most of the narrowing was induced by the CT acquisition.

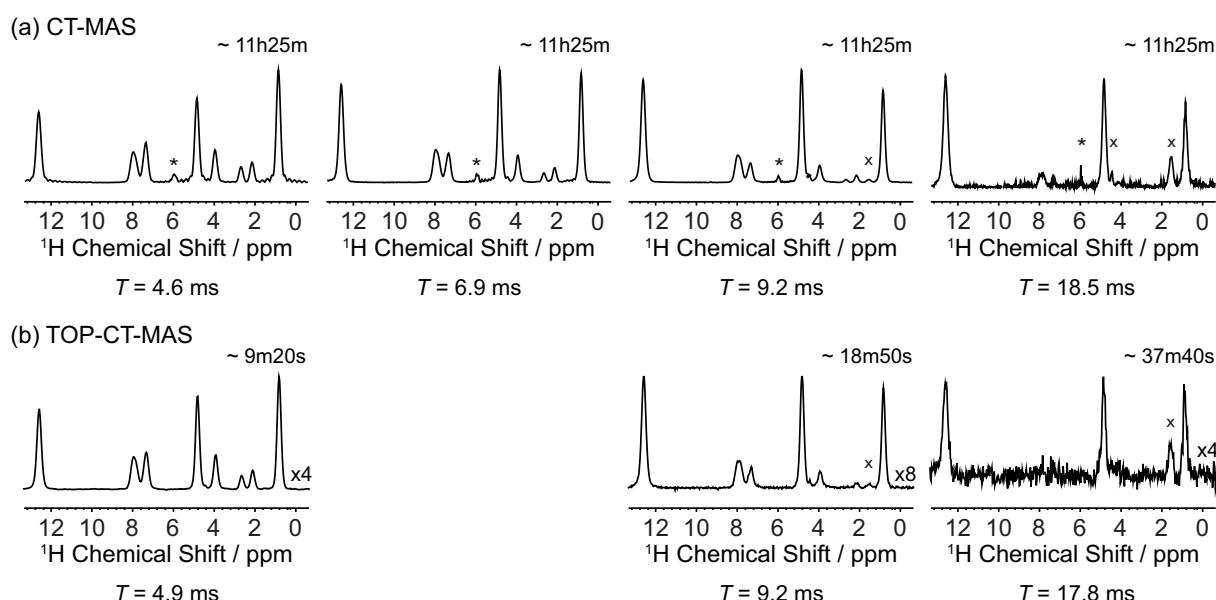


Figure 3.4. Comparison between different 1D ^1H spectra of β -AspAla acquired with (a) the CT-MAS experiment and (b) the TOP-CT-MAS experiments. CT-MAS spectra were acquired using constant-time periods (T) of 4.6, 6.9, 9.2 and 18.5 ms, while TOP-CT-MAS using T of 4.9, 9.2 and 17.8 ms. CT-MAS spectra are the vertical projection of the two-dimensional experiments, while TOP-CT-MAS spectra are the extraction of the central row. The transmitter peak is indicated with an asterisk, which is better suppressed in the TOP-CT-MAS experiments due to a superior phase cycle (the pulse programs with the detailed phase cycles are given in <https://doi.org/10.1016/j.jmr.2019.06.015>). Peaks due to impurities are marked with an x. On the top-right corner of each spectrum is reported the corresponding acquisition time for comparison.

TOP-CT-MAS

The most significant difference between conventional CT-MAS and the TOP-CT-MAS implementation is that the indirect dimension of the latter is sampled at 0.925 kHz, well below the Nyquist rate of the ^1H chemical shift spectrum (about 12 kHz for β -AspAla at 21.14 T). Because of the correspondingly large indirect sampling interval (1.081 ms), the entire constant-time signal envelope can be collected with far fewer discrete samples than the conventional approach, making the TOP-CT-MAS implementation potentially far more economical in terms of experiment time. The obvious disadvantage is that the Fourier transform of the indirect dimension, as acquired, presents a useless ^1H shift spectrum because of severe spectral aliasing resulting from the undersampling.

The affine transformation specified by **Equation 3.1** nullifies this disadvantage by reconstructing the unaliased, high-resolution constant-time ^1H spectrum in the large spectral window of the directly acquired dimension from a CT-MAS experiment with an undersampled ϵ_A coordinate. The sampling rate of the indirect dimension using the TOP transformation needs only to be small enough to accommodate the frequency components which evolve along the vertical grid axis *after* transformation of the dataset. This is controlled by the design of the experiment and processing protocol. Here, the TOP transformation maps the residual non-refocusable broadening onto the vertical axis. For each site in our β -AspAla sample linewidths are no greater than 160 Hz, as shown in **Figure 3.3b**.

A sampling rate of less than 1 kHz suffices for sampling the ϵ_A dimension in the TOP approach, for a potential reduction in experiment time *by one to two orders of magnitude* when TOP processing principles are leveraged. Note that if the indirect sampling rate dimension is so low that the t_B coordinate is undersampled, the TOP transformation will introduce sidebands into the spectrum along the t_A coordinate at integer multiples of the indirect sampling rate. Generally, the decay of the non-refocusable contributions is exponential, so these sidebands manifest more strongly among the undesired long-tailed dispersive signal components, as shown in **Figure 3.10** in **Section 3.5**. This improves the robustness of the approach in the face of a poor estimation of an appropriate indirect sampling interval.

In **Figure 3.4a** that has an indirect sampling rate of 27.75 kHz, conventional CT-MAS experiments took more than 11 h to complete. Note that an unnecessarily long 32-step phase cycle was used, but even if the conventional CT-MAS experiments had used the 14-step phase cycle to select the symmetry pathway in **Figure 3.1b** directly, each experiment would still have taken nearly six hours to complete. In **Figure 3.4b**, we see comparable high-resolution constant-time ^1H spectra constructed from undersampled ϵ_A coordinates using the TOP approach. When $T = 4.9$ ms, just eight points in the indirect dimension are enough to obtain the narrowed proton spectrum, in an experiment lasting less than ten minutes. Furthermore, in the TOP-CT-MAS experiments with constant time intervals of 4.9 ms and 17.8 ms no loss in signal-to-noise per unit time was observed when compared to the CT-MAS experiment at 4.6 and 18.3 ms, while the experiment with $T=9.2$ ms shows about 30% less sensitivity when the TOP transformation is applied. Here, the overall reduction in experiment time for experiments at comparable T normalised to the number of scans was between a factor 8 and 32 using the TOP implementation.

Undersampling the indirect dimension of the CT-MAS experiment does not quantitatively affect the line width measurements when TOP processing is used. We show in **Appendix** that linewidths in the constant-time ^1H spectra are comparable between conventional CT-MAS and the TOP-CT-MAS implementation. In fact, we see from **Figure 3.2c** that the final 2D TOP-CT-MAS dataset correlates evolution along the t_A and t_B coordinates along orthogonal dimensions. We can therefore measure site-specific non-refocusable contributions to linewidths in the dimension orthogonal to the constant-time ^1H spectrum. We also show in **Appendix** that the non-refocusable contributions determined from the TOP-CT-MAS experiments are comparable to those determined from the T_2' experiments shown in **Figure 3.3d**.

At first sight it might appear that the TOP transformation circumvents a seemingly fundamental limitation on the bandwidth of linearly sampled signals. This is not the case. First, the Nyquist criterion should only be expected to apply strictly to directly acquired signals, since acquisition actually proceeds along the indirect dimension by concatenation of separate experiments. More importantly, however, is that interactions “A” and “B” coevolve in the directly acquired signal. Signals acquired along the indirect dimension need then only supply the information necessary to allow deconvolution of the two interactions. When, as in 2D PASS and CT-MAS, the experiment is designed such that interactions “A” and “B” evolve independently along oblique coordinates in the initial 2D coordinate system, an affine transformation exists for separating and correlating them along orthogonal dimensions in the final 2D coordinate system. For the classes of 2D experiments that do not possess this property, however, TOP processing cannot effectively swap the direct and indirect sampling intervals. This is true, for example, of the two shift spectra axes in a 2D heteronuclear correlation experiment.

Finally, it is worth noting that when using the strategy of direct pathway selection, use of the z-filter with hypercomplex acquisition is unnecessary to obtain pure absorption mode lineshapes provided that the pulse is moved from the beginning to the end of the constant-time interval. Thus, only the first two pulses are necessary for the minimal CT-MAS experiment, allowing a phase cycle with as few as 3 steps. If sensitivity is not a limiting factor, this would permit acquisition of constant time ^1H spectra in less than two minutes.

3.3 The anti-z-COSY Experiment

We now move to introduce a conceptually new approach to homonuclear decoupling in solids. We hypothesize that in the regime of ultra-fast MAS, where the residual linewidths are less than the chemical shift differences, that the homogeneous contribution to the residual broadening will be, at least in part, of the following character:

$$\hat{H}_{res}^{MAS} \propto \sum_{i,j} A_{ij} I_{iz} I_{jz}, \quad (3.2)$$

where the residual coupling strengths A_{ij} will potentially depend on many factors that are difficult to predict,^{162,255} and may be orientation dependent. (Here we neglect the case of equivalent spins). If this is true, then this part of

the residual will have character analogous to homonuclear scalar couplings. As a result, homonuclear decoupling approaches developed for scalar couplings should produce ^1H line narrowing in solids in the fast MAS regime.

The anti-z-COSY pulse sequence (**Figure 3.1c**) was first introduced by Oschkinat *et al.*²³⁹ to probe correlations between connected transitions, and then implemented by Pell *et al.*²⁴⁰ for the purpose of broadband homonuclear J decoupling in solution-state NMR. The experiment produces phase-sensitive COSY-like 2D spectra in which, in the limit where the pulse angle β tends to zero, the fine structure of the diagonal peaks only contains correlations between transitions in which the passive spins (the coupling partners) have inverted their spin state between t_1 and t_2 . As shown schematically in **Figure 3.5**, these peaks are arranged only along the 45° anti-diagonal of the diagonal peak, perpendicular to the main diagonal. In this limit, on-diagonal and counter-diagonal peaks are absent. Double shearing of the spectrum, followed by a projection as illustrated in **Figure 3.5** yields a spectrum where only pure chemical shift peaks are retained while broadening due to residual coupling of the type in **Equation 3.2** is suppressed.

Figure 3.6 shows the one-dimensional ^1H spectrum obtained from the 45° projection of the diagonal peaks in an anti-z-COSY experiment on a sample of powdered thymol, in comparison to the conventional echo-detected MAS experiment. Both spectra were obtained at 100.0 kHz MAS. The anti-z-COSY spectrum yields resonances that are a factor two narrower.

If this is correct, the suppression of the unwanted off-diagonal and of the on-diagonal peaks in the multiplet, and therefore the level of line-narrowing, depends on the choice of β angle, and the β dependence can be predicted.²⁴⁰ First, we note that the intensity of the unwanted correlations is proportional to $\cos^2\beta$ and the intensity of the wanted anti-diagonal peaks in the multiplet is proportional to $\sin^2\beta$. Usually a β angle between 10° and 20° gives the best compromise between spectral purity and sensitivity, since reducing the flip angle leads to better line narrowing but also to lower intensities of the wanted peaks. The spectrum in **Figure 3.6** was obtained with $\beta = 5^\circ$.

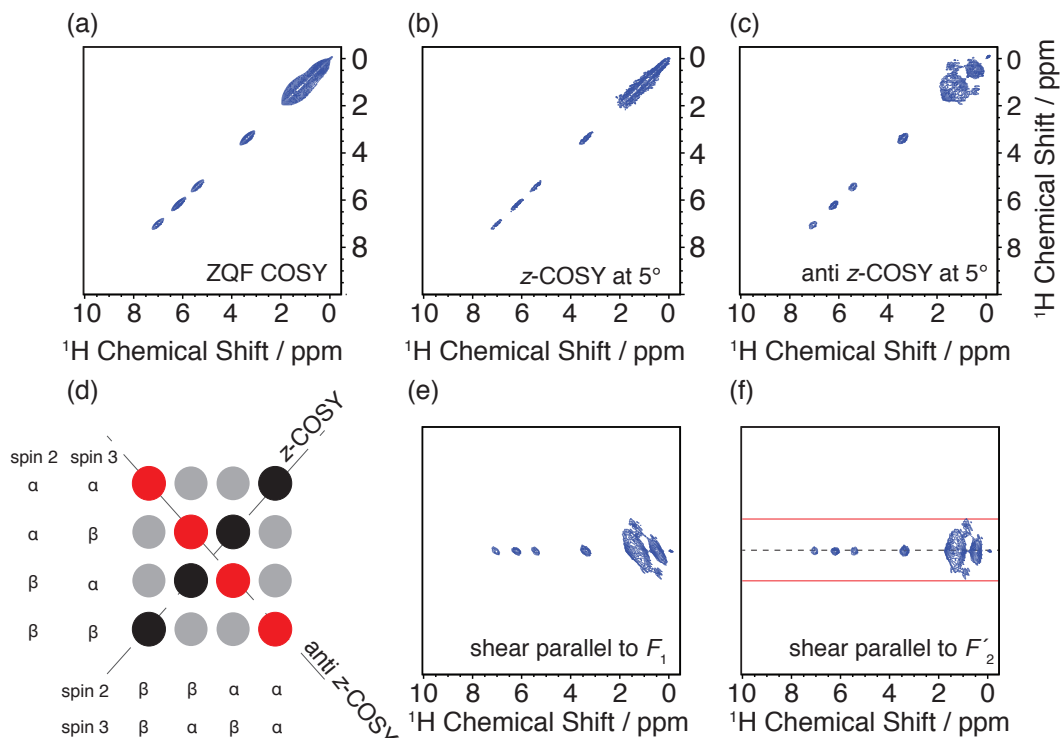


Figure 3.5. (a) ZQF COSY, (b) z-COSY, and (c) anti-z-COSY of thymol acquired at a MAS rate of 100.0 kHz. (b) and (c) were acquired with a β flip angle of 5° . (d) Schematic of the diagonal-peak multiplet of a spin with a coupling to two other spins showing the on-diagonal (black circles), off-diagonal (grey circles), and counter-diagonal peaks (red circles). The labels correspond to the spin states of the two coupled spins. (e) First shear of the spectrum in (c), parallel to F_1 (shear factor $\kappa_t = -1$), and (f) second shear of the spectrum in (c), parallel to the new F_2' dimension (shear factor $\kappa_t = +\frac{1}{2}$), where the shear factor κ_t is given by the shear angle θ according to the equation $\tan\theta = \kappa_t$. The shear angle θ is the angle by which the perpendicular lines to the shear axis tilt.

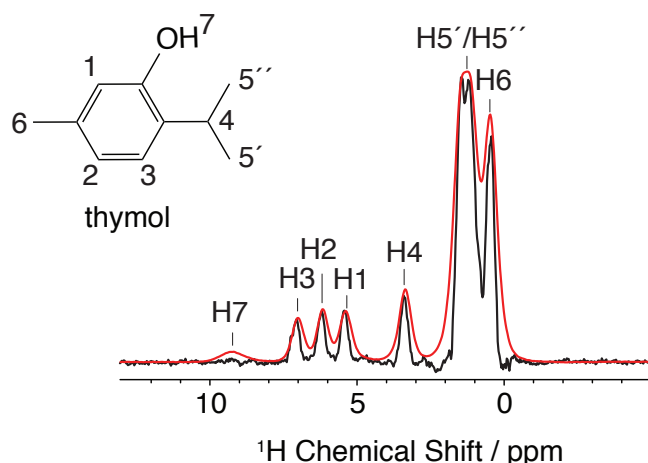


Figure 3.6. Echo-detected spectrum (red) anti-z-COSY spectrum (black) of thymol acquired at 100.0 kHz MAS rate. For the anti-z-COSY experiment a β value of 5° was used.

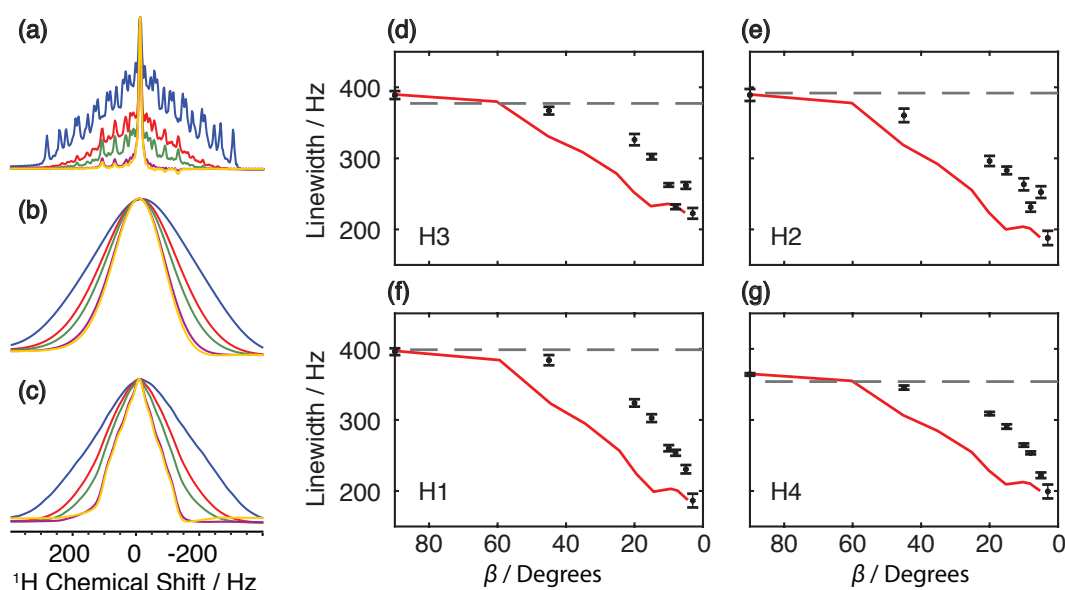


Figure 3.7. (a-c) Simulations of anti-z-COSY experiments (flip angles 90° , 35° , 25° , 15° , and 5°) with 5 coupling partners and a distribution of A values of 40, 60, 90, 160, and 240 Hz, using additional Gaussian line broadening of 145 Hz in (b) and Gaussian line broadening of 29 Hz and chemical shift distribution of 200 Hz in (c). (d-e) Measured apparent linewidth of protons H1-H4 as a function of β . The dashed lines denote the proton linewidths measured in the conventional spin echo spectrum and the red lines are plots of the scaled linewidths of the Gaussian distributions of (b).

In **Figure 3.7d-g** we show the linewidths for obtained from anti-z-COSY experiments on thymol for β angles ranging from 90° to 3° . In the spectrum with $\beta = 3^\circ$ for H1, H2, H3, and H4 of thymol we measure linewidths of 186, 188, 223 and 199 Hz, as compared to 396, 389, 389, and 364 Hz in the $\beta = 90^\circ$ spectra. This corresponds to a factor of between 1.8 to 2.1 increase in resolution. **Figure 3.7a-c** also shows simulations of the lineshapes we might expect using a very simple model where the lineshape is due to a coupling of the kind given in **Equation 3.2**, with 5 coupling partners and a distribution of A values (here 40, 60, 90, 160, and 240 Hz) (the simulations were performed using the SPINACH package.²⁵⁶ Further details are given in **Section 3.5**). We see in **Figure 3.7a** how the lineshape is clearly made up of a “foreground” resonance centered at the isotropic chemical shift, due to the anti-diagonal, with constant width, and a constant “background” coupling pattern that yields the broadening which decreases in relative intensity as β decreases. This is entirely equivalent to the liquid-state scenario. However, **Figure 3.7b** and **c** show how when we introduce more line broadening (**Figure 3.7b**) and a distribution of chemical shifts (**Figure 3.7c**), while the underlying behavior still consists in a constant foreground anti-diagonal and a decaying background broad function, the apparent resulting overall lineshapes now appear like a continuous narrowing

that can be fitted to a single apparent Gaussian at any given value of β . Given the large number of residual couplings, their anisotropy, and the presence of a chemical shift distribution that we expect for the solid-state case, we therefore expect lineshapes of the kind shown in **Figure 3.7b** and **c**. The experimental linewidths in **Figure 3.7d-g** are then compared to predictions obtained by fitting the model lineshapes of **Figure 3.7b** and **c** to a single Gaussian at each value of β (red lines). Given the very simple model used, the agreement is highly satisfactory.

The method is also demonstrated for microcrystalline powder samples of β -AspAla and strychnine. **Figure 3.8** shows projections extracted from anti-z-COSY spectra of β -AspAla acquired with different β angles and which clearly show the same decoupling effect as observed for thymol above. Again in this sample the linewidths are reduced by almost a factor of two for most of the protons. **Figure 3.9** shows the resolution gained from the application of the anti-z-COSY experiment for a powdered sample of strychnine. Once again, as compared with the conventional echo MAS spectrum, the 1D spectrum extracted from the 2D anti-z-COSY spectrum acquired with $\beta = 8^\circ$ is significantly narrower.

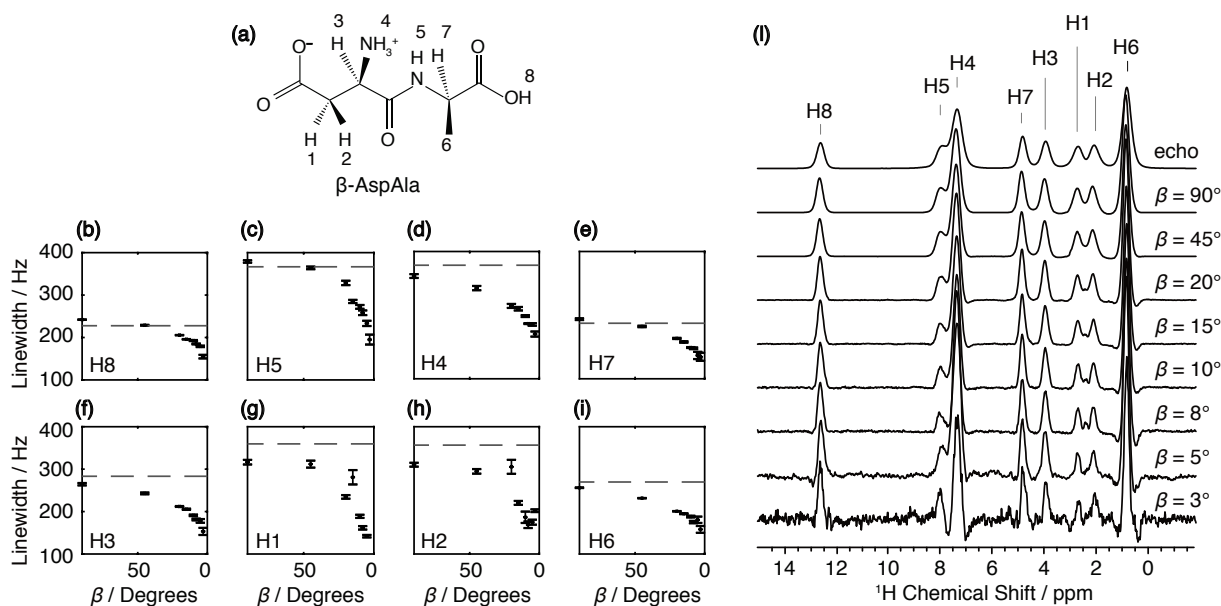


Figure 3.8. (a) Chemical structure of β -AspAla. (b-i) Measured linewidth of protons H1-H8 as a function of β , used in the mixing element of the anti-z-COSY experiment. The dashed lines denote the proton linewidths measured in the conventional spin echo spectrum. (f) Spectra of β -AspAla obtained from integral projections from the sheared 2D anti-z-COSY spectra acquired with β from 90° to 5° at 100.0 kHz MAS. The top spectrum was acquired with the conventional spin echo sequence at the same sample spinning rate.

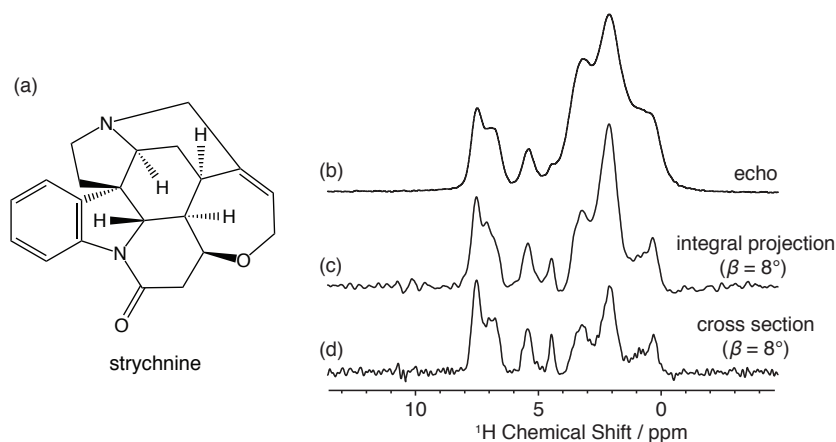


Figure 3.9. (a) Echo-detected 100. kHz MAS spectrum of strychnine. (b) and (c) show 1D spectra extracted as an integral projection (b) and a cross section (c) from a sheared 100.0 kHz MAS 2D anti-z-COSY spectrum acquired with $\beta = 8^\circ$.

All three samples used here yield very clean anti-z-COSY spectra. In solution, these experiments can be compromised by artifacts due to the appearance of cross-peaks in the spectrum, as well as zero-quantum transfers within the diagonal peaks, and additional measures are required to actively suppress them. However, we note that both of these unwanted peaks appear in anti-phase with respect to the couplings, and have no net integral.^{240,257} The numerous couplings and the comparatively broader lines expected in the solid-state mean that these unwanted peaks are likely to largely cancel themselves out, with no further action required. Indeed the cross-peaks we observe here are extremely weak in intensity (see **Figure 3.12** in **Section 3.5**).

We note that for the anti-z-COSY concept to work properly it is important that any z magnetization generated after the first β pulse remains unchanged during the τ_z period and changes spin state only after the application of the second mixing element ($180^\circ + \beta$). Any effect that changes any of the spin during the mixing scheme would then lead to a reduction in resolution. Notably, spin diffusion during the z-filter delay must be avoided since it will induce spin flip-flops. While the fast spinning rate at which the experiments were performed here slows down homonuclear spin diffusion, it is also important to keep the τ_z delay as short as possible. Additionally, any dipolar recoupling effects caused by the mixing pulses will also induce spin flips, and again lead to a loss in decoupling efficiency. As a result, we use a mixing time equal to a single rotor period. Finally, any character to the residual broadening that is not of the kind in **Equation 3.2** will not be removed by the anti-z-COSY experiment. These effects, and others, leading to limiting resolution will be explored further in the future.

We also note that according to our initial hypothesis, the faster the sample spinning is, the more likely we are to be in the regime of **Equation 3.2**. Thus, we expect this strategy to also lead to better resolution at lower spinning speeds, but with less efficiency. This is confirmed in **Figure 3.16**. Echo-detected spectrum (a) and integral projections of the sheared 2D anti-z-COSY spectra (b-e) of β -AspAla at 62.5 kHz. Anti-z-COSY experiments were acquired using β values of 90° and 5° . Spectra were acquired both with and without rotor synchronization for the z-filter. **Figure 3.16** in **Section 3.5**, where we show the results of an anti-z-COSY experiment applied to β -AspAla using a 1.3 mm MAS probe and a sample spinning rate of 62.5 kHz. In this case we obtain a gain in resolution of a factor between 1.19 and 1.97, yielding linewidths of 185-304 Hz.

3.4 Conclusions

In this chapter, we have focused on the development of approaches to further increase the resolution of ^1H spectra at ultra-fast MAS. First, we investigate the residual line broadening in β -AspAla at 111.0 kHz MAS using T_2' measurements, and we find that the contribution of non-refocusable interactions to residual line broadening is around 30-40%. We show that this contribution can be removed in 2D constant time experiments, leading to linewidths up to 40% narrower than the one-pulse experiment. We observe that this narrowing is due to both the CT effect, and the effect of the delayed acquisition caused by the constant-time echo. Despite the line narrowing effect, this experiment requires long experimental times (typically of the order of a few hours) to obtain high-resolution in the indirect dimension. We overcome this drawback by adapting the experiment to the TOP transformation. The resulting TOP-CT-MAS experiment provides comparable results to the conventional CT-MAS experiment, but in a fraction of the experimental time. For β -AspAla we obtain ^1H spectra with up to 30% improvement in resolution as compared to the one-pulse experiment, in less than 10 minutes.

We also introduced another simple approach to achieve homonuclear decoupling in organic solids at fast MAS. Based on the assumption that part of the residual homonuclear dipolar coupling in the ultra-fast MAS regime has a J-coupling like character, we adapt the anti-z-COSY experiment, which is used in solution-state NMR for decoupling homonuclear scalar couplings, to the solid-state. We show that this experiment offers up to a factor of two gain in resolution as compared to conventional echo-detected MAS experiments. Smaller flip angles produce spectra with higher resolution, despite they also lead to significant sensitivity losses. For thymol, $\beta = 3^\circ$ leads to linewidths of 186, 188, 223 and 199 Hz, as compared to 396, 389, 389, and 364 Hz observed in the $\beta = 90^\circ$ spectra.

3.5 Supplementary Information and Discussion

3.5.1 Supplementary Information Section 3.2

TOP Transformation Sidebands

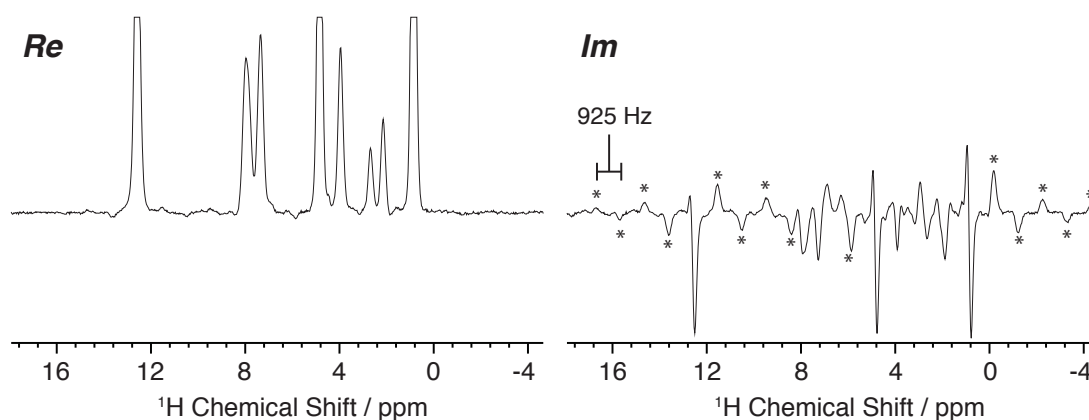


Figure 3.10. Complex form of the constant-time ^1H spectrum obtained by TOP-CT-MAS process with $T = 4.6$ ms. The long dispersion mode tails of the spectrum of the nonrefocusable signal components are folded because the spectral window corresponding to the t_B coordinate (sampling interval 1.081 ms) is insufficient to contain them. Consequently, processing sidebands (with several prominent ones indicated by asterisks), appearing at integer multiples of $1/1.081$ ms = 925 Hz, flank the signals in the imaginary channel. These sidebands are not prominent in the real channel since the absorption mode spectrum of the nonrefocusable signal components is much narrower, such that the folding in this channel is limited

3.5.1 Supplementary Information Section 3.3

Thymol spectra acquired with different flip angles

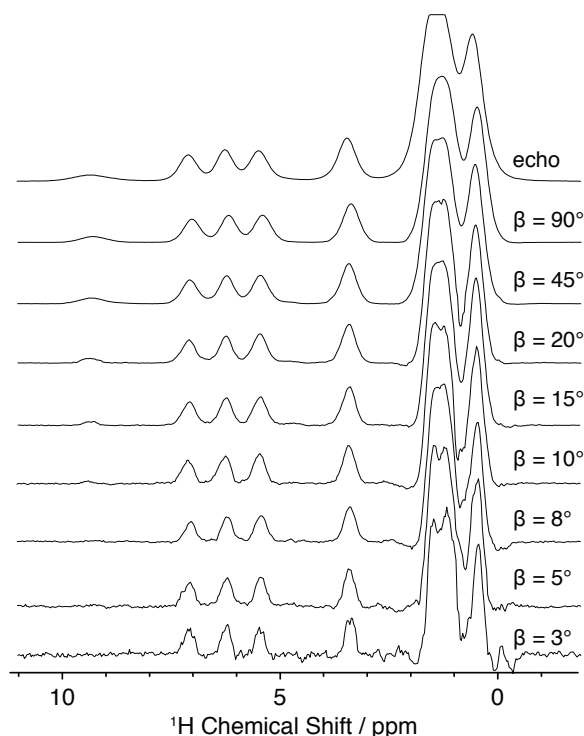


Figure 3.11. Spectra of thymol obtained with a conventional echo experiment (top) and as integral projections from the sheared 2D anti-z-COSY spectra acquired with β of 90° , 45° , 20° , 15° , 10° , 8° , 5° , and 3° . All spectra were acquired at 100.0 kHz MAS.

2D plots of anti-z-COSY spectra

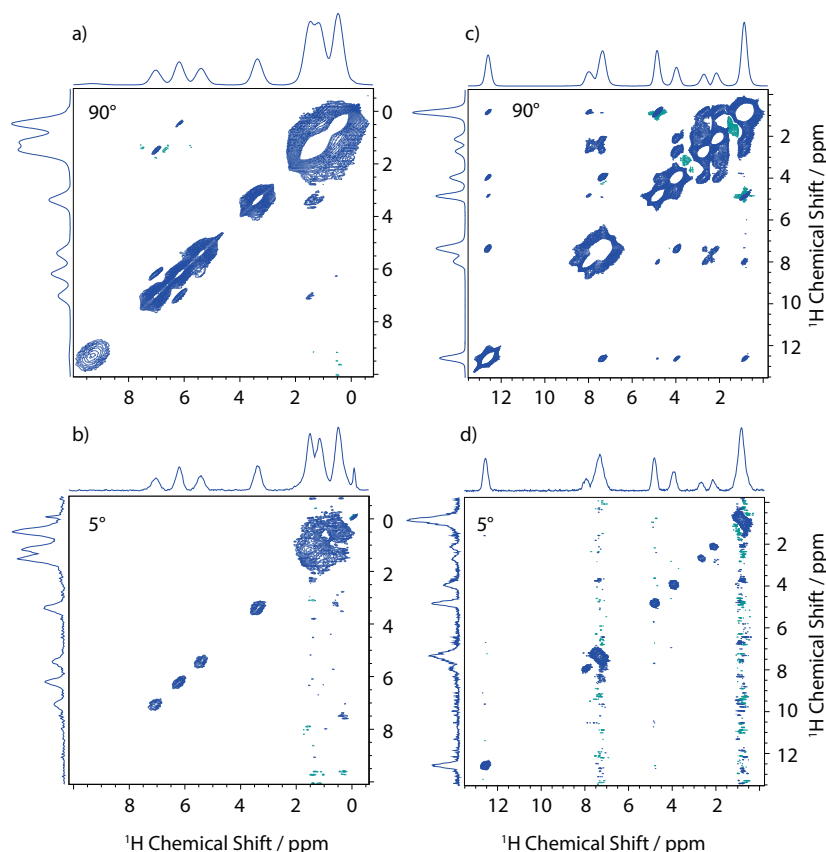


Figure 3.12. 2D anti-z-COSY spectra of thymol (a,b) and β -AspAla (c,d) acquired with flip angles of 90° (a,c) and 5° (b,d).

Simulations

Simulations were carried out using the SPINACH package for MATLAB,²⁵⁶ and then processed using Topspin 4.0.6. For all simulated spectra, we used a spin system composed of 6 inequivalent spins. The residual dipolar coupling of the form shown in **Equation 3.2** was simulated by coupling one of the spins with the other five through J-couplings with of 40, 60, 90, 160, and 240 Hz. A T_1 relaxation time of 3 s and a T_2 relaxation time of 0.04 s were used for all spins. The magnetic field was equal to 14.1 T. We used an offset of 4000 Hz, F_1 and F_2 spectral widths of 10000 Hz and 512 complex points in both dimensions. With these parameters, we have simulated the anti-z-COSY spectra using β equal to 90° , 35° , 25° , 15° , and 5° . The results are shown in the figures below.

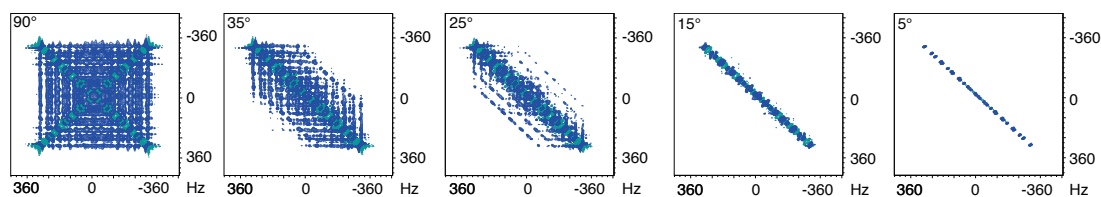


Figure 3.13. Simulations of **Figure 3.7a**. No further processing was used

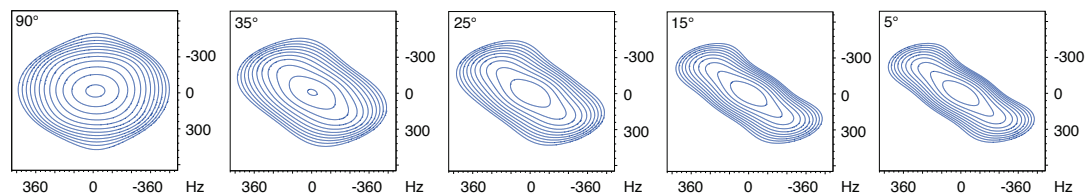


Figure 3.14. Simulations of **Figure 3.7b**. Here, Gaussian line broadening of 145 Hz was applied to simulate the residual couplings typical of molecular solid spinning at about 100 kHz MAS.

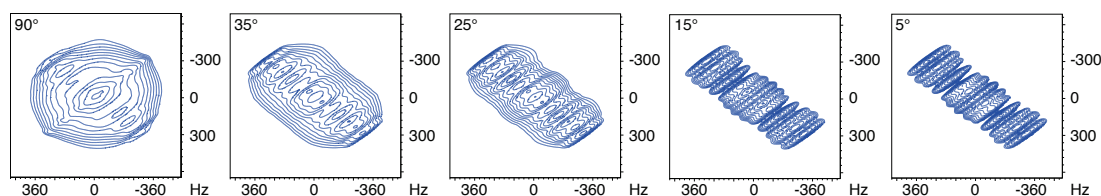


Figure 3.15. Simulations of **Figure 3.7c**. Here, Gaussian line broadening of 29 Hz and a chemical shift distribution of 198 Hz were used. Chemical shift distribution was simulated by carrying out seven different simulations for each flip angle, each of which had a chemical shift offset for the detected spin of 33 Hz from the previous one. Then, these spectra were added using a Gaussian distribution (ratio of 0.3:0.5:0.7:1:0.7: 0.5:0.3).

Anti-z-COSY at 62.5 kHz MAS

The anti-z-COSY and the echo experiments at 62.5 kHz MAS were carried out on a 19.4 T Bruker Avance Neo spectrometer equipped with a 1.3 mm C/N/H CPMAS probe. The sample was β -AspAla. In the anti-z-COSY experiments, an echo of one ten rotor periods was added before the t_1 evolution period in order to suppress the strong ^1H background of the probe. The echo experiment was carried out as well using an echo delays of ten rotor periods. The spectra and the measured linewidths are given below.

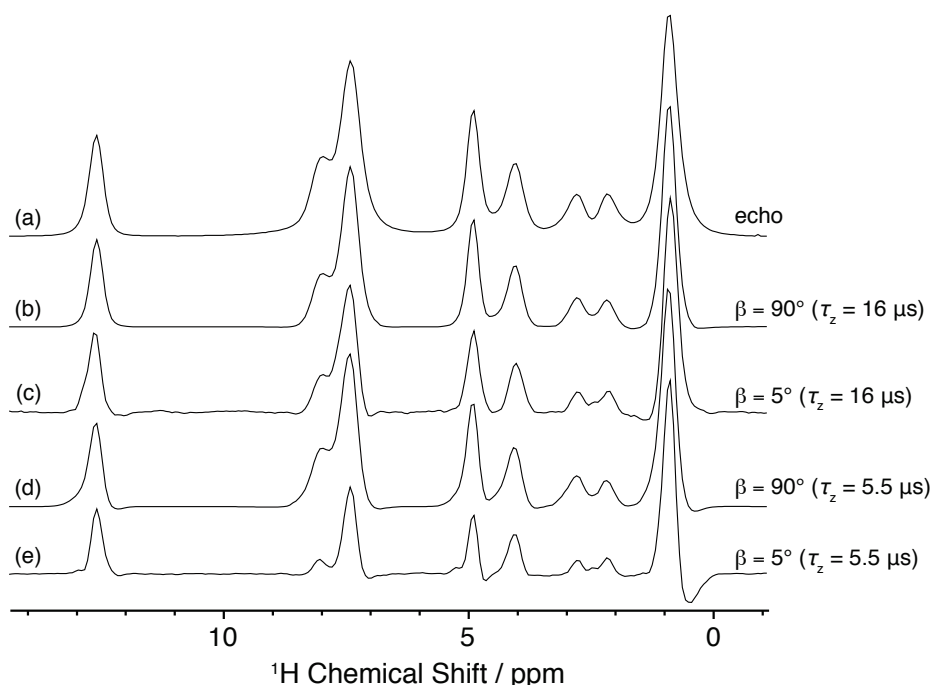


Figure 3.16. Echo-detected spectrum (a) and integral projections of the sheared 2D anti-z-COSY spectra (b-e) of β -AspAla at 62.5 kHz. Anti-z-COSY experiments were acquired using β values of 90° and 5° . Spectra were acquired both with and without rotor synchronization for the z-filter.

Chapter 4 Machine Learning Chemical Shifts

This chapter has been adapted with permission from:

*Paruzzo, F. M., Hofstetter, A., Musil, F., De, S., Ceriotti, M. and Emsley, L., "Chemical shifts in molecular solids by machine learning". Nature Communications **2018**, 9 (1), 4501. (post-print)*

A revolution in solid-state NMR has occurred with the introduction of accurate methods to calculate chemical shifts,²⁵⁸⁻²⁶⁰ in particular using plane wave DFT methods developed for periodic systems based on the PAW/GIPAW approach.^{104,105,261} As mentioned already in **Chapter 1.4**, this, in combination with the strong dependence of chemical shifts on local atomic environments, has enabled the development of chemical shift based NMR crystallography.

The power of the method arises from the fact that plane wave DFT with the GIPAW method is accurate enough to reproduce the exquisite sensitivity of chemical shifts to changes in local atomic environments. However, this approach also has severe limitations. The cubic scaling of the computational cost with system size prevents the application to larger and more complex crystals, or non-equilibrium structures. If one wanted to use more accurate ab initio calculations, the expense is prohibitive.

Machine learning (ML) is emerging as a new tool in many areas of chemical and physical science, and potentially provides a method to bridge the gap between the need for high accuracy calculations and limited computational power.²⁶²⁻²⁶⁶ Notably, prediction of chemical shifts for the specific case of proteins in solution using methods based on large experimental databases, using traditional²⁶⁷⁻²⁷⁴ or machine learning approaches,²⁷⁵⁻²⁷⁷ have met with considerable success in predicting shifts based on local sequence and structural motifs, and are widely used today. While there are some examples of machine learned experimental and ab-initio chemical shifts of liquid and gas phase molecules,²⁷⁸⁻²⁸² when this work has been published there was only one example of machine learning being applied to calculations of chemical shifts in solids, which deals with the specific case of silicas.²⁸³ Molecular solids are characterized by the combinatorial complexity and diversity of organic chemistry, the subtle dependence on conformations, and the long and short range effects of crystal packing, which leads to a considerably broader range of chemical environments and possible chemical shieldings than found e.g. in proteins. All of these aspects, compounded by the fact that there is no extensive database of experimental chemical shifts for molecular solids, make this class of systems particularly challenging for machine learning.

In this chapter, we develop a machine learning framework to predict chemical shifts in molecular solids, which we refer to in the following also as ShiftML, which is based on capturing the local environments of individual atoms, and which makes it well suited for the prediction of such local properties. We then show that the model has sufficient accuracy to be used in the isotropic chemical shift driven NMR crystallography protocol to correctly determine the crystal structure of cocaine, and AZD8329.

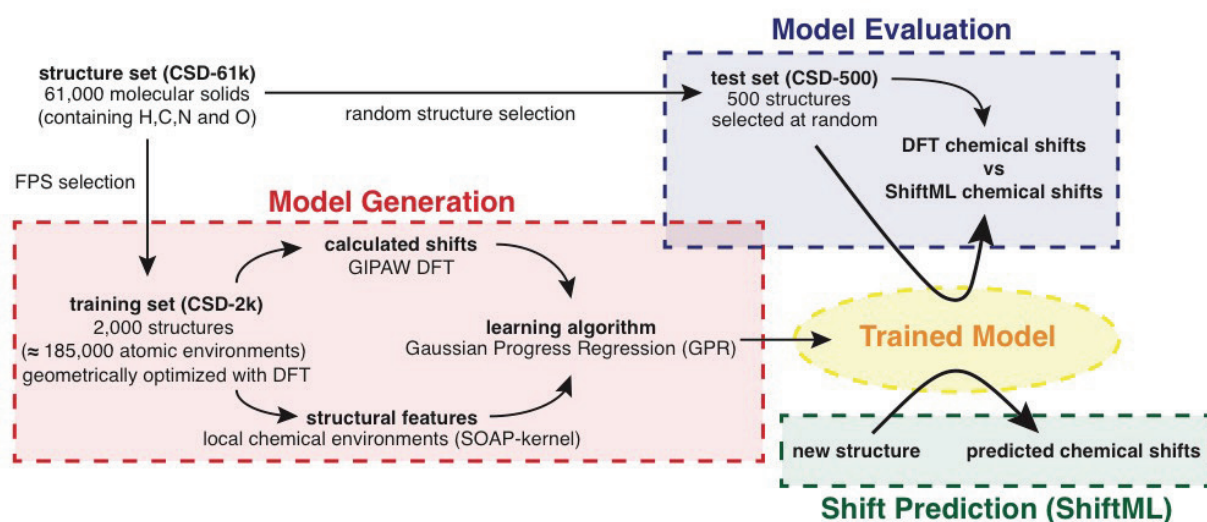


Figure 4.1. Scheme of the machine learning model used for the chemical shift predictions.

4.1 Materials and Methods

The protocol of the machine learning model is schematically illustrated in **Figure 4.1**. In the absence of a database of experimental shifts, and given that experiments alone do not provide a 1:1 mapping between chemical shifts and a single atomic configuration, we train the model on DFT calculated chemical shifts for structures taken from the Cambridge Structural Database (CSD).²⁸⁴

We first extract from the CSD a large set of about 61,000 structures, corresponding to all the structures in the CSD with fewer than 200 atoms, in order to make DFT chemical shift calculation affordable, and containing C and H and allowing for N and/or O, to reduce the space to organic molecular crystals (we call this set CSD-61k, see **Section 4.4** for details on the structures selection). Given that performing a GIPAW calculation for all of these structures would be prohibitively demanding, we then select a random subset of 500 structures (CSD-500) that are representative of the chemical diversity in the CSD, and we use it to test the accuracy of our model.

For cross-validation and training, instead, we select 2,000 structures (corresponding to about 185,000 atomic environments) out of the CSD-61k using a farthest point sampling (FPS) algorithm^{285,286} (CSD-2k). This step ensures near-uniform sampling of the conformational space, improving the quality of the model when using a relatively small number of reference calculations (see **Section 4.4** for more details).

To avoid including spurious environments in the model, e.g. environments that might not be well described by DFT, we also automatically detect and discard from the training set atomic environments with values of the DFT calculated shifts that are anomalous based on a cross validation procedure described **Section 4.4**. Note that using this unbiased statistical analysis we detected only a small fraction of environments as outliers (e.g. 211 out of 76,214 for ^1H , or 0.3%). We observe that the performance of the model degrades noticeably if one does not use this procedure. This pruning as well as the parameter optimization procedure, described below, were done exclusively using cross validation on the CSD-2k set. (Notably the test sets were not subject to any curation).

In order to reduce the computational cost of the training and testing procedures we then finally remove from the training set all the symmetrically equivalent environments. In case of ^1H , this reduced the size of the training set from 70,000 to about 35,000 different atomic environments.

All the atomic positions of the structures in the training and testing sets were geometrically optimized with DFT, using the Quantum Espresso suite,²⁸⁷⁻²⁸⁹ prior to calculation of the chemical shieldings using the GIPAW DFT method.^{104,105} Note that the DFT geometric optimization ensures “reasonable” geometries will be used even for crystal structures containing errors (e.g. improbable ^1H positions). Parameters for the DFT calculations are given

in **Section 4.4**. The calculated chemical shieldings σ are converted to the corresponding chemical shifts δ through the relationship $\delta = \sigma_{ref} - \sigma$. Here, we used a σ_{ref} of 30.8 ppm (for ^1H) and 169.5 ppm (for ^{13}C), found through linear regression between the calculated and experimental chemical shifts for cocaine.

We use the Gaussian Process Regression (GPR) framework²⁹⁰ to predict the chemical shift of a new atomic configuration based on a statistical model that identifies the correlations between structure and shift for a reference set of training configurations, for which the chemical shifts have been determined by a GIPAW DFT calculation. The predicted chemical shielding for a given atom is given by

$$\sigma(X) = \sum_i \alpha_i k(X, X_i), \quad (4.1)$$

where X and X_i correspond respectively to a description of the chemical environment of the atom for which we are making a prediction, and that of one of the training configurations. The weights α_i are obtained by requiring that **Equation 4.1** is consistent with the values computed by DFT for the reference structures. The essential ingredient that differentiates one GPR-based framework from another is the kernel function $k(X, X_i)$ which describes and assesses the similarity between atomic environments, and provides basis functions to approximate the target properties.

Here our model relies on the Smooth Overlap of Atomic Positions (SOAP) kernel. For the SOAP kernels,^{291,292} each atomic environment is represented as a three dimensional neighbourhood density given by a superposition of Gaussians, one centred at each of the atom positions in a spherical neighbourhood within a cut-off radius r_c from the core atom. The Gaussians have a variance ζ^2 , and a separate density is built for each atomic species. The kernel is then constructed as the symmetrized overlap between the amplitudes representing X and X' . This degree of overlap thus measures the similarity between the environments X and X' .

The SOAP and GPR parameters are given in **Appendix**. SOAP-based structural kernels contain several adjustable hyper-parameters, which are discussed in Ref. 292. However, we have not systematically explored the full parametric space here, instead we chose reasonable values of the parameters without extensive fine-tuning, based on previous experience²⁶³ and with some optimization by cross-validation on the CSD-2k training set. We also combine kernels computed for different cutoff radii to capture the contributions to shifts from different length scales,²⁶³ as is described in detail below. The calculations of the local environment, the similarity kernel and the weighted correlations were done using the glosim2 package.²⁹³

4.2 ShiftML

Training and Validation

Note that machine learning models must by definition be trained on the property that is to be predicted. Here that corresponds to experimental chemical shifts. However, for molecular solids there are currently only around 100 compounds with reliable crystal structures and for which assigned ^1H or ^{13}C shifts have been published, despite the rapidly increasing activity of NMR in crystal structure determination. This is at least an order of magnitude too few structures to hope to determine a reliable prediction model. In this light, we note that today GIPAW chemical shift calculations can accurately reproduce experimental shifts.^{116,294} Thus we propose to develop a machine learning model to predict chemical shifts by training the model on a database made up of GIPAW calculated shifts from a large and diverse set of reference crystal structures. *If the model can then accurately predict GIPAW chemical shifts, we hypothesise that it should also be in good agreement with experimental shifts.* We also note in this context that even if there was a database of experimental shifts, there would be a challenge to machine learning related to the fact that the experiment reports on structures that include dynamics or distributions, making the connection between shifts and environments ambiguous. Learning using GIPAW calculated shifts does not suffer from this problem.

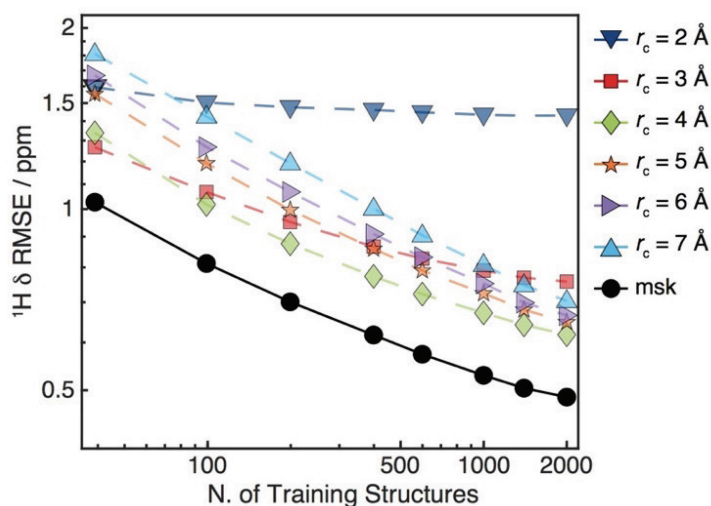


Figure 4.2. ^1H chemical shift prediction error of the trained model for the CSD-500 set. The RMSE prediction error between chemical shifts calculated with ShiftML and GIPAW DFT is shown for different local environment cut-off radii, and for the multi-kernel (labelled as msk), as a function of the training set size.

As mentioned above, our model is built using a SOAP kernel.^{291,292} This framework, combined with GPR, has been used to model the stability and properties of a number of different systems,^{263,292,295} and has been extended to the prediction of tensorial properties.²⁹⁶ We can see that this choice of kernel should be *particularly well adapted to predicting chemical shifts*, since it describes the local environments around each atom without any simplification, and this is indeed what the chemical shift also probes, as it is determined by the screening of the nucleus from the main magnetic field by the electron density at the nucleus. Note that it should be possible to tune and train other ML methods to accurately predict chemical shifts of molecular crystals. While these possibilities will be explored in future work, the model we present here is already accurate enough to substitute for DFT calculations in NMR crystallography.

Figure 4.2 shows the chemical shift error between the DFT calculations and the ML predictions for the CSD-500 set, which is representative of the expected accuracy for the entire CSD-61k. The figure shows the overall prediction accuracy for ^1H chemical shifts as RMSE in ppm between the shifts calculated with DFT and with ShiftML as a function of the cut-off radius (r_c) and as a function of the number of training structures included from CSD-2k. The effect of the different cut-off radii is clearly visible. For example, for $r_c = 2 \text{ \AA}$ the prediction error for a small training set (< 10 structures or < 100 atomic environments) can be smaller than for the larger radii, but does not improve significantly with increasing size of the training set. On the contrary, for $r_c = 7 \text{ \AA}$ we observe a relatively large prediction error for a small training set, but even with 2,000 structures (35,000 environments), the prediction error is still decreasing. A similar behaviour is observed for the prediction errors of the ^{13}C , ^{15}N and ^{17}O chemical shifts (see **Figure 4.12** in **Section 4.4**).

The observed differences in the behaviour of the prediction error with respect to r_c clearly indicates the influence of the different extents of the local environment on the chemical shift. Short range interactions are sufficient to explain the rough order of magnitude of the shift, but long range interactions are required to learn about the higher order influences of next-nearest neighbours on shifts. However, for long range interactions, a much larger number of environments is needed in order to determine the correlation between environment and shift.

We exploit these differences to generate a combined SOAP kernel consisting of a linear combination of the single local environment kernels,²⁶³ with weightings of 256 ($r_c = 2 \text{ \AA}$), 128 ($r_c = 3 \text{ \AA}$), 32 ($r_c = 4 \text{ \AA}$), 8 ($r_c = 5 \text{ \AA}$ and $r_c = 6 \text{ \AA}$) and 1 ($r_c = 7 \text{ \AA}$). This weighting was determined by rough optimization around values inspired by previous experience,²⁶³ and by cross-validation on the CSD-2k training set. It is clear that learning with the combined kernel leads consistently to lower prediction errors than any of the single kernels, although the improvement in performance varies between nuclei (see **Figure 4.12**).

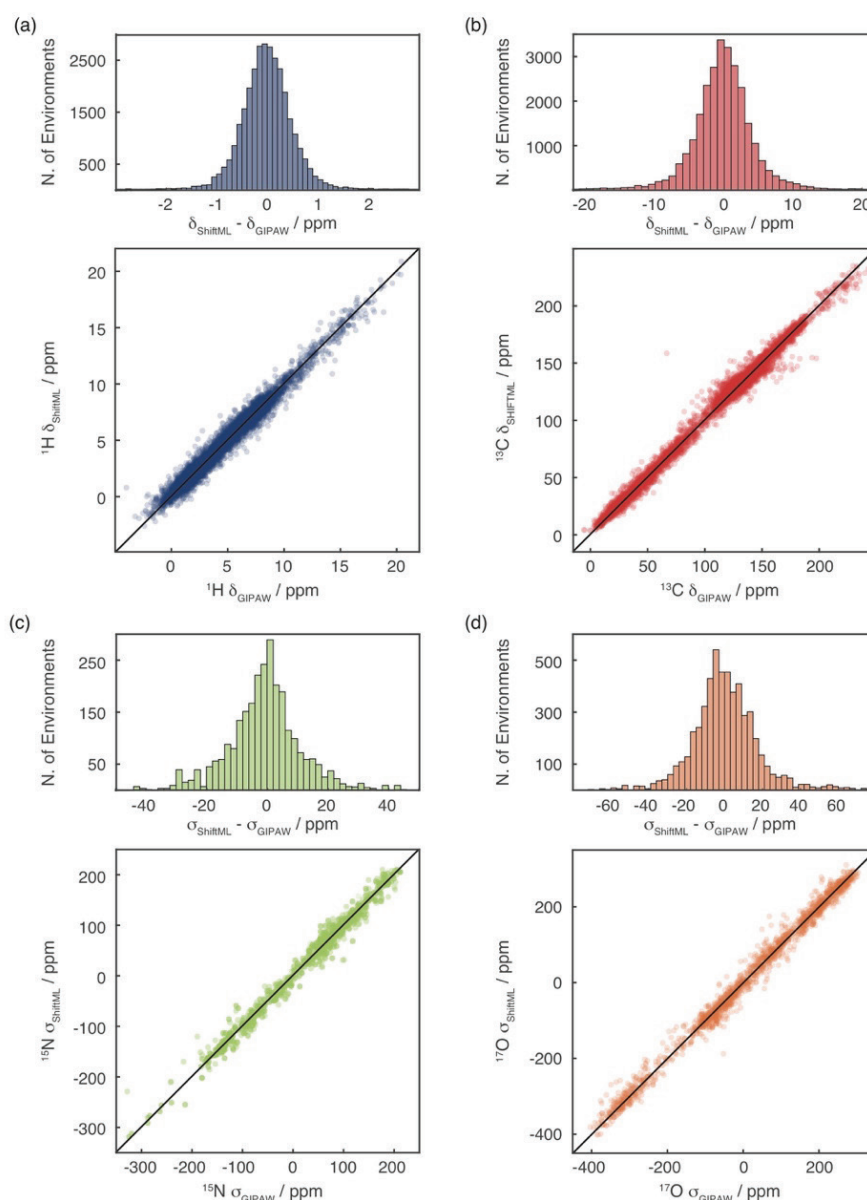


Figure 4.3. Comparison of predictions from ShiftML and GIPAW DFT. Histograms and scatterplots showing the correlation between ^1H (a), ^{13}C (b), ^{15}N (c) and ^{17}O (d) chemical shifts (shieldings) calculated with GIPAW and ShiftML. The black lines indicate a perfect correlation.

Figure 4.3a-d shows correlation plots between ^1H , ^{13}C , ^{15}N and ^{17}O chemical shifts calculated by DFT and by ShiftML for the CSD-500 set trained on the whole CSD-2k combined kernel. Using the combined kernel, we reach an error between ShiftML and DFT calculated chemical shifts of 0.49 ppm for ^1H (4.3 ppm for ^{13}C , 13.3 ppm for ^{15}N and 17.7 ppm for ^{17}O). This is very comparable with reported DFT chemical shift accuracy for ^1H of 0.33-0.43 ppm,^{116,294} while requiring a fraction of the computational time and cost: *less than 1 CPU minute compared to ~62-150 CPU hours for DFT chemical shift calculation on structures containing 86 atoms* (around 350 valence electrons) (see **Figure 4.10** in **Section 4.4**). For the other nuclei, the ML accuracy is slightly lower than reported values (1.9-2.2 ppm for ^{13}C , 5.4 ppm for ^{15}N and 7.2 ppm for ^{17}O),^{116,294} which is not surprising as there are (currently) significantly less training environments for the heteronuclei than for ^1H .

The R^2 coefficients between the chemical shifts calculated with DFT and with ShiftML are 0.97 for ^1H , 0.99 for ^{13}C , 0.99 for ^{15}N , and 0.99 for ^{17}O .

Note that the CSD-500 set used for testing is selected randomly from CSD-61k and not curated. Indeed, we find that many of the atomic environments in the CSD-500 set with a relatively high prediction RMSE possess either

unusual cavities inside their crystal structure, possibly indicating an organic cage surrounding non-crystalline solvent or other atoms, or exhibit strongly delocalised π -bonding networks. While there is no theoretical reason preventing the machine learning model from correctly describing such environments, they are rare and not well represented within the training set. CSD-500 thus constitutes a fairly demanding test set.

Predicting Shifts for Polymorphs

Having evaluated the power of the trained model to predict the diverse CSD-500 set, we now look at the capacity to predict potentially subtler differences by looking at a set of polymorphs of a given structure. **Figure 4.4a** and **b** show the correlation between the ^1H shifts calculated by GIPAW DFT and by ShiftML for 30 polymorphs of cocaine and 14 polymorphs of AZD8329, all of which were previously generated with a CSP procedure.^{120,121} The figure clearly shows that *ShiftML* is able to accurately predict the differences in ^1H chemical shift for different polymorphs.

We find a chemical shift prediction error (RMSE) between GIPAW DFT and ShiftML for ^1H for the cocaine polymorphs of 0.37 ppm and for AZD8329 of 0.46 ppm. Note that these values are slightly less than for the CSD-500 set, which might be expected when looking at these two fairly typical organic structures, and suggesting that the randomly selected CSD-500 indeed provides a good overall benchmark.

Note that for these cases the DFT structure optimization and GIPAW chemical shift calculation were done with a different DFT program (CASTEP),²⁹⁷ which suggests that ShiftML is robust with respect to small deviations from the fully optimized structures. (As shown in **Figure 4.9** in **Section 4.4**, performing the prediction using Quantum Espresso consistently leads to comparable prediction accuracy.)

For the heteronuclei we obtain an RMSE between GIPAW DFT and ShiftML for cocaine of 3.8 ppm for ^{13}C , 12.1 ppm for ^{15}N and 15.7 ppm for ^{17}O . For AZD8329 the ^{15}N and ^{17}O RMSEs are proportionally larger (17.7 and 54.7 ppm), and we attribute this to the fact that the molecule contains a rather unusual C-O...H-N / C-O...H-O H-bonded dimer structure, for which the learning is thus even sparser than for the heteronuclei in general. To illustrate the unusual nature of this motif, we note that the calculated ^{17}O shifts using DFT also change by up to 50 ppm for structures geometrically optimized either by the CASTEP protocol used in Ref. 121, or the Quantum Espresso protocol used here (the RMSE between ML and DFT for the Quantum Espresso geometrically optimized structures is reduced to 10.9 and 11.5 ppm for ^{15}N and ^{17}O !). The RMSE of 4.0 ppm for ^{13}C for AZD8329 is in line with the other systems.

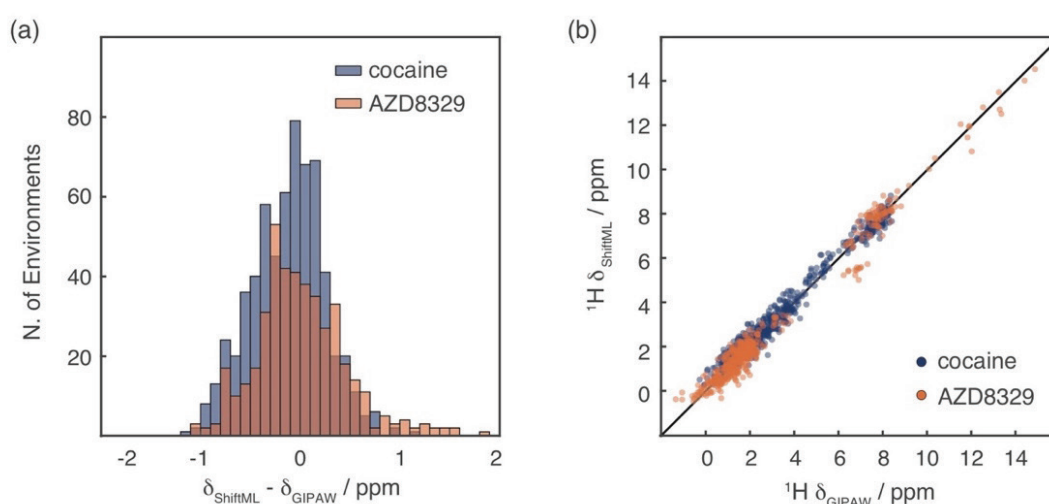


Figure 4.4. Comparison of predictions from ShiftML and GIPAW DFT for polymorphs of cocaine and AZD8329. (a) Histogram showing the distribution of the differences between ^1H chemical shifts calculated with GIPAW and with ShiftML for the polymorphs of cocaine (blue), and the polymorphs of AZD8329 (orange). (b) Scatterplot showing the correlation between ^1H chemical shifts calculated with GIPAW and ShiftML for cocaine (blue) and AZD8329 (orange). The black line indicates a perfect correlation.

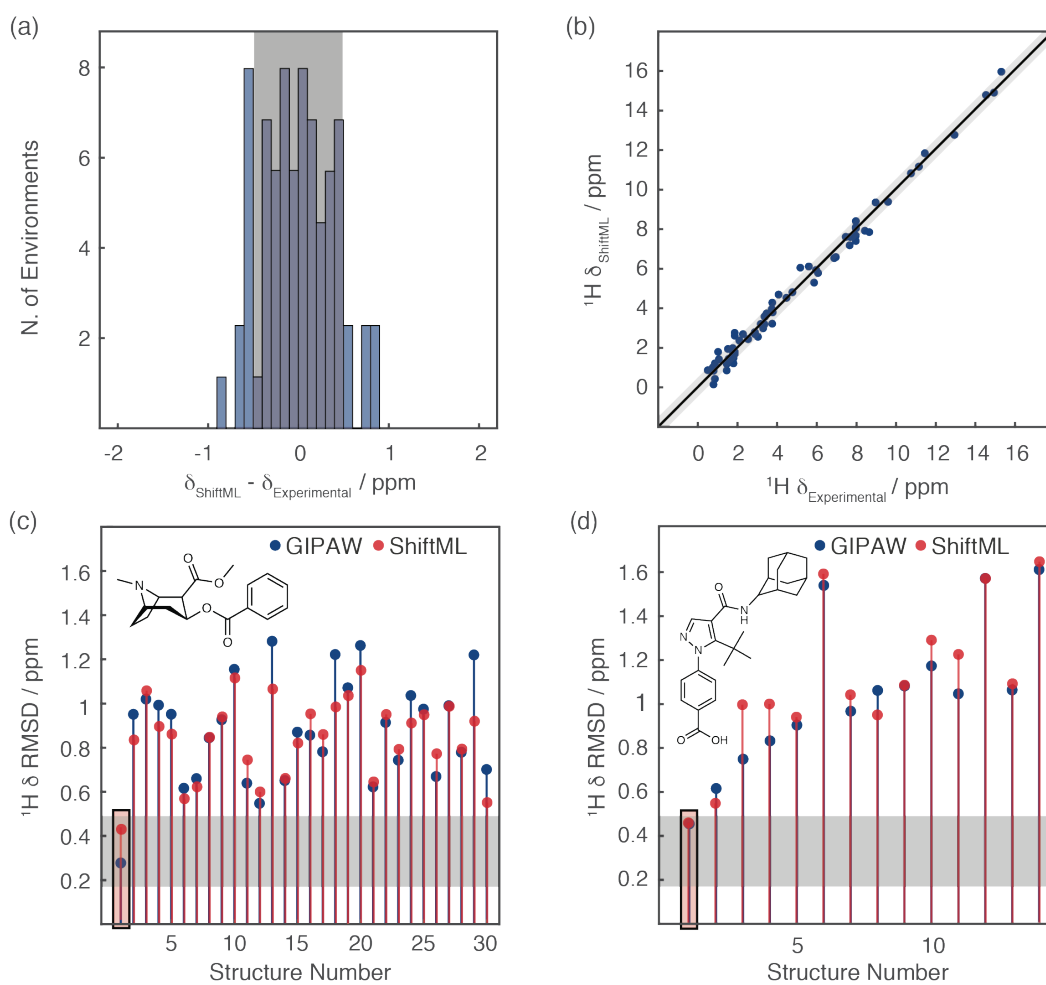


Figure 4.5. Comparison of ShiftML to experimentally measured shifts. (a) Histogram showing the distribution of differences between experimentally measured ^1H chemical shifts and ^1H chemical shifts calculated with ShiftML for six different crystal structures (see **Appendix** for the structures and numerical values of the shifts). (b) Scatterplot showing the correlation between these experimentally measured ^1H chemical shifts and shifts calculated with ShiftML. (c-d) Comparison between calculated and experimental ^1H chemical shifts for the most stable structures obtained with CSP for cocaine (c) and AZD8329 (d). These structures were obtained from literature.^{120,121} For each candidate structure an aggregate RMSE is shown between experimentally measured shifts and shifts calculated using either GIPAW (blue) or ShiftML (red). The grey zones represent the confidence intervals of the GIPAW DFT ^1H chemical shift RMSE, as described in the text,¹¹⁶ and candidates (in c and d) that have RMSEs within this range would be determined as correct crystal structures using a chemical shift driven solid-state NMR crystallography protocol.

Predicting Experimental Shifts and Structure Determination

Further, the significance of the method is illustrated by comparison to experimentally measured shifts. This comparison is particularly important since the training protocol did not involve any experimentally measured chemical shifts. We find that *the predicted shifts are accurate enough to allow crystal structure determination for both cocaine and AZD8329 from powder samples using the isotropic chemical shift based NMRX approach.*

Figure 4.5a and **b** show the correlation between experimentally measured ^1H chemical shifts and the ^1H chemical shifts calculated by ShiftML for crystal structures of the six molecules shown in **Figure 4.6** (numerical values of the experimental chemical shifts and the crystal structures are given in **Appendix**). The comparison between experimental and calculated ^1H chemical shifts for all crystal structures (for a total of 68 shifts) gives an error (RMSE) of 0.39 ppm and a R^2 coefficient of 0.99. This compares very favourably to the equivalent agreement found between GIPAW DFT and experiment which for this set of structures is an RMSE of 0.38 ppm.

Figure 4.5c and **d** show in blue the RMSE between DFT calculated and experimental ^1H chemical shifts for the 30 polymorphs predicted by CSP to have the lowest energy for cocaine and the 14 *cis* polymorphs of AZD8329. For both molecules the only structure in agreement with the GIPAW DFT calculations, to below a ^1H DFT chemical shift

confidence interval of 0.49 ppm,¹¹⁶ is the correct crystal structure. In the same plots we overlay the result where the experimental shifts are now compared to shifts predicted with ShiftML. Note that the RMSE between experiment and the predicted chemical shifts follows the same trends as for the DFT calculated shifts, and that here again the only structures below the confidence interval of 0.49 ppm are the two correct crystal structures. Note, that the cut-off of 0.49 ppm with respect to experiment has been evaluated for GIPAW DFT chemical shifts^{116,294} and to rigorously repeat the CSP procedure for the ML method, the accuracy should be re-evaluated using more extensive benchmarking of ShiftML to experiment, which will be the subject of further work.

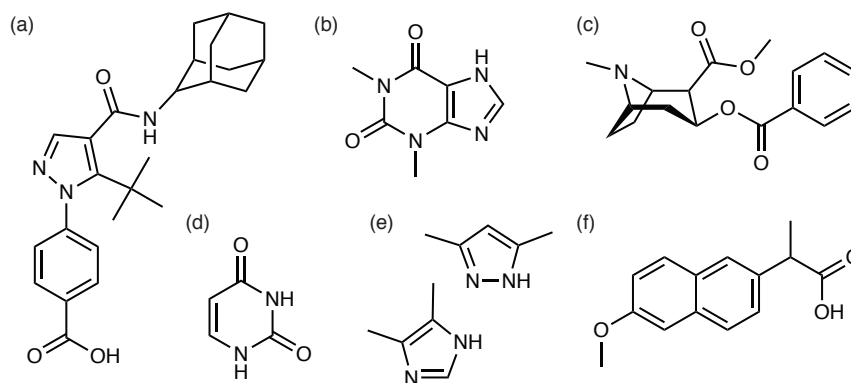


Figure 4.6. Chemical structures of the six molecules used to evaluate the correlation between experimentally measured ^1H chemical shifts and the shifts calculated by ShiftML. The structures are given as AZD8329 (a), theophylline (b), cocaine (c), uracil (d), 3,5-dimethylimidazole and 4,5-dimethylimidazole (e) and naproxen (f).

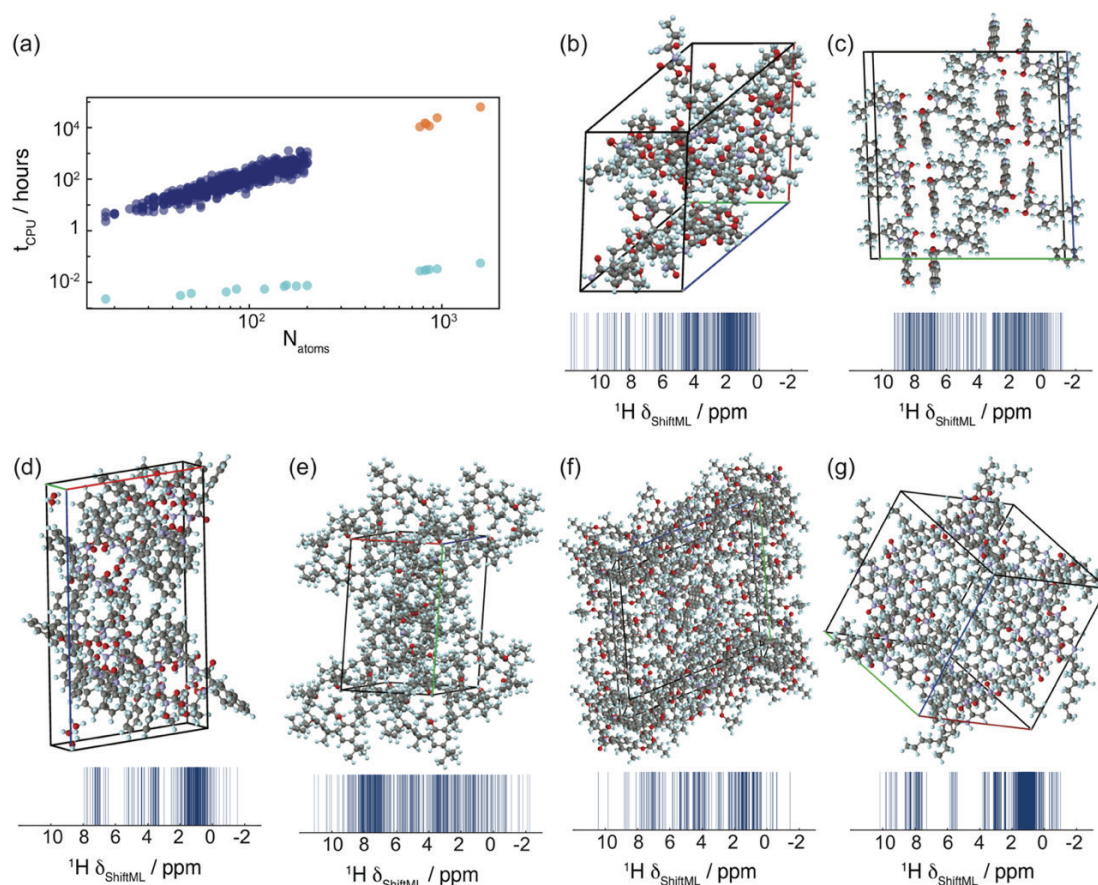


Figure 4.7. Chemical shift calculation times and large structures. (a) DFT GIPAW calculation time (blue) and ShiftML prediction time (turquoise) for different system sizes. The GIPAW DFT calculation time for the six large structures (orange) is estimated from a cubic dependence on the number of valence electrons in the structure (see **Section 4.4**). (b-g) 3D-schemes and ^1H NMR spectra predicted with ShiftML, of the six large molecular crystals with CSD Refcodes: (b) CAJYUH,²⁹⁸ $N_{\text{atoms}} = 828$, (c) RUKTOI,²⁹⁹ $N_{\text{atoms}} = 768$, (d) EMEMUE,³⁰⁰ $N_{\text{atoms}} = 860$, (e) GOKXOV,³⁰¹ $N_{\text{atoms}} = 945$, (f) HEJBUW,³⁰² $N_{\text{atoms}} = 816$, (g) RAYFEF,³⁰³ $N_{\text{atoms}} = 1,584$.

Predicting Shifts for Large Structures.

Finally, we note that the accuracy of the method does not depend on the size of the structure, and that the prediction time is linear in the number of atoms. For the structures we calculate here the prediction time actually appears nearly constant, because it is dominated by the loading time of the reference SOAP vector (see **Figure 4.7a**). We have used this method to calculate the NMR spectra (shown in **Figure 4.7b-g**) for six structures from the CSD having among the largest numbers of atoms per unit cell (containing only H, C, N, O), with between 768 and 1,584 atoms per unit cell. (See **Figure 4.13** in **Section 4.4** for the chemical formula). The values of the predicted chemical shifts are shown in **Figure 4.13**. **Figure 4.7a** shows the comparison between the GIPAW calculation time and the required ML prediction time. *We estimate that the whole calculation would require around 16 CPU years by GIPAW. ShiftML requires less than 6 CPU minutes to calculate the shifts for all the compounds.*

4.3 Conclusions

In this chapter, we have presented a ML model based on local environments to predict chemical shifts of molecular solids containing HCNO to within current DFT accuracy. The R^2 coefficients between the chemical shifts calculated with DFT and with ShiftML are 0.97 for ^1H , 0.99 for ^{13}C , 0.99 for ^{15}N , and 0.99 for ^{17}O , corresponding to RMSEs of 0.49 ppm for ^1H , 4.3 ppm for ^{13}C , 13.3 ppm for ^{15}N , and 17.7 ppm for ^{17}O . The approach allows the calculation of chemical shifts for structures with ~ 100 atoms in less than 1 minute, reducing the computational cost of chemical shift predictions in solids by a factor of between 5 to 10 thousand compared to current DFT chemical shift calculations, and thereby relieves a major bottleneck in the use of calculated chemical shifts for structure determination in solids.

Far from being just a benchmark of a machine-learning scheme, the method is accurate enough to be used to determine structures by comparison to experimental shifts in chemical shift based NMRX approaches to structure determination, as shown here for cocaine and AZD8329. The ML model only scales linearly with the number of atoms and, for the prediction of individual structures, is dominated by a constant I/O overhead. Here it allows the calculation of chemical shifts for a set of six structures with between 768 and 1584 atoms in their unit cells in less than six minutes (an acceleration of a factor 10^6 for the largest structure).

The accuracy of the method is likely to increase further with the size of the training set, and subsequently with the future evolution of the accuracy of the method used to calculate the reference shifts used in training (here DFT), or by using experimental shifts if a large enough set were available. A web version based of ShiftML is publicly available at <http://shiftml.epfl.ch>. The model used here can easily be extended to organic solids including halides or other nuclei, and to network materials such as oxides, and these will be the subject of further work.

Note that the current version of ShiftML available at <http://shiftml.epfl.ch> has already been updated (see **Chapter 5**). New features includes: (i) the extension of the training set to 3564 structures, (ii) the opportunity to upload structures containing H and C, and allowing for N, O and also S, (iii) the replacement of the seven multi-scale SOAP kernel with a radially-scaled SOAP kernel^{290,304,305} (which reduces memory requirements while maintaining comparable accuracy), (iv) the implementation of a strategy to calculate the uncertainties associated with the chemical shifts predictions³⁰⁶ and (v) the sparsification of the SOAP fingerprints with the FPS algorithm³⁰⁷ (which further accelerate the ML predictions). Finally, note also that after the publication of the first ShiftML paper, a few other ML models have been developed for the calculation of chemical shifts in solids.^{308,309}

4.4 Supplementary Information and Discussion

Selection of the Training and Test Sets

All crystal structures contained in the CSD-2k and CSD-500 sets were obtained from the Cambridge Structural Database (CSD).²⁸⁴ A total of 88,648 structures was downloaded from the CSD, using two different selection criteria: the maximum number and the type of atoms contained in the unit-cell. We selected only structures with

fewer than 200 atoms, and containing C and H and allowing for N and/or O. From this set we extracted a subset of 61,012 (CSD-61k) structures by removing (i) structures with missing protons, and (ii) structures where the distance of at least one pair of atoms was smaller than the sum of their covalent radii minus 0.3 Å. In addition, structures containing partial occupancy were resolved by keeping only the first of the atoms with partial occupancy. If we were not able to resolve the disorder, the entire structure was not included. The disorder was assumed to be removed, if the number of atoms, for each atom type, was an integer multiple of the number of atoms given in the chemical formula. Note that, as we sorted through more than 60,000 structures, the whole procedure was automatized and we did not manually select the most stable structure for a given disorder. However, here we are not looking for ground state structures but instead only for physically reasonable structures to expand our data-set. The remaining structures were then used to create both the training (CSD-2k) and the test set (CSD-500) for ShiftML. The CSD-500 set was created by randomly picking 500 structures from the CSD-61k, excluding the structures already selected for the training set.

DFT Calculations.

All the DFT calculations were carried out using the DFT program Quantum ESPRESSO.^{287,288} For all structures in the CSD-2k and CSD-500 databases we first carried out geometry optimization using plane wave DFT. We used ultrasoft pseudopotentials with GIPAW^{104,105} reconstruction, H.pbe-kjpaw_psl.0.1.UPF, C.pbe-n-kjpaw_psl.0.1.UPF, N.pbe-n-kjpaw_psl.0.1.UPF and O.pbe-n-kjpaw_psl.0.1.UPF from the USSP pseudopotential database [<http://www.quantum-espresso.org/pseudopotentials>].³¹⁰ The optimizations were carried out using the generalized-gradient-approximation (GGA) density functional PBE,³¹¹ with a wave-function energy cut-off of 60 Ry, a charge density energy cut-off of 240 Ry and without k-points. The Grimme van der Waals dispersion correction³¹² was included in order to account for van der Waals interactions. The geometry optimizations were done relaxing all atomic positions while keeping the lattice parameters fixed.

The single-point energy was then computed for each geometrically optimized structure, using higher wave-function and charge density energy cut-offs of 100 Ry and 400 Ry respectively. For these calculations we also used a Monkhorst-Pack grid of k -points³¹³ corresponding to a maximum spacing of $2\pi \times 0.06 \text{ \AA}^{-1}$ in the reciprocal space. The k -points and energy cut-off values were optimized to ensure convergence of the electron density. Finally, we calculated the chemical shieldings σ_{DFT} using the GIPAW method, with the same parameters as used for the single-point energy calculation.

Note that using a convergence threshold in the single-point energy calculation of 10^{-8} Ry leads to a residual random error on the macroscopic contribution to the shifts of the order of 0.1 ppm. Fully converged results can be achieved with a threshold of 10^{-12} - 10^{-14} Ry.

Machine Learning.

We model the isotropic chemical shielding as a function of the local environment A using a Gaussian Process Regression (GPR) framework, that assumes that chemical shift values predicted by the model can be written as

$$\sigma(A) = f(A) + \varepsilon, \quad (4.2)$$

where the function f is a Gaussian Process²⁹⁰ and ε represents the error of the prediction, which is modeled as independent identically distributed Gaussian variates, with variance σ_n^2 . Following the GPR framework, the isotropic chemical shielding function becomes:

$$\sigma(A) = \sum_{i=1}^N \alpha_i k(A, X_i)^\zeta, \quad (4.3)$$

where $\{X_i\}_{i=1}^N$ is a training set of N reference local environments for which the isotropic chemical shieldings are known, k is a kernel function measuring the covariance between local environments and ζ is a hyperparameter controlling the sensitivity of the kernel.

The weights α_i can be computed by inverting the kernel matrix $K_{ij} = k(X_i, X_j)$ computed between the reference configurations, including a regularization that depends on an estimate of the intrinsic uncertainty in the fit (which is due to errors in the training set, the limitations of the model or the reduced number of training configurations)

$$\alpha_i = \sum_j [K^\zeta + \sigma_n^2 \mathbf{1}]^{-1}_{ij} \sigma(X_j). \quad (4.4)$$

To assess the correlation between local atomic environments A and B , we use the SOAP kernel²⁹¹ defined by the rotationally invariant overlap between smooth representations of their atomic density:

$$k(A, B) = \int_{SO(3)} \left| \int_{\mathbb{R}^3} \rho_A(\vec{r}) \rho_B(\hat{R}\vec{r}) d\vec{r} \right|^2 d\hat{R}, \quad (4.5)$$

where the density is built as a superimposition of Gaussians having width ς , centered on the atoms within a cutoff distance of the central atom in the environment

$$\rho_A(\vec{r}) = \sum_{i \in A} \exp[-\|\vec{r} - (\vec{r}_i - \vec{r}_A)\|^2 / 2\varsigma^2] f_c(|\vec{r}_i - \vec{r}_A|). \quad (4.6)$$

The details of the construction, and the extension to the case with many atomic species, are given in Refs. 263 and 292.

Farthest Point Sampling Algorithm.

Given that a GPR model is essentially an interpolation procedure between the reference configurations, it is crucial that training points are chosen to cover as uniformly as possible the space of structures for which one wants to perform predictions. To achieve this uniform sampling we use a FPS algorithm^{285,286} to sort the CSD-61k in descending order of “diversity”. Essentially, we select the first structure at random, and then pick the others in the sequence such that

$$k = \underset{k \in \text{CSD-61k}}{\operatorname{argmax}} \min_{j \in \text{selection}} |X_k - X_j|, \quad (4.7)$$

where the distance is the kernel-induced distance associated with an average SOAP kernel for the entire structure.²⁹² The CSD-2k set corresponds to the first 2,000 configurations identified with this procedure.

Detection of Unusual Environments.

The quality of the training set is essential to ensure the optimal performance of a machine learning algorithm. However, the individual curation of the 2,000 molecular crystals of the CSD-2k dataset would be very time consuming and cumbersome. Note, that the 2,000 molecular crystals correspond to around 35,000 symmetrically non-equivalent atomic environments for ^1H alone and the following detection procedure is applied directly to the individual atomic environments instead of the whole molecular crystals.

We automate this detection procedure by assessing the “instability” of the prediction of the shielding of a given local environment using the difference between the predictions of several GPR models and the reference DFT-shielding.

We define this indicator as:

$$\varepsilon(X) = \frac{1}{M} \sum_{i=1}^M (y_i(X) - y(X)), \quad (4.8)$$

where each of the M models is made using a 2-fold split of the shuffled training set that does not include the structure X . In total we generate $M=40$ models, where each is generated using a different random shuffling of the data. Environments with a large value of $|\varepsilon(X)|$ are not well-described by the rest of the training set within the SOAP-GPR framework. Note, that the error would cancel out in the case of random noise within the prediction, while a large value of $|\varepsilon(X)|$ corresponds to a systematic error in the predicted chemical shielding, that could be associated to the limitations listed below. We define local environments to be unusual when $|\varepsilon(X)|$ is larger than three times the standard deviation of $|\varepsilon(X)|$ over the whole training set, and we then do not use them for training.

We perform this elimination procedure on the CSD-2k dataset using a single kernel for each element ($r_c = 4.5$ Å for ^1H , 4 Å for ^{13}C , 4 Å for ^{15}N and 3 Å for ^{17}O). The hyperparameters of the single kernels used in the elimination procedure were determined using a grid search and 3-fold cross validation on the uncleaned CSD-2k training set. The ^1H environments excluded with this approach are shown in **Figure 4.8**.

It is interesting to see that in several cases we can trace the unusual behavior of the environment to subtle errors in the DFT calculations, or to physical phenomena that are ill described within our DFT model (metallic systems, zwitterions,...). However, note that we are not systematically removing such structures and that the training set still contains many structures with the listed features. More details on the environments detected as “unusual” are given in <https://doi.org/10.1038/s41467-018-06972-x>.

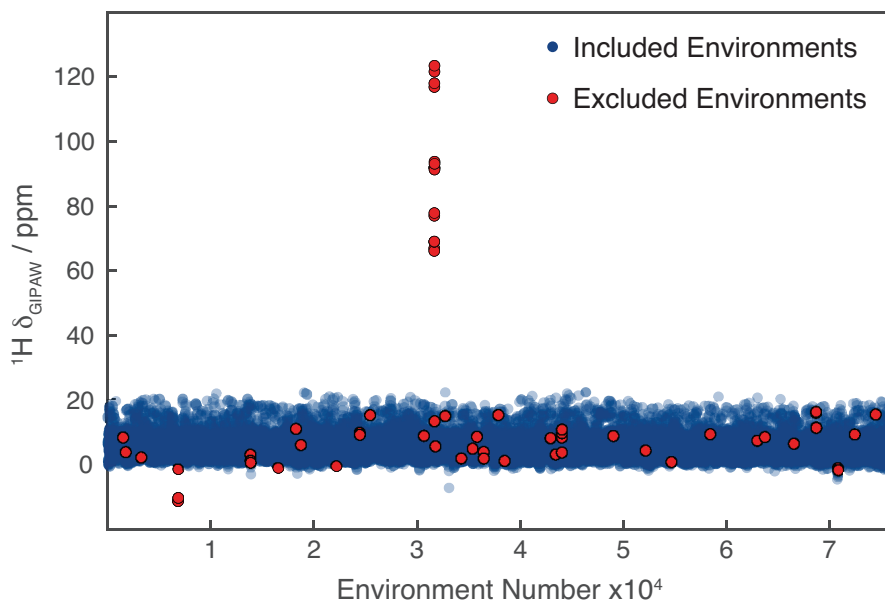


Figure 4.8. ^1H chemical shifts of the 76,214 environments in the CSD-2k set. The environments excluded using the unusual structures detection procedure described in the text are shown in red.

NMR Crystallography.

To validate the accuracy of the chemical shifts calculated with ShiftML, we replicated the last step of the protocol for isotropic chemical shift based NMRX.^{116,120} We perform this analysis for cocaine and form 4 of AZD8329.^{120,121} The value σ_{ref} for the conversion between chemical shieldings to chemical shifts is calculated for each structure with a linear regression between calculated and experimental shifts, imposing a slope equal to 1. This procedure is done independently for the ^1H chemical shieldings calculated with DFT and ShiftML. The geometry of the structures predicted with CSP, as well as their chemical shift values calculated with GIPAW and the experimental chemical shifts of the observed polymorphs were obtained from Refs. 120 and 121. Remarkably, the high accuracy shown in **Figure 4.5** was obtained using crystal structures with only ^1H positions geometrically optimized and DFT

chemical shift calculations carried out using a different program (CASTEP) to the one we used to build our training set (Quantum Espresso). **Figure 4.9** shows the results obtained for cocaine and AZD8329 after all-atom optimization and calculation of GIPAW chemical shifts with Quantum Espresso. Here we show fewer structures compared to **Figure 4.5**, due to the fact that we limit ourselves to calculate DFT chemical shifts of structures with less than 250 atoms. This selection removes structures 15 for cocaine and structures 2, 11 and 14 for AZD8329. The accuracy is consistent with that reported in **Figure 4.5**, although the all-atom optimization leads to some significant structural differences compared to the only ^1H geometrically optimized structures, especially for AZD8329. We find a chemical shift prediction error (RMSE) for ^1H for cocaine of 0.40 ppm and for AZD8329 of 0.51 ppm, which is very comparable to the expected GIPAW DFT accuracy. For the heteronuclei we obtain, for cocaine and AZD8329 respectively, 3.5 and 3.4 ppm for ^{13}C , 9.3 ppm and 11.0 ppm for ^{15}N and 12.2 ppm and 11.5 ppm for ^{17}O . Details on the chemical shift values used for this analysis are given in **Appendix**.

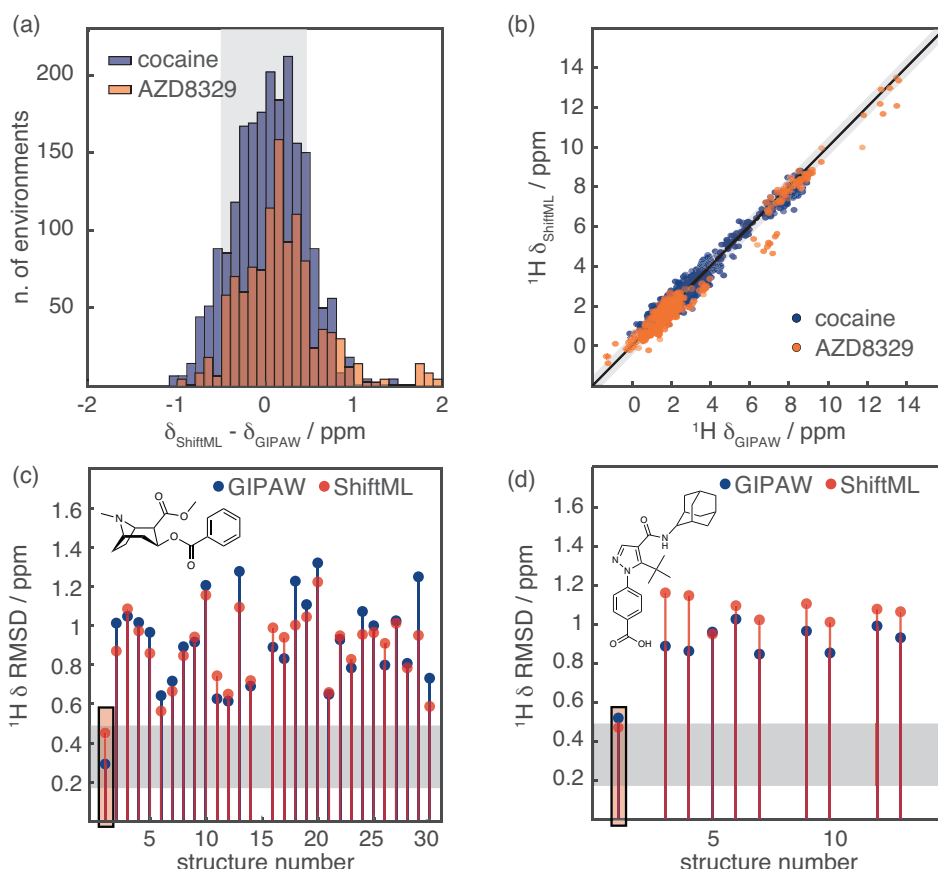


Figure 4.9. NMR crystallography of cocaine and the form 4 of AZD8329. (a) Histogram showing the distribution of the differences between chemical shifts calculated with GIPAW and ShiftML. The blue bars were calculated for the polymorphs of cocaine, and the orange ones for the polymorphs of AZD8329. (b) Scatterplot showing the correlation between GIPAW and ShiftML chemical shifts for cocaine (blue) and AZD8329 (orange). The black line indicates a perfect correlation. (c-d) Comparison between calculated and experimental ^1H chemical shifts for the most stable structures obtained with CSP for cocaine (c) and form 4 of AZD8329 (d). Chemical shifts were calculated using GIPAW (blue) and ShiftML (red). The highlighted bars correspond to the candidates that would be selected as correct crystal structures using the chemical shift based solid-state NMR crystallography protocol. In (a-d) the grey zones represent the confidence intervals of the ^1H chemical shift RMSD, as described in the text.¹¹⁶

DFT Calculation Times

Figure 4.10 shows the CPU time needed for part of the GIPAW DFT calculations done for this work. The calculations shown in **Figure 4.10a** were done on polymorph 1 of the cocaine dataset, which contains 86 atoms per unit-cell, while the ones in **Figure 4.10b** were done on 500 structures of the CSD-2k set. In **Figure 4.10a** the calculation time is plotted as a function of the number of Monkhorst-pack k-points³¹³ per axis for three different energy-cut-off (E_{cutoff}) values: 40 Ry (blue), 70 Ry (red), 100 Ry (yellow). When increased, these two parameters improve the accuracy of the calculation, but at the same time they drastically increase the computational time needed to carry

out the calculation. **Figure 4.10b** shows the CPU time for the GIPAW chemical shift calculations (blue dots) and for the DFT structure optimizations (green squares) as a function of the number of valence electrons (N_e) per unit-cell. For the GIPAW chemical shift calculations the energy-cut-off was 100 Ry, using a Monkhorst-pack grid with a k-point spacing of $2\pi \times 0.06 \text{ \AA}^{-1}$. For the DFT structure optimizations the energy-cut-off was 60 Ry and no k-points were used. The red line shows the best fit between the number of valence electrons and the required CPU time for the GIPAW chemical shift calculations as $t_{\text{CPU}} = aN_e^2 + bN_e^3$, where the N_e^3 scaling accounts for the general DFT scaling and the N_e^2 describes the scaling of the matrix inversion, which dominates for small system sizes. The best fit parameters are given as 8.83e^{-04} (a) and 1.02e^{-06} (b).

Currently the machine learning model has only been rigorously tested and applied for structures optimized with DFT. Also slight structural changes away from the equilibrium geometry of a molecular crystal have been shown to result in significant changes in the chemical shifts.¹⁸ For this reason, the predictive accuracy of ShiftML for non-equilibrium structures has not yet been quantified. However, **Figure 4.10b** clearly shows that the computational cost for the structure optimization is negligible compared to the computational cost of the GIPAW chemical shift calculations.

For structures with $N_e \approx 100$ the GIPAW shift calculations require around 10x more CPU time as the DFT structure optimization, and for $N_e \approx 1,000$, 80x more CPU time is required.

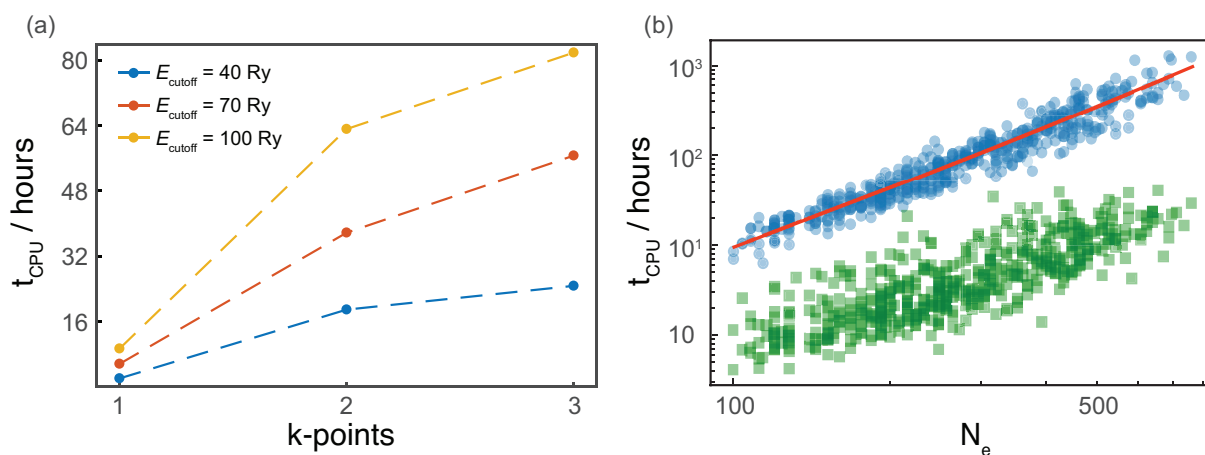


Figure 4.10. CPU time for NMR chemical shift calculations using the GIPAW method. (a) The CPU time is shown as a function of the DFT accuracy, determined by the plane-wave cutoff energy E_{cutoff} and the number of k-points in each dimension for polymorph 1 of cocaine. The charge density energy cut-offs were set to $E_\rho = 4E_{\text{cutoff}}$. (b) The CPU time is shown as function of increasing system size in CSD-2k. The green squares and blue dots show individual geometry optimization and GIPAW chemical shift DFT calculations, respectively. The red line shows the best fit between the number of valence electrons and the required CPU time as $t_{\text{CPU}} = aN_e^2 + bN_e^3$ (8), with $a = 0.0162$ and $b = 5.91\text{e}^{-06}$.

ShiftML Prediction Times

The ShiftML run-times are shown in **Figure 4.7**. The prediction times scale linearly with the number of atoms per unit cell. However, for all the structures investigated here (from 20 to 1,500 atoms per unit-cell) the required prediction time is dominated by a constant pre-factor associated with the used training set.

Prior to the prediction step, the SOAP reference vector between the test and the training structures is created. This step should be linear in the size of the test-structures, but is currently dominated by the size of the training set. As a result, this takes around one CPU minute for any of the investigated structures here. The actual subsequent chemical shift prediction, which is linear in the number of atoms within the test-structure, requires at most 10-20 CPU seconds for the large investigated structures.

Note that prior to the chemical shift predictions, the single kernels for all the atomic species must be loaded into virtual memory and the multiscale kernel created. On one CPU this currently takes around 45 minutes. Note, that

this has to be done only once, independently of the number and size of the test-structures that are subsequently calculated.

Learning and Evaluation Curves

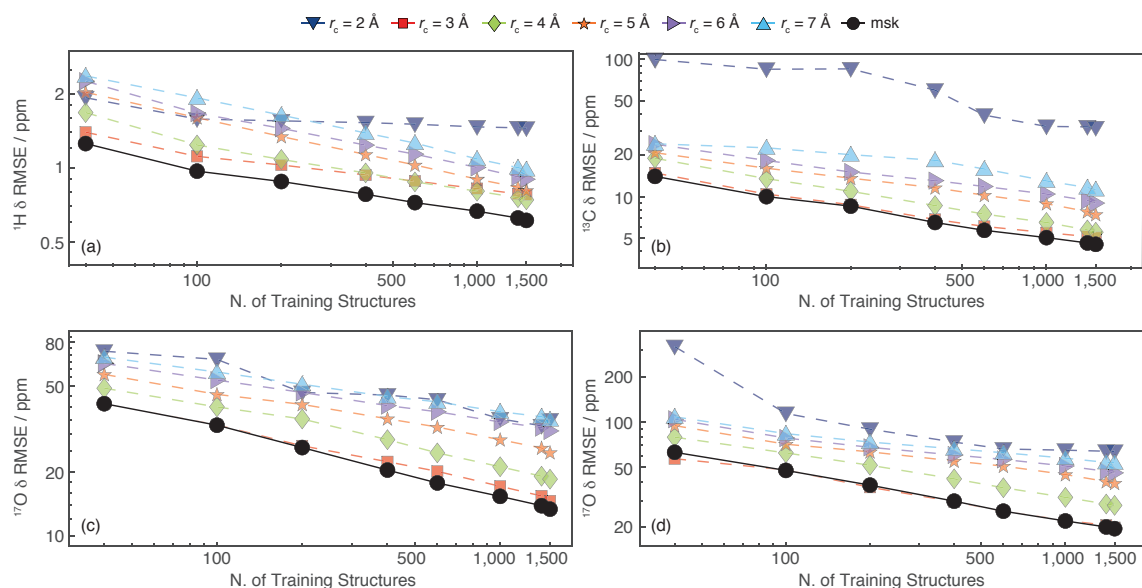


Figure 4.11. RMSE learning curves showing the error between chemical shifts calculated with DFT and ShiftML. The curves are for ^1H (a), ^{13}C (b), ^{15}N (c) and ^{17}O (d) chemical shieldings. The multi-kernel learning-curve is labelled as msk.

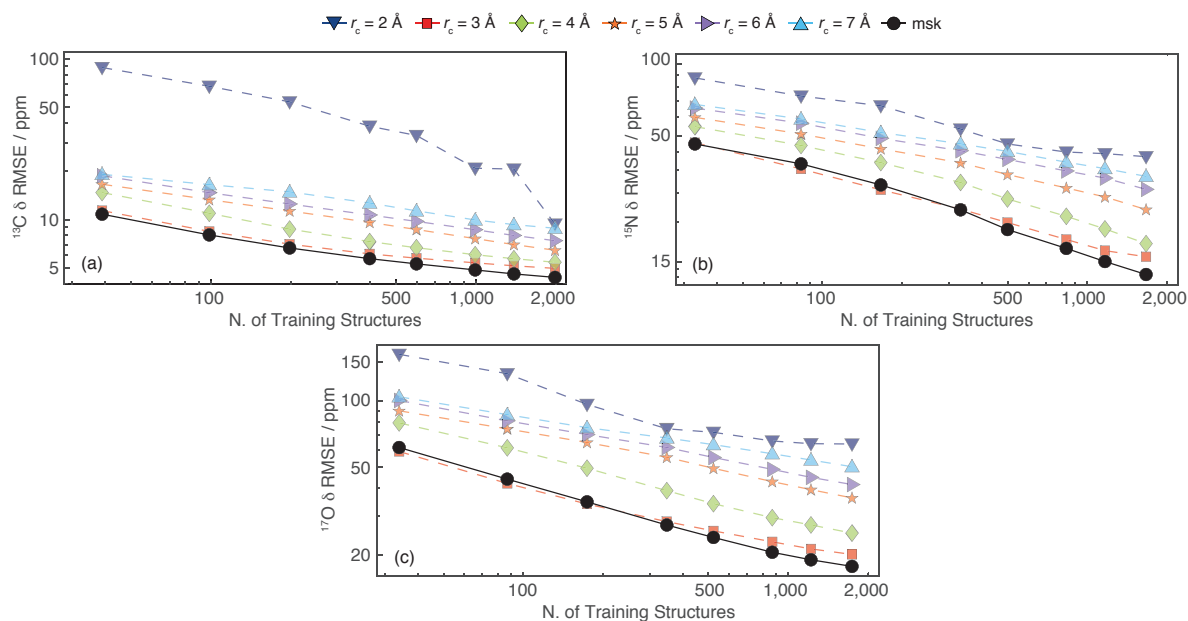


Figure 4.12. RMSE evaluation curves showing the error between chemical shifts calculated with DFT and ShiftML. The curves are relative to ^{13}C (a), ^{15}N (b) and ^{17}O (c) chemical shieldings. The errors were measured for different training set sizes, and evaluated on the CSD-500 test set. The RMSE evaluation curves were acquired as described in the paper. The multi-kernel learning-curve is labelled as msk.

Figure 4.11 shows the RMSE learning curves for ^1H , ^{13}C , ^{15}N and ^{17}O for the different local environment cut-off radii, and for the multi-kernel. The training was done on up to 1500 randomly selected frames, while testing on 400 structures selected randomly from the CSD-2k set excluding the structures already selected for the training set. For each point, the random sampling was repeated N times (where N is equal to 300, 255, 215, 170, 130, 85, 45 and 5 respectively for training set sizes of 40, 100, 200, 400, 600, 1000, 1400 and 1500 structures)

Figure 4.12 shows the results of the predictions of the chemical shifts of the CSD-500 set as a function of the cut-off value and the size of the training set. The parameters for the multi-scale kernel prediction were optimized using 3-fold cross validation on the CSD-2k set. More details are given in <https://doi.org/10.1038/s41467-018-06972-x>.

Structures and Chemical Shifts of the CSD-6 Set

For all of the structures in CSD-6 we removed atoms with partial occupations, following the same procedure as for the CSD-61k set, leaving only one conformation in the structure file. Missing hydrogen atoms were added with the program IQmol. Prior to the chemical shift calculations all the coordinates of the structures were DFT optimized using the same parameters as for the CSD-2k set. Chemical shieldings predicted for each structure are given in <https://doi.org/10.1038/s41467-018-06972-x>.

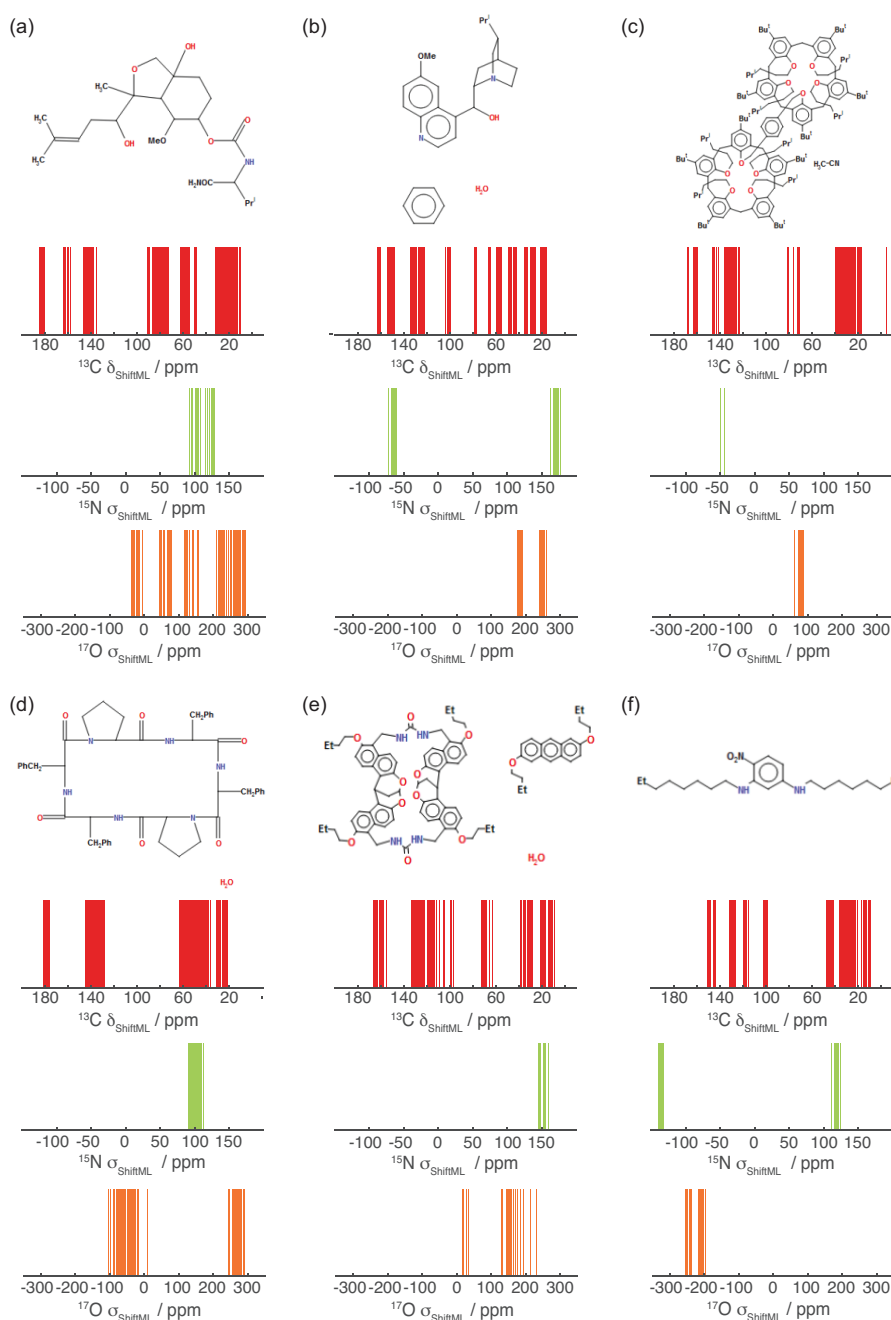


Figure 4.13. Chemical formula and corresponding ^{13}C , ^{15}N and ^{17}O NMR spectra predicted using ShiftML of the six large molecular crystals with CSD Refcodes. (a) CAJYUH,²⁹⁸ $N_{\text{atoms}} = 828$, (b) RUKTOI,²⁹⁹ $N_{\text{atoms}} = 768$, (c) EMEMUE,³⁰⁰ $N_{\text{atoms}} = 860$, (d) GOKXOV,³⁰¹ $N_{\text{atoms}} = 945$, (e) HEJBWU,³⁰² $N_{\text{atoms}} = 816$, (f) RAYFEF,³⁰³ $N_{\text{atoms}} = 1,584$.

Chapter 5 Rethinking NMR Crystallography

This chapter has been adapted with permission from:

Hofstetter, A., Balodis, M., Paruzzo, F.M., Widdifield, C.M., Stevanato, G., Pinon, A.C., Bygrave, P.J., Day, G.M. and Emsley, L., "Rapid structure determination of molecular solids using chemical shifts directed by unambiguous prior constraints". *Journal of the American Chemical Society* **2019**, 141, 42, 16624-16634. (**post-print**)

and

Engel E.A., Anelli, A., Hofstetter, A., Paruzzo, F., Emsley, L. and Ceriotti, M., "A Bayesian approach to NMR crystal structure determination", *Physical Chemistry Chemical Physics* **2019**, 21, 23385-23400. (**pre-print**)

We have already introduced methods to address some of the drawbacks of the isotropic chemical shift based NMRX protocol in the previous chapters. In this chapter, we discuss drawbacks intrinsic to the CSP based protocol, and we propose a modification of the NMRX protocol in order to overcome them.

The first limitation of the current NMRX protocol that we address here is related to the CSP step. Because CSP can require considerable computational resources, which increases rapidly with the structural degrees of freedom, the CSP based NMRX method for *de novo* structure determination is currently limited to systems with up to about 10 degrees of torsional freedom within the molecule,³¹⁴ and going beyond this requires some prior knowledge or intuition.^{122,150} Indeed, in order to circumvent these limitations CSP methods often make assumptions (for example based on space groups or predicted conformational energies) to limit the search space of possible structures. However, this leads to failure of the NMRX method to determine the correct crystal structure, when the correct structure is excluded from the search space.

Our protocol so far exploits geometric constraints from solid-state NMR only in the final step, in order to select the correct crystal structure from an ensemble of predicted structures. Introducing experimental constraints earlier in the CSP process would be an obvious way to guide and accelerate structure determination. The bottleneck for CSP of flexible molecules usually relates to the size of the molecular conformational space, so guidance to constrain the size of the search space would be most valuable if it relates to single molecule conformations. However, it is not immediately clear how experimental measurements on the crystalline samples would be relevant to restrict the single molecule conformational space. Note that an example has been given in which structural information was included at earlier stages of the CSP by biasing the search using semi-empirical molecule specific pseudo-forces derived from chemical shifts.¹³⁶

Here, we introduce an NMRX method in which we use unambiguous constraints from solid-state NMR to restrict the CSP search space to relevant regions of conformational space. The approach directs the determination procedure from the first steps towards the correct crystal structure, without the need for assumptions. We parametrize the approach on the crystal structures of cocaine, flutamide, and flufenamic acid and demonstrate a significant acceleration in computational times for these compounds. Most significantly, using chemical shifts calculated with both DFT and machine learning (see **Chapter 4**), we correctly determine the crystal structure of powdered ampicillin, for which the traditional NMRX approach would have failed.

The second limitation we address here is the estimation of the confidence in the identification of the experimental crystal structure using the NMRX approach. Indeed, while usually sufficiently accurate, GIPAW (or machine learned) shifts are not exact and the underlying atomic structures of candidates is subject to the accuracy of the level of theory at which they are described, leading to uncertainties in predicted NMR shifts. In this chapter, we investigate how the chemical shift prediction uncertainty within a set of candidate structures can be translated into a quantitative probability that one of the candidate structures corresponds to the experimental structure.

Traditionally, a structure is considered to be consistent with experiment if the RMSD of its shifts from the experimentally measured values falls within the currently expected chemical shift accuracy.^{116,294,315} However, this approach is again limited. It neither allows determination of the experimental structure when multiple candidates exhibit similar RMSDs within the “confidence interval”, nor does it provide a means of quantifying how likely different candidates are to match the experimental structure in any but the most clear-cut cases.

Here, we propose a Bayesian framework to quantitatively evaluate the probability that a structure among a given set corresponds to the experiment, on a continuous scale from 0 to 100% confidence. Using this approach, we find that the structures of flufenamic acid, cocaine, and AZD8329 can be identified with very high confidence (between 91% and 100%). In contrast, we show that the determination of the structure of flutamide is substantially less certain (82% confidence) and confirms the low confidence (13%) in the capability to determine the structure of theophylline.¹²⁰ We further introduce a method to visualise the Bayesian probabilities of the candidate structures in combination with a low-dimensional representation of their similarity, computed according to their chemical shifts or their geometry. Finally, we find that for the compounds considered here the errors in the calculated ^{13}C shifts are substantially larger than literature estimates of the uncertainty in ^{13}C shifts, and that with self-consistently determined uncertainties the inclusion of ^{13}C shifts (in addition to ^1H shifts) leads to consistently more reliable structure determinations.

5.1 Materials and Methods

Experiments of **Section 5.2** were performed on a Bruker 500 wide-bore Avance III and a Bruker 900 US² wide-bore Avance Neo NMR spectrometers, equipped with H/X/Y 3.2 mm and H/C/N/D 1.3 mm CPMAS probes. All experiments were carried out at room temperature.

The 2D ^1H - ^{13}C dipolar heteronuclear correlation (HETCOR) experiments were performed using a 12.5 kHz MAS frequency for ampicillin, flutamide and cocaine and at 24.0 kHz MAS frequency for flufenamic acid. In all experiments, SPINAL-64 was used for heteronuclear decoupling during direct acquisition and eDUMBO-1₂₂ for homonuclear dipolar decoupling in the indirect dimension. 256 t_1 increments with 16 and 128 transients were acquired for ampicillin, 256 t_1 increments with 64 transients for flutamide, 64 t_1 increments with 4 transients for flufenamic acid and 256 t_1 increments with 16 transients for cocaine.

For the assignment of ampicillin, additional NMR experiments were required. These experiments were performed on either a standard-bore Bruker 700 Avance III or a wide-bore Bruker 500 Avance III, equipped with a 3.2 mm HCN 4.0 mm HX CPMAS probes respectively.

The 2D ^{13}C - ^{13}C refocused incredible natural abundance double quantum transfer experiment (INADEQUATE) was performed at 13.0 kHz MAS. The polarization was transferred from ^1H to ^{13}C prior to the indirect evolution period, using a CP 2.5 ms long. SPINAL-64 (100 kHz nutation frequency) was used for heteronuclear decoupling during both evolution dimensions. For this experiment, we collected 128 t_1 increments with 1760 transients each. τ delays during the indirect dimension evolution were set to 3.84 ms and the length of the z-filter was 1.0 ms.

The ^1H - ^{15}N CP-HETCOR NMR experiment was carried out at 15.0 kHz MAS. SPINAL-64 was used during direct acquisition for heteronuclear decoupling (83 kHz nutation frequency), and eDUMBO-1₂₂ was used for

homonuclear decoupling in the indirect dimension. The ^1H - ^{13}C CP prior to the indirect evolution period was 300 μs long. For this experiment we collected 64 t_1 increments with 1440 transients each.

The ^{15}N magic-angle-turning (MAT) experiment was performed at 1.90 kHz MAS. SPINAL-64 was used for heteronuclear decoupling during both t_1 and t_2 (100 kHz nutation frequency). The ^1H - ^{13}C CP prior to the indirect evolution period was 5.5 ms long. For this experiment we collected 125 t_1 increments with 1024 transients each.

The assignment of ^{13}C and ^1H chemical shifts for flutamide, flufenamic acid and cocaine was taken from the work done by Baiaş and co-workers.¹²⁰

The assignment of the ^{13}C spectra of ampicillin has been done by Clayden *et al.*³¹⁶ and then revised by Antzutkin *et al.*,³¹⁷ but as mentioned in these works, the assignment is ambiguous, and so we revised it. To assign the ^{13}C NMR spectra at natural abundance a ^{13}C - ^{13}C INADEQUATE experiment was carried out. To assign the ^1H directly attached to ^{13}C , the ^1H - ^{13}C HETCOR spectra were used. To assign the ^1H directly attached to ^{15}N , a ^1H - ^{15}N HETCOR experiment was done, which also helped for the assignment of ^{15}N resonances. To distinguish the ^{15}N chemical shifts belonging to NH and NH_3 resonances, a ^{15}N CP-MAT experiment was done. The NH_3 resonance was assigned to the peak with negligible chemical shift anisotropy due to the fast exchange of the three attached ^1H atoms. The assignment was cross-validated by comparing the experimental chemical shifts to shifts calculated with the GIPAW DFT method using the XRD crystal structure, albeit with optimized hydrogen positions. Spectra are shown in **Appendix**.

The ^1H and ^{13}C chemical shifts were referenced indirectly to tetramethylsilane using the methyl signals of alanine at 1.3 ppm (^1H) and 20.5 ppm (^{13}C),³¹⁸ while ^{15}N chemical shifts were referenced using glycine at -347.54 ppm. ^1H chemical shifts were corrected for the scaling factor due to homonuclear decoupling, which was determined using ^1H 1D spectra acquired on a Bruker 900 US² wide-bore Avance Neo NMR spectrometer. Post-processing was done using Topspin 3.5 or 3.6.1. The cluster generation and analysis were performed with home-written Python and MATLAB codes and using the sketch-map package.^{292,319-321}

For **Section 5.3** no new NMR experiments were carried out. The chemical shift values, the assignment and candidate crystal structures were obtained from the original publications, where they are described in detail.^{120,121,322} 45 theophylline, 50 flufenamic acid, 21 flutamide, and 30 cocaine candidates within 10 kJ/mol of the respective lowest-energy structure were retained and are considered in this work. They can be found in the supplementary information of Ref. 120. 11 AZD8329 candidates within 30 kJ/mol of the most stable predicted crystal structure for a given conformation were retained and are considered in this work. They can be found in the supplementary information of Ref. 121.

The GIPAW DFT calculations for flutamide and theophylline were performed using CASTEP v5.0 with the PBE exchange-correlation functional³¹¹ without dispersion correction, an equivalent plane-wave energy cut-off of 40.42 Ry and a Monkhorst-Pack k-point grid³¹³ with a maximum spacing of $2\pi \times 0.05 \text{ \AA}^{-1}$. The GIPAW DFT calculations for flufenamic acid were performed using CASTEP v5.5 with the PBE exchange-correlation functional³¹¹ with a Tkatchenko-Scheffler semi-empirical dispersion correction,³²³ an equivalent plane-wave energy cut-off of 51.45 Ry and a Monkhorst-Pack k-point grid with a maximum spacing of $2\pi \times 0.05 \text{ \AA}^{-1}$. For all these structures, the calculations used on-the-fly generated GIPAW pseudopotentials.¹⁰⁴ The GIPAW DFT calculations for AZD8329 and cocaine were performed using Quantum Espresso v6.3 with the PBE exchange-correlation functional³¹¹ with a Grimme D2 semi-empirical dispersion correction³¹² and an equivalent plane-wave energy cut-off of 100 and 400 Ry for the wavefunction and density, respectively. The calculations used pseudopotentials from the PSLibrary.³²⁴

The machine-learning model used to predict the ^1H chemical shifts in both **Section 5.2** and **Section 5.3** is version 1.1 of ShiftML. More details on this version can be found in **Chapter 4.3**, **Section 5.5** and at <https://shiftml.epfl.ch>.

The accuracy of this model in predicting ^1H chemical shifts is estimated to be 0.44 ppm with respect to GIPAW DFT chemical shifts.

5.2 NMRX Directed by Unambiguous Prior Constraints

A Novel Conformer Selection

Figure 5.1a illustrates a case where the current NMRX method fails. Analogously to a successful NMRX case (schematically illustrated in **Figure 1.1**), a large ensemble of single molecule conformers is generated and sorted by conformational energy. However, here the molecular conformer present in the crystal structure is of very high relative energy in the gas phase, and thus does not pass the selection criterion by energy. An illustrative example of this case could be when intra-molecular hydrogen bonds stabilize the most stable conformations in the gas phase, while the crystal structure conformation is stabilized through inter-molecular hydrogen bonds or other interactions only present in the solid phase. Thus, following the normal selection steps based on the conformational energy, the correct conformer is not included in the crystal packing and lattice energy calculation steps, and as a consequence is not present in the trial crystal structures that are compared to the experimental data.

Taking this into account, one could extend the crystal structure determination procedure, and we consider two ways below. One option is to loosen the initial selection criteria, thus allowing more conformers to proceed to the following steps. This approach will increase the computational cost, often prohibitively. Even moderately flexible molecules can have hundreds of conformations, each requiring significant computing; for example, for ampicillin, one of the molecules studied here, the CSP procedure required, on average, just over 3 days of computing on 200 dedicated CPUs per conformer (yielding a total of 54 days for all the conformers). Thus, this approach either involves very long timescales or requires access to very large-scale computing. The second option is to use a different initial selection criterion including information from experiment.

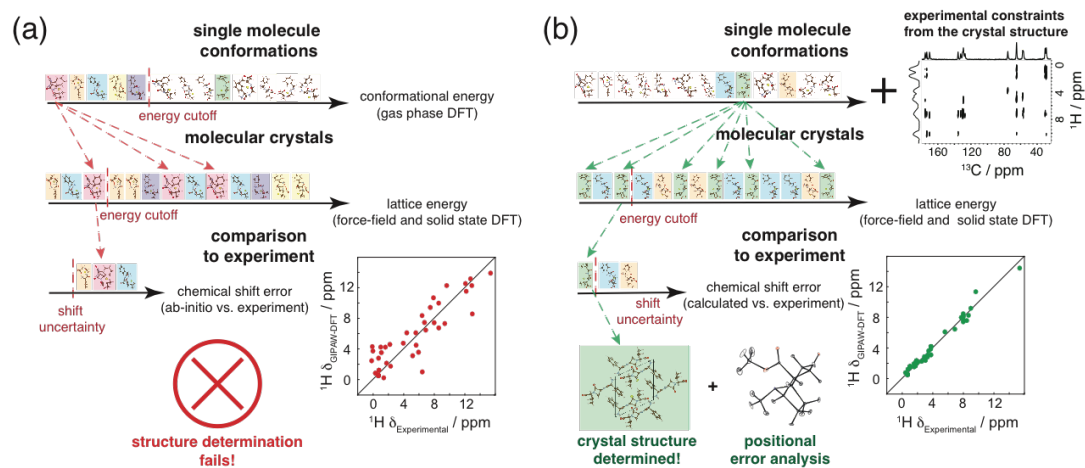


Figure 5.1. Schematic of current and proposed CSP-NMRX methods. (a) An example of a failed structure prediction using the current NMRX method. (b) An example of the proposed experimentally constrained NMRX method, which successfully overcomes the failure of the current NMRX method shown in panel (a). In each panel the structures in the first line depict single molecule gas phase conformations sorted by their conformational energy. After applying a given selection criterion a reduced conformer set is used to generate an ensemble of possible crystal structures (represented by the 2nd line in each panel). The coloured boxes are intended as a guide to the eye, as to which conformer results in which crystal structures. The 3rd line in each panel represents crystal structures picked from the 2nd line after a further selection criterion is applied. This final set of structures is then compared to the experimental chemical shifts, to determine the correct crystal structure. In each panel the scatterplot shows the experimental ^1H chemical shift plotted against the DFT-calculated ^1H chemical shift for the trial structure with the lowest error between DFT and experimental chemical shifts. An example of a successful structure prediction using the current NMRX method is shown in **Figure 1.1**.

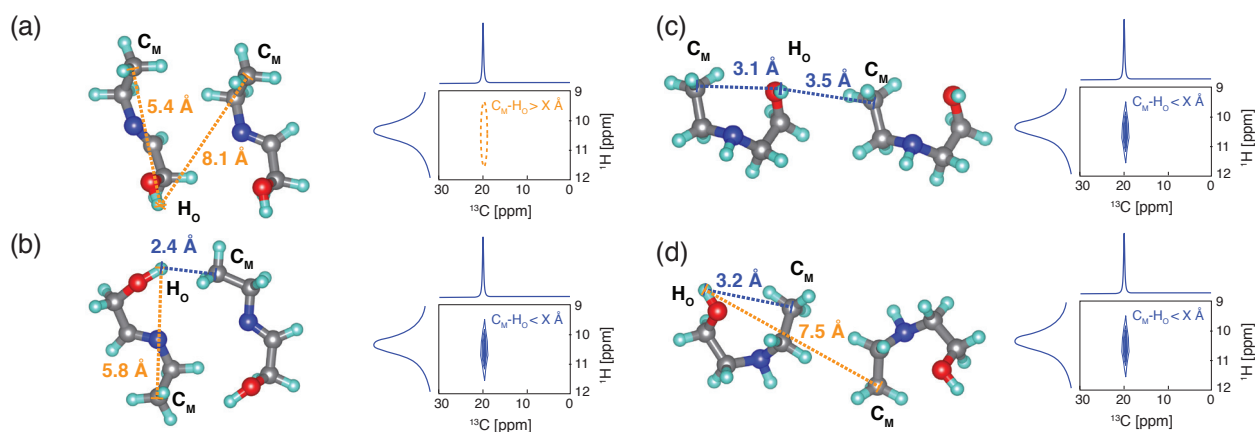


Figure 5.2. Schematic illustrations of ^1H - ^{13}C HETCOR spectra (right) for four different structural fragments (left) and the derived constraints. Structures (a) and (b) contain an “open” conformer. Structures (c) and (d) contain a “closed” conformer. Blue dotted lines are sufficiently short C-H distances between C_M and H_O to generate peaks in the spectra. Orange dotted lines are too long to generate peaks. After applying the constraints with a threshold distance of $X = 3.5$ Å, we see that the absence of a peak in fragment (a) is the only unambiguous constraint.

Figure 5.1b illustrates this second approach, which we introduce here. Contrary to the standard CSP methods, no assumptions based on calculated energy are made in the initial conformer selection process. Instead a sub-ensemble of conformers is selected using experimental constraints from solid-state NMR experiments on the powdered microcrystalline sample. This approach guides the conformational sub-ensemble selection towards the correct crystal conformer, and thus reduces the chances that the structure determination is limited by possibly erroneous assumptions.

However, experimentally we only have access to the full crystal structures and cannot probe the underlying “virtual” gas phase conformations independently. Thus, we need to measure experimentally accessible constraints that would be unambiguously fulfilled *both* in the crystal structure as well as in the gas phase conformations. Note that commonly used solid-state NMR constraints, such as the presence of (dipolar-coupling mediated) cross peaks in NMR correlation experiments^{60,103,124,128,325–332} due to internuclear proximity, do not contain unambiguous information about the gas phase conformations. This is because a cross peak could arise either from intra- or inter-molecular proximity.

Here we introduce a novel approach that extracts unambiguous conformational constraints on the single molecule conformations present in crystalline samples. The approach is schematically illustrated in **Figure 5.2**, where we differentiate between two conformers (“open” and “closed”) by analysing a ^1H - ^{13}C HETCOR spectrum.

The ^1H - ^{13}C HETCOR spectrum contains two different types of information. First, cross-peaks that are present indicate atoms that are close in space. Second, absent cross-peaks contain information about atoms that are more than a certain distance “X” apart, where “X” possibly depends on the CP contact time, experimental setup and the investigated system. **Figure 5.2** shows that only the information from the absent cross-peaks in the solid-state spectra can be directly transferred to constraints on the single molecule conformations. This is demonstrated with a thought experiment. If the heteroatoms C_M and H_O are close in space, the cross-peak at $\text{C}_\text{M}-\text{H}_\text{O}$ will be present in the HETCOR spectra. However, the cross-peak can result either from a short intra-molecular $\text{C}_\text{M}-\text{H}_\text{O}$ distance (i.e. the “closed” conformer) (**Figure 5.2c-d**) or from a short inter-molecular interatomic distance (which can be from the “closed” or the “open” conformer) (**Figure 5.2b-c**). Thus, the presence of a cross peak does not contain unambiguous information about the single molecule conformer, as the fragments in **Figure 5.2b-d** contain both possible conformations.

An absent cross-peak for $\text{C}_\text{M}-\text{H}_\text{O}$ however indicates that C_M and H_O are at least “X” angstroms apart, for both intra- and inter-molecular $\text{C}_\text{M}-\text{H}_\text{O}$ distances (**Figure 5.2a**). This can only happen for the “open” conformer. Thus,

information from the absent cross-peaks is unambiguous regarding the single molecule conformation and can be used as a constraint on trial structure generation.

Note that the fragment in **Figure 5.2b** also contains the “open” conformation, but would be expected to contain a cross-peak for C_M-H_O and thus will not result in a constraint on the distance between C_M and H_O . However, such cases only result in fewer constraints on the single molecule conformer but do not induce any incorrect constraints.

Note also that it is not *a priori* clear what the threshold distance “X” is. In general, we expect to reliably see all 1H - ^{13}C HETCOR cross-peaks at least up to 3.0 Å.³³³ To establish a reliable value for the threshold distance “X”, accessible in the 1H - ^{13}C HETCOR experiments used here, we investigate the correlation between interatomic 1H - ^{13}C distances and signal intensities of the cross-peaks in the HETCOR experiments recorded for cocaine, flutamide, and flufenamic acid.

For these three compounds the experiments were performed at different contact-times, magic-angle spinning-rates and on different spectrometers. **Figure 5.3** shows that for cocaine we have signal-to-noise ratios (SNR) of up to 80, while flufenamic acid has a maximum SNR of around 10. Additionally, for a 1H - ^{13}C HETCOR experiment, where the signal is transferred from the 1H to the ^{13}C , the SNR also depends on the number of protons involved in the transfer, as well as the number of protons overlapping at a given frequency.

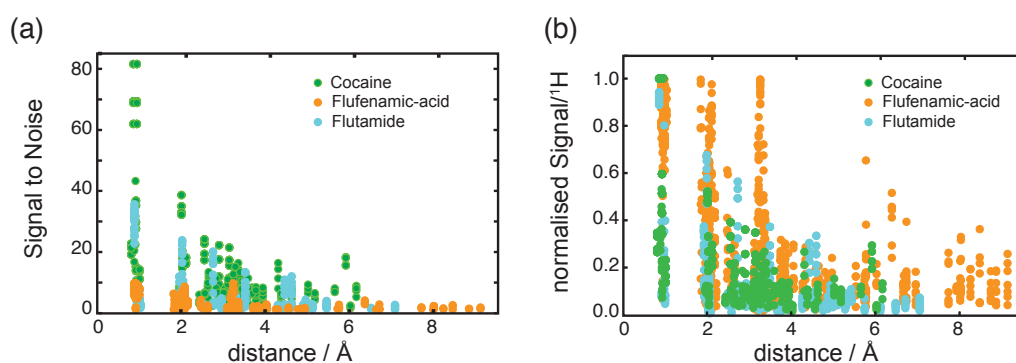


Figure 5.3. Signal intensity of 1H - ^{13}C HETCOR cross-peaks plotted against the corresponding interatomic distance for cocaine (green), flufenamic acid (orange) and flutamide (cyan). (a) The SNR is extracted directly for all 1H - ^{13}C HETCOR at different contact-times and different experimental setups. (b) The normalised SNR per 1H allows a direct comparison across different experimental setups and for cross-peaks corresponding to a different number of protons.

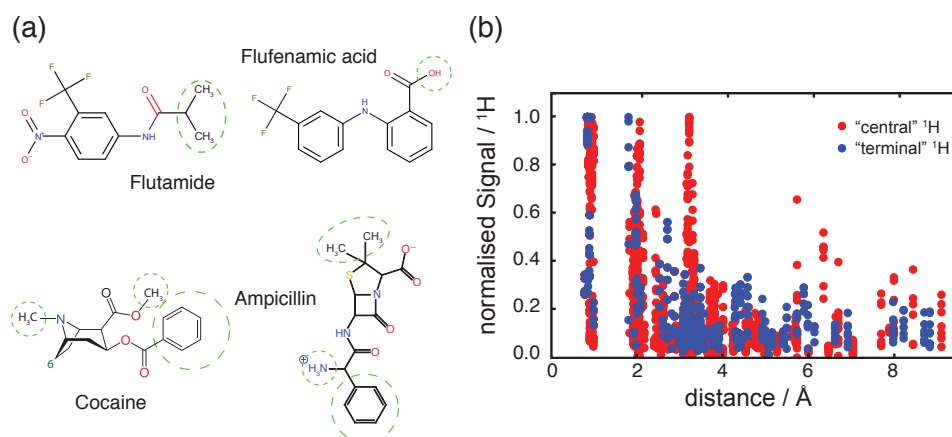


Figure 5.4. (a) Illustration of “terminal” protons, which contribute to eventual conformational constraints. (b) normalised SNR of 1H - ^{13}C HETCOR cross-peaks plotted against the corresponding interatomic distance for “central” protons (red) and “terminal” protons (blue), which are used to generate conformational constraints.

To make different spectra comparable, we first estimate that the number of active protons for a given cross-peak in a spectrum is proportional to the maximum signal intensity at a given frequency in ω_1 . The signal intensity of each cross-peak is then re-normalised by this number of protons. Then, we consider the difference in overall SNR between spectra by re-normalising each cross-peak with respect to the maximum proton-normalised SNR per spectra. This leads to a normalised SNR per ^1H , which is comparable across all experiments and which is shown in **Figure 5.3**.

Once we have selected a reliable threshold distance $X \text{ \AA}$ for a given SNR cut-off (this process is described below), the selected threshold distance in combination with each absent HETCOR cross peak is transformed into a constraint on the conformer space as, *“if the HETCOR cross peak between C_x and H_y is below the SNR cut-off it is classified as absent and so the distance between the atoms C_x and H_y must exceed $X \text{ \AA}$.”*

For each single molecule conformer all of the generated constraints were checked and the conformers were sorted according to the number of constraints violated. This procedure allows us to select conformers for the subsequent CSP procedure. If we are confident in the extracted constraints, it is sufficient to only select the sub-ensemble with the lowest number of constraint violations. However, if this sub-ensemble is very small or if additional computational resources are available, the selected sub-ensemble can easily be extended to include structures with a progressively higher number of constraint violations. Accepting conformations with a small number of constraint violations can allow for moderate changes in molecular geometry between the gas phase and crystal structure.

Parametrization Using Known Structures

We first establish the range of reliable threshold distances “ X ” for a given SNR cut off S_{norm} . For this we investigate the correlation between S_{norm} and the corresponding inter-atomic distances for the three trial compounds cocaine, flufenamic acid and flutamide.

^1H - ^{13}C HETCOR experiments were performed with ^1H - ^{13}C contact times of 0.5, 0.75, 1.0 and 1.5 ms; 0.1, 0.5, 1.5, 2.0, 3.0 and 3.5 ms; and 0.1, 0.3, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75 and 2.0 ms respectively for cocaine, flufenamic acid and flutamide. We re-normalised the spectra as described in **Section 5.5**. The resulting normalised SNR per ^1H is then comparable between compounds, see **Figure 5.3**.

However, **Figure 5.3** shows that although there is a correlation between the normalised SNR and the corresponding inter-atomic distance, there are significant fluctuations. This is expected since the HETCOR experiment is quite simple but is subject to spin diffusion relayed transfer, among others. We find that the effect of these fluctuations can be minimised by only considering correlations/distances from protons that are situated towards the extremities of the molecules. (We note here, that this currently results in a reduced number of extracted constraints. If the constraints could be extracted in a more quantitative manner, e.g. by accounting for changes in peak intensities due to ^1H - ^1H spin diffusion, then the selection criteria could be made stronger). However, these distances are the most information-rich in terms of the overall molecular conformations. We thus only consider cross-peaks resulting from the “terminal” protons marked with a green dotted-line in **Figure 5.4a**. This results in a clearer correlation between normalised SNR and the corresponding inter-atomic distances, as shown in **Figure 5.4b**.

To aid our interpretation of the selection procedure we apply a sketch-map^{292,319-321} analysis to the gas-phase conformer ensembles. The details of the sketch-map analysis including an interpretation of the underlying conformational changes for cocaine, flutamide and flufenamic acid are given in **Section 5.5**, in **Figure 5.16**, **Figure 5.17** and **Figure 5.18**.

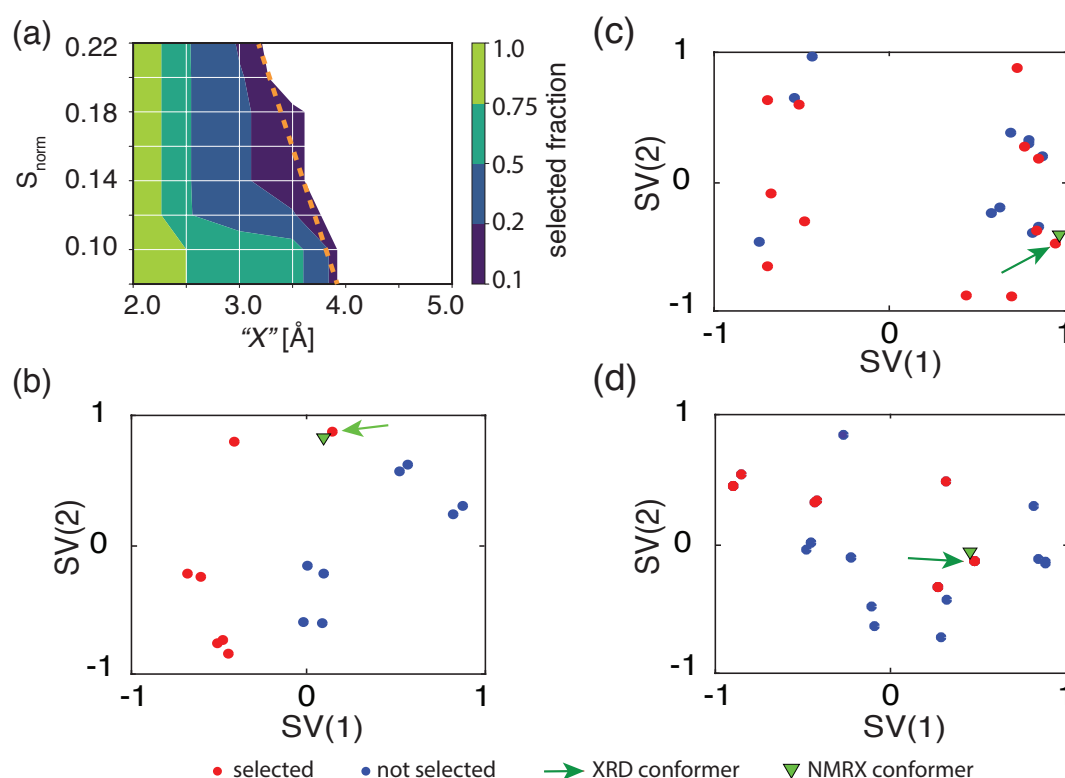


Figure 5.5. (a) Grid search results of the threshold distance “X” and S_{norm} cut-off values for flutamide, cocaine and flufenamic acid. The colour-map shows the percentage of selected structures from within the conformer ensemble. The white area indicates the region where the correct conformer is not selected. Optimal selection parameters should select the smallest conformer ensemble that still contains the correct structure. This corresponds to the dark blue regions within the different panels. The dashed orange line denotes the boundary at which the selection process starts to fail. (b-d) Conformer selection for flutamide (b), flufenamic acid (c) and cocaine (d). The panels show the sketch-map projections of the gas-phase ensembles. Red dots represent the structures that are selected when a threshold distance of 3.5 Å and a S_{norm} cut-off value of 0.14 were used. The green triangle shows the conformer found in the XRD-generated crystal structure. The green arrow points to the gas-phase conformer which results in the correct crystal structure after the CSP procedure.

Flutamide. The initial gas-phase ensemble of flutamide conformers generated in the first step of CSP contains 15 conformers,¹²⁰ of which 7 are in the *trans* and 8 are in the *cis* conformation with respect to the amide group. All the absent cross-peaks in a series of ^1H - ^{13}C HETCOR spectra were used to generate the conformational constraints shown in **Figure 5.6a**. **Figure 5.5b** shows the selected sub-ensemble of gas-phase conformers in the sketch map that fulfil the most constraints. The sub-ensembles with the lowest number of violations (2 of 10 total constraints) are selected for the subsequent CSP procedure.

Note, that these two constraints are violated for all conformers and do not correspond to significant changes in the conformation, as the involved atoms are not separated by more than 2 bonds. The reduced ensemble contains the gas-phase conformer that led to the correct crystal structure during the subsequent CSP procedure,¹²⁰ while being able to reduce the gas-phase conformer ensemble from 15 to 7 conformations. This significantly reduces the computational cost of the following CSP steps by approximately 54% (assuming that all conformers lead to similar numbers of putative crystal structures), while still including the correct gas-phase conformer that leads to the observed crystal structure. Additionally, the constraints from the absent cross-peaks uniformly selected the *trans* form in all 7 conformers.

Flufenamic acid. The gas-phase conformer ensemble for flufenamic acid contains 26 molecular conformations.¹²⁰ **Figure 5.5c** shows the selection of the sub-ensembles with the lowest number of violations (0 of 2 total constraints) using ^1H - ^{13}C HETCOR. The extracted constraints are shown in **Figure 5.6c**. Note that, for flufenamic acid, there are only two non-aromatic protons and that the cross-peaks from the aromatic protons are not distinguishable due to overlap in the ^1H dimension. However, the distance constraints extracted solely from the carboxyl proton (see

Figure 5.6c) were sufficient to reduce the number of relevant conformers by 46% (from 26 to 14 conformers), while still selecting the correct conformer, leading to the observed crystal structure.

Cocaine. The initial CSP conformer ensemble for cocaine contains 27 conformers.¹²⁰ **Figure 5.5d** shows the selection of the sub-ensembles with the lowest number of constraint violations (2 out of 10 total constraints) extracted from the ^1H - ^{13}C HETCOR NMR spectra (**Figure 5.6b**). As with flutamide, these two constraints were violated for all conformers and do not correspond to significant changes in the conformation, as the involved atoms are separated by only 3 bonds. **Figure 5.22** (Section 5.5) shows that the ^1H - ^{13}C HETCOR constraints were able to distinguish between the folding (closed and bent form) and stretching of the cocaine molecule with respect to the aromatic group as well as a flip in the methylamine group. Here, the relevant conformer ensemble is reduced by around 55% (from 27 to 12 conformations), while retaining the conformer that leads to the correct crystal structure.

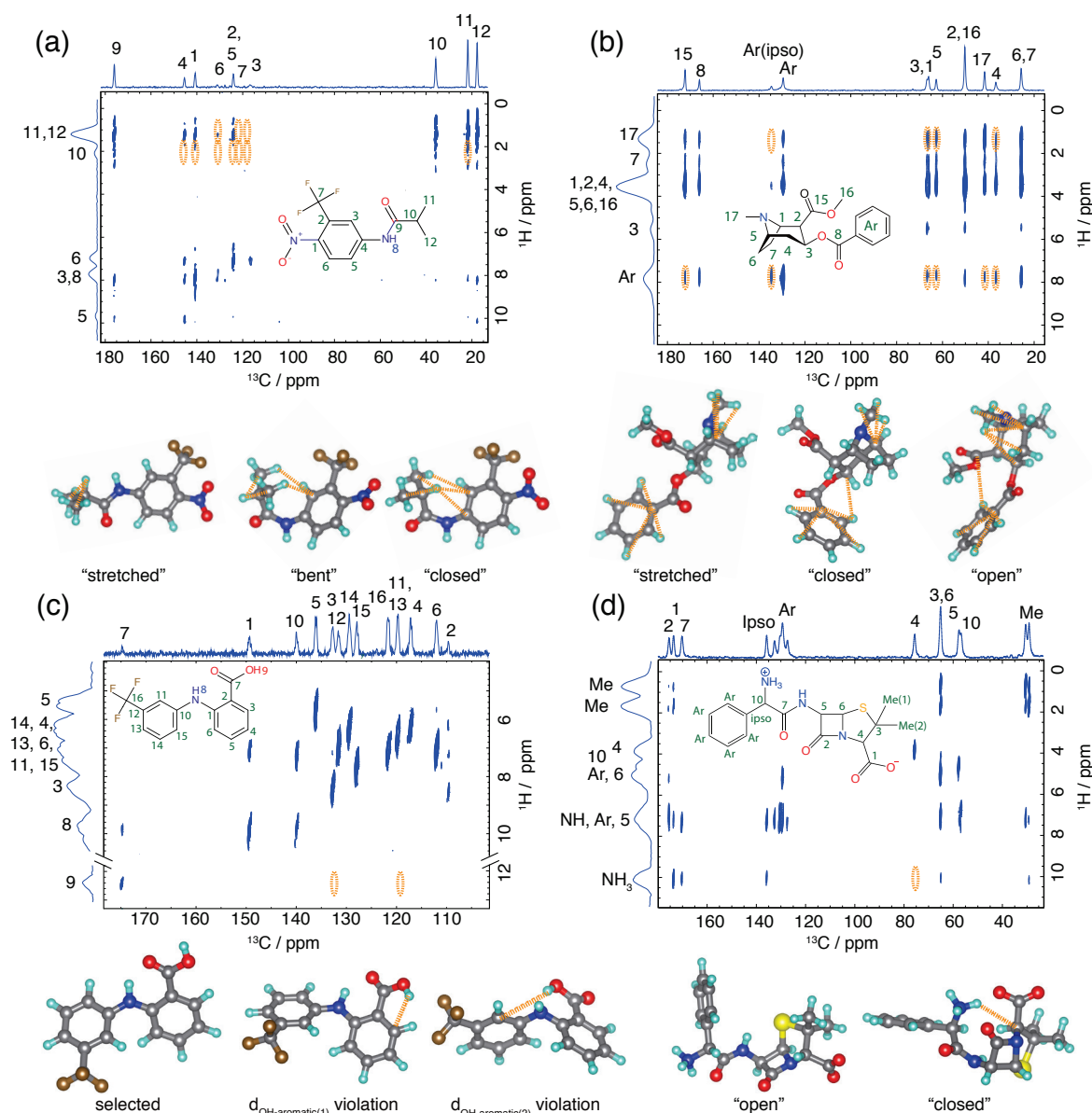


Figure 5.6. The top part in each panel shows the ^1H - ^{13}C HETCOR spectrum of: (a) flutamide with a 1.25 ms ^1H - ^{13}C cross-polarization contact time, (b) cocaine with a 1.0 ms contact time, (c) flufenamic acid with a 1.5 ms contact time and (d) ampicillin with a 1.5 ms contact time. ^{13}C peaks are assigned based on the literature³¹⁷ and ^1H peaks are assigned from HETCOR spectra and DFT chemical shift calculations (see Section 5.5). The cross-peaks from the terminal protons (**Figure 5.4a**) below a S_{norm} of 0.14 were used as constraints on the conformer ensembles, and are indicated as orange ellipsoids. The lower part of each panel shows the violated constraints extracted from all of the ^1H - ^{13}C HETCOR cross-peaks for different example conformers within the ensembles.

Crystal Structure Determination of Ampicillin

In contrast to the three cases above, the crystal structure determination of ampicillin would have failed using the usual NMRX protocol. In the first step, an ensemble of 16 locally stable gas-phase conformers was generated (for details, see **Section 5.5**) and the ensemble was then sorted according to the isolated molecule conformational energy. **Figure 5.7b** shows that all the conformers within 25 kJ·mol⁻¹ of the lowest energy structure are stabilized through an intra-molecular hydrogen bond between the amino nitrogen and oxygen atoms of the carboxyl group, whose strength is enhanced by the zwitterionic nature of the molecule. However, in the known single-crystal XRD structure, these intra-molecular hydrogen bonds between charged ends of the molecule are sacrificed to allow the formation of strong, charge-assisted inter-molecular hydrogen bonds, with the molecule adopting a more extended, open conformation.

Figure 5.7b also shows that the single molecule conformation closest to the crystal conformer is one of the highest energy gas phase conformers, nearly 100 kJ·mol⁻¹ higher in energy than the lowest energy single molecule conformer. In the normal CSP method a cut-off of around 20-25 kJ·mol⁻¹ would typically be applied to the conformational ensemble^{120,151} to limit the number of conformers that must be considered during the time-consuming crystal packing search. The correct conformer falls well outside this energy range and, thus, would be eliminated at this stage, preventing the successful generation of the observed crystal structure. To successfully determine the correct crystal structure, the subsequent CSP steps would have had to proceed without applying any energetic cutoff on the single-molecule conformers. This would be possible for the 16 conformers of ampicillin and use of large scale computing to perform the searches in parallel, but is problematic as a general method, as the conformational space of even moderately flexible molecules can often include hundreds of individual conformers.¹⁵¹

To address this problem, we apply experimental constraints extracted from ¹H-¹³C HETCOR NMR spectra at different contact times (i.e. 0.1, 0.3, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0 and 2.25 ms, detailed in **Section 5.5**). **Figure 5.6d** shows the assigned ¹H-¹³C HETCOR NMR spectra of ampicillin at a 1.5 ms contact time together with the labelled 2D structure. Following the protocol established for cocaine, flutamide and flufenamic acid, the SNR was then normalised over all experimental setups as described above. As we did for the other three molecules, we only consider cross-peaks resulting from *terminal*-protons, see **Figure 5.4a**. Using a S_{norm} of 0.14, and an X of 3.5 Å, that were parametrized on the reference compounds above, the extracted constraints are circled in orange and are shown on three example conformers below the spectra. **Figure 5.7a** shows the sub-ensembles with no violations (0 out of 1 total constraint). **Figure 5.6d** shows that only conformers without an intra-molecular hydrogen bond are selected. Also, from **Figure 5.7b** it is clear that the energetically high conformers are preferentially selected. Note, that in a classical CSP-NMRX approach these conformers would have not been selected. For the next step in the CSP procedure, only 7 out of the original 16 structures were considered. This reduces the computational cost by approximately 55%.

For each conformer remaining within this reduced gas-phase ensemble, we generated a crystal structure ensemble using a quasi-random sampling³³⁴ of lattice parameters, molecular positions and orientations within the commonly observed space groups. All 154,000 generated crystal structures were first optimized using an atomic-multipole based force field,³³⁵ followed by dispersion corrected DFT-D re-optimization of the lowest energy crystal structures, producing a final set of 75 candidate crystal structures. The full procedure is detailed in **Section 5.5**.

¹H chemical shift values were then calculated with GIPAW DFT and a machine learned method (ShiftML) for each candidate structure and compared to the experimental chemical shifts (details are given in **Section 5.5**). **Figure 5.8** shows the RMSE between DFT calculated and measured ¹H chemical shifts together with the calculated relative lattice energies for the candidate set. Based on currently-accepted metrics we expect a valid structure to have a ¹H RMSE of 0.33 ppm (± 0.16 ppm) or lower.¹¹⁶ This is indicated as the grey zone in **Figure 5.8**. Predicted structures with ¹H chemical shift difference within this zone are thus considered to be indistinguishable from experiment with a confidence of 1σ.

Figure 5.9 shows the RMSE between ShiftML calculated and measured ^1H chemical shifts together with the DFT calculated relative lattice energies for the candidate set. Using a benchmark set of 11 molecular crystal structures with around 150 experimental ^1H chemical shifts (as described in **Appendix**) we expect a correct structure to have a ^1H RMSE of 0.346 ppm (± 0.195 ppm) or lower. Note that the RMSE between experiment and the predicted chemical shifts follows broadly similar relative to the DFT-calculated shifts (**Figure 5.8**).

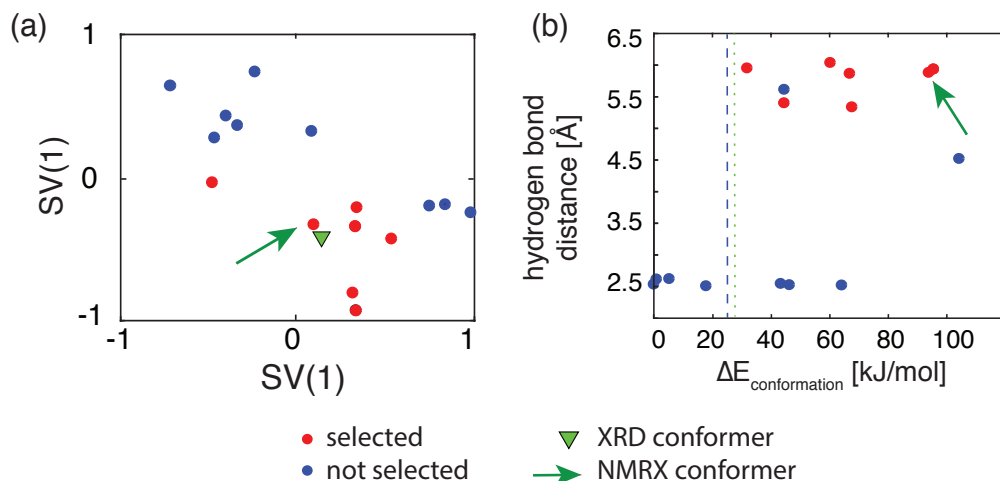


Figure 5.7. Conformer selection for ampicillin. (a) The panel shows the sketch-map projections of the gas-phase conformer ensemble. Red dots represent the structures which are selected using a threshold distance “X” of 3.5 Å and a S_{norm} of 0.14. The green triangle denotes the conformer found in the XRD-determined crystal structure. The green arrow points to the gas-phase conformer which results in the correct crystal structure after the CSP procedure. (b) Scatterplot showing the relative difference in the energy (ΔE) for the single molecule conformers of ampicillin against the shortest intra-molecular hydrogen-bond distance (N-O distance). The blue dashed line is the typical cut off energy (25 kJ/mol) used for conformer selection in CSP. The green dotted line is a guide to the eye to show at which ΔE the conformers with inter-molecular hydrogen bonds become accessible. The green arrow shows the conformer which results in the correct crystal structure.

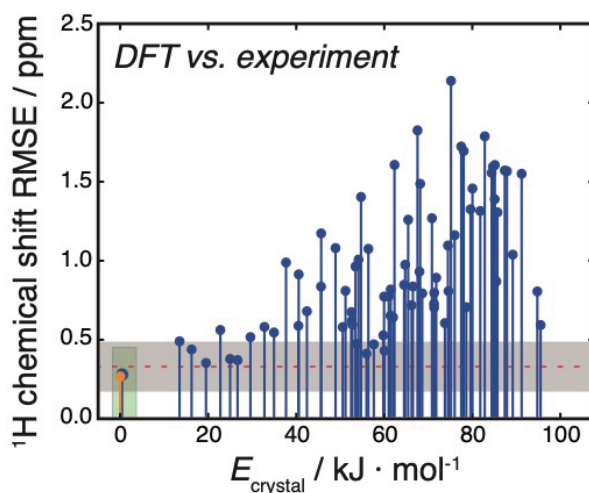


Figure 5.8. Comparison of crystal structure candidates. The structures are sorted according to their relative lattice energy, as specified on the horizontal axis. The vertical axis shows ^1H chemical shift RMSE between DFT calculated and experimental chemical shifts. The orange marker shows the ^1H chemical shift RMSE for the single-crystal XRD structure. The red line shows the mean of the current difference between experimental and DFT calculated ^1H chemical shifts with the distinguishability limits (at the 1σ level) indicated as grey shaded zone, as described in the main text.

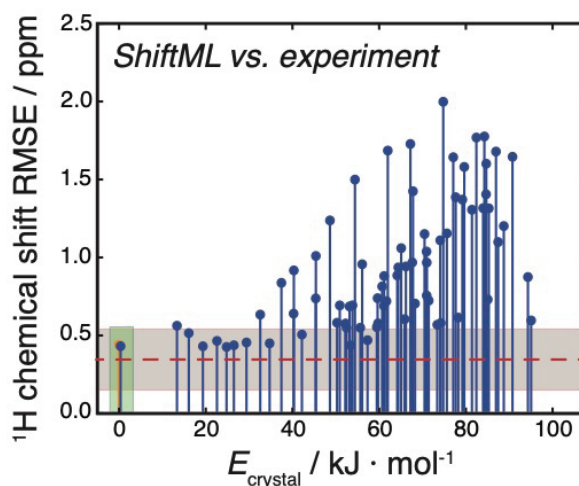


Figure 5.9. Comparison of crystal structure candidates. The structures are sorted according to their relative lattice energy, horizontal axis. The vertical axis shows ^1H chemical shift RMSE between ShiftML calculated and experimental chemical shifts. The orange marker shows the ^1H chemical shift RMSE for the single-crystal XRD structure. The red line shows the mean of the current error (0.346 ppm) between experimental and ShiftML calculated ^1H chemical shifts with the limits at one standard deviation (0.195 ppm) indicated as grey shaded zone, as described below.

Based on the agreement between experimental and calculated ^1H chemical shifts, both for ShiftML and DFT, we find that the crystal structure lowest in lattice energy, with a large gap in energy to the next predicted structure, also best produces the experimental NMR chemical shifts from the powdered microcrystalline sample used in the present study (**Figure 5.8**, **Figure 5.9**). Thus, we identify this structure as the correct candidate structure. Using chemical shifts calculated either directly from DFT or using ShiftML, several higher energy putative crystal structures produce ^1H chemical shifts within the acceptable error bounds. However, none of these alternative structures falls within the usual energy range of observed polymorphism (typically up to 7–8 kJ/mol)³³⁶ above the best candidate structure. Thus, our final structure selection relies on both the chemical shifts and calculated lattice energies.

The structure determined here agrees very well with the known reference structure determined by single-crystal XRD,³³⁷ as illustrated in **Figure 5.10a**. The deviation in atomic positions in the NMR structure from the powder is 0.278 Å, measured as the RMSD of all heavy atoms (excluding protons) in a 20-molecule cluster taken from the two structures. The single-molecule heavy atom RMSD is 0.068 Å. The largest deviation in the lattice parameters is a contraction of 6.8% in the *b* lattice parameter, and a unit cell volume of the CSP-NMRX structure 7.4% smaller than the single-crystal structure (see **Appendix**). This difference in volume is not unexpected as the NMRX structure is a temperature-free structure resulting from lattice energy minimization, while the single-crystal structure was determined at room temperature. The slightly shorter lattice parameters in the NMRX structure are in line with the expected thermal expansion of an organic molecular crystal.

Finally, we proceed with a positional error analysis that leads to the fully determined structure shown in **Figure 5.10c**. The positional error analysis is performed using the DFT-calculated ^1H chemical shifts following the procedure outlined by Hofstetter and Emsley¹⁸ and is detailed in **Section 5.5** (using DFT-MD here). The average positional RMSE on the NMR powder structure is $\langle r_{av} \rangle = 0.176$ Å, which corresponds to an average equivalent displacement parameter $U_{eq} = 0.0103$ Å². This compares with $\langle r_{av} \rangle = 0.149$ Å and $U_{eq} = 0.0074$ Å² for the single-crystal XRD structure.³³⁷ Note that the positional RMSE on the single-crystal XRD structure only considers the heavy atoms, while the positional RMSE on the NMR powder structure also includes the ^1H atoms.

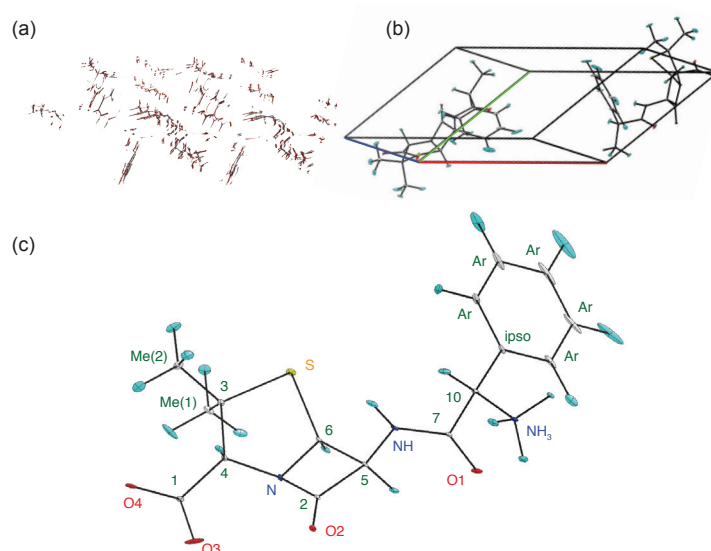


Figure 5.10. (a) Comparison between the structure of ampicillin as determined by the constrained powder ^1H CSP-NMRX and the single-crystal XRD determined structure.³³⁷ (b-c) ORTEP plot of the ampicillin (b) crystal and (c) single molecule structure drawn at the 90% probability level. The anisotropic ellipsoids correspond to a ^1H chemical shift RMSE of 0.49 ppm and to an average positional RMSE of $\langle r_{\text{av}} \rangle = 0.144 \text{ \AA}$.

5.3 A Probabilistic Crystal Structure Selection

Figure 5.11 shows examples of the analysis that is traditionally performed in chemical shift driven NMRX for AZD8329 and ampicillin. The RMSE between the experimental shifts and those predicted for multiple CSP candidates is computed using assigned ^1H chemical shifts, and compared to the typical uncertainty of DFT (or ML) predictions. The structure with the lowest RMSE is deemed to be the best candidate and identified as the experimental structure, provided the RMSE is consistent with the inherent uncertainty in the predicted shifts. In the case AZD8329, only one structure is consistent with experiment, making the structure determination conclusive. In the case of ampicillin, although the correct candidate has the lowest RMSE, several others are consistent with experiment within the inherent uncertainty in their predicted shifts. Based on this analysis, it is consequently impossible to assess how trustworthy identifying the best candidate as the experimental structure would be. In practice, energetic considerations strongly favour the correct candidate and facilitate determining the correct crystal structure.

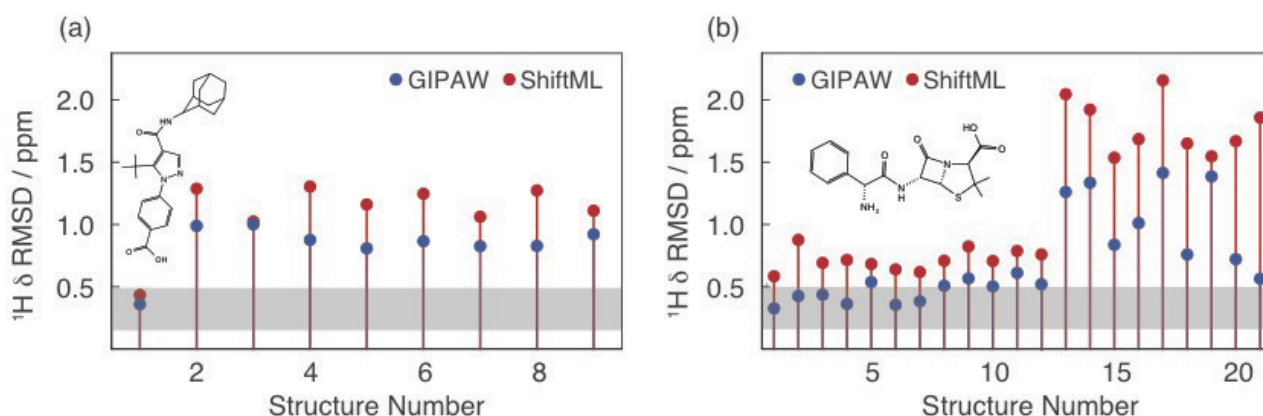


Figure 5.11. RMSEs of the GIPAW (blue) and ML (red) ^1H chemical shifts of the (a) and (b) AZD8329/ampicillin CSP candidates with respect to experiment. The grey area indicates the one sigma confidence interval for the GIPAW-DFT ^1H chemical shifts as determined by the typical error of GIPAW-DFT predictions with respect to the experimentally measured shifts for a set of benchmark compounds of known atomic structure.

Cases such as ampicillin, in which chemical shift based NMRX is complicated by the presence of two or more candidates in close agreement with experimental NMR shift data, are the primary reason for developing the Bayesian framework.

Theoretical Aspects

In the probabilistic approach to isotropic chemical shift based NMRX that we now develop each candidate structure constitutes a “model”, M , for which we determine the posterior probability, $p(M|\mathbf{y}^*)$, of corresponding to the experimental structure, given experimentally determined shifts, \mathbf{y}^* . The experimental shifts may originate from a single or multiple chemical species and may or may not have been partially or fully assigned to particular nuclei within the compound of interest. For each model the prior probability of matching the experimental structure is denoted by $p(M)$ and can in principle incorporate information regarding the thermodynamic stability of different candidates. Noting that stability estimates are often not accurate on the scale of differences between models, here we choose to set aside such considerations and assume uniform priors for all n_M models, $p(M) = 1/n_M$.

We denote the probability of observing shifts \mathbf{y} for a given model M as $p(\mathbf{y}|M)$ and the probability of observing a shift \mathbf{y} before we run the experiment as $p(\mathbf{y}) = \sum_M p(\mathbf{y}|M)p(M)$. Bayes theorem dictates that

$$p(M|\mathbf{y}^*) = \frac{p(\mathbf{y}^*|M)p(M)}{p(\mathbf{y}^*)} = \frac{p(\mathbf{y}^*|M)p(M)}{\sum_{M'} p(\mathbf{y}^*|M')p(M')}. \quad (5.1)$$

Clearly, in order to evaluate the posterior $p(M|\mathbf{y}^*)$, the conditional probability distribution $p(\mathbf{y}|M)$ must be defined. Given GIPAW or ML estimates of the shifts \mathbf{y}^M for each model M , the simplest model for the conditional distribution of the shift associated with a particular nucleus j takes the form of a normal distribution.

$$p_j(y|M) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{1}{2}\left(\frac{y - y_j^M}{\sigma_j}\right)^2\right) \quad (5.2)$$

The width σ_j represents an estimate of the typical error in the calculated shift with respect to experiment. Note that it is very important to have an accurate σ_j for this approach to work. We discuss different approaches to determining σ_j in **Section 5.5**.

The way **Equation 5.2** is translated into a posterior $p(M|\mathbf{y}^*)$, which quantifies the confidence in designating the model M as the experimental structure, depends on the level of assignment of the chemical shifts. In **Section 5.5** we analyse the scenarios with full chemical shift assignment, without chemical shift assignment and with partial chemical shift assignment. However, the results shown in this thesis focus only on the fully assigned case, which corresponds to the typical NMRX case. Examples and a discussion of chemical shift based NMRX with partial assignments or without assignments of shifts are given in the original publication (*Engel et al., Phys. Chem. Chem. Phys* 2019, just accepted).

Benchmark Systems

In order to demonstrate the Bayesian approach to NMRX, we use it to quantify the confidence in the structure determination of six molecular crystals (see **Figure 5.12**). We also demonstrate the use of two-dimensional visualisations of the similarity between candidate structures, both in terms of their structural features and in terms of their predicted chemical shifts, following the recipe given in **Section 5.5**.

Ampicillin, AZD8329, cocaine, theophylline, flufenamic acid, and flutamide (**Figure 5.12**) have all previously been studied using NMRX.^{120,121,322} In each case the experimental NMR shifts have been fully assigned to nuclei, the corresponding crystal structures are known, and DFT shifts for a pool of CSP candidates are available. Furthermore,

for all six compounds the CSP candidates include a representative of the experimental structure, which is referred to as the “correct candidate” in the following. The full assignments of the experimentally measured shifts to particular nuclei in the compounds used in the following are detailed in **Appendix**.

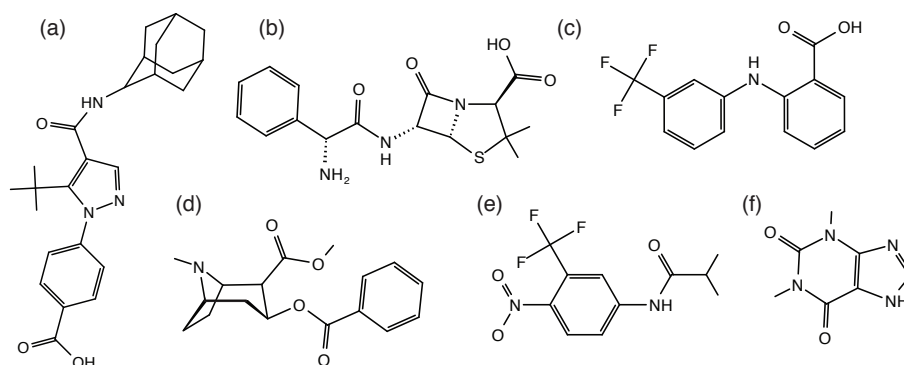


Figure 5.12. Chemical structures of (a) AZD8329, (b) ampicillin, (c) flufenamic acid, (d) cocaine, (e) flutamide and (f) theophylline.

Quantitative Structure Determination and Visualisation

From **Figure 5.13** we see that on the basis of the same ^1H shifts from GIPAW calculations, we estimate that the correct structure is identified with high confidence in 4 out of the 6 benchmark cases (75% ampicillin, 88% for flutamide, and 100% for AZD8329 and cocaine), and with some uncertainty in the case of flufenamic acid (60%). In the case of theophylline, the analysis confirms that the experimental structure cannot be distinguished.¹²⁰

In order to elucidate why the level of confidence in the structural determination varies among the benchmark problems, we generate a two dimensional visualisation based on principal component analysis (PCA) in which the CSP candidates for each compound are arranged such that pairwise distances reflect their dissimilarity, and which simultaneously shows the probability with which each candidate matches experiment (more details in **Section 5.5**). **Figure 5.14** shows the representations of the similarity of the CSP candidates for each of the six compounds.

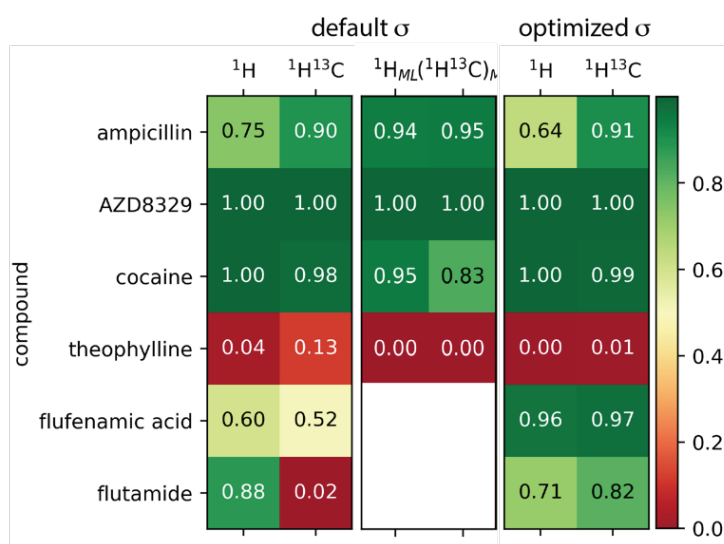


Figure 5.13. Overview of the results of NMR crystal structure determinations for the five benchmark compounds using ^1H and ^{13}C shifts calculated with ML or GIPAW, respectively. Each cell is colored and labeled according to the Bayesian probability of matching experiment assigned to the representative of the experimental structure among the CSP candidates -- this probability provides the key indicator of the reliability of the structure determination. The left and middle panels show the Bayesian probabilities of matching experiment calculated on the basis of the default global uncertainties of $\sigma_{\text{H}}^{\text{DFT}} = 0.33 \pm 0.16$ ppm and $\sigma_{\text{C}}^{\text{DFT}} = 1.9 \pm 0.4$ ppm. The right panel shows the Bayesian probabilities based on uncertainties estimated for each individual compound under consideration by maximizing $p(\mathbf{y}^*)$ with respect to $\{\sigma_j\}$. The estimated uncertainties are given as, $\sigma_{\text{H}}^{\text{DFT}} = 0.28 \pm 0.09$ ppm, and $\sigma_{\text{C}}^{\text{DFT}} = 2.7 \pm 0.9$ ppm.

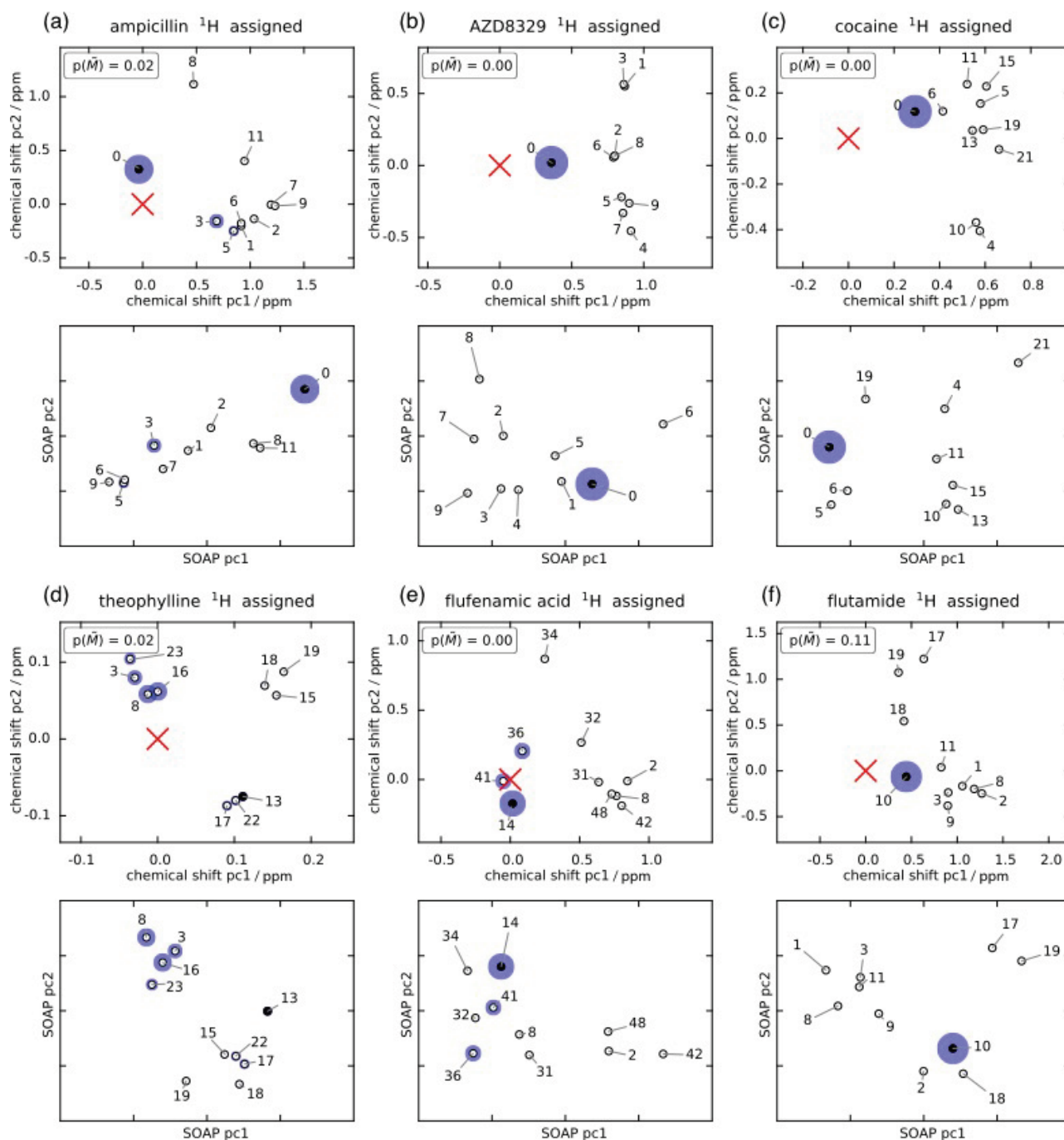


Figure 5.14. Evaluation of the top 10 (a) ampicillin, (b) AZD8329, (c) cocaine, (d) theophylline, (e) flufenamic acid and (f) flutamide CSP candidates. The correct candidates are shown as filled circles and the others as empty circles. For each candidate the probability of matching experiment $p(\vec{M}|\mathbf{y}^*)$ is indicated by the area of the blue disk. The candidates are labelled according to their rank in terms of configurational energy with zero indicating the energetically most favourable candidate. The respective upper panels show the similarity of the candidates to each other and to the (out-of-sample embedded) experimental data (shown as a red cross) in terms of their fully assigned ^1H GIPAW-DFT shifts. Their positions are determined by the first two chemical shift principal components (chemical shift pc1 and pc2), which can be understood, for the case of full assignment, as the two orthogonal linear combinations of their shifts for which the pairwise distances in the associated two-dimensional space most accurately reproduce the (dis-)similarity between candidates in terms of the Euclidean distance between the full vectors of ^1H GIPAW shifts. $p(\vec{M})$ denotes the probability that the virtual candidate, which represents structures potentially missing from the CSP candidate pool, matches experiment. The respective lower panels show the structural similarity of the candidates in terms of their SOAP features. While the relative distances of structures are a measure of their (dis-)similarity, the absolute value of the principal components (pc) from the (K)PCA constructions described in **Section 5.5** has no intuitive physical meaning and is therefore not shown.

For each compound we show the similarity in terms of ^1H chemical shifts (top panels) and in terms of structure (lower panels). The similarity in terms of chemical shifts reflects the resolving power of NMR. The similarity in terms of their structural features reflects how distinct the geometries of different candidates are. By embedding experiment, i.e. the experimentally measured shifts, in the representations of shift similarity one can also assess how closely (or not) the shifts of different candidates agree with experiment. First, by looking at the similarity as seen through the chemical shifts one can tell whether failure to identify conclusively the correct structure is due to lack of resolving power of NMR, or to the inaccuracy of the predicted shifts. For example, the case of theophylline (**Figure 5.14d**) shows that structures 8 and 16, which are identified as the most likely candidates, exhibit very distinct ^1H chemical shifts from structure 13, which is the correct candidate. Hence, even though there are only four ^1H shifts, this analysis suggests that more accurate predictions of the ^1H shifts would probably suffice to correctly determine the structure. In contrast, in the case of flufenamic acid (**Figure 5.14e**) the three structures with non-zero probability are all similarly close to experiment as they are to each other. In this case, it seems that shifts from additional chemical species, or a dramatic increase in the accuracy of shift predictions, would be needed to resolve the ambiguity.

Whenever two or more structures are close together in the shift-based representation, it would be hard to distinguish them by means of an NMR experiment. For instance, this is the case for structures 13, 17 and 22 of theophylline, as can be seen in **Figure 5.14d**. Meanwhile, the geometry-based representation, which is also shown in **Figure 5.14d** clearly shows that structure 13 is actually distinct. This geometric difference is not reflected in the value of the shifts, which is at least in part due to the small number of hydrogen atoms in a theophylline molecule. For comparison, the similarity of the structures 3, 8, 16, 23 in terms of chemical shifts clearly reflects an underlying geometric similarity.

NMRX Using ML Predictions of Chemical Shifts

Above we have made use of extensive preexisting GIPAW NMR calculations. In practice GIPAW shift predictions come at substantial cost, if the size and complexity of the system of interest permit them in the first place. Fortunately, ML shift predictions prove sufficiently reliable to determine structures. This is demonstrated by reconstructing the Bayesian models on ML shifts for all systems except flufenamic acid and flutamide. The latter two contain fluorine, leaving them outside the scope of the current ML model. The results are shown in **Figure 5.13** and demonstrate that ML-based NMRX almost matches the resolving power achieved with GIPAW predictions of NMR shifts.

Adding ^{13}C Chemical Shifts

Irrespective of whether NMR chemical shifts are predicted using GIPAW-DFT calculations or ML methods, ^1H shifts do not always suffice to pin down the experimental structure. Many studies have included other additional information, such as elements of geometry from dipolar couplings, and the chemical shifts of other nuclei such as ^{13}C or ^{15}N , either singly or jointly. The cases of flufenamic acid and theophylline highlight the limits of ^1H NMRX for compounds with few distinct hydrogen atoms, with a low, 60 % confidence in the structure determination in the former case, and the determination of the experimental structure being simply impossible in the latter. This makes it tempting to turn to ^{13}C chemical shift data in search for more information to exploit in distinguishing the experimental structure. However, in agreement with current wisdom,¹²⁰ **Figure 5.13** suggests that the inclusion of ^{13}C shifts reduces the confidence in the identification of the experimental structure.

The fact that the resolving power of NMRX appears to deteriorate upon inclusion of ^{13}C shifts warrants further discussion. Note that, in a Bayesian framework, adding more information should never degrade the prediction accuracy, unless the accuracy of such information is overestimated. The degradation of prediction accuracy therefore indicates that the value $\sigma_C^{DFT} = 1.9 \pm 0.4$ ppm based on benchmark data^{116,294,315} substantially underestimates the actual error for the compounds considered here.

Following the strategy of maximizing $p(\mathbf{y}^*)$ with respect to $\{\sigma_j\}$ proposed in **Section 5.5**, the typical error in ^{13}C shifts can be estimated to a substantially larger $\sigma_C^{DFT} = 2.7 \pm 0.9$ ppm. This is substantiated by the RMSD of the GIPAW shifts of the correct candidates with respect to the respective experimentally measured shifts of 2.6 ± 1.4 ppm. For comparison, the corresponding RMSD of the ^1H GIPAW shifts is 0.28 ± 0.09 ppm and thus entirely consistent with the global estimate of $\sigma_H^{DFT} = 0.33 \pm 0.16$.

Figure 5.13 demonstrates that, provided the compound-dependent, data-driven estimate of the errors in GIPAW ^{13}C chemical shifts derived here is used, the inclusion of ^{13}C shifts in the analysis indeed tends to improve rather than impair the resolving power of NMRX.

For instance, for flufenamic acid the structure determination is not limited by the accuracy of the predicted ^1H (and indeed ^{13}C) shifts, but rather by the accuracy of the estimates of the typical errors in those shifts. Accordingly, its structure can be determined with almost complete confidence (96%) provided accurate estimates of the typical errors in ^1H (and ^{13}C) shifts (see **Figure 5.13**).

5.4 Conclusions

In this chapter, we have developed a modification of our NMRX protocol. First, we have proposed a new method to perform the conformer selection in CSP. Indeed, the most severe limitations of CSP based NMR crystallography are encountered when a molecule has many possible conformers and the molecular conformation adopted in the crystal could be significantly higher in energy than the most stable gas-phase conformation. In such cases, the usual energetic thresholds applied to the conformational ensemble used to generate candidate crystal structures create a risk of missing the true conformer as well as the crystal packing. As an example, the traditional NMRX approach fails to determine the structure of ampicillin due to excluding the required conformer in the first step of CSP. At the same time, it is not affordable to remove any conformer selection and including all possible conformers during crystal structure generation, due to the prohibitively high computational costs.

To overcome this, we have introduced a modified NMRX method which includes unambiguous prior NMR constraints, in this case ^1H - ^{13}C correlations, at the conformer search stage within CSP. The key development is a novel approach that extracts unambiguous conformational constraints on the single molecule conformations present in crystalline samples. We parametrised the proposed method on the crystal structure determinations of three flexible molecules that were previously studied using NMRX: cocaine, flutamide and flufenamic acid. For all of these compounds we found that the method reproduces NMRX results and determines the correct crystal structure, while reducing the computational cost by between 46 and 55%. Note that these three molecules are relatively small and the savings in computational expense will be greater for larger molecules with more conformational degrees of freedom.

We also demonstrated the capability of the novel constrained NMRX method by successfully determining the crystal structure of powdered ampicillin. By constraining the structural search space, we were able to more than halve the time for the full crystal structure determination, while ensuring that the correct conformer is not excluded. We also emphasize that the large reduction in computational resources, demonstrated here, paves the way for the NMRX based determination of larger and more flexible molecules, which would previously have been out of the scope of the NMRX approach.

We note that the experimentally guided CSP method demonstrated here is not limited to pure NMRX applications but that the derived constraints can be used in any crystal structure determination methodology, which needs to limit the number of investigated conformations in order to reduce its computational cost.

We believe that the method is robust and we have chosen the experimental constraints, based on ^1H - ^{13}C NMR correlation experiments, for their relative simplicity and ease of access. However, we note that ^1H - ^{13}C correlation-based experiments are not the only ones that can give conformational constraints. Also, here the extraction of the

constraints was performed in a fairly basic and straightforward manner. We believe that if the constraints could be extracted in a more quantitative manner, e.g. by accounting for changes in peak intensities due to ^1H - ^1H spin diffusion, the selection criteria can be made stronger, further reducing the conformational space and improving the computational efficiency and reliability of the methodology.

The second modification we proposed is the introduction of a Bayesian framework to quantify the confidence in identifications of experimental structures using NMRX. Using this analysis framework, we demonstrate that definitive identifications are sometimes possible even if the corresponding shift RMSE does not fall within the traditional confidence interval. This relies on exploiting all available information, much of which the traditional RMSE measure of agreement with experiment is blind to.

We also notably use this approach to conclude that literature benchmarks for the accuracy in the prediction of ^{13}C chemical shifts underestimate the uncertainties. We find that ^{13}C errors for GIPAW-DFT predicted shifts for the compounds used here are 2.7 ± 0.9 ppm, as opposed to previous estimates of 1.9 ± 0.4 ppm. If we use our corrected error estimates, incorporating ^{13}C shifts into the analysis improves the reliability of structure determination. In the case of flufenamic acid the use of self-consistently computed uncertainties lifts the ambiguity on the structure determination.

We also introduce a visual representation of the crystal structure landscape based on a low-dimensional projection that reflects the similarity between the structure of the candidates, or directly on their NMR shifts. These visualisations help determine whether lack of structural diversity, insufficient resolving power of the experiment, or uncertainties in the computationally-determined shifts are involved in inconclusive structural determinations.

In combination, the Bayesian framework and the low-dimensional representations of candidate similarity provide an integrated way of identifying among a pool of candidate structures which most closely approximates the experimental one, performing sanity checks of the comprehensiveness of the pool, the associated predicted NMR shifts, and the initial identification, quantifying the confidence in the identification assuming the sanity checks have provided satisfactory results, analyzing what factors limit the confidence or, when definitive identification of the experimental structure is not possible, the resolving power of the crystal structure determination.

5.5 Supplementary Information and Discussion

5.5.1 Supplementary Information Section 5.2

Signal to Noise Analysis

The signal to noise ratio (SNR) extraction and analysis were done using the Signals extracted directly from TopSpin 4.0.5 in text file format together with a home-written python script.

The SNR was extracted as:

$$SNR = \frac{\maxval(S)}{2 * noise}, \quad (5.3)$$

where $\maxval(S)$ is the maximum intensity at a given ^1H and ^{13}C chemical shifts coordinate ± 0.2 ppm.- Note, that after a first extraction of $\maxval(S)$ the ^1H and ^{13}C coordinates were centered above $\maxval(S)$ and a refined $\maxval(S)$ was extracted.

The noise was extracted as the variance of the intensity for 100 areas (0.4×0.4 ppm) within the spectra. The initial 10 noise-areas were chosen manually, as to not contain any cross-peaks. The subsequent 90 noise-areas were chosen at random and were included in the noise intensity if the maximum signal intensity within the random area was less-than or-equal to two times the maximum signal intensity in an area previously selected. **Figure 5.3a**

shows the extracted SNR of all ^1H - ^{13}C HETCOR spectra for cocaine, flufenamic acid and flutamide against the corresponding inter-atomic distance.

First, we normalise each cross-peak by the number of protons. For this we estimate the number of protons for a given cross-peak in a spectrum by the maximum signal intensity at the given frequency, which is given from the maximum SNR at a given ^1H coordinate. In a next step, we take into account the difference in sensitivity between the spectra, due to the specific experimental setups, by normalising each cross-peak with respect to the maximal proton-normalised SNR per spectrum. This leads to a normalised SNR per ^1H , which is comparable across all experiments and is shown in **Figure 5.3b**.

Gas-Phase Conformer Generation

For cocaine, flutamide and flufenamic acid, the CSP conformers and crystal structures were generated as described by Baiaş and co-workers.¹²⁰

For ampicillin, we generated as complete and unbiased a set of gas phase conformers as possible using a low-mode conformational search (LMCS) method,^{338,339} as implemented in MacroModel.³⁴⁰ Energies were calculated during the conformer search using the OPLS3 force field.³⁴¹ The only prior knowledge used was that the molecule was present in the zwitterionic configuration throughout the conformer search. Minimum and maximum move distances of 3 and 6 Å were applied and 12,000 search steps were performed (2,000 per flexible dihedral angle). Duplicate molecular geometries were identified and removed using an all-atom RMS deviation of atomic positions, with a 0.05 Å tolerance.

All conformers were re-optimized in Gaussian09 using dispersion-corrected density functional theory (DFT-D) at the B3LYP/6-311G** level of theory with the D3BJ dispersion correction.³⁴² The N-H bond lengths at the amino nitrogen atom were constrained to 1.035 Å to keep the molecule in its zwitterionic form. Without this constraint, a hydrogen atom transfers from the amino to the carboxyl group during DFT reoptimization of many of the conformers. Importantly, the resulting non-zwitterionic conformers are not relevant to the known polymorphs of ampicillin.

In analysing the conformers resulting from the search, we found that the configuration around chiral centres could be reversed during the LMCS search. Therefore, all possible diastereomers of ampicillin were found to be present in the results. All conformers of a different diastereomer to that of interest were removed from the conformational ensemble before CSP was undertaken.

Sketch-Map Analysis

The cluster generation and analysis was performed with home-written Python and MATLAB codes and using the sketch-map package.^{292,319-321} The sketch-map parameters are given **Table 5.1**. They were chosen following the procedure described in Ceriotti *et al.*³²⁰ and the tutorial on <https://sketchmap.org>. The sketch-map analysis was not sensitive to small variations in the chosen parameters, as it is noted in the reference papers.^{292,320,321} As starting point for the sketch-map analysis we used all dihedral angles, not containing protons, over the full 2π range. This gives 55, 47, 31 and 35 dihedral angles for ampicillin, cocaine, flutamide and flufenamic acid, within a range of $-\pi$ to π .

Structure	$\Sigma = \sigma$	A	B	a	b
Cocaine	13	4	4	1	2
Ampicillin	6	2	2	1	1
Flutamide	6	3	3	1	1
Flufenamic Acid	6	2	2	1	1

Table 5.1. Sketch-map parameters

Ampicillin. The gas-phase CSP conformer ensemble of ampicillin contains 16 locally stable conformations (after DFT-D geometry optimization). The conformers are labeled according to increasing force-field energy. Conformer 14 is the most similar to the conformer in the crystal and resulted in the correct crystal structure after the remaining CSP procedure. **Figure 5.15** shows the sketch-map analysis of the ampicillin gas-phase ensemble.

Cocaine. The gas-phase CSP conformer ensemble of cocaine contains 27 locally stable conformations (after DFT-D geometry optimization). The conformers are labeled according to increasing force-field energy. Conformer 2 resulted in the correct crystal structure after the remaining CSP procedure.¹²⁰ **Figure 5.16** shows the sketch-map representation of the locally stable cocaine conformers. The main changes along the sketch-map principle components are rotations of the ester group (along SV(1)) and rotations within the methylamine group (along SV(2)).

Flutamide. The gas-phase CSP conformer ensemble of flutamide contains 15 locally stable conformations (after DFT-D geometry optimization). Of those, 7 are in the trans and 8 in the cis conformation with respect to the amide group. The conformers are labeled according to increasing force-field energy. Conformer 1 resulted in the correct crystal structure after the remaining CSP procedure.¹²⁰ **Figure 5.17** shows the sketch-map representation of the locally stable flutamide conformers. The sketch-map representation shows a relatively distinct clustering along the sketch-map axes, which correspond to the cis and trans conformations and rotations of the methyl groups. The SV(2) axis also partially corresponds to rotations of the aromatic ring.

Flufenamic acid. The initial CSP conformer ensemble of flufenamic acid contains 26 locally stable conformations (after DFT-D geometry optimization). The conformer 3 resulted in the correct crystal structure after the remaining CSP procedure.¹²⁰ **Figure 5.18** shows the sketch-map representation of the flutamide gas-phase conformer ensemble. The main changes along the sketch-map principle components correspond to rotations of the carboxyl group (along SV(1)) and rotations of the two aromatic groups (along SV(2)).

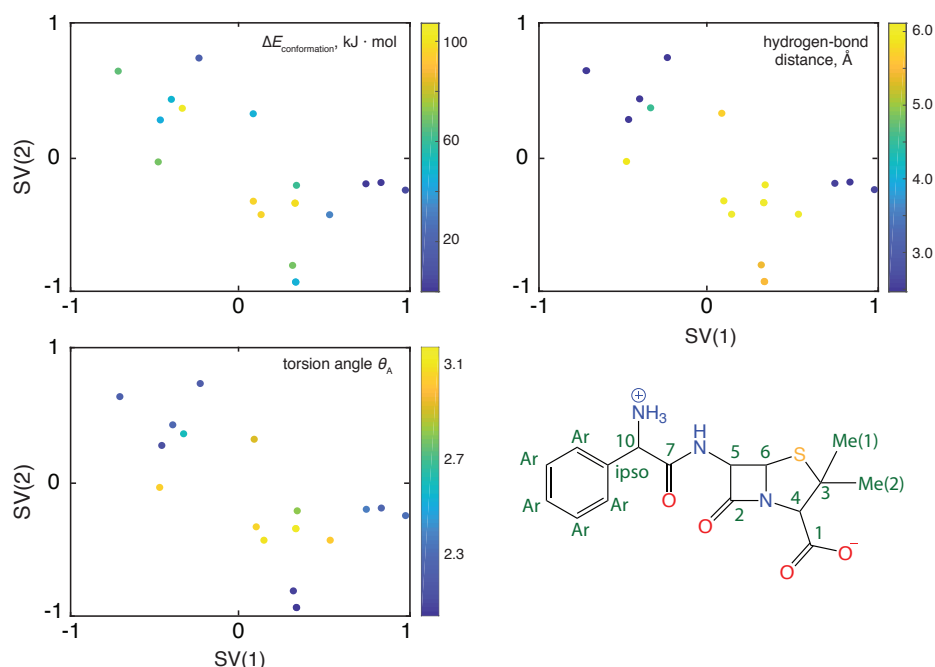


Figure 5.15. Sketch-map representation of the locally stable ampicillin conformers and the conformer from the single-crystal XRD structure. To show the extent of the sub-clustering the panels are coloured according to different molecular properties. Top left shows the difference in conformational energy ($\Delta E_{\text{conformation}}$). Top right shows the shortest intra-molecular hydrogen-bond distance between either NH_3 or NH and the carboxyl group. Bottom left shows the torsion angle θ_A , which is defined as the torsion angle between $\text{C}_{10}\text{-C}_7\text{-N(H)-(N)H}$. In general, the clustering seems to correspond to conformational changes along the $\text{C}_{\text{ipso}}\text{-C}_{10}\text{-C}_7\text{-N(H)-C}_5$ chain and to relative changes between the methyl and carboxyl groups. Bottom right, shows the 2D structure of ampicillin with the labelling scheme used

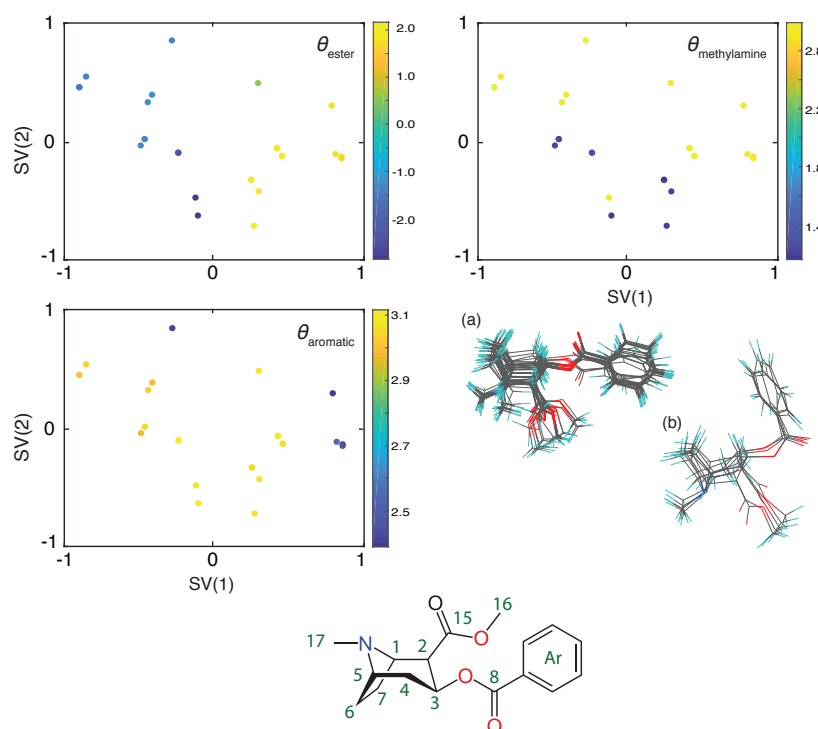


Figure 5.16. (Top) Sketch-map representation of the locally stable cocaine conformations. To show the extent of the sub-clustering, the panels are coloured according to different torsion angles reporting on different torsion angle values in the molecule. θ_{ester} is defined as the C1-C2-C15-O4 torsion angle and reports on rotations of the ester group. $\theta_{\text{methylamine}}$ is defined as the C2-C1-N-C17 torsion angle and reports on rotations of the methyl group attached to the nitrogen. θ_{aromatic} is defined as the C(ortho)-C(ipso)-C8-O2 torsion angle and specifies the orientation of the phenyl ring plane relative to the nearby carboxylate group plane. The lower right panel shows the overlapped conformation with the aromatic ring roughly in the plane defined by the carboxylate group (a) and roughly perpendicular to the carboxylate group plane (b). (Bottom) 2D structure of cocaine with the labelling scheme used.

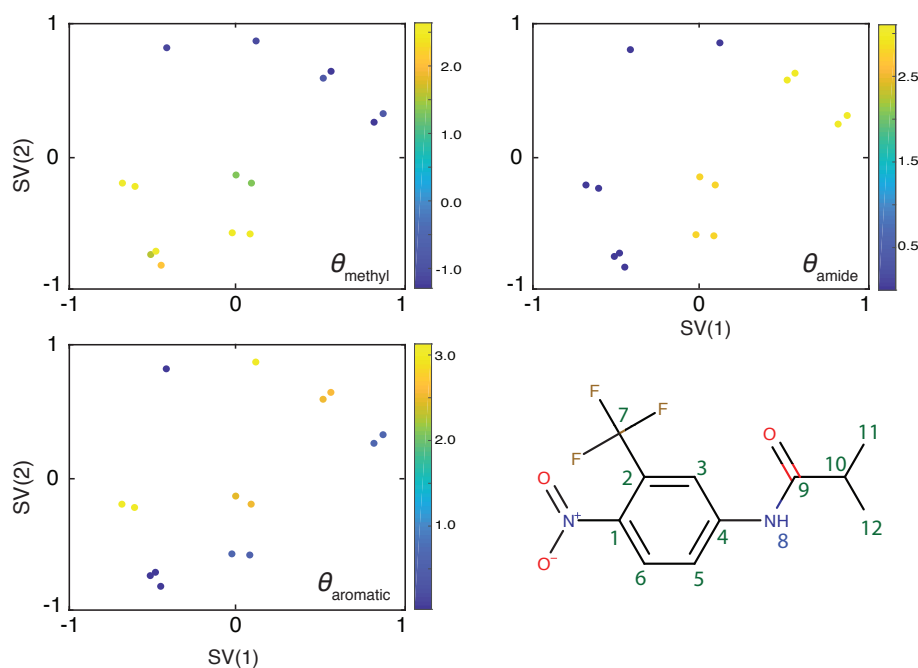


Figure 5.17. Sketch-map representation of the gas-phase flutamide conformer ensemble. To show the extent of the sub-clustering, the panels are coloured according to different torsion angles reporting on different torsion angle values in the molecule. θ_{methyl} is defined as the C11-C10-C9-N(H) torsion angle and reports on rotations of the methyl carbons. θ_{amide} is defined as the C4-N(H)-C9-O1 torsion angle and reports on the amide conformation. θ_{aromatic} is defined as the C3-C4-N(H)-C9 torsion angle and reports on rotations of the aromatic group. The lower right panel shows the 2D structure of flutamide with the labelling scheme used.

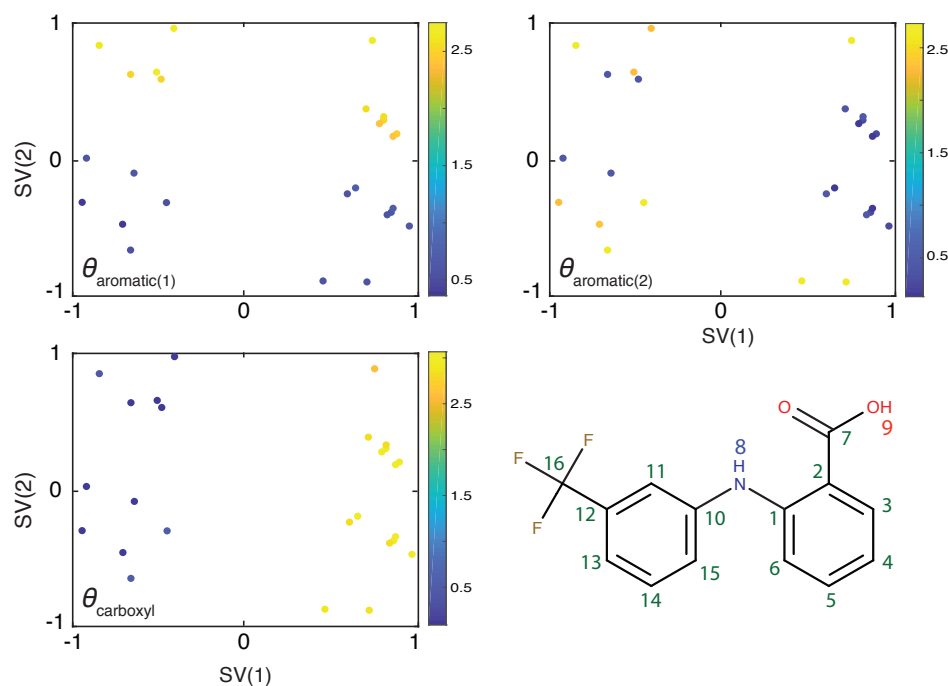


Figure 5.18. Sketch-map representation of the gas-phase flufenamic acid conformers. To show the extent of the sub-clustering, the panels are coloured according to different torsion angles reporting on different torsion angle values in the molecule. $\theta_{\text{aromatic}(1)}$ is defined as the C15-C10-N(H)-H(N) torsion angle and reports on rotations of aromatic ring with the attached trifluoromethyl. $\theta_{\text{aromatic}(2)}$ is defined as the C7-C2-N(H)-H(N) torsion angle and reports on rotations of aromatic ring with the attached carboxyl. θ_{carboxyl} is defined as the C1-C2-C7-O(H) torsion angle and reports on rotations of the carboxyl group. The lower right panel shows the 2D structure of flufenamic acid with the labelling scheme used.

Parametrization of the Constraints

It is not a priori clear as to what the threshold distance X should but in general we expect ^1H - ^{13}C HETCOR cross-peaks in solid-state NMR spectra to appear for all interatomic distances of up to 3.5 Å. Here, we investigate the use of threshold distances (X) from 2.0 to 5.0 Å in steps of 0.5 Å and for S_{norm} cut-off values from 0.08 to 0.22 in steps of 0.02 for the polymorphs of cocaine, flutamide, flufenamic acid. **Figure 5.19** shows the set of successful parameters for each molecule individually.

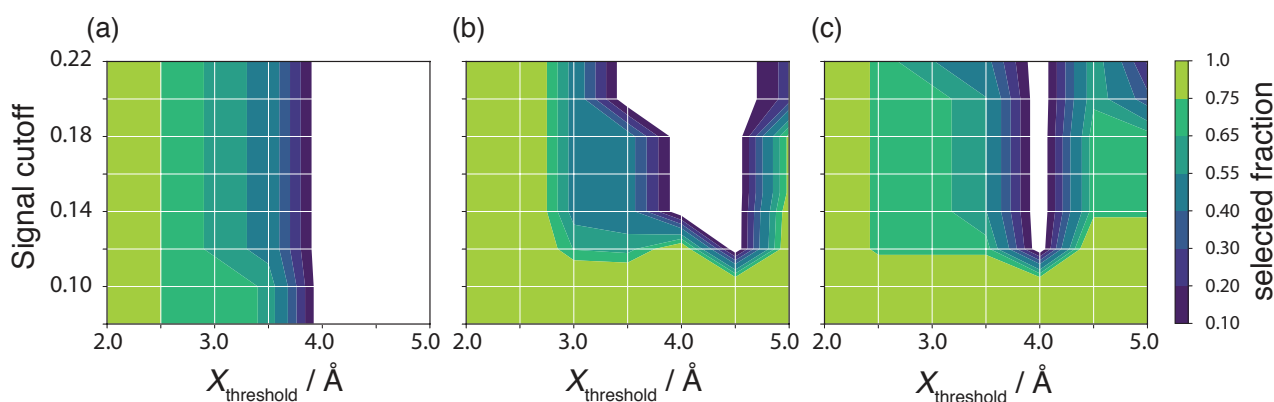


Figure 5.19. Grid search results of the threshold distance (X) and normalized SNR cut-off values (S_{norm}) for (a) flutamide, (b) cocaine and (c) flufenamic acid. The colour-map shows the percentage of selected structures from within the conformer ensemble. The white area indicates the region where the correct conformer was not selected. Optimal selection parameters should select the smallest conformer sub-ensemble, while still containing the correct structure. This corresponds to the dark blue regions within the different panels.

Conformer Selection

The ensemble selection was done with home-written Python codes. The HETCOR cross peaks below a S_{norm} value of 0.14 were interpreted as hydrogen-carbon distances greater than 3.5 Å (the “threshold” distance, X). For each conformation, the number of fulfilled constraints was counted and the conformers were sorted in decreasing order.

Flutamide. Conformer selection for flutamide was done based on constraints from multiple HETCOR NMR experiments, with variable contact times of 0.1, 0.3, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75 and 2.0 ms. The ^1H and ^{13}C cross-peaks from the two methyl groups could not be distinguished. Also, the ^1H cross peaks from H3 and H8 as well as the ^{13}C cross peaks from C5 and C2 are too close and thus indistinguishable. Therefore, if a cross-peak was observed it was attributed to all atoms in the given group. All the conformers, sorted by the number of experimental constraints satisfied are given at <https://doi.org/10.1021/jacs.9b03908>.

Cocaine. The ^1H - ^{13}C HETCOR NMR experiments were performed on cocaine at variable contact times of 0.5, 0.75, 1.0 and 1.5 ms. The ^1H and ^{13}C cross-peaks from the aromatic group could not be distinguished. Also, the ^{13}C cross-peaks from C6 and C7, the ^{13}C cross-peaks from C2 and C16, as well as the ^1H cross-peaks from H1, H2, H4, H5, and H6 were too close and hence indistinguishable. Therefore, if a cross-peak was observed it was attributed to all atoms in the given group. All the conformers sorted by number of experimental constraints satisfied are given at <https://doi.org/10.1021/jacs.9b03908>.

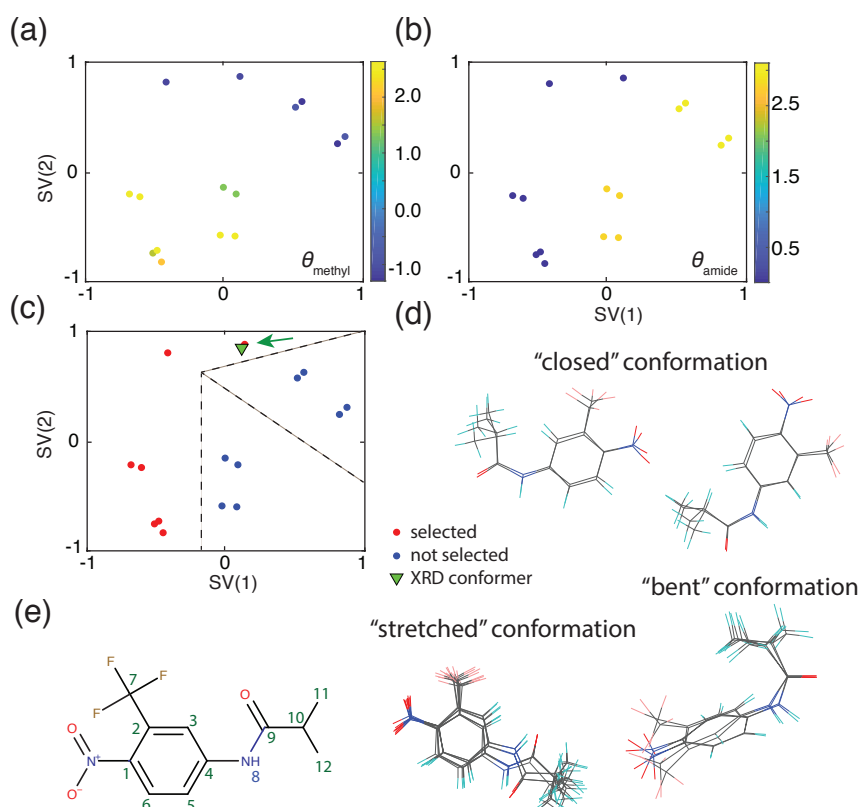


Figure 5.20. (a-b) Sketch-map representation of the gas-phase flutamide conformer ensemble. To show the extent of the sub-clustering, the panels are coloured according to different torsion angles reporting on different torsion angle values in the molecule. θ_{methyl} is defined as the C11-C10-C9-N(H) torsion angle and reports on rotations of the methyl groups. θ_{amide} is defined as the C4-N(H)-C9-O1 torsion angle and reports on the amide conformation. (c) Sketch-map projection of the gas-phase flutamide ensemble. Red dots represent the conformers with the lowest number of constraint violations, and are thus selected. The green triangle shows the conformer present in the XRD-determined crystal structure. The green arrow points to the gas-phase conformer, which resulted in the correct crystal structure after the CSP procedure. The black dashed lines indicate the regions where the different conformer sub-ensembles, shown in (d) are located. (d) Overlay of the structures within the different sketch-map clusters. The “stretched” conformations correspond to the trans conformers and are all selected. The “bent” and “closed” conformations correspond to the cis conformers and are not selected. (e) 2D structure of flutamide with the labelling scheme used.

Flufenamic Acid. Conformer selection for flufenamic acid was done based on constraints from multiple HETCOR NMR experiments, with variable contact times of 0.1, 0.5, 1.0, 1.5, 3.0 and 3.5 ms. The ^1H cross-peaks from H4, H13 and H14 as well as the ^1H cross-peaks from H6, H11 and H15 are too close and thus indistinguishable. Therefore, if a cross-peak was seen it was attributed to all atoms in the given group. All the conformers, sorted by the number of experimental constraints satisfied, are given at <https://doi.org/10.1021/jacs.9b03908>.

Ampicillin. The conformer selection for ampicillin was done using constraints from ^1H - ^{13}C HETCOR NMR experiments with variable contact times of 0.1, 0.3, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75 and 2.25 ms. The ^1H and ^{13}C cross-peaks from the two methyl groups could not be distinguished. Also, the ^1H cross-peaks from Ar2-6, H5 and NH, the ^1H cross-peaks from Ar1, H10 and H6, the ^{13}C cross-peaks from C3 and C6 as well as the ^{13}C cross-peaks from Ar1-5 were too close and hence indistinguishable. Therefore, if a cross-peak was observed it was attributed to all atoms in the given group. All the conformers, sorted by the number of experimental constraints satisfied, are given at <https://doi.org/10.1021/jacs.9b03908>.

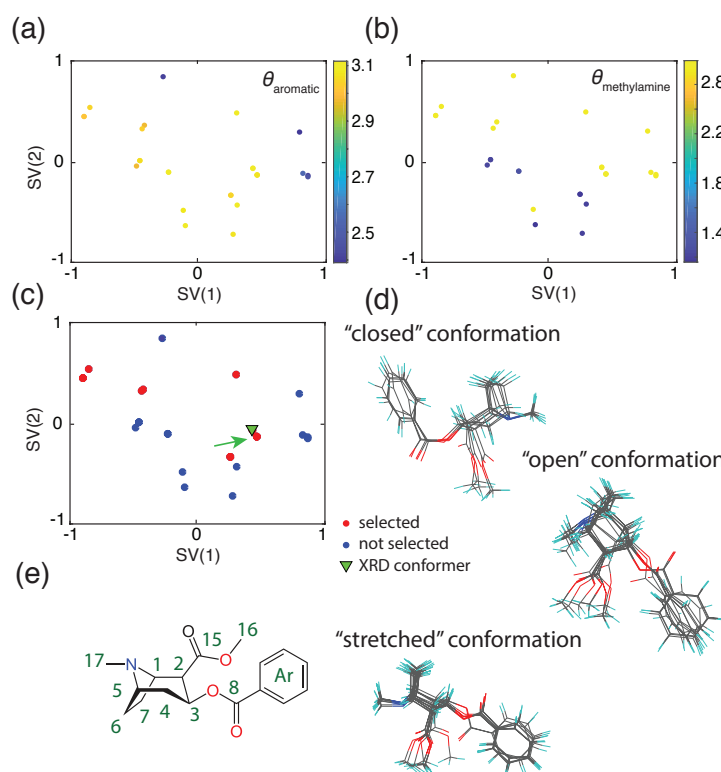


Figure 5.21. (a-b) Sketch-map representation of locally stable cocaine conformers. To show the extent of the sub-clustering, the panels are coloured according to different torsion angles reporting on different torsion angle values in the molecule. $\theta_{\text{methylamine}}$ is defined as the C2-C1-N-C17 torsion angle and reports on rotations of the methyl group attached to the nitrogen. θ_{aromatic} is defined as the C(ortho)-C(ipso)-C8-O2 torsion angle and reports on flips of the aromatic group. (c) Sketch-map projection of the gas-phase cocaine ensemble. Red dots represent the structures with the lowest number of constraint violations, and are thus selected. The green triangle shows the conformer present in the XRD-determined crystal structure. The green arrow points to the gas-phase conformer, which resulted in the correct crystal structure after the CSP procedure. (d) Overlay of the structures within the different sketch-map clusters. The “stretched” conformations correspond to the selected conformers. The “closed” conformations contain different θ_{aromatic} torsional angle and are not selected. The “open” conformation contains a different $\theta_{\text{methylamine}}$ torsional angle and are not selected. (e) 2D structure of cocaine with the labelling scheme used.

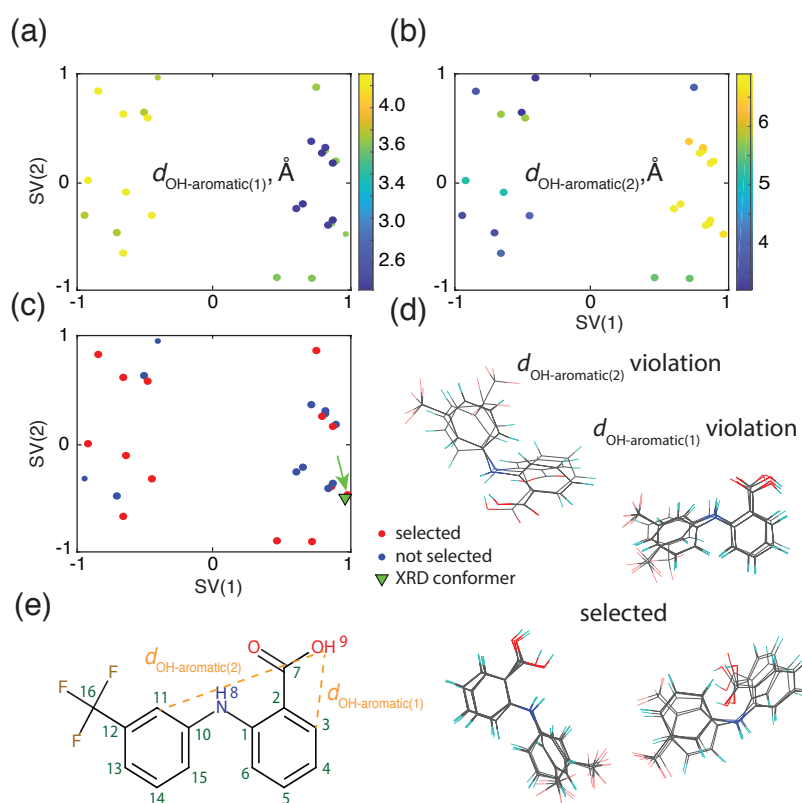


Figure 5.22. (a-b) Sketch-map representation of the gas-phase flufenamic acid conformer ensemble. To show the extent of the sub-clustering, the panels are coloured according to the distance [Å] between the OH group and the two aromatic rings. The distance is expressed as the distance between the carboxyl proton and C3/C11 (as shown in e). (c) Sketch-map projection of the gas-phase flufenamic acid conformer ensemble. Red dots represent the conformers with the lowest number of constraint violations, and are thus selected. The green triangle shows the conformer present in the XRD-determined crystal structure. The green arrow points to the gas-phase conformer, which resulted in the correct crystal structure after the CSP procedure. (d) Overlay of the structures within the different sketch-map clusters. (e) 2D structure of flufenamic acid with the used labelling scheme.

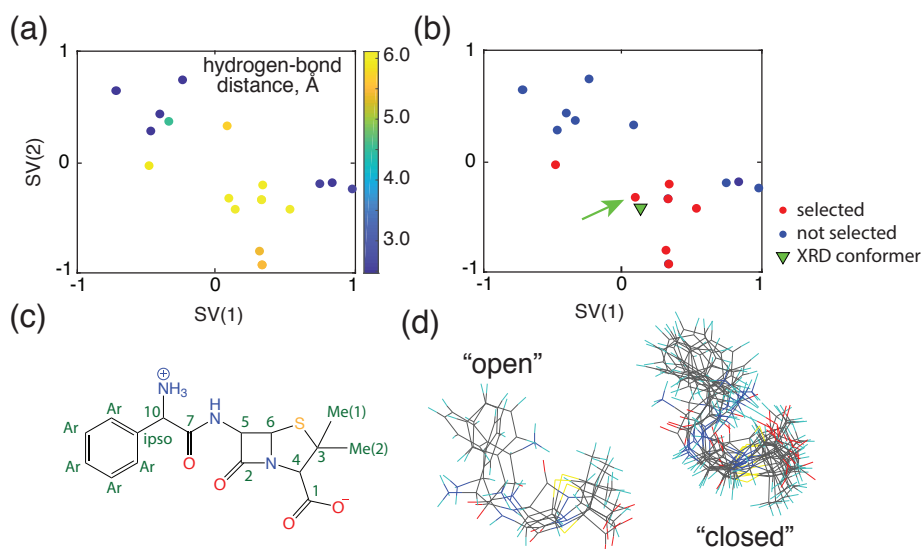


Figure 5.23. (a) Sketch-map representation of the locally stable ampicillin conformations. To show the extent of the sub-clustering, the panel is coloured according to the shortest intra-molecular hydrogen-bond distance [Å] between either NH_3 or NH and the carboxyl group. (b) Sketch-map projection of the gas-phase ampicillin ensemble. Red dots represent structures with the lowest number of constraint violations, and are therefore selected. The green triangle shows the conformer found in the XRD-determined crystal structure. The green arrow points to the gas-phase conformer that resulted in the correct crystal structure after the CSP procedure. (c) 2D structure of ampicillin with the labelling scheme used. (d) Overlay of the structures within the different sketch-map clusters. The “open” conformations correspond conformers without an intra-molecular hydrogen bond and were selected. The “closed” conformations mostly contain an intra-molecular hydrogen bond and were not selected.

Ampicillin Crystal Structure Generation

From the 7 selected conformations (AMCILL_OPLS3_5, AMCILL_OPLS3_7, AMCILL_OPLS3_10, AMCILL_OPLS3_12, AMCILL_OPLS3_13, AMCILL_OPLS3_14 and AMCILL_OPLS3_15, provided at <https://doi.org/10.1021/jacs.9b03908>) a set of crystal structures was generated using a low-discrepancy, quasi-random search of crystal packing variables using the GLEE (Global Lattice Energy Explorer) code.³³⁴ Each space group considered is sampled separately, by generating trial structures with unit cell dimensions, molecular positions and orientations sampled using a low discrepancy method. Crystal structures were generated in the 11 most commonly observed Sohncke space groups (1, 4, 5, 18, 19, 76, 78, 92, 96, 144, 145) until 2000 valid (successfully lattice energy minimized) crystal structures were generated in each space group, for each conformer.

All generated trial crystal structures were geometry-optimized using the crystal structure modelling code DMACRYS³³⁵ with the molecular geometry fixed at the gas-phase geometry. Intermolecular interactions were evaluated using the s using atomic multipoles, up to hexadecapole on each atom, derived using a distributed multipole analysis³⁴³ of the B3LYP/6-311G** charge density.

All predicted crystal structures within 20 kJ mol⁻¹ in total (intermolecular + conformational) energy of the lowest energy structure were then re-optimized using dispersion-corrected DFT. To ensure that all selected conformers were represented in the final crystal structures, a minimum of 5 crystal structures were taken from each conformer (whether or not they fell within the lowest 20 kJ mol⁻¹). This selection resulted in a total of 75 crystal structures. These were geometrically optimized with DFT using the Castep suite,²⁹⁷ using the PBE functional, the D2 dispersion correction, a 500 eV basis set cutoff and *k*-points sampled on a Monkhorst-Pack grid to provide a maximum reciprocal point spacing of 0.04 Å⁻¹. Each crystal structure was optimized in two stages: first, with the unit cell fixed from the force field predicted crystal structure, then fully geometrically optimized, including the unit cell and all atomic positions. The resulting structures were used as starting points for the chemical shift modelling (see below). All the geometrically optimized structures are given at <https://doi.org/10.1021/jacs.9b03908>.

Ampicillin Chemical Shift Calculations and Crystal Structure Selection

Structure modelling. Prior to the chemical shift calculations, all the trial structures and the single-crystal XRD structure of ampicillin³³⁷ were fully geometrically optimized, including the unit cell and all atomic positions, using the same DFT parametrization as outlined for the chemical shift calculations below.

DFT chemical shift calculation. For the ampicillin crystal structure selection, the magnetic shielding at the ¹H and ¹³C nuclei in the 75 trial crystal structures were calculated with plane-wave DFT using the GIPAW formalism^{104,105} and the Quantum ESPRESSO suite.²⁸⁸ For the GIPAW DFT calculations, the generalized-gradient-approximation (GGA) density functional PBE³¹¹ was used. We used the ultrasoft pseudopotentials with GIPAW reconstruction,^{104,105} C.pbe-n-kjpaw_psl.1.0.0.UPF, N.pbe-n-kjpaw_psl.1.0.0.UPF, H.pbe-kjpaw_psl.1.0.0.UPF, O.pbe-nl-kjpaw_psl.1.0.0.UPF and S.pbe-nl-kjpaw_psl.1.0.0.UPF from the pslibrary database [<https://dalcorso.github.io/pslibrary/>]<http://theosrv1.epfl.ch/Main/Pseudopotentials>. A wave-function energy cut-off of 100 Ry, a charge density energy cut-off of 400 Ry and a Monkhorst-Pack grid of *k*-points corresponding to a maximum spacing of 0.04 Å⁻¹ in the reciprocal space was used.³¹³ The electron density self-consistency convergence threshold was set to 10⁻¹² Ry. All Quantum ESPRESSO input and output files are given at <https://doi.org/10.1021/jacs.9b03908>.

ShiftML chemical shift calculation. For the ampicillin crystal structure selection, the magnetic shieldings of the 75 trial crystal structures were calculated using the ShiftML version described above. **Figure 5.9** shows the RMSE between ShiftML calculated and measured ¹H chemical shifts together with the DFT calculated relative lattice energies for the candidate set. Note that the RMSE between experiment and the ShiftML predicted chemical shifts follows the same trends as the RMSE between experiment and the GIPAW DFT calculated shifts (**Figure 5.8**).

Shielding to shift conversion and RMSE calculation. The calculated magnetic shielding was referenced to the experimental chemical shifts using the linear relationship $\sigma_{exp} = a - b\delta_{DFT}$, where the slope was fixed (i.e., $b = 1$) and the offset (a) was fit for each trial structure individually. For the ^1H chemical shift RMSE calculation, the methyl protons of each methyl group and the NH_3 protons were averaged. As it was not possible to distinguish the aromatic protons experimentally, as well as to distinguish the 2 methyl groups experimentally, the chemical shifts within each group were sorted, both for experimental and DFT chemical shifts, and then compared to each other. This was done for each crystal structure individually. The RMSE was calculated as,

$$RMSE = \sqrt{\sum_{i=1}^N \frac{(\sigma_{i,exp} - \sigma_{i,calc})^2}{N}}, \quad (5.4)$$

where σ_{exp} denotes the experimental chemical shift, σ_{calc} denotes the calculated chemical shift and the index i runs over all protons (N) within the asymmetric unit.

Positional error estimation

The positional error estimation, using DFT calculated chemical shifts, is done following the procedure introduced by Hofstetter and co-workers.¹⁸ First, we generate an ensemble of slightly perturbed crystal structures using a set of molecular dynamics (MD) simulations at finite temperatures. By “slightly perturbed” we refer to structures that remain within the same local minima and do not undergo any significant conformational shifts. The MD simulations are done at the DFT level using the universal force engine i-PI³⁴⁴ together with the Quantum ESPRESSO suite.²⁸⁸ During the MD simulations the crystal structures were kept at a constant temperature using the NVT ensemble and a GLE thermostat.³⁴⁵ The used temperatures are given as 1 K, 5 to 50 K in steps of 5 K and 60 to 240 K in steps of 10 K. For each temperature a MD simulation was run during 20 ps and with 1 step per fs. From each temperature we then extract 10 structures at random (5 from the first 10 ps and 5 from the last 10 ps), leading to 300 structures in total with a maximal positional displacement of 1.75 Å, for which the ^1H chemical shifts are calculated. This leads to a maximal chemical shift RMSD for ^1H of 1.99 ppm. **Figure 5.24** shows the correlation between positional deviations and the ^1H chemical shift RMSD.

For the MD DFT calculations the generalized-gradient-approximation (GGA) density functional PBE was used.³¹¹ We used the ultrasoft pseudopotentials with GIPAW reconstruction,^{104,105} C.pbe-n-kjpaw_psl.1.0.0.UPF, N.pbe-n-kjpaw_psl.1.0.0.UPF, H.pbe-n-kjpaw_psl.1.0.0.UPF, O.pbe-nl-kjpaw_psl.1.0.0.UPF and S.pbe-nl-kjpaw_psl.1.0.0.UPF from the pslibrary database [<https://dalcorso.github.io/pslibrary/>]. A wave-function energy cut-off of 60 Ry, a charge density energy cut-off of 240 Ry and no k -points. The electron density self-consistency convergence threshold was set to 10^{-8} Ry. For the GIPAW DFT calculations the same parametrization as for the chemical shift calculations of the trial crystal structures was used.

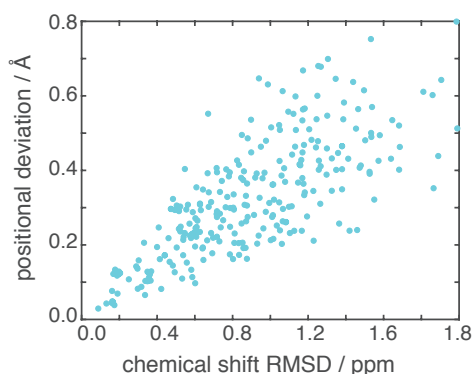


Figure 5.24. Correlation between positional RMSD (Å) and ^1H chemical shift RMSD (ppm) for an ensembles of perturbed crystal structures of ampicillin generated by MD.

5.5.2 Supplementary Information Section 5.3

Calculation of the Posterior Probability

With Full Chemical Shift Assignments. In order to evaluate $p(M|\mathbf{y}^*)$, one needs to combine information from all experimental shifts $\mathbf{y}^* = \{y_j^*\}$, determining the conditional probability $p(\mathbf{y}^*|M)$ based on the probabilities for individual shifts in **Equation 5.2**. In the simplest case a full assignment of the experimental shifts to the nuclei in the compound has been determined, for example through methods such as those described by Baiaes and co-workers.¹²⁰ Assuming independent errors on shifts from distinct nuclei, $p(\mathbf{y}^*|M)$ becomes,

$$p(\mathbf{y}^*|M) = \prod_j p_j(y_j^*|M). \quad (5.5)$$

Without Chemical Shift Assignments. Although the default scenario will involve full assignments of experimental shifts to particular nuclei, in rare cases definitive assignments may not be available. One must then consider the different ways of assigning the experimental shifts. If the permutation vector that describes one such assignment is denoted as \mathbf{a} , the conditional probability may be written as a sum over assignments,

$$p(\mathbf{y}^*|M) = \sum_{\mathbf{a}} p(\mathbf{y}^*|M, \mathbf{a}) p(\mathbf{a}|M), \quad (5.6)$$

where one can define the conditional probability for a given assignment as,

$$p(\mathbf{y}|M, \mathbf{a}) = \prod_j p_{a_j}(y_j|M). \quad (5.7)$$

If there is no heuristic way to determine the likelihood of a given assignment, $p(\mathbf{a}|M)$ has to be set to a constant. In this case, if one defines the matrix of conditional probabilities $P_{ij} = p_i(y_j|M)$, $p(\mathbf{y}^*|M)$ is proportional to the permanent of the matrix, $p(\mathbf{y}^*|M) = \text{perm } \mathbf{P}/n!$.

With Partial Chemical Shift Assignments. Cases in which none of the experimental shifts can be assigned are rare. In most cases the sum in **Equation 5.6** only needs to be evaluated over a subset of all the possible permutations of indices \mathbf{a} . In practice this means that \mathbf{P} can be made block-diagonal, each block \mathbf{P}_k corresponding to a group of nuclei that are distinct from the rest, but for which assignments among them are not available. The overall conditional probability can be written as a product between the permanents of the blocks,

$$p(\mathbf{y}^*|M) = \prod_k \text{perm } \mathbf{P}_k / n_k! \quad (5.8)$$

Where n_k indicates the size of the k -th block. While evaluating the permanent has a cost that grows combinatorially with the size of \mathbf{y}^* , algorithms with a low prefactor make its evaluation affordable up to a few tens of nuclei (per block k). In extraordinary cases where its evaluation is not possible, a pragmatic but generally inaccurate alternative is to assume **Equations 5.6** and **Equation 5.8** to be dominated by the contribution from the assignment producing the best-match between \mathbf{y}^M and \mathbf{y}^* .

Estimate of the Reference Errors

Clearly, the evaluation of $p(M|\mathbf{y}^*)$ requires an estimate of the uncertainties σ_j in calculated shifts. Assuming that any errors in the experimental determination of the shifts can be neglected, there are still multiple sources of errors to consider. First, experimental shifts average over thermal and quantum fluctuations, while GIPAW shifts are usually calculated for the nearest local energetic minimum. Second, approximations in the description of the electronic structure lead to errors in the predicted shifts. Third, errors are incurred by the conversion of the chemical shieldings obtained from GIPAW calculations (and ML models trained thereon) into chemical shifts.

Finally, when using a ML model, an environment-dependent statistical error relative to the GIPAW reference is added on top of the underlying theory/experiment discrepancy.

The statistical error σ_j^{ML} , can be characterised efficiently and accurately (more details can be found in the original publication *Engel et al., Phys. Chem. Chem. Phys.* 2019, *just accepted*), but estimating the error of the underlying GIPAW shifts with respect to experiment σ_j^{DFT} , usually requires extensive benchmarks. Existing datasets^{116,294,315} suggest that the typical errors are of the order of $\sigma_H^{DFT} = 0.33 \pm 0.16$ ppm and $\sigma_C^{DFT} = 1.9 \pm 0.4$ ppm. As an alternative to these estimates, one can assess σ_j for a specific molecule by considering $p_j(y|M)$ to depend parametrically on the uncertainty σ_j and maximizing $p(y^*)$ with respect to $\{\sigma_j\}$. Notably, this kind of maximum-likelihood approach usually requires large amounts of data. Consequently, one should either use a single, global value of σ_j for all environments in the crystal, or use the benchmark values to define a prior distribution for σ_j . In **Section 5.3** we discuss results obtained using a single, global value of σ_j per chemical species. The uncertainty in the predicted shifts arising from the conversion of the chemical shieldings is generally insignificant and will henceforth be neglected.

Accounting for Missing Structures

Chemical shift based NMRX relies strongly on CSP to generate candidate structures. Although CSP is constantly improving in thoroughness and energetic accuracy,¹¹⁹ one cannot entirely rule out the possibility that the experimental structure is not among the proposed candidates. We account for this scenario by adding a virtual structure \tilde{M} to the ensemble of CSP candidates, which represents the “neglected” structures. While its properties are largely an arbitrary choice, it makes sense to use a Gaussian with a mean and width corresponding to the mean and standard deviation of the shifts of the CSP candidates. If \tilde{M} has a substantial probability of matching experiment, one should question the comprehensiveness of the CSP candidate pool.

Visualizing the NMR Structural Landscape

Particularly in cases in which the Bayesian analysis does not allow the conclusive identification of the experimental structure, it is useful to gather further insights into the reasons why NMRX has reached the limits of its resolving power, and into whether and how it might be possible to reach a clearer assignment. A PCA of the shifts of all models provides a means of generating a low-dimensional representation that reflects the similarity of the different models in terms of their NMR shifts, in which one can then embed experiment. Unfortunately, prior assignments of shifts are required and one is limited to considering shifts from one chemical species.

We thus instead introduce a universally applicable approach, based on the definition of a kernel $k(M; M')$ (see paragraph below) and which reflects the probability that two models could be confused with each other when seen through the lens of their chemical shifts and the available degree of shift-structure assignment. A kernel PCA (KPCA) extracts a principal component projection of the models (and experiment). This approach owes its universal applicability to the availability of meaningful estimates of $p(y|M)$ in the presence of shifts from multiple chemical species and irrespective of whether shift assignments are available or not. Note that, if assignments are indeed available, i.e. when $p(y|M)$ is defined by **Equations 5.2** and **Equation 5.5**, and a global uncertainty is used, the distances in the KPCA representation again become a direct measure of the shift RMSDs – with the caveat that distortions can be introduced by the low-dimensional projection.

Embedding the experimentally measured shifts in a low-dimensional representation of the shift similarity provides a scale to the (dis)similarity of CSP candidates. In cases in which the experimental structure cannot be uniquely identified, it further provides a means of assessing whether two or more models are viable representatives of the experimental structure because they are indistinguishable in terms of their shifts, or because their predicted shifts are too inaccurate to resolve which one agrees with experiment despite distinct shift signatures.

We further perform a PCA on the structural features of all models as described within SOAP framework.^{291,292} Loosely speaking, atomic configurations are represented in terms of an atom-density, which distinguishes the different involved chemical species.³⁴⁶ It is constructed as the sum of Gaussian distributions centered on the atomic positions and symmetrised with respect to global translations and rigid rotations of the atomic configuration. The SOAP features correspond to coefficients obtained by expanding this atom-density description of atomic configurations in spherical harmonics and a set of orthogonal radial basis functions. This structural PCA allows us to generate a low-dimensional representation of the structural similarity of the different models. This provides complementary information to the KPCA representation of shift similarity, and permits distinguishing whether NMRX has reached the limits of its resolving power (a) because structurally dissimilar models produce similar shifts, (b) because the distinction between structurally very similar models is impossible, or (c) because the distinction between structurally dissimilar models with dissimilar shifts cannot be made due to the uncertainties in the predicted (and measured) shifts. It is worth noting that constructing the measure of structural similarity on a SOAP representation of the models is only one particular choice. In general, any metric of structural (dis)similarity for example the single molecule RMSE³⁴⁷ – can be used as a basis for a KPCA projection of structural similarity.

NMR-Based Similarity Kernel

We construct a matrix of pairwise distances between models (one of which may be experiment) $d(M, M') = -\ln p(M, M')$, where $p(M, M')$ is the probability of mistaking M for M' on the basis of shifts measurements. Momentarily setting aside normalisation, $p(M, M')$ can be calculated as,

$$p(M, M') = \int d\mathbf{y} p(M|\mathbf{y})p(\mathbf{y}|M') = \int d\mathbf{y} \frac{p(\mathbf{y}|M)p(\mathbf{y}|M')}{p(\mathbf{y}|M) + p(\mathbf{y}|M')}. \quad (5.9)$$

In the limit of infinitesimal uncertainties in the reference shifts, $\mathbf{y}^{M'}$, this simplifies to,

$$\lim_{\sigma^{M'} \rightarrow \varepsilon} p(M, M') \propto \varepsilon p(\mathbf{y}^{M'}|M), \quad (5.10)$$

which is then symmetrised and normalised, giving

$$p(M, M') = \frac{p(\mathbf{y}^{M'}|M) + p(\mathbf{y}^M|M')}{2\sqrt{p(\mathbf{y}^M|M)p(\mathbf{y}^{M'}|M')}}. \quad (5.11)$$

In the case, in which the probability is constructed from fully-assigned shifts, the resulting distance function is proportional to the squared Euclidean distance between the vectors containing chemical shifts of the various nuclei. A similarity kernel is then constructed by centering the associated distance matrix d ,

$$k(M, M') = \sum_{M'', M'''} h(M, M'')d(M'', M''')h(M''', M'), \quad (5.12)$$

$$h(M, M') = \delta_{M, M'} - 1/N_M, \quad (5.13)$$

and is then used in a KPCA scheme to identify the two principal components on which to represent structural diversity.

The Machine Learning Model

Here we discuss chemical shifts predicted using a ML model, which extends the GPR model built around the SOAP framework^{291,292} presented in **Chapter 4** by (i) training set sparsification via a projected process (PP) strategy,^{290,304,305} (ii) the efficient estimation of the uncertainty in predictions using a resampling approach,³⁰⁶ and (iii) the radial scaling approach,³⁴⁶ which drastically improves the computational performance compared to the

original multi-scale approach. Sparsification of the SOAP descriptions of atomic environments further speeds up predictions. The new model extends the original model presented in **Chapter 4** by incorporating sulfur-containing compounds thereby increasing the training set from 2000 to 3564 structures, and (slightly) outperforming it. Crucially, the expected errors of 0.48 ppm for out-of-sample predictions of ^1H shifts are comparable to the inherent error of the underlying GIPAW-DFT predictions with respect to experiment of around 0.33 ± 0.16 ppm.^{116,294,315}

It is worth noting that Liu *et al.*³⁰⁹ have recently demonstrated that, despite replacing the SOAP description of atomic densities with a non-symmetry-adapted real-space discretised equivalent, a sufficiently complex neural network architecture can tease out improvements of up to around 20% in prediction accuracy using the original training data. We nonetheless here choose a SOAP-GPR framework noting that the statistical ML uncertainties are uncorrelated with the inherent errors of the reference GIPAW data and must therefore be added to the GIPAW error(s) in quadrature. In consequence, reductions in ML errors at this point reap insignificant improvements to the resolving power of ML-based NMR crystallography without accompanying reductions in the underlying GIPAW errors with respect to experiment. The SOAP-GPR framework is robust, easily trained, has recently been generalised to the prediction of tensorial properties such as (anisotropic) chemical shielding tensors.²⁹⁶ Furthermore, it provides accurate estimates of prediction uncertainty.³⁰⁶ These are particularly important in this context, not only to estimate the reliability of assignments, but also because DFT calculations can at times yield unreliable results, and the ML model can be improved by automatically discarding problematic training data.

Chapter 6 Conclusions

In summary, the work shown in this thesis addresses some of the weak points of the isotropic chemical shift based NMRX approach for molecular solids. This approach is based on the combination of experimental NMR, CSP and calculations of NMR parameters, and over the last decade it has shown to be a powerful tool to obtain atomic level information from powdered molecular solids (allowing also for *de novo* crystal structure determination!). Given the experimental challenges related to the use of XRD with powdered samples, and given that often molecular solids are only available in the form of a powder (as it is increasingly the case in the pharmaceutical industry for example), the strengthen of NMR-based crystallographic methods has a significant impact in many areas of chemistry, life and material science.

We can divide the drawbacks of the isotropic chemical shift based NMRX method into two main categories: the experimental and the computational problems. Experimentally, the main difficulties are related to the acquisition of high-resolution ^1H NMR spectra in the solid-state (which is most of the times the nucleus of choice for this method). Computationally, the main disadvantages of this method are the computational costs associated with CSP and with the calculations of NMR parameters. In CSP, computational costs are usually cut down with severe cuts in the candidate crystal structures, which can lead to failure of the NMRX protocol when the correct structure is excluded from the CSP set.

From the experimental side, we have worked on methods to improve the resolution of ^1H spectra. We have first focused on the use of homonuclear dipolar decoupling sequences. In this frame, we have made a broad comparison of the experimental performance of CRAMPS experiments, comparing the “first-generation” sequences (developed in the 60s-70s) with newest and more commonly used phase-modulated schemes. Furthermore, we have written step by step guidelines for the optimization of these pulse sequences. Finally, we have probed the limits on resolution using these sequences by measuring transverse dephasing times. We have found that coherence lifetimes are limited by the appearance of a coherent oscillatory behaviour that leads to a residual anisotropic splitting, and that this oscillation can be completely removed in a double spin-echo experiment.

Next, we have focused on the development of methods to further improve the resolution at ultra-fast MAS ($\nu_R > 100$ kHz). In this spinning regime, we have shown by measuring T_2' values that non-refocusable interactions contribute to about the 30-40% of the residual line broadening in β -AspAla. To remove this contribution, we have developed the CT-MAS sequence, which makes use of a constant-time to remove the spectral broadening due to non-refocusable interactions. We have shown that this sequence leads to linewidths up to 40% narrower than the one-pulse experiment for β -AspAla. We have also shown that by combining the CT-MAS sequence with the TOP transformation we can obtain similar line narrowing at a fraction of the experimental time.

Then, we have hypothesized that in this spinning regime, where the residual linewidths are less than the chemical shift differences of the coupled spins, the homogeneous contribution to the residual broadening will be, at least in part, of a character similar to the scalar coupling ($\propto I_{1z}I_{2z}$). We have verified this hypothesis by using the anti-z-COSY pulse sequence (which is used for broadband homonuclear J decoupling in solution-state NMR) to reduce the residual line broadening of ^1H NMR spectra of powdered organic solids. We have demonstrated on three

samples that this experiment offers up to a factor of two gain in resolution as compared to conventional echo-detected MAS experiments.

From the computational side, we have first worked on reducing the computational costs associated with calculating NMR parameters by developing a machine learning method to accurately predict chemical shifts of molecular solids. This method, which was trained with the chemical shifts calculated with GIPAW DFT of 2000 crystal structures of molecular solids, is able to predict chemical shifts thousands of times faster than traditional DFT methods, while keeping comparable accuracy. The R^2 coefficients between the chemical shifts calculated with DFT and with ShiftML are 0.97 for ^1H , 0.99 for ^{13}C , 0.99 for ^{15}N , and 0.99 for ^{17}O , corresponding to RMSEs of 0.49 ppm for ^1H , 4.3 ppm for ^{13}C , 13.3 ppm for ^{15}N , and 17.7 ppm for ^{17}O . The calculation times are of the order of a few tens of seconds. The error (RMSE) between experimental and predicted ^1H chemical shifts for a set of six molecular crystals (for a total of 68 shifts) was 0.39 ppm. We have also shown that this protocol can be used in the NMRX approach to correctly determine the crystal structure of cocaine and AZD8329.

For reducing the CSP computational costs we developed a method to restrict the CSP search space to relevant regions of the conformational space. In the current NMRX protocol, the CSP conformer selection is based on single-molecule energies. This leads to failure of the method when the molecular conformer present in the crystal structure is of very high relative energy in the gas phase, which can be the case, for example, when the crystal structure conformation is stabilized through inter-molecular hydrogen bonds or other interactions only present in the solid phase. We have introduced a method that selects the conformer based on unambiguous experimental constraints (in our case, H-C distances) extracted from solid-state NMR experiments. We have shown that this method considerably decreases the number of conformers selected in multiple NMRX cases, saving a large amount of computational costs related to the CSP protocol, while ensuring that the correct conformer is not excluded. The large reduction in computational costs associated with the introduction of this method and of ShiftML is key for the application of the NMRX approach to larger and more flexible molecules. More importantly, we have shown that this method is able to correctly determine the crystal structure of ampicillin, which would have failed using the traditional approach because it adopts a high energy conformer in its crystal structure.

Finally, we have introduced a Bayesian framework to determine the confidence in the identification of the experimental crystal structure. This approach allows to unambiguously clarify NMRX cases where more than one candidate crystal structure falls within the confidence interval. Furthermore, we have shown that using this framework it is consistently possible to use ^{13}C shifts to improve the accuracy of the structure determination (if the estimated ^{13}C errors for GIPAW-DFT calculated shifts are correct). For instance, for flufenamic acid the addition of ^{13}C chemical shifts using our framework increases the confidence in the determination of the correct crystal structure from 80 to 96%. We have also introduced a visualization approach for the similarity between candidates in terms of their chemical shifts and in terms of their structures. This kind of visualization allows to critically assess the reliability of NMRX, giving information on whether lack of structural diversity, insufficient resolving power of the experiment, or uncertainties in the computationally-determined shifts are involved in inconclusive structural determinations.

Appendix

Chapter 2

Additional Section 2.2 and 2.3 Information

Experimental Parameters

The following tables show the experimental parameters used for the acquisition of the 1D ^1H CRAMPS spectra and the T_2' measurements show in **Section 2.2**. ω_r is the rotor spinning rate, ω_{RF} is the ^1H rf power, τ_c is the length of a single decoupling cycle (the length of the decoupling element is $m \cdot \tau_c$), τ_{acq} is the length of the acquisition window, m is the number of time the decoupling cycle is repeated within the decoupling element, ps is the number of phase steps used to describe the phase modulated pulses, α is the LG4 rotation axis, θ_1 is the length of the flip angle used at the end of the 2τ evolution in $E_y E_y$ experiments to measure the T_2' , θ_2 is the length of the pre-pulse, θ_{c1} and θ_{c2} are the lengths of the compensation pulses. All experiments can be found in the folder “Experiments” in the supporting information.

Sequence	LG								
Sample	Alanine			Glycine			β -AspAla		
$(\omega_r/2\pi)$ / kHz	3.0	12.5	22.0	3.0	12.5	22.0	3.0	12.5	22.0
$(\omega_{\text{RF}}/2\pi)$ / kHz	94	83	77	80	83	83	80	83	83
τ_c / μs	11	9.6	10.5	6.5	9.6	9.6	7	9.6	9.6
τ_{acq} / μs	3	3.1	3.1	3	3.1	3.1	3	3.1	3.1
m	1	1	1	1	1	1	1	1	1
ps	64	64	64	64	64	64	64	64	64
θ_1 / μs	1	1	1	1	1	1	1	1	1
θ_2 / μs	1.3	0.6	1.3	1.3	0.6	1.2	1	0.6	1.1
θ_{c1} / μs	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
θ_{c2} / μs	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
LG Rotation	cc	cc	cc	cc	cc	cc	cc	cc	cc

Table A.1. Experimental parameters used for the acquisition of the 1D ^1H CRAMPS experiment and T_2' measurements using the LG sequence for homonuclear dipolar decoupling. In LG rotation, “cc” stands for “counter-clockwise”.

Sequence	WHH-4								
Sample	Alanine			Glycine			β -AspAla		
$(\omega_r/2\pi)$ / kHz	3.0	12.5	22.0	3.0	12.5	22.0	3.0	12.5	22.0
$(\omega_{\text{RF}}/2\pi)$ / kHz	156.25	156.25	156.25	156.25	156.25	156.25	156.25	156.25	156.25
τ_c / μs	16.5	15.9	16.5	18.3	16.8	16.2	17.4	16.8	16.5
τ_{acq} / μs	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7

m	1	1	1	1	1	1	1	1	1
ps	-	-	-	-	-	-	-	-	-
$\theta_1 / \mu\text{s}$	1	1	1	1	1	1	1	1	1
$\theta_2 / \mu\text{s}$	1.4	1.3	1.3	1.35	1.35	1.3	1.35	1.35	1.35
$\theta_{c1} / \mu\text{s}$	-	-	-	-	-	-	-	-	-
$\theta_{c2} / \mu\text{s}$	-	-	-	-	-	-	-	-	-

Table A.2. Experimental parameters used for the acquisition of the 1D ^1H CRAMPS experiment and T_2' measurements using the WHH-4 sequence for homonuclear dipolar decoupling.

Sequence	MREV-8								
Sample	Alanine			Glycine			β -AspAla		
$(\omega_r/2\pi) / \text{kHz}$	3.0	12.5	22.0	3.0	12.5	22.0	3.0	12.5	22.0
$(\omega_{\text{RF}}/2\pi) / \text{kHz}$	156.25	156.25	156.25	156.25	156.25	156.25	156.25	156.25	156.25
$\tau_c / \mu\text{s}$	33	33	31.8	31.8	33	32.4	32.4	33	33
$\tau_{\text{acq}} / \mu\text{s}$	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
m	1	1	1	1	1	1	1	1	1
ps	-	-	-	-	-	-	-	-	-
$\theta_1 / \mu\text{s}$	1	1	1	1	1	1	1	1	1
$\theta_2 / \mu\text{s}$	1.28	1.28	1.1	1.1	1.15	1.05	1.1	1.15	1.1
$\theta_{c1} / \mu\text{s}$	-	-	-	-	-	-	-	-	-
$\theta_{c2} / \mu\text{s}$	-	-	-	-	-	-	-	-	-

Table A.3. Experimental parameters used for the acquisition of the 1D ^1H CRAMPS experiment and T_2' measurements using the MREV-8 sequence for homonuclear dipolar decoupling.

Sequence	BR-24								
Sample	Alanine			Glycine			β -AspAla		
$(\omega_r/2\pi) / \text{kHz}$	3.0	12.5	22.0	3.0	12.5	22.0	3.0	12.5	22.0
$(\omega_{\text{RF}}/2\pi) / \text{kHz}$	156.25	156.25	156.25	156.25	156.25	156.25	156.25	156.25	156.25
$\tau_c / \mu\text{s}$	99	103.5	100.8	95.4	102.6	102.6	95.4	102.6	102.6
$\tau_{\text{acq}} / \mu\text{s}$	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
m	1	1	1	1	1	1	1	1	1
ps	-	-	-	-	-	-	-	-	-
$\theta_1 / \mu\text{s}$	1	1	1	1	1	1	1	1	1
$\theta_2 / \mu\text{s}$	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3
$\theta_{c1} / \mu\text{s}$	-	-	-	-	-	-	-	-	-
$\theta_{c2} / \mu\text{s}$	-	-	-	-	-	-	-	-	-

Table A.4. Experimental parameters used for the acquisition of the 1D ^1H CRAMPS experiment and T_2' measurements using the BR-24 sequence for homonuclear dipolar decoupling.

Sequence	PMLG								
Sample	Alanine			Glycine			β -AspAla		
$(\omega_r/2\pi) / \text{kHz}$	3.0	12.5	22.0	3.0	12.5	22.0	3.0	12.5	22.0
$(\omega_{\text{RF}}/2\pi) / \text{kHz}$	72	72	80	70	84	83	71	80	77
$\tau_c / \mu\text{s}$	19.3	18.5	27	19.8	16.2	23.8	20.4	16.2	26.4

$\tau_{\text{acq}} / \mu\text{s}$	3	3	3.05	3.25	2.95	3.05	3.2	2.95	3.05
m	1	1	1	1	1	1	1	1	1
ps	64	64	64	64	64	64	64	64	64
$\theta_1 / \mu\text{s}$	1	1	1	1	1	1	1	1	1
$\theta_2 / \mu\text{s}$	0.5	0.4	0.9	0.65	0.65	0.8	1	0.5	0.8
$\theta_{c1} / \mu\text{s}$	0.9	0.85	0.85	0.45	1	0.95	0.35	0.95	0.95
$\theta_{c2} / \mu\text{s}$	0.05	0.15	0.15	0.55	0.15	0.15	0.55	0.15	0.15
PMLG Rotation	cc	cc	cc	cc	cc	cc	cc	cc	cc

Table A.5. Experimental parameters used for the acquisition of the 1D ^1H CRAMPS experiment and T_2' measurements using the PMLG sequence for homonuclear dipolar decoupling. In PMLG rotation, “cc” stands for “counter-clockwise”.

Sequence	DUMBO-1								
Sample	Alanine			Glycine			β -AspAla		
$(\omega_r/2\pi) / \text{kHz}$	3.0	12.5	22.0	3.0	12.5	22.0	3.0	12.5	22.0
$(\omega_{\text{RF}}/2\pi) / \text{kHz}$	100	110	100	100	120	119	100	110	112
$\tau_c / \mu\text{s}$	32.25	28.5	26.5	32.25	25	27.5	32.25	28.5	28
$\tau_{\text{acq}} / \mu\text{s}$	3.65	3.6	3.8	3.65	3.55	3.75	3.65	3.55	3.75
m	2	2	2	2	2	2	2	2	2
ps	64	64	64	64	64	64	64	64	64
$\theta_1 / \mu\text{s}$	1	1	1	1	1	1	1	1	1
$\theta_2 / \mu\text{s}$	0.8	1	0.8	0.8	0.35	0.8	0.7	1	1.35
$\theta_{c1} / \mu\text{s}$	0.8	0.8	0.55	0.75	0.9	0.85	0.75	0.9	0.85
$\theta_{c2} / \mu\text{s}$	0.33	0.38	0.45	0.4	0.25	0.45	0.4	0.35	0.45

Table A.6. Experimental parameters used for the acquisition of the 1D ^1H CRAMPS experiment and T_2' measurements using the DUMBO-1 sequence for homonuclear dipolar decoupling.

Sequence	eDUMBO-1 ₂₂								
Sample	Alanine			Glycine			β -AspAla		
$(\omega_r/2\pi) / \text{kHz}$	3.0	12.5	22.0	3.0	12.5	22.0	3.0	12.5	22.0
$(\omega_{\text{RF}}/2\pi) / \text{kHz}$	125	125	125	125	125	125	125	125	125
$\tau_c / \mu\text{s}$	30.2	27.25	27.25	29.5	25	28	30	25	28
$\tau_{\text{acq}} / \mu\text{s}$	3.6	3.9	3.8	3.2	3.65	3.8	3.65	3.9	3.8
m	2	2	2	2	2	2	2	2	2
ps	64	64	64	64	64	64	64	64	64
$\theta_1 / \mu\text{s}$	1	1	1	1	1	1	1	1	1
$\theta_2 / \mu\text{s}$	0.4	0.4	0.5	0.6	0.4	0.45	0.7	0.4	0.5
$\theta_{c1} / \mu\text{s}$	0.7	0.65	0.55	0.6	0.65	0.55	0.6	0.65	0.55
$\theta_{c2} / \mu\text{s}$	0.8	0.35	0.45	0.8	0.35	0.45	0.9	0.35	0.45

Table A.7. Experimental parameters used for the acquisition of the 1D ^1H CRAMPS experiment and T_2' measurements using the eDUMBO-1₂₂ sequence for homonuclear dipolar decoupling.

Sequence	LG4								
Sample	Alanine			Glycine			β -AspAla		
$(\omega_r/2\pi) / \text{kHz}$	3.0	12.5	22.0	3.0	12.5	22.0	3.0	12.5	22.0

$(\omega_{\text{RF}}/2\pi) / \text{kHz}$	80	125	125	80	125	125	80	125	125
$\tau_c / \mu\text{s}$	25	15.33	17.67	24.5	16	17	25	15.33	16.8
$\tau_{\text{acq}} / \mu\text{s}$	4.35	4.6	4.6	3.35	4.6	4.55	3.35	4.6	4.6
m	3	3	3	3	3	3	3	3	3
$\alpha / ^\circ$	70	70	70	70	70	70	70	70	70
ps	120	120	120	120	120	120	120	120	120
$\theta_1 / \mu\text{s}$	1	1	1	1	1	1	1	1	1
$\theta_2 / \mu\text{s}$	0.3	0.4	0.1	0.1	0.4	0.55	0.1	0.4	0.3
$\theta_{c1} / \mu\text{s}$	0.5	0.5	0.45	0.2	0.6	0.5	0.3	0.5	0.5
$\theta_{c2} / \mu\text{s}$	0.5	1	1	0.55	0.8	0.8	0.55	0.9	0.9

Table A.8. Experimental parameters used for the acquisition of the 1D ^1H CRAMPS experiment and T_2' measurements using the LG4 sequence for homonuclear dipolar decoupling.

Linewidth Measurements

The following tables show the linewidths of the resonances observed in 1D ^1H CRAMPS experiments shown in **Section 2.2**. All linewidths values are corrected for the scaling factor.

Sample		Alanine		
MAS / kHz	Sequence	Linewidth / Hz		
		CH_3	CH	NH_3^+
3.0	WHH-4	189	221	1318
	MREV-8	162	193	1363
	BR-24	115	180	1530
	PMLG	482	358	1677
	DUMBO-1	136	208	1371
	eDUMBO-1 ₂₂	371	365	1224
	LG4	199	237	1493
12.5	MAS	1425	1671	1813
	LG	1472	1584	1635
	WHH-4	207	196	873
	MREV-8	213	189	1065
	BR-24	170	273	1199
	PMLG	228	180	814
	DUMBO-1	151	169	914
	eDUMBO-1 ₂₂	104	149	878
	LG4	167	177	1010
22.0	MAS	1102	1014	1214
	LG	1255	1770	1297
	WHH-4	264	221	724
	MREV-8	272	196	669
	BR-24	584	461	1893
	PMLG	213	171	629
	DUMBO-1	205	195	592

	eDUMBO-1 ₂₂	113	117	399
	LG4	190	161	499

Table A.9. Linewidths of the resonances observed in the 1D ¹H CRAMPS and MAS experiments of powdered alanine shown in **Section 2.2**.

Sample		Glycine		
MAS / kHz	Sequence	Linewidth / Hz		
		CH _{2a}	CH _{2b}	NH ₃ ⁺
3.0	WHH-4	547	264	492
	MREV-8	230	245	427
	BR-24	176	210	241
	PMLG	572	560	708
	DUMBO-1	172	211	315
	eDUMBO-1 ₂₂	421	383	519
	LG4	236	242	332
12.5	WHH-4	358	289	429
	MREV-8	343	346	537
	BR-24	375	340	441
	PMLG	256	228	336
	DUMBO-1	189	191	282
	eDUMBO-1 ₂₂	195	204	309
	LG4	241	237	309
22.0	LG	1701	1485	1574
	WHH-4	653	327	656
	MREV-8	593	401	629
	BR-24	688	533	782
	PMLG	236	217	337
	DUMBO-1	186	191	316
	eDUMBO-1 ₂₂	180	204	325
	LG4	318	256	389

Table A.10. Linewidths of the resonances observed in the 1D ¹H CRAMPS experiments of powdered glycine shown in **Section 2.2**.

Sample		β-AspAla							
MAS / kHz	Sequence	Linewidth / Hz							
		CH ₃	CH _{2a}	CH _{2b}	CH _(Asp)	CH _(Ala)	NH ₃ ⁺	NH	OH
3.0	WHH-4	220	184	426	217	262	386	441	291
	MREV-8	190	192	198	187	202	415	343	260
	BR-24	125	180	196	220	186	214	170	206
	PMLG	440	653	182	377	315	366	404	432
	DUMBO-1	133	171	195	166	163	317	250	260
	eDUMBO-1 ₂₂	251	371	545	520	192	325	423	307
	LG4	157	163	203	177	123	357	-	321
12.5	WHH-4	221	242	204	234	217	456	229	244

	MREV-8	221	414	346	267	221	504	467	420
	BR-24	292	227	290	216	187	417	300	287
	PMLG	166	201	222	138	183	253	-	179
	DUMBO-1	119	130	163	160	176	342	205	248
	eDUMBO-1 ₂₂	119	135	158	175	157	330	264	295
	LG4	155	123	178	136	149	289	325	220
22.0	WHH-4	267	-	-	155	322	647	-	388
	MREV-8	272	237	497	329	275	435	531	327
	BR-24	585	-	-	1075	525	808	426	286
	PMLG	178	171	333	228	156	382	347	264
	DUMBO-1	121	120	150	158	156	179	-	132
	eDUMBO-1 ₂₂	109	104	130	124	160	208	287	126
	LG4	203	172	201	206	155	366	253	301

Table A.11. Linewidths of the resonances observed in the 1D ¹H CRAMPS experiments of powdered β-AspAla shown in **Section 2.2**.

Peak Positions

The following tables show the peak positions and the scaling factor observed in the 1D ¹H CRAMPS experiments shown in **Section 2.2**. Scaling factors were calculated as detailed in **Section 2.1**.

Sample		Alanine				
MAS / kHz		Sequence	Peak Position / ppm			Experimental Scaling Factor
			CH ₃	CH	NH ₃ ⁺	
3.0	WHH-4	1.13	3.61	8.54	0.55	
	MREV-8	1.13	3.60	8.54	0.47	
	BR-24	1.13	3.60	8.54	0.39	
	PMLG	1.13	3.38	8.54	0.58	
	DUMBO-1	1.13	3.49	8.54	0.47	
	eDUMBO-1 ₂₂	1.13	3.52	8.54	0.53	
	LG4	1.13	3.65	8.54	0.42	
12.5	MAS	1.13	3.62	8.54	-	
	LG	1.13	3.80	8.54	0.67	
	WHH-4	1.13	3.67	8.54	0.58	
	MREV-8	1.13	3.61	8.54	0.46	
	BR-24	1.13	3.58	8.54	0.38	
	PMLG	1.13	3.46	8.54	0.56	
	DUMBO-1	1.13	3.49	8.54	0.49	
	eDUMBO-1 ₂₂	1.13	3.48	8.54	0.53	
	LG4	1.13	3.66	8.54	0.44	
22.0	MAS	1.13	3.62	8.54	-	
	LG	1.13	3.78	8.54	0.65	
	WHH-4	1.13	3.67	8.54	0.54	
	MREV-8	1.13	3.68	8.54	0.45	

	BR-24	1.13	3.75	8.54	0.36
	PMLG	1.13	3.69	8.54	0.50
	DUMBO-1	1.13	3.64	8.54	0.53
	eDUMBO-1 ₂₂	1.13	3.51	8.54	0.55
	LG4	1.13	3.65	8.54	0.41
Reference Position		1.13	3.62	8.54	

Table A.12. Peak position of the resonances observed in the 1D ¹H CRAMPS and MAS experiments of powdered alanine shown in **Section 2.2**. The reference peak positions were obtained from the 22.0 kHz MAS spectrum of powdered alanine.

Sample		Glycine			
MAS / kHz Sequence		Peak Position / ppm			Experimental Scaling Factor
		CH _{2a}	CH _{2b}	NH ₃ ⁺	
3.0	WHH-4	2.30	3.41	8.00	0.56
	MREV-8	2.30	3.40	8.00	0.46
	BR-24	2.30	3.41	8.00	0.38
	PMLG	2.30	3.44	8.00	0.57
	DUMBO-1	2.30	3.36	8.00	0.45
	eDUMBO-1 ₂₂	2.30	3.50	8.00	0.52
	LG4	2.30	3.40	8.00	0.38
12.5	WHH-4	2.30	3.45	8.00	0.58
	MREV-8	2.30	3.47	8.00	0.45
	BR-24	2.30	3.36	8.00	0.38
	PMLG	2.30	3.42	8.00	0.57
	DUMBO-1	2.30	3.34	8.00	0.47
	eDUMBO-1 ₂₂	2.30	3.35	8.00	0.53
	LG4	2.30	3.41	8.00	0.42
22.0	LG	2.30	3.96	8.00	0.53
	WHH-4	2.30	3.49	8.00	0.58
	MREV-8	2.30	3.52	8.00	0.45
	BR-24	2.30	3.38	8.00	0.36
	PMLG	2.30	3.39	8.00	0.50
	DUMBO-1	2.30	3.30	8.00	0.43
	eDUMBO-1 ₂₂	2.30	3.33	8.00	0.50
	LG4	2.30	3.34	8.00	0.38
Reference Position		2.30	3.80	8.00	

Table A.13. Peak position of the resonances observed in the 1D ¹H CRAMPS experiments of powdered glycine shown in **Section 2.2**. The reference positions were obtained from Ref. 348. Note that the reference position of the CH_{2b} peak is different from the measured position. For the RMSE calculations we used a reference value of 3.4 ppm for this resonance, which is in line with the values we observe experimentally.

Sample		β-AspAla								
MAS / kHz Sequence		Peak Position / ppm								Experimental Scaling Factor
		CH ₃	CH _{2a}	CH _{2b}	CH _(Asp)	CH _(Ala)	NH ₃ ⁺	NH	OH	
3.0	WHH-4	0.86	1.92	2.55	3.88	4.79	7.34	7.99	12.59	0.59
	MREV-8	0.86	2.18	2.74	4.00	4.89	7.43	8.08	12.59	0.49
	BR-24	0.86	2.21	2.75	4.04	4.99	7.58	8.36	12.59	0.39
	PMLG	0.86	2.24	2.89	3.70	4.83	7.17	7.67	12.59	0.58
	DUMBO-1	0.86	1.99	2.59	3.79	4.65	7.15	7.82	12.59	0.49
	eDUMBO-1 ₂₂	0.86	1.99	2.67	3.88	4.86	7.37	7.84	12.59	0.55
	LG4	0.86	2.11	2.72	3.99	4.85	7.47	-	12.59	0.40
12.5	WHH-4	0.86	2.05	2.61	3.91	4.87	7.42	8.16	12.59	0.60
	MREV-8	0.86	2.17	2.78	4.01	4.93	7.45	8.14	12.59	0.48
	BR-24	0.86	2.11	2.70	4.04	4.90	7.47	8.38	12.59	0.40
	PMLG	0.86	1.99	2.55	3.78	4.68	7.19	-	12.59	0.60
	DUMBO-1	0.86	1.95	2.48	3.71	4.59	7.18	7.74	12.59	0.49
	eDUMBO-1 ₂₂	0.86	2.03	2.59	3.81	4.72	7.32	7.88	12.59	0.57
	LG4	0.86	2.15	2.71	4.00	4.90	7.50	8.17	12.59	0.44
22.0	WHH-4	0.86	-	-	3.97	4.86	7.54	-	12.59	0.60
	MREV-8	0.86	1.99	2.77	4.00	4.90	7.43	8.04	12.59	0.48
	BR-24	0.86	-	-	4.06	5.23	7.71	8.64	12.59	0.36
	PMLG	0.86	1.97	2.51	3.92	4.90	7.33	7.79	12.59	0.51
	DUMBO-1	0.86	1.90	2.40	3.63	4.55	7.08	-	12.59	0.48
	eDUMBO-1 ₂₂	0.86	2.07	2.62	3.88	4.78	7.26	7.75	12.59	0.56
	LG4	0.86	2.14	2.70	4.01	4.91	7.50	8.19	12.59	0.42
Reference Position		0.86	2.13	2.71	3.96	4.85	7.36	7.94	12.59	

Table A.14. Peak position of the resonances observed in the 1D ¹H CRAMPS experiments of powdered β -AspAla shown in **Section 2.2**. Reference positions were obtained from the 111.0 kHz ¹H MAS spectrum of powdered β -AspAla.

T_2' Measurements

The following table show the T_2' values measured by fitting the dephasing curves shown in **Figure 2.8**, **Figure 2.9** and **Figure 2.10** of **Section 2.2**. The curves were fit using an exponential decay $S(2\tau_d) = a \cdot \exp(-2\tau_d/T_2')$.

Sequence	T_2' / ms								
	Alanine			Glycine			β -AspAla		
	3.0 kHz MAS	12.5 kHz MAS	22.0 kHz MAS	3.0 kHz MAS	12.5 kHz MAS	22.0 kHz MAS	3.0 kHz MAS	12.5 kHz MAS	22.0 kHz MAS
LG	0.9	1.4	1.2	0.8	1.4	1.0	1.2	0.9	1.2
WHH-4	4.3	1.8	2.5	3.2	2.8	1.7	2.8	4.2	2.9
MREV-8	5.8	3.1	2.5	3.7	2.6	1.1	6.6	4.4	2.9

BR-24	11.2	2.3	2.8	10.6	4.2	1.5	15.3	5.4	2.8
PMLG	1.3	6.7	5.6	0.4	2.8	-	2.6	4.5	5.0
DUMBO-1	9.4	8.4	6.3	6.8	7.3	6.0	11.3	12.2	11.8
eDUMBO-1₂₂	1.4	10.2	9.7	1.8	7.0	7.9	2.4	10.4	14.9
LG4	5.0	13.4	8.4	4.2	4.1	6.9	6.6	9.7	8.8

Table A.15. T_2' values obtained by fitting the dephasing curves shown in **Section 2.2**.

WHH-4 Pulse Program

```
;whh4.fmp
;
;1D CRAMPS experiment using the WHH-4 sequences for homonuclear dipolar decoupling.
;The FID is sampled in acquisition windows within the WHH-4 sequence.
;
;For a detailed description on the parameters optimization of this pulse program, see:
F.M. Paruzzo and L. Emsley (2019), High-Resolution 1H NMR of Powdered Solids by
Homonuclear Dipolar Decoupling
;
;Tested on Bruker Avance III HD / TopSpin 3.5
;Last Modification: F.M. Paruzzo, 16th July 2019
;
;$CLASS=Solids
;$DIM=1D
;$TYPE=homonuclear dipolar decoupling
;$COMMENT=1D CRAMPS experiment using WHH-4 sequences
;
;-----
;Parameters:
;d1 : recycle delay
;d2 : WHH-4 tau delay
;d4 : WHH-4 2tau delay
;d5 : WHH-4 delay during acquisition window
;d9 : time for acquisition of FID points during each acquisition window
;
;p11 : not used (=0)
;p12 : power level for initial 90 degree and pre-pulse
;p13 : power level for WHH-4 pulses
;
;p1 : 1H 90 degree pulse length at p13 (WHH-4 pulses)
;p9 : length of the acquisition window
;p10: tau (us)
;p12 : 1H 90 degree pulse length at p12
;p14: pre-pulse length (typically between 0.5 and 1.5 us)
;
;l11: number of data points to be detected each acquisition window
;
;cnst10 : additional phase of WHH-4 pulses
;cnst25 : additional phase of pre-pulse
;cnst31 : spinning rate (Hz)
;
;o1 : 1H offset, best when at the right side of the resonances, close to the region to
decouple
;ns : n*4
;tau : tau value for WHH-4 decoupling
;trtc : calculated as ratio between rotor period/cycle time
;dead : receiver dead time in s
;cycle : decoupling cycle time
;count : total number of acquisition window/decoupling pulses
;-----

"acqt0=0"

dwellmode auto
#include <Avancesolids.incl>
#include <Delayssolids.incl>
```



```

    "d9=0.1u*(l11)"                ;set the sampling window in
Avancesolids.incl as l11 multiples of 100 ns
define delay dead
    "dead=1.6u"
    "p9=dead+d9+0.1u+1.4u"
define delay tau
    "tau=p10"
    "d2=tau-p1"
    "d4=2*tau-p1"
    "d5=(2*tau-p1-p9)/2"
define delay cycle
    "cycle=6*tau"
define loopcounter count
    "count=aq/cycle-1"
define delay rest ;
    "rest=aq-(count*cycle)"
define delay trot
    "trot=1/cnst31"
define delay trtc
    "trtc=trot/cycle"

1 ze
    1u ip4+cnst25                ;increment the phase of the pre-pulse
    1u ip10+cnst10               ;increment the WHH-4 phases
    1u ip11+cnst10
    1u ip12+cnst10
    1u ip13+cnst10
2 d1 pl12:f1                    ;recycle delay
    10u reset1:f1                ;synchronise pulse and detection RF
    STARTADC                     ;prepare adc for sampling, set
reference frequency, defined in Avancedru.incl
    RESETPHASE                   ;reset reference phase
    1u rpp10                     ;reset phase list pointer
    (p12 pl12 ph1):f1            ;first 90 at pl12
    (p14 pl12 ph4):f1            ;pre-pulse at pl12
    .1u DWL_CLK_ON
4 dead pl13:f1
    d9 RG_ON                     ;acquire l11 data points
    .1u RG_OFF
    1.4u
    d5                           ;WHH-4 decoupling
    p1:f1 ph10
    d2
    p1:f1 ph11
    d4
    p1:f1 ph12
    d2
    p1:f1 ph13
    d5                           ;end of WHH-4 decoupling
    lo to 4 times count
    rest
    1u DWL_CLK_OFF
    rcyc=2
5 100m wr #0
6 exit

ph0=2
ph1=1 2 3 0
ph4=0
ph10= 2
ph11= 1
ph12= 3
ph13= 0
ph30=1
ph31=0 1 2 3

```

MREV-8 Pulse Program

```

;mrev8.fmp
;
;1D CRAMPS experiment using the MREV-8 sequences for homonuclear dipolar decoupling.
;The FID is sampled in acquisition windows within the MREV-8 sequence.
;
;For a detailed description on the parameters optimization of this pulse program, see:
F.M. Paruzzo and L. Emsley (2019), High-Resolution 1H NMR of Powdered Solids by
Homonuclear Dipolar Decoupling
;
;Tested on Bruker Avance III HD / TopSpin 3.5
;Last Modification: F.M. Paruzzo, 16th July 2019
;
;$CLASS=Solids
;$DIM=1D
;$TYPE=homonuclear dipolar decoupling
;$COMMENT=1D CRAMPS experiment using MREV-8 sequences
;
;-----
;Parameters:
;d1 : recycle delay
;d2 : MREV-8 tau delay
;d4 : MREV-8 2tau delay
;d5 : MREV-8 delay during acquisition window
;d9 : time for acquisition of FID points during each acquisition window
;
;pl1 : not used (=0)
;pl12 : power level for initial 90 degree and pre-pulse
;pl13 : power level for MREV-8 pulses
;
;p1 : 1H 90 degree pulse lenght at pl13 (MREV-8 pulses)
;p9 : lenght of the acquisition window
;p10: tau (us)
;p12 : 1H 90 degree pulse lenght at pl12
;p14: pre-pulse length (typically between 0.5 and 1.5 us)
;
;l11: number of data points to be detected each acquisition window
;
;cnst10 : additional phase of MREV-8 pulses
;cnst25 : additional phase of pre-pulse
;cnst31 : spinning rate (Hz)
;
;o1 : 1H offset, best when at the right side of the resonances, close to the region to
decouple
;ns : n*4
;tau : tau value for MREV-8 decoupling
;trtc : calculated as ratio between rotor period/cycle time
;dead : receiver dead time in s
;cycle : decoupling cycle time
;count : total number of acquisition window/decoupling pulses
;-----

"acqt0=0"

dwellmode auto
#include <Avancesolids.incl>
#include <Delayssolids.incl>

"d9=0.1u*(l11)" ;set the sampling window in
Avancesolids.incl as l11 multiples of 100 ns
define delay dead
"dead=1.6u"
"p9=dead+d9+0.1u+1.4u"
define delay tau
"tau=p10"
"d2=tau-p1"
"d4=2*tau-p1"
"d5=(2*tau-p1-p9)/2"

```

```

define delay cycle
  "cycle=12*tau"
define loopcounter count
  "count=aq/cycle-1"
define delay rest
  "rest=aq-(count*cycle)"
define delay trot
  "trot=1/cnst31"
define delay trtc
  "trtc=trot/cycle"

1 ze
  1u ip4+cnst25          ;increment the phase of the pre-pulse
  1u ip10+cnst10         ;increment the MREV-8 phases
  1u ip11+cnst10
  1u ip12+cnst10
  1u ip13+cnst10

2 d1 pl12:f1              ;recycle delay
  10u reset1:f1           ;synchronise pulse and detection RF
  STARTADC                ;prepare adc for sampling, set
reference frequency, defined in Avancedru.incl
  RESETPHASE              ;reset reference phase
  1u rpp10                 ;reset phase list pointer
  (p12 pl12 ph1):f1       ;first 90 at pl12
  (p14 pl12 ph4):f1       ;pre-pulse at pl12
  .1u DWL_CLK_ON
4 dead pl13:f1
  d9 RG_ON                 ;acquire l11 data points
  .1u RG_OFF
  1.4u
  d5                       ;MREV-8 decoupling
  p1:f1 ph10
  d2
  p1:f1 ph11
  d4
  p1:f1 ph12
  d2
  p1:f1 ph13
  d4
  p1:f1 ph13
  d2
  p1:f1 ph11
  d4
  p1:f1 ph12
  d2
  p1:f1 ph10
  d5                       ;end of MREV-8
decoupling
  lo to 4 times count     ;make sure td points are sampled
  rest
  1u DWL_CLK_OFF
  rcyc=2
5 100m wr #0
6 exit

ph0=0
ph1=1 2 3 0
ph4=0
ph10= 0
ph11= 1
ph12= 3
ph13= 2
ph30=1
ph31=0 1 2 3

```

BR-24 Pulse Program

```

;br24.fmp
;
;1D CRAMPS experiment using the BR-24 sequences for homonuclear dipolar decoupling.
;The FID is sampled in acquisition windows within the BR-24 sequence.
;
;For a detailed description on the parameters optimization of this pulse program, see:
F.M. Paruzzo and L. Emsley (2019), High-Resolution 1H NMR of Powdered Solids by
Homonuclear Dipolar Decoupling
;
;Tested on Bruker Avance III HD / TopSpin 3.5
;Last Modification: F.M. Paruzzo, 16th July 2019
;
;$CLASS=Solids
;$DIM=1D
;$TYPE=homonuclear dipolar decoupling
;$COMMENT=1D CRAMPS experiment using BR-24 sequences
;
;-----
;Parameters:
;d1 : recycle delay
;d2 : BR-24 tau delay
;d4 : BR-24 2tau delay
;d5 : BR-24 delay during acquisition window
;d9 : time for acquisition of FID points during each acquisition window
;
;pl1 : not used (=0)
;pl12 : power level for initial 90 degree and pre-pulse
;pl13 : power level for BR-24 pulses
;
;p1 : 1H 90 degree pulse length at pl13 (BR-24 pulses)
;p9 : length of the acquisition window
;p10: tau (us)
;p12 : 1H 90 degree pulse length at pl12
;p14: pre-pulse length (typically between 0.5 and 1.5 us)
;
;l11: number of data points to be detected each acquisition window
;
;cnst10 : additional phase of BR-24 pulses
;cnst25 : additional phase of pre-pulse
;cnst31 : spinning rate (Hz)
;
;o1 : 1H offset, best when at the right side of the resonances, close to the region to
decouple
;ns : n*4
;tau : tau value for BR-24 decoupling
;trtc : calculated as ratio between rotor period/cycle time
;dead : receiver dead time in s
;cycle : decoupling cycle time
;count : total number of acquisition window/decoupling pulses
;-----

"acqt0=0"

dwellmode auto
#include <Avancesolids.incl>
#include <Delayssolids.incl>

"d9=0.1u*(l11)" ;set the sampling window in
Avancesolids.incl as l11 multiples of 100 ns
define delay dead
"dead=1.6u"
"p9=dead+d9+0.1u+1.4u"
define delay tau
"tau=p10"
"d2=tau-p1"
"d4=2*tau-p1"
"d5=(2*tau-p1-p9)/2"

```

```

define delay cycle
    "cycle=36*tau"
define loopcounter count
    "count=aq/cycle-1" ;make sure td datapoints are sampled
define delay rest
    "rest=aq-(count*cycle)"
define delay trot
    "trot=1/cnst31"
define delay trtc
    "trtc=trot/cycle"

1 ze
    1u ip4+cnst25 ;increment the phase of the pre-pulse
    1u ip10+cnst10 ;increment the BR-24 phases
    1u ip11+cnst10
    1u ip12+cnst10
    1u ip13+cnst10

2 d1 pl12:f1 ;recycle delay
    10u reset1:f1 ;synchronise pulse and detection RF
    STARTADC ;prepare adc for sampling, set
reference frequency, defined in Avancedru.incl
    RESETPHASE ;reset reference phase
    1u rpp10 ;reset phase list pointer
    (p12 pl12 ph1):f1 ;first 90 at pl12
    (p14 pl12 ph4):f1 ;pre-pulse at pl12
    .1u DWL_CLK_ON
4 dead pl13:f1
    d9 RG_ON
    .1u RG_OFF ;take l11 complex data points
    1.4u
    d5 ;BR-24 decoupling
    p1:f1 ph12
    d2
    p1:f1 ph11
    d4
    p1:f1 ph13
    d2
    p1:f1 ph10
    d4
    p1:f1 ph10
    d2
    p1:f1 ph11
    d4
    p1:f1 ph13
    d2
    p1:f1 ph12
    d4
    p1:f1 ph11
    d2
    p1:f1 ph12
    d4
    p1:f1 ph10
    d2
    p1:f1 ph13
    d4
    p1:f1 ph13
    d2
    p1:f1 ph12
    d4
    p1:f1 ph11
    d2
    p1:f1 ph12
    d4
    p1:f1 ph10
    d2
    p1:f1 ph13
    d4
    p1:f1 ph13

```

```

d2
p1:f1 ph12
d4
p1:f1 ph10
d2
p1:f1 ph11
d4
p1:f1 ph10
d2
p1:f1 ph11
d5
;end of BR-24
decoupling
  lo to 4 times count      ;make sure td points are sampled
  rest
  1u DWL_CLK_OFF
  rcyc=2
5 100m wr #0
6 exit

ph0=2
ph1=0 1 2 3
ph4= 0
ph10= 0
ph11= 1
ph12= 2
ph13= 3
ph30=1
ph31=3 0 1 2

```

Windowless Sequences Pulse Program

```

;dumbo.fmp
;
;1D CRAMPS experiment using phase-modulated sequences for homonuclear dipolar
decoupling.
;The FID is sampled in acquisition windows between homonuclear dipolar decoupling
pulses.
;This program also contains two additional compensation pulses before and after the
decoupling pulse, which have shown to improve the spectral resolution.
;This program was written based on the pulse program "dumbold.es", written by Elodie
Salager.
;
;For a detailed description on the parameters optimization of this pulse program, see:
F.M. Paruzzo and L. Emsley (2019), High-Resolution 1H NMR of Powdered Solids by
Homonuclear Dipolar Decoupling
;
;Tested on Bruker Avance III HD / TopSpin 3.5
;Last Modification: F.M. Paruzzo, 16th July 2019
;
;$CLASS=Solids
;$DIM=1D
;$TYPE=homonuclear dipolar decoupling
;$COMMENT=1D CRAMPS experiment using phase-modulated sequences
;
;-----
;Parameters:
;d1 : recycle delay
;d9 : time for acquisition of FID points during each acquisition window
;
;p11 : power level for compensation pulses
;p12 : power level for initial 90 degree and pre-pulse
;p13 : power level for decoupling pulse
;
;p8 : delay to adjust the length of the acquisition window
;p9 : length of the acquisition window
;p10: length of the decoupling pulse
;p12 : 1H 90 degree pulse length at p12

```

```

;p14: pre-pulse length (typically between 0.5 and 1.5 us)
;p21: length of the second compensation pulse
;p22: length of the first compensation pulse
;
;l11: number of data points to be detected each acquisition window, set to 2-32
;
;cnst21 : additional phase of second compensation pulse
;cnst22 : additional phase of first compensation pulse
;cnst25 : additional phase of pre-pulse
;cnst31 : spinning rate (Hz)
;
;o1 : 1H offset, best when at the right side of the resonances, close to the region to
decouple
;ns : n*4
;trtc : calculated as ratio between rotor period/cycle time
;dead : receiver dead time in s
;acqu : extra delay within the acquisition window in s
;cycle : decoupling cycle time
;count : total number of acquisition window/decoupling pulses
;spnam1 : file name for homonuclear dipolar decoupling pulse
;-----
"acqt0=0"

dwellmode auto
#include <Avancesolids.incl>
#include <Delayssolids.incl>

"d9=0.1u*(l11)" ;set the sampling window in Avancesolids.incl as l11
multiples of 100 ns
define delay dead
  "dead=1.6u" ;receiver dead time
define delay acqu
  "acqu=p8"
"p9=p8+dead+d9+0.1u+p22+p21" ;total length of the acquisition window, including
compensation pulses (us)
define delay cycle
  "cycle=p9+p10" ;cycle time
define delay trot
  "trot=1/cnst31"
define delay trtc
  "trtc=trot/cycle"
define loopcounter count
  "count=aq/(cycle)-1"
define delay rest
  "rest=aq-(count*cycle)"
"spw1=plw13"

1 ze
  ip4+cnst25 ;increment the phase of the pre-pulse
  ip22+cnst22 ;increment the phase of the first compensation pulse
  ip21+cnst21 ;increment the phase of the second compensation pulse

2 d1 pl12:f1 ;recycle delay
  10u reset1:f1 ;synchronise pulse and detection RF
  STARTADC ;prepare adc for sampling, set reference frequency,
defined in Avancedru.incl
  RESETPHASE ;reset reference phase

  p12:f1 ph1 ;first 90 at pl12
  p14:f1 ph4 ;pre-pulse at pl12
  .1u DWL_CLK_ON

3 dead
  acqu
  d9 RG_ON ;acquire l11 data points
  .1u RG_OFF
  (p22 pl1 ph22):f1 ;first compensation pulse at pl1
  (p10:sp1 ph10):f1

```



```

(p21 pl1 ph21):f1                                ;second compensation pulse at
pl1
  lo to 3 times count
  rest
  1u DWL_CLK_OFF
  rcyc=2
4 100m wr #0
5 exit

ph0= 0
ph1= 1 2 3 0
ph4= 0
ph10= 0
ph21= 3
ph22= 1
ph30= 0
ph31= 0 1 2 3

```

Chapter 3

Additional Section 3.2 Information

Linewidths Measurement

Table A.16 shows the linewidths of the resonances observed in the one-pulse, CT-MAS and TOP-CT-MAS experiments. All linewidths were measured with a home-written MATLAB script. All linewidths values are given in Hz.

Experiment		Linewidths / Hz							
		CH ₃	CH ₂ ^(a)	CH ₂ ^(b)	CH _(Asp)	CH _(Ala)	NH ₃	NH	OH
one-pulse		272	298	299	266	228	337	373	249
CT-MAS	4.6 ms	213	211	212	220	219	266	337	247
	6.8 ms	191	190	194	201	196	248	337	231
	9.2 ms	184	181		205	190	247	343	228
	18.5 ms	171				191			242
TOP-CT-MAS	4.6 ms	205	205	205	212	205	255	336	232
	9.2ms	186			203	192	245	358	228
	18.5ms	158				170			260

Table A.16. Linewidths of the resonances observed in the experiments of powdered β -AspAla shown in **Section 3.2**.

T_2' Measurement

Table A.17 shows the T_2' values obtained from the fittings shown in **Figure 3.3c**. We also report here the contributions to the total linewidth of the interactions that are refocused and not refocused by the application of the π pulse. These contributions are calculated from the T_2' values (and also shown in **Figure 3.3d**).

	CH ₃	CH ₂ ^(a)	CH ₂ ^(b)	CH _(Asp)	CH _(Ala)	NH ₃	NH	OH
T_2' / ms	2.5	3.1	2.8	3.1	4.7	2.1	3.8	5.2
Non-refocusable Broadening / Hz	127	103	112	68	104	152	84	61
Refocusable Broadening / Hz	145	195	187	198	125	185	289	188

Table A.17. T_2' values obtained by fitting the dephasing curves shown in **Figure 3.3c**.

Non-Refocusable Broadening from TOP-CT-MAS Experiments

Table A.18 shows the contributions of non-refocusable interactions to the total linewidths. These values are extracted from TOP-CT-MAS experiments, by measuring the linewidths of the resonances in the dimension orthogonal to the constant-time ^1H spectra in double-sheared TOP-CT-MAS experiment. We also report the contributions determined using T_2' measurements for comparison.

Experiment		Non-refocusable Broadening / Hz							
		CH ₃	CH ₂ ^(a)	CH ₂ ^(b)	CH _(Asp)	CH _(Ala)	NH ₃	NH	OH
T ₂ ¹ Measurements		127	103	112	68	104	152	84	61
TOP-CT-MAS	4.6 ms	176	206	220	193	157	201	176	158
	9.2ms	110			138	106	135	114	103
	18.5ms	86				91			85

Table A.18. Contribution to the non-refocusable interactions to the line broadening extracted from TOP-CT-MAS experiments.

Additional Section 3.3 Information

Apparent linewidths of thymol spectra

Resonance	Linewidth / Hz								
	echo	$\beta = 90^\circ$	$\beta = 45^\circ$	$\beta = 20^\circ$	$\beta = 15^\circ$	$\beta = 10^\circ$	$\beta = 8^\circ$	$\beta = 5^\circ$	$\beta = 3^\circ$
H1	358	396	384	324	303	260	254	230	187
H2	393	389	360	296	283	263	231	252	188
H3	399	389	367	326	302	263	232	262	223
H4	349	364	346	309	291	265	247	222	199
H5'*	672	631	617	412	512	348	478	294	230
H5''*	672	631	617	405	512	318	315	516	396
H6	448	448	422	344	329	317	295	264	225

Table A.19. Measured linewidths of thymol echo-detected spectrum and spectra obtained as integral projections from the sheared 2D ant-z-COSY spectra acquired with β of 90° , 45° , 20° , 15° , 10° , 8° , 5° , and 3° at 100.0 kHz MAS. The linewidths were extracted through gaussian deconvolution. The asterisks denote protons that overlaps in the spectra.

Apparent linewidths of β -AspAla spectra

Resonance	Linewidth / Hz								
	echo	$\beta = 90^\circ$	$\beta = 45^\circ$	$\beta = 20^\circ$	$\beta = 15^\circ$	$\beta = 10^\circ$	$\beta = 8^\circ$	$\beta = 5^\circ$	$\beta = 3^\circ$
H8	284	242	229	206	196	193	185	179	155
H5	473	379	364	329	285	272	258	233	195
H4	445	342	314	272	265	248	230	228	205
H7	284	240	223	195	187	174	172	155	152
H3	342	264	243	212	205	191	181	178	153
H1	416	316	312	235	281	189	162	142	?
H2	409	310	295	305	220	187	169	175	202
H6	337	256	232	201	195	187	175	180.7	158

Table A.20. Measured linewidths of β -AspAla echo-detected spectrum and spectra obtained as integral projections from the sheared 2D anti-z-COSY spectra acquired with β of 90° , 45° , 20° , 15° , 10° , 8° , 5° , and 3° at 100.0 kHz MAS. The linewidths were extracted through gaussian deconvolution.

Anti-z-COSY at 62.5 kHz MAS

Resonances	Linewidths / Hz				
	echo	$\beta = 90^\circ$ rot. sync.	$\beta = 5^\circ$ rot. sync.	$\beta = 90^\circ$ rot. async.	$\beta = 5^\circ$ rot. async.
H8	281	290	280	299	237
H5	526	448	434	531	304
H4	477	436	434	429	269
H7	284	265	265	274	185
H3	423	363	293	353	259
H1	515	390	274	426	216
H2	543	372	375	366	291
H6	346	312	301	322	270

Table A.21. Measured linewidths of β -AspAla echo-detected spectrum and spectra obtained as integral projections from the sheared 2D anti-z-COSY spectra acquired with β of 90° and 5° at 62.5 kHz MAS. The linewidths were extracted through gaussian deconvolution.

Chapter 4

NMR Crystallography.

Experimental chemical shifts were referenced to the ^1H resonance observed for adamantane at 1.87 ppm with respect to TMS. We used assigned chemical shifts values and we account for rotational dynamics of the methyl groups by averaging the chemical shift values of the three ^1H positions to a single value for each methyl group. For AZD8329 the chemical shifts of the CH_2 groups were also averaged. The RMSE calculation was carried out in MATLAB using a home-written script. The chemical structures of cocaine and AZD8329, together with the assignment of the experimental chemical shifts are shown in **Figure A.1** and **Table A.22**. ^1H chemical shieldings calculated with GIPAW and ShiftML are given in <https://doi.org/10.1038/s41467-018-06972-x>.

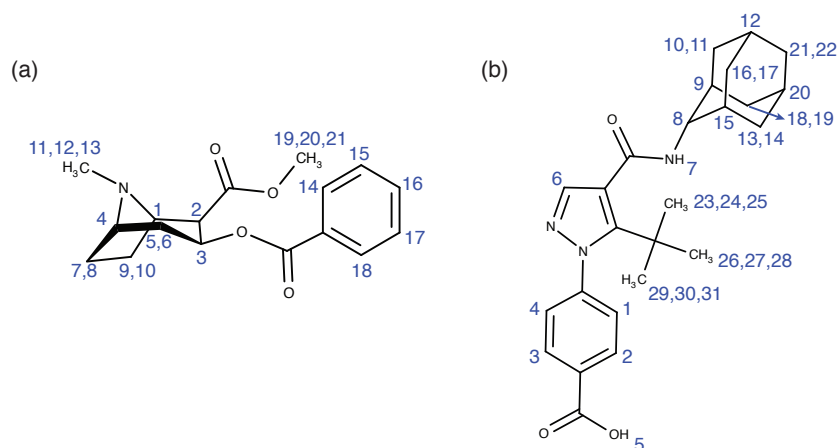


Figure A.1. Chemical structure of cocaine (a) and AZD8329 (b) and the labelling scheme used here

Cocaine		AZD8329	
Atom Label	^1H δ (ppm)	Atom Label	^1H δ (ppm)
1	3.76	1	6.92
2	3.78	2	8.69
3	5.63	3	9.01
4	3.32	4	8.47
5	3.49	5	15.37

6	3.06	6	7.73
7	2.91	7	9.64
8	3.38	8	2.90
9	2.56	9	1.78
10	2.12	10,11	1.88
11,12,13	1.04	12	1.8
14	8.01	13	1.6
15	8.01	14	0.44
16	8.01	15	1.54
17	8.01	16,17	1.88
18	8.01	18,19	0.8
19,20,21	3.78	20	1
		21,22	1.74
		23,24,25, 26,27,28, 29,30,31	0.73

Table A.22. Experimental chemical shifts of cocaine and the form 4 of AZD8329. The labelling scheme is given in **Figure A.1**. When more than one atom corresponds to a single chemical shift value, their values were average

Prediction Parameters

Table A.23 and **Table A.24** show the parameters used for the single and the multi-scale kernel predictions respectively.

Atom	Cut-off (r_c)	Gaussian width (ζ)	l_{\max}	n_{\max}	σ_n	ζ
^1H	2	0.3	9	9	0.1	2
	3	0.3	9	9	0.1	2
	4	0.4	9	9	0.1	2
	5	0.4	9	9	0.1	2
	6	0.5	9	12	0.1	2
	7	0.5	9	12	0.1	2
^{13}C	2	0.3	9	9	0.01	2
	3	0.3	9	9	3.0	2
	4	0.4	9	9	5.0	2
	5	0.4	9	9	3.0	2
	6	0.5	9	12	1.0	2
	7	0.5	9	12	1.0	1
^{15}N	2	0.3	9	9	0.5	2
	3	0.3	9	9	1.0	2
	4	0.4	9	9	0.1	2
	5	0.4	9	9	0.1	2
	6	0.5	9	12	0.1	2
	7	0.5	9	12	0.05	2
^{17}O	2	0.3	9	9	0.5	2
	3	0.3	9	9	5.0	2
	4	0.4	9	9	5.0	2

5	0.4	9	9	5.0	2
6	0.5	9	12	1.0	2
7	0.5	9	12	7.0	2

Table A.23. Kernel and GPR parameters. The GPR parameters (σ_n and ζ) are the ones used in single kernel predictions.

Atom	Multi-Scale Kernel Weights						σ_n	ζ
	$r_c = 2 \text{ \AA}$	$r_c = 3 \text{ \AA}$	$r_c = 4 \text{ \AA}$	$r_c = 5 \text{ \AA}$	$r_c = 6 \text{ \AA}$	$r_c = 7 \text{ \AA}$		
^1H	256	128	32	8	8	1	0.1	2
^{13}C	256	512	64	8	8	1	2.0	2
^{15}N	256	128	32	8	8	1	0.1	2
^{17}O	256	128	32	8	8	1	5.0	2

Table A.24. Kernel weights and GPR parameters used for multi-scale kernel prediction.

Comparison to Experiments

Comparison between ^1H experimental chemical shifts and ^1H chemical shifts calculated with ShiftML were carried out analysing 68 chemical shifts obtained from 6 crystal structures. The names, IUPAC IDs, CSD reference codes (when available) and references to the experimental NMR data of the analysed crystal structures are the following:

- Naproxen, (2S)-2-(6-Methoxy-2-naphthyl)propanoic acid, COYRUD11, Ref. 349
- Uracil, Pyrimidine-2,4(1H,3H)-dione, URACIL, Ref. 350
- Co-crystal of 3,5-dimethylimidazole and 4,5-dimethylimidazole, Ref. 142
- Theophylline, 1,3-Dimethyl-3,7-dihydro-1H-purine-2,6-dione, BAPLOT01, Ref. 120
- Cocaine, methyl (1R,2R,3S,5S)-3- (benzoxyloxy)-8-methyl-8-azabicyclo[3.2.1] octane-2-carboxylate, COCAIN10, Ref. 120
- AZD8329, 4-[4-(2-adamantylcarbamoyl)-5-tert-butylpyrazol-1-yl]benzoic acid, Ref. 121

The crystal structures (i-iv) were obtained from Ref. 294, where the experimentally determined crystal structures were subjected to all-atom geometry optimization with fixed lattice parameters, as described in the reference. Crystal structures (v) and (vi) were obtained from Refs. 120 and 121 respectively.

We used assigned chemical shift values and we account for rotational dynamics of the methyl groups by averaging the chemical shift values of the three ^1H positions to a single value for each methyl group. The calculated chemical shieldings σ are converted to the corresponding chemical shifts δ through the relationship $\delta = \sigma_{ref} - \beta\sigma$. For each structure, we calculated the value of σ_{ref} and β by a linear regression between calculated and experimental shifts. The calculations were carried out in MATLAB using a home-written script. The chemical structures, together with the assigned experimental chemical shifts and the parameters for conversion between shieldings and shifts are shown in **Figure A.2** and **Table A.25**.

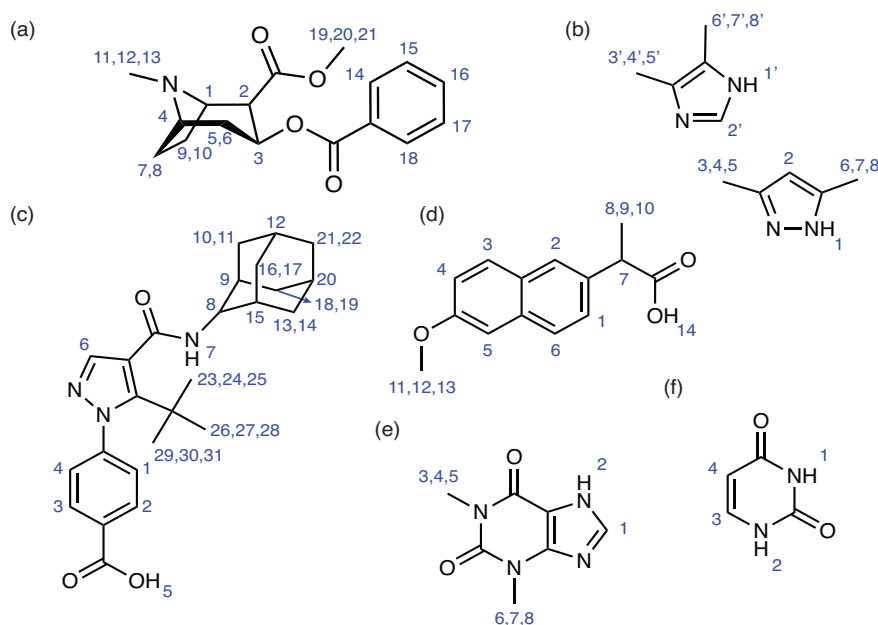


Figure A.2. Chemical structures of the compounds used for experimental comparison. In order, cocaine (a), 3,5-dimethylimidazole and 4,5-dimethylimidazole (b), AZD8329 (c), naproxen (d), theophylline (e) and uracil (f), and the labelling scheme used here.

Naproxen			Uracil		
Atom Label	Experimental ^1H δ (ppm)	ShiftML ^1H δ (ppm)	Atom Label	Experimental ^1H δ (ppm)	ShiftML ^1H δ (ppm)
1	7	6.87	3	7.5	7.76
2	6.1	6.07	2	10.8	10.68
3	3.8	3.74	1	11.2	11.22
4	4.5	4.40	4	6	5.85
5	4.1	4.51			
6	5.9	5.11			
7	3.2	3.15			
8,9,10	1.8	1.98			
11,12,13	2.3	2.63			
14	11.5	11.74			
σ_{ref}	25.38	β 0.81	σ_{ref}	23.71	β 0.74
3,5-dimethylimidazole & 4,5-dimethylimidazole			Theophylline		
Atom Label	Experimental ^1H δ (ppm)	ShiftML ^1H δ (ppm)	Atom Label	Experimental ^1H δ (ppm)	ShiftML ^1H δ (ppm)
2'	4.8	5.17	2	14.6	14.57
6',7',8'	0.7	0.77	1	7.7	7.27
3',4',5'	1.4	0.91	3,4,5	3.4	3.22
1'	13	12.55	6,7,8	3.4	3.52
6',7',8'	1.4	1.20			
3',4',5'	1.5	1.35			
1'	15	14.92			
2'	5.2	6.14			
σ_{ref}	29.91	β 0.99	σ_{ref}	25.98	β 0.83
Cocaine			AZD8329		
Atom Label	Experimental ^1H δ (ppm)	ShiftML ^1H δ (ppm)	Atom Label	Experimental ^1H δ (ppm)	ShiftML ^1H δ (ppm)
1	3.76	3.95	1	6.92	6.53
2	3.78	3.22	2	8.69	7.85
3	5.63	6.11	3	9.01	9.35

4	3.32	3.73	4	8.47	7.91
5	3.06	2.55	5	15.37	15.95
6	3.49	2.99	6	7.73	7.60
7	2.91	2.69	7	9.64	9.37
8	3.38	3.18	8	2.90	2.79
9	2.56	2.44	9	1.78	1.98
10	2.12	2.37	10	1.88	1.79
11,12,13	1.04	1.80	11	1.88	2.61
14	8.01	8.40	12	1.8	1.68
16	8.01	7.39	13	1.6	1.28
15	8.01	7.66	14	0.44	0.87
17	8.01	8.09	15	1.54	1.94
18	8.01	8.03	16	1.88	2.76
19,20,21	3.78	4.28	17	1.88	1.69
			18	0.8	1.21
			19	0.8	0.43
			20	1	1.42
			21	1.74	1.47
			22	1.74	1.21
			23,24,25	0.73	0.84
			26,27,28	0.73	1.02
			29,30,31	0.73	0.14
σ_{ref}	30.04	β 0.96	σ_{ref}	28.39	β 0.91

Table A.25. Experimental and calculated chemical shifts of naproxen, uracil, the co-crystal of 3,5-dimethylimidazole and 4,5-dimethylimidazole, theophylline, cocaine and AZD8329. The labelling scheme is given in **Figure A.2**. When more than one atom corresponds to a single chemical shift value, their values were averaged

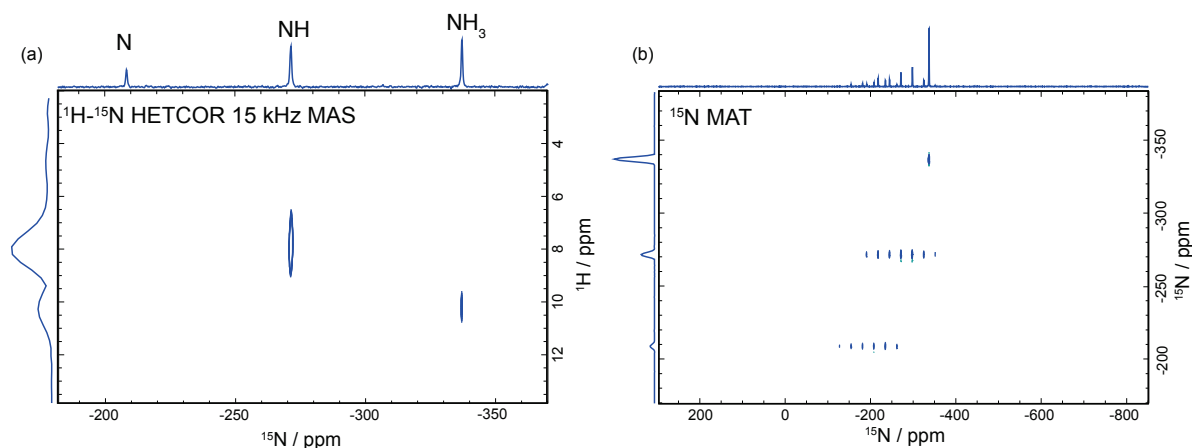
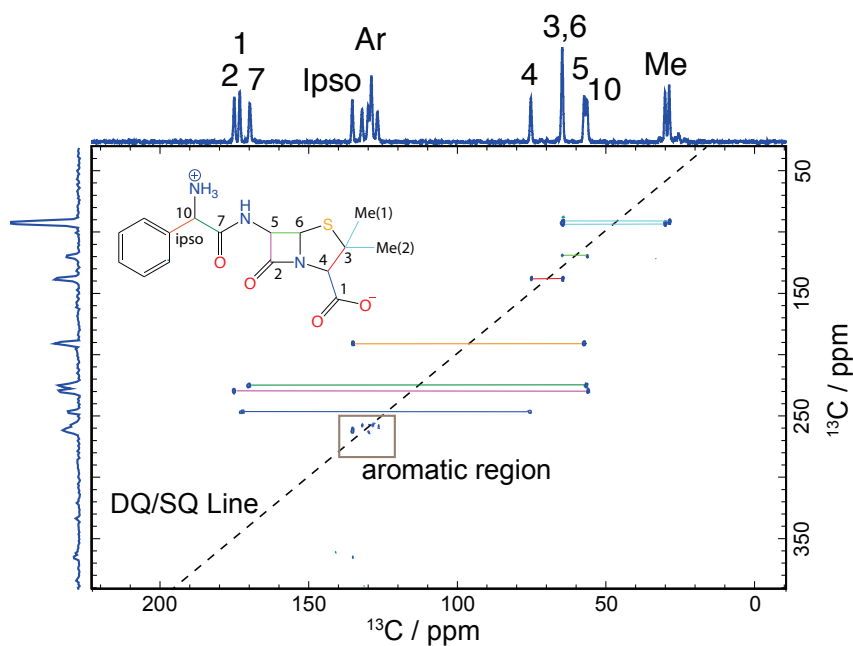
Chapter 5

Additional Section 5.2 Information

INADEQUATE and MAT Spectra of Ampicillin and Assigned Chemical Shifts

Label	^1H , ppm	^{13}C , ppm	^{15}N , ppm
Me ₁	0.6	30.1	-
Me ₂	1.6	28.9	-
4	4.0	75.3	-
10	4.8	57.4	-
6	5.2	64.8	-
Ar	5.4	128.3	-
5	6.6	56.5	-
Ar	7.1	129.0	-
Ar	7.2	132.0	-
Ar	7.3	129.9	-
Ar	7.6	126.9	-
N	-	-	Around -210
NH	7.5	-	Around -270
NH ₃	10	-	Around -340
3	-	64.8	
Ar(ipso)	-	135.4	
7	-	169.8	

1	-	173.2
2	-	175.0

Table A.26. Ampicillin experimental chemical shifts.**Figure A.3.** ^{15}N spectra of ampicillin used for the ^1H and ^{15}N assignments. (a) ^1H - ^{15}N HETCOR spectra of ampicillin measured at 16.4 T and 15.0 kHz MAS. (b) ^{15}N MAT spectra of ampicillin at 16.4 T.**Figure A.4.** ^{13}C - ^{13}C INADEQUATE spectra of ampicillin used for the ^{13}C assignments, measured at 11.7 T and 13.0 kHz MAS.**ShiftML Error Estimation.**

Comparison between ^1H experimental chemical shifts and ^1H chemical shifts calculated with ShiftML were carried out analysing around 150 chemical shifts obtained from 11 crystal structures. In addition to the structures given in the “Comparison to Experiment” paragraph of **Section 4.5**, the following structures were used (for which we list names, IUPAC IDs, CSD reference codes when available and references to the experimental NMR data of the analysed crystal structures):

- Anthranilic acid, 2-aminobenzoic acid, AMBACO05, Refs. 294, 351
- Cimetidine, 1-cyano-2-methyl-3-[2-[(5-methyl-1*H*-imidazol-4-yl)methylsulfanyl]ethyl]guanidine, CIMETD, Refs. 294, 352
- Phenobarbital, 5-Ethyl-5-phenyl-1,3-diazinane-2,4,6-trione, PHBARB06, Refs. 294, 353

- Thymol, 5-Methyl-2-(propan-2-yl)phenol, IPMEPL, Ref. 116
- Terbutaline hemisulfate, 2-t-Butylamino-1-(3,5-dihydroxyphenyl)ethanol, ZIVKAQ, Refs. 112, 294

These crystal structures were obtained from Ref. 294, where the experimentally determined crystal structures were subjected to all-atom geometry optimization with fixed lattice parameters, as described in the reference. We only used the ^1H chemical shifts from the references, which were clearly distinguishable and did not have a broad peak spanning several ppm.

We used assigned chemical shift values and we account for rotational dynamics of the methyl groups by averaging the chemical shift values of the three ^1H positions to a single value for each methyl group. For chemical shifts which could not be assigned unambiguously, such as e.g. shifts from CH_2 protons, we assigned the chemical shifts on a best match basis. The calculated chemical shieldings σ are converted to the corresponding chemical shifts δ through the relationship $\delta = \sigma_{\text{ref}} - \beta\sigma$, where the slope (β) and the offset (σ_{ref}) were fit for each reference structure individually. The chemical structures (of the whole error estimation set), the RMSE between experimental and ShiftML predicted ^1H chemical shifts, together with the assigned experimental chemical shifts and the parameters for conversion between shieldings and shifts are shown in **Figure A.5** and **Table A.27**. For the entire reference set we calculate an average RMSE of 0.346 ppm and a standard deviation of 0.195 ppm.

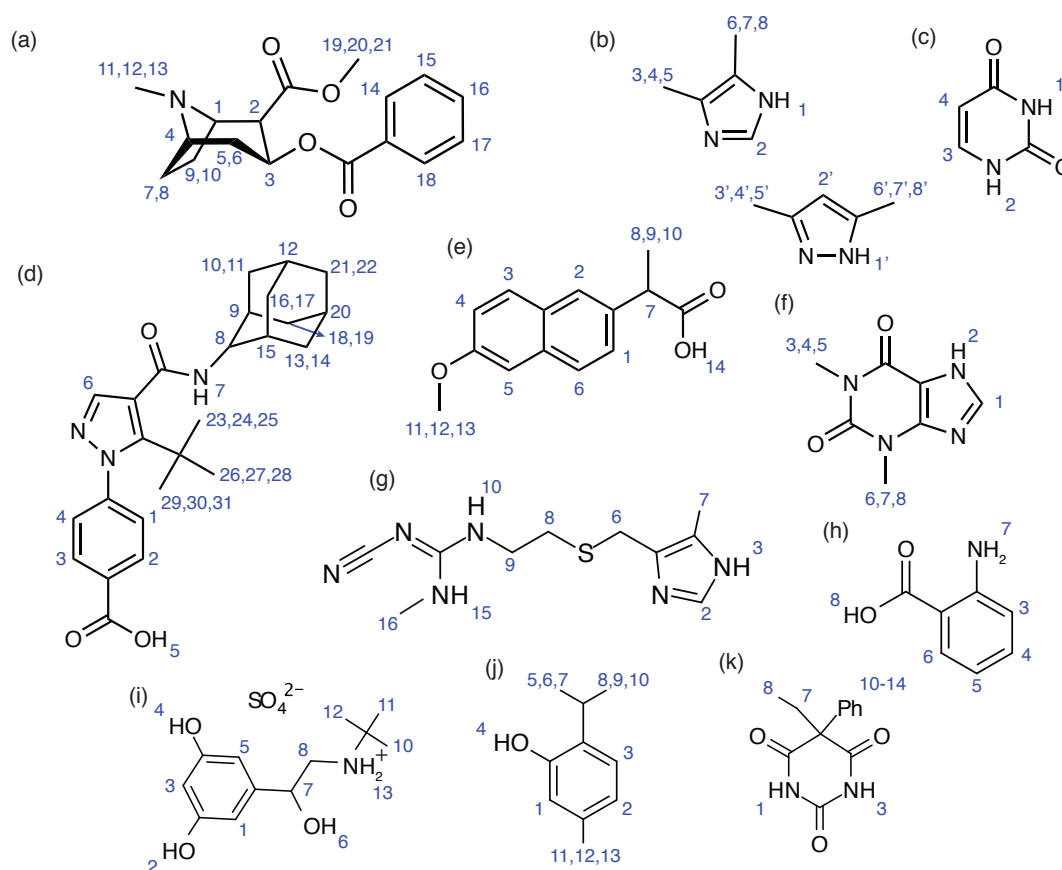


Figure A.5. Chemical structures of the compounds used for the error estimation. In order, cocaine (a), 3,5-dimethylimidazole and 4,5-dimethylimidazole (b), uracil (c), AZD8329 (d), naproxen (e), theophylline (f), cimetidine (g), anthranilic acid (h), terbutaline hemisulfate (i), thymol (j) and phenobarbital (k) and the labelling scheme used here.

Cimetidine			Anthranilic acid		
Atom Label	Experimental ^1H δ (ppm)	ShiftML ^1H δ (ppm)	Atom Label	Experimental ^1H δ (ppm)	ShiftML ^1H δ (ppm)
2	7.64	7.55	Aromatic (1)	5.8	5.74
3	11.84	11.55	Aromatic (2)	6.8	6.66

7	2.24	2.17	NH2	5.4	5.52
10	8.44	9.00	COOH	12.3	12.33
15	9.94	9.86			
16	2.24	2.28			
$\sigma_{ref} = 5.12$ ppm $\beta = 0.86$ RMSE = 0.21 ppm			$\sigma_{ref} = 4.95$ ppm $\beta = 0.79$ RMSE = 0.095 ppm		
Phenobarbital			Thymol		
Atom Label	Experimental ^1H δ (ppm)	ShiftML ^1H δ (ppm)	Atom Label	Experimental ^1H δ (ppm)	ShiftML ^1H δ (ppm)
1	10.3	10.49	1	5.4	5.80
3	8.1	8.34	2	6.19	5.90
7a	2.7	2.69	3	7.08	6.35
7b	1.7	1.63	4	3.38	2.91
8a-c	0.6	0.78	5-7	1.05	0.44
9-14	6.9	6.60	8-10	1.45	1.14
			11-13	0.42	1.68
			14	9.99	10.08
$\sigma_{ref} = 5.08$ ppm $\beta = 0.78$ RMSE = 0.33 ppm			$\sigma_{ref} = 4.93$ ppm $\beta = 0.85$ RMSE = 0.72 ppm		
Terbutaline hemisulfate					
Atom Label	Experimental ^1H δ (ppm)	ShiftML ^1H δ (ppm)			
1	6.83	7.60			
3	6.83	6.50			
4	10.93	10.07			
5	6.83	6.96			
7	4.73	5.26			
10-12	1.33	1.25			
13	7.6	8.22			
$\sigma_{ref} = 5.25$ ppm $\beta = 1.00$ RMSE = 0.44 ppm					

Table A.27. Experimental and calculated chemical shifts of the structures used in the ShiftML benchmarking. The experimental and calculated chemical shifts of the structures that are missing from this table are given in **Table A.25**. The labelling scheme is given in **Figure A.5**. When more than one atom corresponds to a single chemical shift value, their values were averaged.

Ampicillin lattice parameters

A comparison between lattice parameters of the ampicillin crystal structure, as primitive cell, determined with XRD³³⁷ and with NMRX are given in the following table.

Parameter	XRD	NMRX	deviation (%)
a [Å]	12.4	11.7	-5.6
b [Å]	6.2	5.78	-6.8
c [Å]	12.0	12.63	+5.25
α	90.0	90.0	0.0
β	114.5	114.506	<0.1
γ	90.0	90.0	0.0
A [Å]	839.494	777.213	-7.4

Table A.28. Comparison between ampicillin lattice parameter of the crystal structures, as primitive cell, determined with XRD³³⁷ and NMRX.

Additional Section 5.3 Information

Experimental Chemical Shifts

For ampicillin the fully assigned experimental ^1H and ^{13}C chemical shifts with the corresponding labels are given above. For cocaine, flufenamic acid, flutamide, theophylline and AZD8329 the fully assigned experimental shifts were taken from Refs. 121, 322 and for the convenience of the reader are given in the following tables. The labelling scheme is shown in **Figure A.6**.

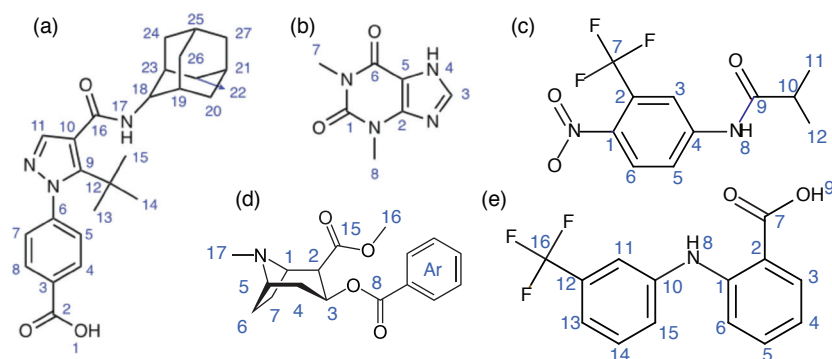


Figure A.6. Chemical structures of theophylline (a) and AZD8329 (b). The distinct ^1H and ^{13}C sites are labeled.

Label	^1H / ppm	^{13}C / ppm
1	15.37	–
2	–	171.04
3	–	131.19
4	8.69	130.48 or 128.05
5	6.92	128.05 or 130.48
6	–	147.31
7	8.47	128.05 or 130.48
8	9.01	130.48 or 128.05
9	–	148.71
10	–	114.10
11	7.73	138.43
12	–	33.42
13	0.73	29.53
14	0.73	29.53
15	0.73	29.53
16	–	172.98
17	9.64	
18	2.90	60.16
19	1.54	34.14
20	0.44 or 1.6	30.80 or 37.41
21	1.00	27.81
22	0.80	36.42 or 30.80
23	1.78	32.45
24	1.88	30.90 or 36.42
25	–	27.81
26	1.88	37.41 or 30.80
27	1.74	37.41

Table A.29. AZD8329 experimental chemical shifts.

Label	¹ H / ppm	¹³ C / ppm
1	–	150.8
2	–	146.1
3	7.7	140.8
4	14.6	–
5	–	105.8
6	–	155.0
7	3.4	29.9
8	3.4	29.9

Table A.30. Theophylline experimental chemical shifts.

Label	¹ H / ppm	¹³ C / ppm
1	3.5	66.0
2	3.5	50.2
3	5.5	66.7
4	3.3	36.7
5	3.4	62.6
6	3.4	25.6
7	2.4	25.6
8	-	165.9
Ar ()	7.8	129.4
Ar (ipso)	-	134.5
15	-	172.2
16	3.5	50.2
17	1.2	41.52

Table A.31. Cocaine experimental chemical shifts.

Label	¹ H / ppm	¹³ C / ppm
1	-	149.3
2	-	109.7
3	8.3	133.0
4	6.0	117.2
5	5.4	136.3
6	6.8	112.0
7	-	175.0
8	9.6	-
9	-6.6	-
10	-	139.9
11	6.9	121.7
12	-	131.7
13	6.2	119.8
14	5.9	129.5
15	7.3	128.1
16	-	124.1

Table A.32. Flufenamic acid experimental chemical shifts.

Label	¹ H / ppm	¹³ C / ppm
1	-	145.4
2	-	124.5
3	7.9	130.9

4	-	140.9
5	9.9	124.5
6	7.1	116.7
7	-	122.0
8	8.0	-
9	-	176.1
10	2.3	35.7
11	1.3	17.7
12	1.3	21.7

Table A.33. Flutamide experimental chemical shifts.

References

- 1 Lunier, M. *Dictionnaire des sciences et des arts contenant*. Vol. 1 (Normant, 1806).
- 2 de L'Isle, J.-B. R. *Cristallographie*. (Paris, 1783).
- 3 IUCr. *Online Dictionary of Crystallography*, 8th February 2019, 09:12 UTC, <http://reference.iucr.org/dictionary/Main_Page> (accessed 9th September 2019).
- 4 Haüy, M. I. A. Essai d'une théorie sur la structure de cristaux appliquée a plusieurs genres de substances cristallisées. *Chez Gogue et Neé de la Rochelle Libraires, Paris* (1784).
- 5 Friedrich, W., Knipping, P. & Laue, M. Interferenzerscheinungen bei röntgenstrahlen. *Ann. Phys.* **346**, 971-988 (1913).
- 6 Bragg, W. H. & Bragg, W. L. The reflection of X-rays by crystals. *Proc. Royal Soc. Lond., Series A, Containing Papers of a Mathematical and Physical Character* **88**, 428-438 (1913).
- 7 Bragg, W. H. & Bragg, W. L. The structure of the diamond. *Proc. Royal Soc. Lond., Series A, Containing Papers of a Mathematical and Physical Character* **89**, 277-291 (1913).
- 8 Wollan, E. & Borst, L. Physics section III monthly report for the period ending December 31, 1944. *Report No.: M-CP-2222, Clinton Laboratories, Oak Ridge, TN, USA* (1945).
- 9 Shull, C. G. & Wollan, E. O. X-Ray, Electron, and Neutron Diffraction. *Science* **108**, 69-75 (1948).
- 10 Gorelik, T. E., Czech, C., Hammer, S. M. & Schmidt, M. U. Crystal structure of disordered nanocrystalline α l-quinacridone determined by electron diffraction. *Crystengcomm* **18**, 529-535 (2016).
- 11 Das, P. P. *et al.* Crystal Structures of Two Important Pharmaceuticals Solved by 3D Precession Electron Diffraction Tomography. *Org. Process. Res. Dev.* **22**, 1365-1372 (2018).
- 12 Gruene, T. *et al.* Rapid Structure Determination of Microcrystalline Molecular Compounds Using Electron Diffraction. *Angew. Chem. Int. Edit.* **57**, 16313-16317 (2018).
- 13 Jones, C. G. *et al.* The CryoEM Method MicroED as a Powerful Tool for Small Molecule Structure Determination. *ACS Cent. Sci.* **4**, 1587-1592 (2018).
- 14 Rabi, I. I., Zacharias, J. R., Millman, S. & Kusch, P. A new method of measuring nuclear magnetic moment. *Phys. Rev.* **53**, 318-318 (1938).
- 15 Bloch, F. Nuclear induction. *Phys. Rev.* **70**, 460 (1946).
- 16 Purcell, E. M., Torrey, H. C. & Pound, R. V. Resonance absorption by nuclear magnetic moments in a solid. *Phys. Rev.* **69**, 37 (1946).
- 17 Harris, K. D. & Cheung, E. Y. How to determine structures when single crystals cannot be grown: opportunities for structure determination of molecular materials using powder diffraction data. *Chem. Soc. Rev.* **33**, 526-538 (2004).
- 18 Hofstetter, A. & Emsley, L. Positional Variance in NMR Crystallography. *J. Am. Chem. Soc.* **139**, 2573-2576 (2017).
- 19 Martineau, C., Senker, J. & Taulelle, F. in *Annual Reports on Nmr Spectroscopy, Vol 82* Vol. 82 *Annu. Repo. NMR Spectro.* (ed G. A. Webb) 1-57 (2014).
- 20 Abragam, A. & Abragam, A. *The principles of nuclear magnetism*. (Oxford university press, 1961).
- 21 Duer, M. J. Solid state NMR spectroscopy. *Blackwell Science Ltd., London, DOI* **10**, 9780470999394 (2002).
- 22 Mehring, M. *High resolution NMR spectroscopy in solids*. Vol. 11 (Springer Science & Business Media, 2012).
- 23 Haeberlen, U. *High Resolution NMR in Solids Selective Averaging: Supplement 1 Advances in Magnetic Resonance*. Vol. 1 (Elsevier, 2012).
- 24 Karplus, M. Contact electron-spin coupling of nuclear magnetic moments. *J. Chem. Phys.* **30**, 11-15 (1959).

- 25 Viger-Gravel, J., Meyer, J. E., Korobkov, I. & Bryce, D. L. Probing halogen bonds with solid-state NMR spectroscopy: observation and interpretation of J (77 Se, 31 P) coupling in halogen-bonded P [double bond, length as m-dash] Se... I motifs. *Crystengcomm* **16**, 7285-7297 (2014).
- 26 Massiot, D., Fayon, F., Alonso, B., Trebosc, J. & Amoureux, J. P. Chemical bonding differences evidenced from J-coupling in solid state NMR experiments involving quadrupolar nuclei. *J. Magn. Reson.* **164**, 160-164 (2003).
- 27 Benedict, H. *et al.* Nuclear scalar spin-spin couplings and geometries of hydrogen bonds. *J. Am. Chem. Soc.* **122**, 1979-1988 (2000).
- 28 Brown, S. P., Perez-Torralba, M., Sanz, D., Claramunt, R. M. & Emsley, L. Determining hydrogen-bond strengths in the solid state by NMR: the quantitative measurement of homonuclear J couplings. *Chem. Commun.*, 1852-1853 (2002).
- 29 Del Bene, J. E., Alkorta, I. & Elguero, J. Spin-spin coupling across intermolecular F-Cl...N halogen bonds. *J. Phys. Chem. A* **112**, 7925-7929 (2008).
- 30 Plevin, M. J., Bryce, D. L. & Boisbouvier, J. Direct detection of CH/ π interactions in proteins. *Nat. Chem.* **2**, 466-471 (2010).
- 31 Pake, G. E. Nuclear Resonance Absorption in Hydrated Crystals - Fine Structure of the Proton Line. *J. Chem. Phys.* **16** (1948).
- 32 Harris, R. K. NMR Crystallography. *IUCr Newsletter* **13** (2005).
- 33 Mueller, L. J. & Dunn, M. F. NMR crystallography of enzyme active sites: probing chemically detailed, three-dimensional structure in tryptophan synthase. *Acc. Chem. Res.* **46**, 2008-2017 (2013).
- 34 Bryce, D. L. NMR crystallography: structure and properties of materials from solid-state nuclear magnetic resonance observables. *IUCrJ* **4**, 350-359 (2017).
- 35 Martineau-Corcus, C. NMR Crystallography: A tool for the characterization of microporous hybrid solids. *Curr. Opin. Colloid Interface Sci.* **33**, 35-43 (2018).
- 36 Franks, W. T., Linden, A. H., Kunert, B., van Rossum, B. J. & Oschkinat, H. Solid-state magic-angle spinning NMR of membrane proteins and protein-ligand interactions. *Eur. J. Cell. Biol.* **91**, 340-348 (2012).
- 37 Hamilton, J. A. NMR reveals molecular interactions and dynamics of fatty acid binding to albumin. *Biochim. Biophys. Acta* **1830**, 5418-5426 (2013).
- 38 Radoicic, J., Lu, G. J. & Opella, S. J. NMR structures of membrane proteins in phospholipid bilayers. *Q. Rev. Biophys.* **47**, 249-283 (2014).
- 39 Dudek, M. K., Kazmierski, S., Kostrzewa, M. & Potrzebowski, M. J. in *Annual Reports on Nmr Spectroscopy, Vol 95* Vol. 95 *Annu. Repo. NMR Spectro.* (ed G. A. Webb) 1-81 (2018).
- 40 Urbanova, M., Czernek, J. & Brus, J. Structural Characterization of the New Generation of Drugs and Biomaterials: Application of NMR Crystallography. *Chem. Listy* **112**, 522-530 (2018).
- 41 Zhao, L., Pinon, A. C., Emsley, L. & Rossini, A. J. DNP-enhanced solid-state NMR spectroscopy of active pharmaceutical ingredients. *Magn. Reson. Chem.* **56**, 583-609 (2018).
- 42 Bryce, D. L. New frontiers for solid-state NMR across the periodic table: a snapshot of modern techniques and instrumentation. *Dalton Trans.* **48**, 8014-8020 (2019).
- 43 Harris, R. K. NMR studies of organic polymorphs and solvates. *Analyst* **131**, 351-373 (2006).
- 44 Lesage, A. Recent advances in solid-state NMR spectroscopy of spin I=1/2 nuclei. *Phys. Chem. Chem. Phys.* **11**, 6876-6891 (2009).
- 45 Carlomagno, T. NMR in natural products: understanding conformation, configuration and receptor interactions. *Nat. Prod. Rep.* **29**, 536-554 (2012).
- 46 Koller, H. & Weiss, M. in *Solid State Nmr* Vol. 306 *Topics in Current Chemistry* (ed J. C. C. Chan) 189-227 (2012).
- 47 Harris, R. K., Wasylishen, R. E. & Duer, M. J. *NMR crystallography*. Vol. 4 (John Wiley & Sons, 2009).
- 48 Ashbrook, S. E., Griffin, J. M. & Johnston, K. E. in *Annual Review of Analytical Chemistry, Vol 11* Vol. 11 *Annu. Rev. Anal. Chem.* (eds P. W. Bohn & J. E. Pemberton) 485-508 (2018).
- 49 Harris, R. K. Applications of solid-state NMR to pharmaceutical polymorphism and related matters. *J. Pharm. Pharmacol.* **59**, 225-239 (2007).
- 50 Zhou, D. H. *et al.* Solid-state NMR analysis of membrane proteins and protein aggregates by proton detected spectroscopy. *J. Biomol. Nmr* **54**, 291-305 (2012).
- 51 Andronesi, O. C. *et al.* Characterization of Alzheimer's-like paired helical filaments from the core domain of tau protein using solid-state NMR spectroscopy. *J. Am. Chem. Soc.* **130**, 5922-5928 (2008).
- 52 Li, S., Zhang, Y. & Hong, M. 3D ^{13}C - ^{13}C - ^{13}C correlation NMR for de novo distance determination of solid proteins and application to a human α -defensin. *J. Magn. Res.* **202**, 203-210 (2010).
- 53 Agarwal, V. *et al.* De Novo 3D Structure Determination from Sub-milligram Protein Samples by Solid-State 100 kHz MAS NMR Spectroscopy. *Angew. Chem. Int. Ed.* **53**, 12253-12256 (2014).

- 54 Rienstra, C. M. *et al.* De novo determination of peptide structure with solid-state magic-angle spinning NMR spectroscopy. *P. Natl. Acad. Sci. USA* **99**, 10260-10265 (2002).
- 55 Habenstein, B. *et al.* Extensive de novo solid-state NMR assignments of the 33 kDa C-terminal domain of the Ure2 prion. *J. Biomol. Nmr* **51**, 235-243 (2011).
- 56 Im, W. & Brooks, C. L., 3rd. De novo folding of membrane proteins: an exploration of the structure and NMR properties of the fd coat protein. *J. Mol. Biol.* **337**, 513-519 (2004).
- 57 Huber, M. *et al.* A Proton-Detected 4D Solid-State NMR Experiment for Protein Structure Determination. *ChemPhysChem* **12**, 915-918 (2011).
- 58 Pintacuda, G., John, M., Su, X. C. & Otting, G. NMR structure determination of protein-ligand complexes by lanthanide labeling. *Acc. Chem. Res.* **40**, 206-212 (2007).
- 59 Xu, J. *et al.* Identifying the critical role of Li substitution in $P_2-Na_x[Li_yNi_zMn_{1-y-z}]O_2$ ($0 < x, y, z < 1$) intercalation cathode materials for high-energy Na-ion batteries. *Chem. Mater.* **26**, 1260-1269 (2014).
- 60 Castellani, F. *et al.* Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. *Nature* **420**, 98-102 (2002).
- 61 Lange, S. *et al.* Structural analysis of a signal peptide inside the ribosome tunnel by DNP MAS NMR. *Sci. Adv.* **2**, e1600379 (2016).
- 62 Retel, J. S. *et al.* Structure of outer membrane protein G in lipid bilayers. *Nat. Commun.* **8**, 2073 (2017).
- 63 Han, Y. *et al.* Solid-state NMR studies of HIV-1 capsid protein assemblies. *J. Am. Chem. Soc.* **132**, 1976-1987 (2010).
- 64 McDermott, A. & Polenova, T. Solid state NMR: new tools for insight into enzyme function. *Curr. Opin. Struct. Biol.* **17**, 617-622 (2007).
- 65 Wong, Y. T. A. & Bryce, D. L. in *Annual Reports on Nmr Spectroscopy, Vol 93* Vol. 93 *Annu. Rep. NMR Spectro.* (ed G. A. Webb) 213-279 (2018).
- 66 Fernandes, A. *et al.* Phase Composition and Disorder in $La_2(Sn,Ti)_2O_7$ Ceramics: New Insights from NMR Crystallography. *J. Phys. Chem. C Nanomater Interfaces* **120**, 20288-20296 (2016).
- 67 Broom, L. K. *et al.* A gel aging effect in the synthesis of open-framework gallium phosphates: structure solution and solid-state NMR of a large-pore, open-framework material. *Dalton Trans.* **46**, 16895-16904 (2017).
- 68 Dawson, D. M. *et al.* A Multinuclear NMR Study of Six Forms of AlPO-34: Structure and Motional Broadening. *J. Phys. Chem. C* **121**, 1781-1793 (2017).
- 69 Dawson, D. M., Moran, R. F., Sneddon, S. & Ashbrook, S. E. Is the ^{31}P chemical shift anisotropy of aluminophosphates a useful parameter for NMR crystallography? *Magn. Reson. Chem.* **57**, 176-190 (2019).
- 70 McKay, D. *et al.* A Picture of Disorder in Hydrous Wadsleyite-Under the Combined Microscope of Solid-State NMR Spectroscopy and Ab Initio Random Structure Searching. *J. Am. Chem. Soc.* **141**, 3024-3036 (2019).
- 71 Mafra, L., Vidal-Moya, J. A. & Blasco, T. in *Annual Reports on Nmr Spectroscopy, Vol 77* Vol. 77 *Annu. Rep. NMR Spectro.* (ed G. A. Webb) 259-351 (2012).
- 72 Mafra, L. *et al.* Structure of Chemisorbed CO_2 Species in Amine-Functionalized Mesoporous Silicas Studied by Solid-State NMR and Computer Modeling. *J. Am. Chem. Soc.* **139**, 389-408 (2017).
- 73 Rocha, J., Paz, F. A. A., Sardo, M. & Mafra, L. Revisiting the crystal structure of dickite: X-ray diffraction, solid-state NMR, and DFT calculations study. *Am. Min.* **103**, 812-818 (2018).
- 74 Afonso, R., Sardo, M., Mafra, L. & Gomes, J. R. B. Unravelling the Structure of Chemisorbed CO_2 Species in Mesoporous Aminosilicas: A Critical Survey. *Environ. Sci. Technol.* **53**, 2758-2767 (2019).
- 75 Kinyanjui, F. G. *et al.* Crystal structure and proton conductivity of $BaSn_{0.6}Sc_{0.4}O_{3-\delta}$: insights from neutron powder diffraction and solid-state NMR spectroscopy. *J. Mater. Chem. A Mater.* **4**, 5088-5101 (2016).
- 76 Pigliapochi, R. *et al.* Structural Characterization of the Li-Ion Battery Cathode Materials $LiTi_xMn_{2-x}O_4$ ($0.2 \leq x \leq 1.5$): A Combined Experimental Li-7 NMR and First-Principles Study. *Chem. Mater.* **30**, 817-829 (2018).
- 77 Marker, K., Reeves, P. J., Xu, C., Griffith, K. J. & Grey, C. P. Evolution of Structure and Lithium Dynamics in $LiNi_{0.8}Mn_{0.1}Co_{0.1}O_2$ (NMC811) Cathodes during Electrochemical Cycling. *Chem. Mater.* **31**, 2545-2554 (2019).
- 78 Reeves, P. J., Seymour, I. D., Griffith, K. J. & Grey, C. P. Characterizing the Structure and Phase Transition of Li_2RuO_3 Using Variable-Temperature O-17 and Li-7 NMR Spectroscopy. *Chem. Mater.* **31**, 2814-2821 (2019).
- 79 Romao, C. P. *et al.* Zero Thermal Expansion in $ZrMgMo_3O_{12}$: NMR Crystallography Reveals Origins of Thermoelastic Properties. *Chem. Mater.* **27**, 2633-2646 (2015).

- Widdifield, C. M., Perras, F. A. & Bryce, D. L. Solid-state $^{185/187}\text{Re}$ NMR and GIPAW DFT study of perrhenates and $\text{Re}_2(\text{CO})_{10}$: chemical shift anisotropy, NMR crystallography, and a metal-metal bond. *Phys. Chem. Chem. Phys.* **17** (2015).
- Reimer, J. NMR assessment of structure and dynamics within metal organic frameworks. *Abstracts of Papers of the American Chemical Society* **255** (2018).
- Kanca, A., Dodd, M., Reimer, J. A. & Uner, D. Following the structure and reactivity of Tuncbilek lignite during pyrolysis and hydrogenation. *Fuel Process. Technol.* **152**, 266-273 (2016).
- Kubicki, D. J. *et al.* Phase Segregation in Cs-, Rb- and K-Doped Mixed-Cation $(\text{MA})_x(\text{FA})_{1-x}\text{PbI}_3$ Hybrid Perovskites from Solid-State NMR. *J. Am. Chem. Soc.* **139**, 14173-14180 (2017).
- Malekshah, R. E., Salehi, M., Kubicki, M. & Khaleghian, A. Crystal structure, molecular docking, and biological activity of the zinc complexes with 2-thenoyltrifluoroacetone and N-donor heterocyclic ligands. *J. Mol. Struct.* **1150**, 155-165 (2017).
- Kubicki, D. J. *et al.* Formation of Stable Mixed Guanidinium-Methylammonium Phases with Exceptionally Long Carrier Lifetimes for High-Efficiency Lead Iodide-Based Perovskite Photovoltaics. *J. Am. Chem. Soc.* **140**, 3345-3351 (2018).
- Balimann, G. *et al.* Chemical applications of high-resolution ^{13}C NMR spectra for solids. *Philos. Trans. Royal Soc. A, Mathematical and Physical Sciences* **299**, 643-663 (1981).
- Fletton, R. A. *et al.* A Comparative Spectroscopic Investigation of 3 Pseudopolymorphs of Testosterone Using Solid-State Ir and High-Resolution Solid-State Nmr. *Spectrochim. Acta A - Molecular and Biomolecular Spectroscopy* **43**, 1111-1120 (1987).
- Fletton, R. A. *et al.* A Comparative Spectroscopic Investigation of 2 Polymorphs of 4'-Methyl-2'-Nitroacetanilide Using Solid-State Infrared and High-Resolution Solid-State Nuclear-Magnetic-Resonance Spectroscopy. *J. Chem. Soc., Perkin Trans. 2*, 1705-1709 (1986).
- Harris, R. K., Say, B. J., Yeung, R. R., Fletton, R. A. & Lancaster, R. W. Cross-Polarization Magic-Angle Spinning Nmr-Studies of Polymorphism - Androstanolone. *Spectrochim. Acta A -Molecular and Biomolecular Spectroscopy* **45**, 465-469 (1989).
- Harris, R. K. *et al.* Cross-Polarization Magic-Angle Spinning Nmr-Studies of Polymorphism - Cortisone-Acetate. *Spectrochim. Acta A -Molecular and Biomolecular Spectroscopy* **46**, 927-935 (1990).
- Christopher, E. A., Harris, R. K. & Fletton, R. A. Assignments of solid-state C-13 resonances for polymorphs of cortisone acetate using shielding tensor components. *Solid State Nucl. Magn.* **1**, 93-101 (1992).
- McGeorge, G. *et al.* Analysis of a solid-state conformational rearrangement using ^{15}N NMR and X-ray crystallography. *J. Phys. Chem. A* **102**, 3505-3513 (1998).
- McGeorge, G., Harris, R. K., Chippendale, A. M. & Bullock, J. F. Conformational analysis by magic-angle spinning NMR spectroscopy for a series of polymorphs of a disperse azobenzene dyestuff. *Journal of the J. Chem. Soc., Perkin Trans. 2*, 1733-1738 (1996).
- Meejoo, S. *et al.* Structural Aspects of the β -Polymorph of (E)-4-Formylcinnamic Acid: Structure Determination Directly from Powder Diffraction Data and Elucidation of Structural Disorder from Solid-State NMR. *Helv. Chim. Acta* **86**, 1467-1477 (2003).
- Harris, K. D., Tremayne, M. & Kariuki, B. M. Contemporary advances in the use of powder X-ray diffraction for structure determination. *Angew. Chem. Int. Ed.* **40**, 1626-1651 (2001).
- Harris, R. K. NMR crystallography: the use of chemical shifts. *Solid State Sci.* **6**, 1025-1037 (2004).
- Tremayne, M., Kariuki, B. M. & Harris, K. D. Structure Determination of a Complex Organic Solid from X-Ray Powder Diffraction Data by a Generalized Monte Carlo Method: The Crystal Structure of Red Fluorescein. *Angew. Chem. Int. Ed.* **36**, 770-772 (1997).
- Middleton, D. A. *et al.* Conformational analysis by solid-state NMR and its application to restrained structure determination from powder diffraction data. *Chem. Commun.*, 1976-1977 (2002).
- Harris, K. D. & Xu, M. Crystal Structure Determination by the Combined Analysis of NMR and Powder Diffraction Data. *eMagRes* (2007).
- Harris, R. K. *et al.* Solid-state NMR and computational studies of 4-methyl-2-nitroacetanilide. *Magn. Reson. Chem.* **44**, 325-333 (2006).
- Taulelle, F. NMR crystallography: crystallochemical formula and space group selection. *Solid State Sci.* **6**, 1053-1057 (2004).
- Elena, B. & Emsley, L. Powder crystallography by proton solid-state NMR spectroscopy. *J. Am. Chem. Soc.* **127**, 9140-9146 (2005).
- Elena, B., Pintacuda, G., Mifsud, N. & Emsley, L. Molecular structure determination in powders by NMR crystallography from proton spin diffusion. *J. Am. Chem. Soc.* **128**, 9555-9560 (2006).

- 104 Pickard, C. J. & Mauri, F. All-electron magnetic response with pseudopotentials: NMR chemical shifts. *Phys. Rev. B* **63**, 245101 (2001).
- 105 Yates, J. R., Pickard, C. J. & Mauri, F. Calculation of NMR chemical shifts for extended systems using ultrasoft pseudopotentials. *Phys. Rev. B* **76**, 024401 (2007).
- 106 Ashbrook, S. E. & McKay, D. Combining solid-state NMR spectroscopy with first-principles calculations - a guide to NMR crystallography. *Chem. Commun.* **52**, 7186-7204 (2016).
- 107 Bonhomme, C. *et al.* First-Principles Calculation of NMR Parameters Using the Gauge Including Projector Augmented Wave Method: A Chemist's Point of View. *Chem. Rev.* **112**, 5733-5779 (2012).
- 108 Yates, J. R. *et al.* A combined first principles computational and solid-state NMR study of a molecular crystal: flurbiprofen. *Phys. Chem. Chem. Phys.* **7**, 1402-1407 (2005).
- 109 Harris, R. K., Joyce, S. A., Pickard, C. J., Cadars, S. & Emsley, L. Assigning carbon-13 NMR spectra to crystal structures by the INADEQUATE pulse sequence and first principles computation: a case study of two forms of testosterone. *Phys. Chem. Chem. Phys.* **8**, 137-143 (2006).
- 110 Harris, R. K. *et al.* Characterization of polymorphs and solvates of terbutaline sulfate. *Cryst. Growth Des.* **8**, 80-90 (2008).
- 111 Harris, R. K. *et al.* NMR crystallography of oxybuprocaine hydrochloride, Modification II degrees. *Phys. Chem. Chem. Phys.* **9**, 360-368 (2007).
- 112 Harris, R. K. *et al.* Computation and NMR crystallography of terbutaline sulfate. *Magn. Reson. Chem.* **48 Suppl 1**, S103-112 (2010).
- 113 Harper, J. K. & Grant, D. M. Enhancing crystal-structure prediction with NMR tensor data. *Cryst. Growth Des.* **6**, 2315-2321 (2006).
- 114 Pickard, C. J., Salager, E., Pintacuda, G., Elena, B. & Emsley, L. Resolving structures from powders by NMR crystallography using combined proton spin diffusion and plane wave DFT calculations. *J. Am. Chem. Soc.* **129**, 8932-+ (2007).
- 115 Salager, E., Stein, R. S., Pickard, C. J., Elena, B. & Emsley, L. Powder NMR crystallography of thymol. *Phys. Chem. Chem. Phys.* **11**, 2610-2621 (2009).
- 116 Salager, E. *et al.* Powder Crystallography by Combined Crystal Structure Prediction and High-Resolution H-1 Solid-State NMR Spectroscopy. *J. Am. Chem. Soc.* **132**, 2564-+ (2010).
- 117 Day, G. M., WD, S. M. & Jones, W. A strategy for predicting the crystal structures of flexible molecules: the polymorphism of phenobarbital. *Phys. Chem. Chem. Phys.* **9**, 1693-1704 (2007).
- 118 Day, G. M. & Cooper, T. G. Crystal packing predictions of the alpha-amino acids: methods assessment and structural observations. *Crystengcomm* **12**, 2443-2453 (2010).
- 119 Reilly, A. M. *et al.* Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.* **72**, 439-459 (2016).
- 120 Baías, M. *et al.* Powder crystallography of pharmaceutical materials by combined crystal structure prediction and solid-state ¹H NMR spectroscopy. *Phys. Chem. Chem. Phys.* **15**, 8069-8069 (2013).
- 121 Baías, M. *et al.* De novo determination of the crystal structure of a large drug molecule by crystal structure prediction-based powder NMR crystallography. *J. Am. Chem. Soc.* **135**, 17501-17507 (2013).
- 122 Widdifield, C. M. *et al.* Does Z' equal 1 or 2? Enhanced powder NMR crystallography verification of a disordered room temperature crystal structure of a p38 inhibitor for chronic obstructive pulmonary disease. *Phys. Chem. Chem. Phys.* **19**, 16650-16661 (2017).
- 123 Widdifield, C. M., Robson, H. & Hodgkinson, P. Furosemide's one little hydrogen atom: NMR crystallography structure verification of powdered molecular organics. *Chem. Commun.* **52**, 6685-6688 (2016).
- 124 Leclaire, J. *et al.* Structure elucidation of a complex CO₂-based organic framework material by NMR crystallography. *Chem. Sci.* **7**, 4379-4390 (2016).
- 125 Baías, M. *et al.* Superstructure of a substituted zeolitic imidazolate metal-organic framework determined by combining proton solid-state NMR spectroscopy and DFT calculations. *Angew. Chem. Int. Edit.* **54**, 5971-5976 (2015).
- 126 Brus, J. *et al.* Predicting the Crystal Structure of Decitabine by Powder NMR Crystallography: Influence of Long-Range Molecular Packing Symmetry on NMR Parameters. *Cryst. Growth Des.* **16**, 7102-7111 (2016).
- 127 Brauckmann, J. O. *et al.* Structural Studies of Polyaramid Fibers: Solid-State NMR and First-Principles Modeling. *Macromolecules* **49**, 5548-5560 (2016).
- 128 Mollica, G., Dekhil, M., Ziarelli, F., Thureau, P. & Viel, S. Quantitative Structural Constraints for Organic Powders at Natural Isotopic Abundance Using Dynamic Nuclear Polarization Solid-State NMR Spectroscopy. *Angew. Chem. Int. Edit.* **54**, 6028-6031 (2015).

- 129 Bradley, J. P., Velaga, S. P., Antzutkin, O. N. & Brown, S. P. Probing intermolecular crystal packing in γ -indomethacin by high-resolution ^1H solid-state NMR spectroscopy. *Cryst. Growth Des.* **11**, 3463-3471 (2011).
- 130 Chierotti, M. R. & Gobetto, R. NMR crystallography: the use of dipolar interactions in polymorph and co-crystal investigation. *Crystengcomm* **15**, 8599-8612 (2013).
- 131 Kalakewich, K., Iuliucci, R. & Harper, J. K. Establishing Accurate High-Resolution Crystal Structures in the Absence of Diffraction Data and Single Crystals-An NMR Approach. *Cryst. Growth Des.* **13**, 5391-5396 (2013).
- 132 Harper, J. K., Doebbler, J. A., Jacques, E., Grant, D. M. & Von Dreele, R. B. A combined solid-state NMR and synchrotron X-ray diffraction powder study on the structure of the antioxidant (+)-catechin 4.5-hydrate. *J. Am. Chem. Soc.* **132**, 2928-2937 (2010).
- 133 Harper, J. K. *et al.* Solid-state NMR characterization of the molecular conformation in disordered methyl α -L-rhamnofuranoside. *J. Phys. Chem. A* **117**, 5534-5541 (2013).
- 134 Kalakewich, K., Iuliucci, R., Mueller, K. T., Eloranta, H. & Harper, J. K. Monitoring the refinement of crystal structures with ^{15}N solid-state NMR shift tensor data. *J. Chem. Phys.* **143**, 194702 (2015).
- 135 Harper, J. K., Iuliucci, R., Gruber, M. & Kalakewich, K. Refining crystal structures with experimental ^{13}C NMR shift tensors and lattice-including electronic structure methods. *Crystengcomm* **15**, 8693-8704 (2013).
- 136 Santos, S. M., Rocha, J. & Mafra, L. NMR Crystallography: Toward Chemical Shift-Driven Crystal Structure Determination of the β -Lactam Antibiotic Amoxicillin Trihydrate. *Cryst. Growth Des.* **13**, 2390-2395 (2013).
- 137 Mafra, L. *et al.* Packing interactions in hydrated and anhydrous forms of the antibiotic Ciprofloxacin: a solid-state NMR, X-ray diffraction, and computer simulation study. *J. Am. Chem. Soc.* **134**, 71-74 (2012).
- 138 Dudenko, D. V. *et al.* Exploiting the Synergy of Powder X-ray Diffraction and Solid-State NMR Spectroscopy in Structure Determination of Organic Molecular Solids. *J. Phys. Chem. C Nanomater Interfaces* **117**, 12258-12265 (2013).
- 139 Ludeker, D. & Brunklaus, G. NMR crystallography of ezetimibe co-crystals. *Solid State Nucl. Magn. Res.* **65**, 29-40 (2015).
- 140 Watts, A. E., Maruyoshi, K., Hughes, C. E., Brown, S. P. & Harris, K. D. M. Combining the Advantages of Powder X-ray Diffraction and NMR Crystallography in Structure Determination of the Pharmaceutical Material Cimetidine Hydrochloride. *Cryst. Growth Des.* **16**, 1798-1804 (2016).
- 141 Filip, X., Borodi, G. & Filip, C. Testing the limits of sensitivity in a solid-state structural investigation by combined X-ray powder diffraction, solid-state NMR, and molecular modelling. *Phys. Chem. Chem. Phys.* **13**, 17978-17986 (2011).
- 142 Sardo, M. *et al.* Diazole-based powdered cocrystal featuring a helical hydrogen-bonded network: structure determination from PXRD, solid-state NMR and computer modeling. *Solid State Nucl. Magn. Reson.* **65**, 49-63 (2015).
- 143 Li, P. *et al.* Structure determination of the theophylline-nicotinamide cocrystal: a combined powder XRD, 1D solid-state NMR, and theoretical calculation study. *Crystengcomm* **16**, 3141-3147 (2014).
- 144 Reddy, G. N. *et al.* An NMR crystallography study of the hemihydrate of 2', 3'-O-isopropylidineguanosine. *Solid State Nucl. Magn. Reson.* **65**, 41-48 (2015).
- 145 Hughes, C. E. *et al.* Determination of a complex crystal structure in the absence of single crystals: analysis of powder X-ray diffraction data, guided by solid-state NMR and periodic DFT calculations, reveals a new 2'-deoxyguanosine structural motif. *Chem. Sci.* **8**, 3971-3979 (2017).
- 146 Rossi, F. *et al.* Unraveling the Hydrogen Bond Network in a Theophylline-Pyridoxine Salt Cocrystal by a Combined X-ray Diffraction, Solid State NMR, and Computational Approach. *Cryst. Growth Des.* **18**, 2225-2233 (2018).
- 147 Tatton, A. S. *et al.* Improving confidence in crystal structure solutions using NMR crystallography: The case of β -piroxicam. *Cryst. Growth Des.* **18**, 3339-3351 (2018).
- 148 Paluch, P., Pawlak, T., Oszejka, M., Lasocha, W. & Potrzebowski, M. J. Fine refinement of solid state structure of racemic form of phospho-tyrosine employing NMR Crystallography approach. *Solid State Nucl. Magn. Res.* **65**, 2-11 (2015).
- 149 Fernandes, J. A., Sardo, M., Mafra, L., Choquesillo-Lazarte, D. & Masciocchi, N. X-ray and NMR Crystallography Studies of Novel Theophylline Cocrystals Prepared by Liquid Assisted Grinding. *Cryst. Growth Des.* **15**, 3674-3683 (2015).
- 150 Nilsson Lill, S. O. *et al.* Elucidating an Amorphous Form Stabilization Mechanism for Tenapanor Hydrochloride: Crystal Structure Analysis Using X-ray Diffraction, NMR Crystallography, and Molecular Modeling. *Mol. Pharm.* **15**, 1476-1487 (2018).

- 151 Thompson, H. P. G. & Day, G. M. Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape. *Chem. Sci.* **5**, 3173-3182 (2014).
- 152 Caravatti, P., Bodenhausen, G. & Ernst, R. Heteronuclear solid-state correlation spectroscopy. *Chem. Phys. Lett.* **89**, 363-367 (1982).
- 153 Roberts, J. E., Vega, S. & Griffin, R. G. Two-dimensional heteronuclear chemical shift correlation spectroscopy in rotating solids. *J. Am. Chem. Soc.* **106**, 2506-2512 (1984).
- 154 Bax, A., Freeman, R. & Frenkiel, T. A. An NMR technique for tracing out the carbon skeleton of an organic molecule. *J. Am. Chem. Soc.* **103**, 2102-2104 (1981).
- 155 Lesage, A., Auger, C., Caldarelli, S. & Emsley, L. Determination of Through-Bond Carbon–Carbon Connectivities in Solid-State NMR Using the INADEQUATE Experiment. *J. Am. Chem. Soc.* **119**, 7867-7868 (1997).
- 156 Geen, H., Titman, J. J., Gottwald, J. & Spiess, H. W. Solid-state proton multiple-quantum NMR spectroscopy with fast magic angle spinning. *Chem. Phys. Lett.* **227**, 79-86 (1994).
- 157 Lowe, I. J. Free Induction Decays of Rotating Solids. *Phys. Rev. Lett.* **2**, 285-287 (1959).
- 158 Andrew, E. R., Bradbury, A. & Eades, R. G. Removal of Dipolar Broadening of Nuclear Magnetic Resonance Spectra of Solids by Specimen Rotation. *Nature* **183**, 1802-1803 (1959).
- 159 Maricq, M. M. & Waugh, J. S. Nmr in Rotating Solids. *J. Chem. Phys.* **70**, 3300-3316 (1979).
- 160 Brunner, E., Freude, D., Gerstein, B. & Pfeifer, H. Residual linewidths of NMR spectra of spin-1/2 systems under magic-angle spinning. *J. Magn. Res. (1969)* **90**, 90-99 (1990).
- 161 Levitt, M., Raleigh, D., Creuzet, F. & Griffin, R. Theory and simulations of homonuclear spin pair systems in rotating solids. *J. Chem. Phys.* **92**, 6347-6364 (1990).
- 162 Zorin, V. E., Brown, S. P. & Hodgkinson, P. Origins of linewidth in H-1 magic-angle spinning NMR. *J. Chem. Phys.* **125**, 144508 (2006).
- 163 Sternberg, U. *et al.* ¹H line width dependence on MAS speed in solid state NMR–Comparison of experiment and simulation. *J. Magn. Res.* **291**, 32-39 (2018).
- 164 Gerstein, B. C., Pembleton, R. G., Wilson, R. C. & Ryan, L. M. High resolution NMR in randomly oriented solids with homonuclear dipolar broadening: Combined multiple pulse NMR and magic angle spinning. *J. Chem. Phys.* **66**, 361 (1977).
- 165 Lee, M. & Goldburg, W. I. Nuclear-Magnetic-Resonance Line Narrowing by a Rotating Rf Field. *Phys. Rev.* **140**, 1261-+ (1965).
- 166 Waugh, J. S., Huber, L. M. & Haeberlen, U. Approach to High-Resolution Nmr in Solids. *Phys. Rev. Lett.* **20**, 180-+ (1968).
- 167 Mansfield, P. Symmetrized Pulse Sequences in High Resolution Nmr in Solids. *J. Phys. C: Solid State Phys.* **4**, 1444-+ (1971).
- 168 Burum, D. P. & Rhim, W. K. Analysis of multiple pulse NMR in solids. III. *J. Chem. Phys.* **71**, 944-956 (1979).
- 169 Bielecki, A., Kolbert, A. C. & Levitt, M. H. Frequency-Switched Pulse Sequences - Homonuclear Decoupling and Dilute Spin Nmr in Solids. *Chem. Phys. Lett.* **155**, 341-346 (1989).
- 170 Vinogradov, E., Madhu, P. & Vega, S. High-resolution proton solid-state NMR spectroscopy by phase-modulated Lee–Goldburg experiment. *Chem. Phys. Lett.* **314**, 443-450 (1999).
- 171 Sakellariou, D., Lesage, A., Hodgkinson, P. & Emsley, L. Homonuclear dipolar decoupling in solid-state NMR using continuous phase modulation. *Chem. Phys. Lett.* **319**, 253-260 (2000).
- 172 Elena, B., de Paëpe, G. & Emsley, L. Direct spectral optimisation of proton–proton homonuclear dipolar decoupling in solid-state NMR. *Chem. Phys. Lett.* **398**, 532-538 (2004).
- 173 Halse, M. E. & Emsley, L. Improved phase-modulated homonuclear dipolar decoupling for solid-state NMR spectroscopy from symmetry considerations. *J. Phys. Chem. A* **117**, 5280-5290 (2013).
- 174 Rhim, W. K., Elleman, D. D. & Vaughan, R. W. Analysis of Multiple Pulse Nmr in Solids. *J. Chem. Phys.* **59**, 3740-3749 (1973).
- 175 Leskes, M., Steuernagel, S., Schneider, D., Madhu, P. & Vega, S. Homonuclear dipolar decoupling at magic-angle spinning frequencies up to 65kHz in solid-state nuclear magnetic resonance. *Chem. Phys. Lett.* **466**, 95-99 (2008).
- 176 Salager, E. *et al.* Enhanced sensitivity in high-resolution ¹H solid-state NMR spectroscopy with DUMBO dipolar decoupling under ultra-fast MAS. *Chem. Phys. Lett.* **469**, 336-341 (2009).
- 177 Salager, E. *et al.* Homonuclear dipolar decoupling with very large scaling factors for high-resolution ultrafast magic angle spinning ¹H solid-state NMR spectroscopy. *Chem. Phys. Lett.* **498**, 214-220 (2010).

- 178 Gan, Z. H., Madhu, P. K., Amoureux, J. P., Trebosc, J. & Lafon, O. A tunable homonuclear dipolar decoupling scheme for high-resolution proton NMR of solids from slow to fast magic-angle spinning. *Chem. Phys. Lett.* **503**, 167-170 (2011).
- 179 Nishiyama, Y. *et al.* Practical choice of ^1H - ^1H decoupling schemes in through-bond ^1H - $\{X\}$ HMQC experiments at ultra-fast MAS. *J. Magn. Res.* **214**, 151-158 (2012).
- 180 Nishiyama, Y., Malon, M., Ishii, Y. & Ramamoorthy, A. 3D $^{15}\text{N}/^{15}\text{N}/^1\text{H}$ chemical shift correlation experiment utilizing an RFDR-based $^1\text{H}/^1\text{H}$ mixing period at 100kHz MAS. *J. Magn. Res.* **244**, 1-5 (2014).
- 181 Kobayashi, T. *et al.* Study of Intermolecular Interactions in the Corrole Matrix by Solid-State NMR under 100 kHz MAS and Theoretical Calculations. *Angew. Chem. Int. Ed.* **52**, 14108-14111 (2013).
- 182 Lin, Y.-L. *et al.* Preparation of fibril nuclei of beta-amyloid peptides in reverse micelles. *Chem. Commun.* **54**, 10459-10462 (2018).
- 183 Schledorn, M. *et al.* in *ISMAR EUROMAR Joint Conference*. 25 - 30 August 2019, Berlin (Germany). Abstract No 2240.
- 184 Halse, M. E., Schlagnitweit, J. & Emsley, L. High-Resolution ^1H Solid-State NMR Spectroscopy Using Windowed LG4 Homonuclear Dipolar Decoupling. *Isr. J. Chem.* **54**, 136-146 (2014).
- 185 Böckmann, A., Ernst, M. & Meier, B. H. Spinning proteins, the faster, the better? *J. Magn. Res.* **253**, 71-79 (2015).
- 186 Goldburg, W. I. & Lee, M. Nuclear magnetic resonance line narrowing by a rotating rf field. *Phys. Rev. Lett.* **11**, 255 (1963).
- 187 Haeberlen, U. & Waugh, J. S. Coherent Averaging Effects in Magnetic Resonance. *Phys. Rev.* **175**, 453-+ (1968).
- 188 Rhim, W. K., Pines, A. & Waugh, J. S. Time-Reversal Experiments in Dipolar-Coupled Spin Systems. *Phys. Rev. B* **3**, 684-+ (1971).
- 189 Mehring, M. & Waugh, J. S. Magic-Angle Nmr Experiments in Solids. *Phys. Rev. B* **5**, 3459-& (1972).
- 190 Burum, D. P., Linder, M. & Ernst, R. R. Low-Power Multipulse Line Narrowing in Solid-State Nmr. *J. Magn. Res.* **44**, 173-188 (1981).
- 191 Hohwy, M., Bower, P., Jakobsen, H. J. & Nielsen, N. C. A high-order and broadband CRAMPS experiment using z-rotational decoupling. *Chem. Phys. Lett.* **273**, 297-303 (1997).
- 192 Hohwy, M. & Nielsen, N. C. Elimination of high order terms in multiple pulse nuclear magnetic resonance spectroscopy: Application to homonuclear decoupling in solids. *J. Chem. Phys.* **106**, 7571-7586 (1997).
- 193 Levitt, M. H. Symmetry in the design of NMR multiple-pulse sequences. *J. Chem. Phys.* **128**, 052205 (2008).
- 194 Lesage, A., Bardet, M. & Emsley, L. Through-bond carbon-carbon connectivities in disordered solids by NMR. *J. Am. Chem. Soc.* **121**, 10987-10993 (1999).
- 195 Gerstein, B. C. High-Resolution Nmr in Solids with Strong Homonuclear Dipolar Broadening - Combined Multiple-Pulse Decoupling and Magic Angle Spinning. *Philos. Trans. Royal Soc. A - Mathematical Physical and Engineering Sciences* **299**, 521-546 (1981).
- 196 Jackson, P. & Harris, R. K. A practical guide to combined rotation and multiple-pulse NMR spectroscopy of solids. *Magn. Reson. Chem.* **26**, 1003-1011 (1988).
- 197 Buszko, M. L., Bronnimann, C. & Maciel, G. ^1H CRAMPS based on TREV. *J. Magn. Res. Ser. A* **103**, 183-187 (1993).
- 198 Levitt, M. H., Kolbert, A. C., Bielecki, A. & Ruben, D. J. High-resolution ^1H NMR in solids with frequency-switched multiple-pulse sequences. *Solid State Nucl. Magn. Reson.* **2**, 151-163 (1993).
- 199 Barbara, T. M. & Baltusis, L. Phase-Cycled, Multiple-Window-Acquisition, Multiple-Pulse Nmr-Spectroscopy. *J. Magn. Reson. Ser. A* **106**, 182-187 (1994).
- 200 Cho, H. Tilted-axis precession and phase-sensitive detection of nuclear magnetization. *J. Magn. Reson. Ser. A* **121**, 8-22 (1996).
- 201 Cho, H. Off-resonance multiple-pulse dynamics in solid-state NMR spectroscopy: A revised coherent averaging theory analysis. *J. Magn. Reson.* **141**, 164-179 (1999).
- 202 Mafra, L., Coelho, C., Siegel, R. & Rocha, J. Assessing the performance of windowed ^1H CRAMPS methods, on biological solids, at high-field and MAS up to 35 kHz. *J. Magn. Reson.* **197**, 20-27 (2009).
- 203 Paul, S. *et al.* Supercycled homonuclear dipolar decoupling sequences in solid-state NMR. *J. Magn. Reson.* **197**, 14-19 (2009).
- 204 Halse, M. E. & Emsley, L. A common theory for phase-modulated homonuclear decoupling in solid-state NMR. *Phys. Chem. Chem. Phys.* **14**, 9121-9130 (2012).

- 205 Paul, S. & Madhva, P. Homonuclear dipolar decoupling in solid-state nuclear magnetic resonance under the regime of moderate to high magic-angle spinning frequencies: Astatus report. *J. Indian I. Sci.* **90**, 69-86 (2012).
- 206 Rhim, W. K., Elleman, D. D., Schreiber, L. B. & Vaughan, R. W. Analysis of multiple pulse NMR in solids. II. *J. Chem. Phys.* **60**, 4595-4604 (1974).
- 207 Amoureux, J.-P. *et al.* Homonuclear dipolar decoupling schemes for fast MAS. *Solid State Nucl. Magn. Res.* **35**, 19-24 (2009).
- 208 Haeberlen, U. *High Resolution NMR in Solids: Advances in Magnetic Resonance*. (Academic Press, 1976).
- 209 Iwamiya, J., Sinton, S., Liu, H., Glaser, S. & Drobny, G. Multiple-pulse sequences for homonuclear decoupling. Experimental verification. *J. Magn. Reson.* (1969) **100**, 367-375 (1992).
- 210 Haeberlen, U. in *Nuclear magnetic resonance in solids*, 251-257 (Springer, 1977).
- 211 Vaughan, R., Elleman, D., Stacey, L., Rhim, W. K. & Lee, J. A simple, low power, multiple pulse NMR spectrometer. *Rev. Sci. Instr.* **43**, 1356-1364 (1972).
- 212 Lesage, A. *et al.* Experimental aspects of proton NMR spectroscopy in solids using phase-modulated homonuclear dipolar decoupling. *J. Magn. Reson.* **163**, 105-113 (2003).
- 213 Vinogradov, E., Madhu, P. K. & Vega, S. Proton spectroscopy in solid state nuclear magnetic resonance with windowed phase modulated Lee-Goldburg decoupling sequences. *Chem. Phys. Lett.* **354**, 193-202 (2002).
- 214 Yamauchi, K., Kuroki, S. & Ando, I. The amide proton NMR chemical shift and hydrogen-bonded structure of glycine-containing peptides and polypeptides in the solid state as studied by multi-pulse-associated high-speed MAS ^1H NMR. *J. Mol. Struct.* **602**, 9-16 (2002).
- 215 Leskes, M., Madhu, P. & Vega, S. A broad-banded z-rotation windowed phase-modulated Lee-Goldburg pulse sequence for ^1H spectroscopy in solid-state NMR. *Chem. Phys. Lett.* **447**, 370-374 (2007).
- 216 Leskes, M., Madhu, P. K. & Vega, S. Supercycled homonuclear dipolar decoupling in solid-state NMR: toward cleaner ^1H spectrum and higher spinning rates. *J. Chem. Phys.* **128**, 052309 (2008).
- 217 Tom O'Haver. A pragmatic introduction to signal processing, <<https://terpconnect.umd.edu/~toh/spectrum/>> (2019, July 1).
- 218 Salager, E., Dumez, J. N., Emsley, L. & Levitt, M. H. A scaling factor theorem for homonuclear dipolar decoupling in solid-state NMR spectroscopy. *J. Magn. Reson.* **212**, 11-16 (2011).
- 219 Elena, B., Lesage, A., Steuernagel, S., Böckmann, A. & Emsley, L. Proton to carbon-13 INEPT in solid-state NMR spectroscopy. *J. Am. Chem. Soc.* **127**, 17296-17302 (2005).
- 220 Bosman, L., Madhu, P., Vega, S. & Vinogradov, E. Improvement of homonuclear dipolar decoupling sequences in solid-state nuclear magnetic resonance utilising radiofrequency imperfections. *J. Magn. Res.* **169**, 39-48 (2004).
- 221 Vega, A. J. Controlling the effects of pulse transients and RF inhomogeneity in phase-modulated multiple-pulse sequences for homonuclear decoupling in solid-state proton NMR. *J. Magn. Res.* **170**, 22-41 (2004).
- 222 Leskes, M., Madhu, P. & Vega, S. Proton line narrowing in solid-state nuclear magnetic resonance: New insights from windowed phase-modulated Lee-Goldburg sequence. *J. Chem. Phys.* **125**, 124506 (2006).
- 223 Coelho, C., Rocha, J., Madhu, P. K. & Mafra, L. Practical aspects of Lee-Goldburg based CRAMPS techniques for high-resolution ^1H NMR spectroscopy in solids: implementation and applications. *J. Magn. Res.* **194**, 264-282 (2008).
- 224 Mote, K. R., Agarwal, V. & Madhu, P. K. Five decades of homonuclear dipolar decoupling in solid-state NMR: Status and outlook. *Prog. Nucl. Magn. Reson. Spectrosc.* **97**, 1-39 (2016).
- 225 Spiess, H. W., Zimmermann, H. & Haeberlen, U. Proton magnetic shielding and susceptibility effects in single crystals of ferrocene. *Chem. Phys.* **12**, 123-130 (1976).
- 226 Vinogradov, E., Madhu, P. K. & Vega, S. A bimodal Floquet analysis of phase modulated Lee-Goldburg high resolution proton magic angle spinning NMR experiments. *Chem. Phys. Lett.* **329**, 207-214 (2000).
- 227 Filip, C. & Hafner, S. Analysis of multiple-pulse techniques under fast MAS conditions. *J. Magn. Reson.* **147**, 250-260 (2000).
- 228 Lesage, A., Duma, L., Sakellariou, D. & Emsley, L. Improved resolution in proton NMR spectroscopy of powdered solids. *J. Am. Chem. Soc.* **123**, 5747-5752 (2001).
- 229 Hahn, E. L. Spin echoes. *Phys. Rev.* **80**, 580 (1950).
- 230 Levitt, M. H., Freeman, R. & Frenkiel, T. Broadband heteronuclear decoupling. *J. Magn. Res.* (1969) **47**, 328-330 (1982).

- 231 Waugh, J. Systematic procedure for constructing broadband decoupling sequences. *J. Magn. Res. (1969)* **49**, 517-521 (1982).
- 232 Shaka, A., Keeler, J., Frenkiel, T. & Freeman, R. An improved sequence for broadband decoupling: WALTZ-16. *J. Magn. Res. (1969)* **52**, 335-338 (1983).
- 233 Tycko, R., Pines, A. & Guckenheimer, J. Fixed point theory of iterative excitation schemes in NMR. *Journal Chem. Phys.* **83**, 2775-2802 (1985).
- 234 Shaka, A. & Keeler, J. Broadband spin decoupling in isotropic-liquids. *Prog. Nucl. Magn. Res. Spectrosc.* **19**, 47-129 (1987).
- 235 Meiboom, S. & Gill, D. Modified spin-echo method for measuring nuclear relaxation times. *Rev. Sci. Instr.* **29**, 688-691 (1958).
- 236 Davis, M. C., Shookman, K. M., Sillaman, J. D. & Grandinetti, P. J. TOP-PASS: A processing algorithm to reduce 2D PASS acquisition time. *J. Magn. Res.* **210**, 51-58 (2011).
- 237 Walder, B. J., Dey, K. K., Kaseman, D. C., Baltisberger, J. H. & Grandinetti, P. J. Sideband separation experiments in NMR with phase incremented echo train acquisition. *Journal Chem. Phys.* **138**, 174203 (2013).
- 238 Avalos, C. E., Walder, B. J., Viger-Gravel, J., Magrez, A. & Emsley, L. Chemical exchange at the ferroelectric phase transition of lead germanate revealed by solid state ²⁰⁷Pb nuclear magnetic resonance. *Phys. Chem. Chem. Phys.* **21**, 1100-1109 (2019).
- 239 Oschkinat, H., Pastore, A., Pfändler, P. & Bodenhausen, G. Two-dimensional correlation of directly and remotely connected transitions by z-filtered COSY. *J. Magn. Res. (1969)* **69**, 559-566 (1986).
- 240 Pell, A. J., Edden, R. A. & Keeler, J. Broadband proton-decoupled proton spectra. *Magn. Reson. Chem.* **45**, 296-316 (2007).
- 241 States, D., Haberkorn, R. & Ruben, D. A two-dimensional nuclear Overhauser experiment with pure absorption phase in four quadrants. *J. Magn. Res. (1969)* **48**, 286-292 (1982).
- 242 Grandinetti, P. J., Ash, J. T. & Trease, N. M. Symmetry pathways in solid-state NMR. *Prog. Nucl. Magn. Res. Spectrosc.* **59**, 121 (2011).
- 243 PhySy Lrd, RMN, Version 1.8 (www.physyapps.com, PhySy Ltd., Grandview Heights, OH 43212).
- 244 Schmieder, P., Zimmer, S. & Kessler, H. Increased resolution in proton detected heteronuclear NMR experiments by folding in the hetero-dimension. *Magn. Res. Chem.* **29**, 375-380 (1991).
- 245 Jeannerat, D. Computer optimized spectral aliasing in the indirect dimension of ¹H-¹³C heteronuclear 2D NMR experiments. A new algorithm and examples of applications to small molecules. *J. Magn. Res.* **186**, 112-122 (2007).
- 246 Hoch, J. C., Maciejewski, M. W., Mobli, M., Schuyler, A. D. & Stern, A. S. Nonuniform sampling and maximum entropy reconstruction in multidimensional NMR. *Acc. Chem. Res.* **47**, 708-717 (2014).
- 247 Brüschweiler, R. & Zhang, F. Covariance nuclear magnetic resonance spectroscopy. *Journal Chem. Phys.* **120**, 5253-5260 (2004).
- 248 Bax, A., Mehlkopf, A. & Smidt, J. Absorption spectra from phase-modulated spin echoes. *J. Magn. Res. (1969)* **35**, 373-377 (1979).
- 249 Bax, A. & Freeman, R. Investigation of complex networks of spin-spin coupling by two-dimensional NMR. *J. Magn. Res. (1969)* **44**, 542-561 (1981).
- 250 Rance, M., Wagner, G., Sørensen, O., Wüthrich, K. & Ernst, R. Application of ω_1 -decoupled 2D correlation spectra to the study of proteins. *J. Magn. Res. (1969)* **59**, 250-261 (1984).
- 251 Powers, R., Gronenborn, A. M., Clore, G. M. & Bax, A. Three-dimensional triple-resonance NMR of ¹³C/¹⁵N-enriched proteins using constant-time evolution. *J. Magn. Res. (1969)* **94**, 209-213 (1991).
- 252 Santoro, J. & King, G. C. A constant-time 2D overbodenhausen experiment for inverse correlation of isotopically enriched species. *J. Magn. Res. (1969)* **97**, 202-207 (1992).
- 253 Vuister, G. W. & Bax, A. Resolution enhancement and spectral editing of uniformly ¹³C-enriched proteins by homonuclear broadband ¹³C decoupling. *J. Magn. Res. (1969)* **98**, 428-435 (1992).
- 254 Colaugh, H. & Nishiyama, Y. Resolution enhancement in proton double quantum magic-angle spinning spectra by constant-time acquisition. *Solid State Nucl. Magn. Res.* (2017).
- 255 Malar, A. A. *et al.* Quantifying proton NMR coherent linewidth in proteins under fast MAS conditions: a second moment approach. *Phys. Chem. Chem. Phys.* (2019).
- 256 Hogben, H., Krzystyniak, M., Charnock, G., Hore, P. & Kuprov, I. Spinach—a software library for simulation of spin dynamics in large spin systems. *J. Magn. Res.* **208**, 179-194 (2011).
- 257 Pell, A. J. & Keeler, J. Two-dimensional J-spectra with absorption-mode lineshapes. *J. Magn. Res.* **189**, 293-299 (2007).
- 258 Dedios, A. C., Pearson, J. G. & Oldfield, E. Secondary and Tertiary Structural Effects on Protein Nmr Chemical-Shifts - an Abinitio Approach. *Science* **260**, 1491-1496 (1993).

- 259 Facelli, J. C. & Grant, D. M. Determination of molecular symmetry in crystalline naphthalene using
solid-state NMR. *Nature* **365**, 325-327 (1993).
- 260 Sebastiani, D. & Parrinello, M. A new ab-initio approach for NMR chemical shifts in periodic systems.
J. Phys. Chem. A **105**, 1951-1958 (2001).
- 261 Blochl, P. E. Projector augmented-wave method. *Phys. Rev. B Condens. Matter* **50**, 17953-17979
(1994).
- 262 Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nat. Mater.* **12**,
191-201 (2013).
- 263 Bartok, A. P. *et al.* Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**
(2017).
- 264 Xue, D. *et al.* Accelerated search for materials with targeted properties by adaptive design. *Nat.*
Commun. **7**, 11241 (2016).
- 265 Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework
for predicting properties of inorganic materials. *Npj Comput. Mater.* **2**, (2016).
- 266 Rupp, M., Tkatchenko, A., Muller, K. R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular
atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
- 267 Shen, Y. & Bax, A. Protein backbone chemical shifts predicted from searching a database for torsion
angle and sequence homology. *J. Biomol. Nmr* **38**, 289-302 (2007).
- 268 Neal, S., Nip, A. M., Zhang, H. Y. & Wishart, D. S. Rapid and accurate calculation of protein H-1, C-13
and N-15 chemical shifts. *J. Biomol. Nmr* **26**, 215-240 (2003).
- 269 Wishart, D. S., Watson, M. S., Boyko, R. F. & Sykes, B. D. Automated H-1 and C-13 chemical shift
prediction using the BioMagResBank. *J. Biomol. Nmr* **10**, 329-336 (1997).
- 270 Iwadate, M., Asakura, T. & Williamson, M. P. C alpha and C beta carbon-13 chemical shifts in proteins
from an empirical database. *J. Biomol. Nmr* **13**, 199-211 (1999).
- 271 Xu, X. P. & Case, D. A. Automated prediction of ¹⁵N, ¹³C(alpha), ¹³C(beta) and ¹³C' chemical shifts in
proteins using a density functional database. *J. Biomol. Nmr* **21**, 321-333 (2001).
- 272 Moon, S. & Case, D. A. A new model for chemical shifts of amide hydrogens in proteins. *J. Biomol. Nmr*
38, 139-150 (2007).
- 273 Vila, J. A., Arnautova, Y. A., Martin, O. A. & Scheraga, H. A. Quantum-mechanics-derived ¹³Calpha
chemical shift server (CheShift) for protein structure validation. *P. Natl. Acad. Sci. USA* **106**, 16972-
16977 (2009).
- 274 Kohlhoff, K. J., Robustelli, P., Cavalli, A., Salvatella, X. & Vendruscolo, M. Fast and accurate predictions
of protein NMR chemical shifts from interatomic distances. *J. Am. Chem. Soc.* **131**, 13894-13895
(2009).
- 275 Meiler, J. PROSHIFT: Protein chemical shift prediction using artificial neural networks. *J. Biomol. Nmr*
26, 25-37 (2003).
- 276 Han, B., Liu, Y., Gininger, S. W. & Wishart, D. S. SHIFTX2: significantly improved protein chemical
shift prediction. *J. Biomol. Nmr* **50**, 43-57 (2011).
- 277 Shen, Y. & Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by
means of an artificial neural network. *J. Biomol. Nmr* **48**, 13-22 (2010).
- 278 Rupp, M., Ramakrishnan, R. & von Lilienfeld, O. A. Machine Learning for Quantum Mechanical
Properties of Atoms in Molecules. *J. Phys. Chem. Lett.* **6**, 3309-3313 (2015).
- 279 Blinov, K. *et al.* Performance validation of neural network based ¹³C NMR prediction using a publicly
available data source. *J. Chem. Inf. Model.* **48**, 550-555 (2008).
- 280 Smurnyy, Y. D., Blinov, K. A., Churanova, T. S., Elyashberg, M. E. & Williams, A. J. Toward more reliable
¹³C and ¹H chemical shift prediction: a systematic comparison of neural-network and least-squares
regression based approaches. *J. Chem. Inf. Model.* **48**, 128-134 (2008).
- 281 Aires-de-Sousa, J., Hemmer, M. C. & Gasteiger, J. Prediction of ¹H NMR chemical shifts using neural
networks. *Anal. Chem.* **74**, 80-90 (2002).
- 282 Kuhn, S., Egert, B., Neumann, S. & Steinbeck, C. Building blocks for automated elucidation of
metabolites: machine learning methods for NMR prediction. *BMC Bioinform.* **9**, 400 (2008).
- 283 Cuny, J., Xie, Y., Pickard, C. J. & Hassanali, A. A. Ab Initio Quality NMR Parameters in Solid-State
Materials Using a High-Dimensional Neural-Network Representation. *J. Chem. Theory Comput.* **12**,
765-773 (2016).
- 284 Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge Structural Database. *Acta*
Crystallogr. B Struct. Sci. Cryst. Eng. Mater. **72**, 171-179 (2016).
- 285 Ceriotti, M., Tribello, G. A. & Parrinello, M. Demonstrating the Transferability and the Descriptive
Power of Sketch-Map. *J. Chem. Theory Comput.* **9**, 1521-1532 (2013).

- 286 Campello, R. J. G. B., Moulavi, D., Zimek, A. & Sander, J. Hierarchical Density Estimates for Data
Clustering, Visualization, and Outlier Detection. *Acm T. Knowl. Discov. D* **10**, 5, (2015).
- 287 Giannozzi, P. *et al.* Advanced capabilities for materials modelling with Quantum ESPRESSO. *J. Phys.*
Condens. Matter **29**, 465901 (2017).
- 288 Giannozzi, P. *et al.* QUANTUM ESPRESSO: a modular and open-source software project for quantum
simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).
- 289 Varini, N., Ceresoli, D., Martin-Samos, L., Girotto, I. & Cavazzoni, C. Enhancement of DFT-calculations
at petascale: Nuclear Magnetic Resonance, Hybrid Density Functional Theory and Car–Parrinello
calculations. *Comput. Phys. Commun.* **184**, 1827-1833 (2013).
- 290 Rasmussen, C. E. & Williams, C. K. *Gaussian processes for machine learning*. Vol. 1 (MIT press
Cambridge, 2006).
- 291 Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 1-16
(2013).
- 292 De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and
alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754-13769 (2016).
- 293 F. Musil, S. De & M. Ceriotti. *Glosim2 package*, <<https://github.com/cosmo-epfl/glosim2>> (2017).
- 294 Hartman, J. D., Kudla, R. A., Day, G. M., Mueller, L. J. & Beran, G. J. Benchmark fragment-based ^1H , ^{13}C ,
 ^{15}N and ^{17}O chemical shift predictions in molecular crystals. *Phys. Chem. Chem. Phys.* **18**, 21686-
21709 (2016).
- 295 Bartók, A. P., Gillan, M. J., Manby, F. R. & Csányi, G. Machine-learning approach for one- and two-body
corrections to density functional theory: Applications to molecular and condensed water. *Phys. Rev.*
B **88**, 1-12 (2013).
- 296 Grisafi, A., Wilkins, D. M., Csanyi, G. & Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial
Properties of Atomistic Systems. *Phys. Rev. Lett.* **120**, 036002 (2018).
- 297 Clark, S. J. *et al.* First principles methods using CASTEP. *Z Kristallogr.* **220**, 567-570 (2005).
- 298 Arico-Muendel, C. C. *et al.* Orally active fumagillin analogues: transformations of a reactive warhead
in the gastric environment. *Acs Med. Chem. Lett.* **4**, 381-386 (2013).
- 299 Dao, H. T., Li, C., Michaudel, Q., Maxwell, B. D. & Baran, P. S. Hydromethylation of Unactivated Olefins.
J. Am. Chem. Soc. **137** (2015).
- 300 Garozzo, D. *et al.* Inclusion networks of a calix[5]arene-based exoditopic receptor and long-chain
alkyldiammonium ions. *Org. Lett.* **5**, 4025-4028 (2003).
- 301 Bats, J. W. *CSD Communication* (2010).
- 302 Huang, G. B. *et al.* Selective recognition of aromatic hydrocarbons by endo-functionalized molecular
tubes via C/N-H center dot center dot center dot pi interactions. *Chinese Chem. Lett.* **29**, 91-94 (2018).
- 303 Plater, M. J., Harrison, W. A., Machado de los Toyos, L. & Hendry, L. The consistent hexameric paddle-
wheel crystallisation motif of a family of 2,4-bis(n-alkylamino)nitrobenzenes: alkyl = pentyl, hexyl,
heptyl and octyl. *J. Chem. Res.* **41**, 235-238 (2017).
- 304 Seeger, M., Williams, C. & Lawrence, N. Fast forward selection to speed up sparse Gaussian process
regression. (2003).
- 305 Csató, L. & Oppel, M. Sparse on-line Gaussian processes. *Neural Comput.* **14**, 641-668 (2002).
- 306 Musil, F., Willatt, M. J., Langovoy, M. A. & Ceriotti, M. Fast and Accurate Uncertainty Estimation in
Chemical Machine Learning. *J. Chem. Theory Comput.* **15**, 906-915 (2019).
- 307 Imbalzano, G. *et al.* Automatic selection of atomic fingerprints and reference configurations for
machine-learning potentials. *J. Chem. Phys.* **148**, 241730 (2018).
- 308 Chaker, Z., Salanne, M., Delaye, J.-M. & Charpentier, T. NMR shifts in aluminosilicate glasses via
machine learning. *Phys. Chem. Chem. Phys.* (2019).
- 309 Liu, S. *et al.* A Multi-Resolution 3D-DenseNet for Chemical Shift Prediction in NMR Crystallography.
arXiv preprint arXiv:1906.00102 (2019).
- 310 Lejaeghere, K. *et al.* Reproducibility in density functional theory calculations of solids. *Science* **351**,
aad3000 (2016).
- 311 Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev.*
Lett. **77**, 3865-3868 (1996).
- 312 Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion
correction. *J. Comput. Chem.* **27**, 1787-1799 (2006).
- 313 Monkhorst, H. J. & Pack, J. D. Special Points for Brillouin-Zone Integrations. *Phys. Rev. B* **13**, 5188-
5192 (1976).
- 314 Neumann, M. A., van de Streek, J., Fabbiani, F. P., Hidber, P. & Grassmann, O. Combined crystal
structure prediction and high-pressure crystallization in rational pharmaceutical polymorph
screening. *Nat. Commun.* **6**, 7793 (2015).

- Dracinsky, M., Unzueta, P. & Beran, G. J. O. Improving the accuracy of solid-state nuclear magnetic resonance chemical shift prediction with a simple molecular correction. *Phys. Chem. Chem. Phys.* **21**, 14992-15000 (2019).
- Clayden, N. J., Dobson, C. M., Lian, L.-Y. & Twyman, J. M. A solid-state ^{13}C nuclear magnetic resonance study of the conformational states of penicillins. *J. Chem. Soc. Perkin Trans. 2*, 1933-1940 (1986).
- Antzutkin, O. N., Lee, Y. K. & Levitt, M. H. ^{13}C and ^{15}N -chemical shift anisotropy of ampicillin and penicillin-V studied by 2D-PASS and CP/MAS NMR. *J. Magn. Reson.* **135**, 144-155 (1998).
- Hayashi, S. & Hayamizu, K. Chemical shift standards in high-resolution solid-state NMR ^{13}C , ^{29}Si , and ^1H nuclei. *Bull. Chem. Soc. Jpn* **64**, 685-687 (1991).
- Ceriotti, M., De, S., Meissner, R. H. & Tribello, G. A. *Sketch map package*, <<https://github.com/cosmo-epfl/sketchmap/>> (2011).
- Ceriotti, M., Tribello, G. A. & Parrinello, M. From the Cover: Simplifying the representation of complex free-energy landscapes using sketch-map. *P. Natl. Acad. Sci. USA* **108**, 13023-13028 (2011).
- De, S., Musil, F., Ingram, T., Baldauf, C. & Ceriotti, M. Mapping and classifying molecules from a high-throughput structural database. *J. Cheminform.* **9**, 6 (2017).
- Pinon, A. C., Rossini, A. J., Widdifield, C. M., Gajan, D. & Emsley, L. Polymorphs of Theophylline Characterized by DNP Enhanced Solid-State NMR. *Mol. Pharm.* **12**, 4146-4153 (2015).
- Tkatchenko, A. & Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **102** (2009).
- Dal Corso, A. Pseudopotentials periodic table: From H to Pu. *Comput. Mater. Sci.* **95**, 337-350 (2014).
- Roberts, J. E., Harbison, G. S., Munowitz, M. G., Herzfeld, J. & Griffin, R. G. Measurement of heteronuclear bond distances in polycrystalline solids by solid-state NMR techniques. *J. Am. Chem. Soc.* **109**, 4163-4169 (1987).
- Colombo, M., Meier, B. & Ernst, R. Rotor-driven spin diffusion in natural-abundance ^{13}C spin systems. *Chem. Phys. Lett.* **146**, 189-196 (1988).
- Raleigh, D., Levitt, M. & Griffin, R. Rotational resonance in solid state NMR. *Chem. Phys. Lett.* **146**, 71-76 (1988).
- Van Rossum, B.-J., De Groot, C., Ladizhansky, V., Vega, S. & De Groot, H. A method for measuring heteronuclear ^1H - ^{13}C distances in high speed MAS NMR. *J. Am. Chem. Soc.* **122**, 3465-3472 (2000).
- Seidel, K. *et al.* Studying molecular 3D structure and dynamics by high-resolution solid-state NMR: Application to L-tyrosine-ethylester. *J. Phys. Chem. A* **109**, 2436-2442 (2005).
- Dekhil, M. *et al.* Determining carbon-carbon connectivities in natural abundance organic powders using dipolar couplings. *Chem. Commun.* **52**, 8565-8568 (2016).
- Märker, K. *et al.* Welcoming natural isotopic abundance in solid-state NMR: probing π -stacking and supramolecular structure of organic nanoassemblies using DNP. *Chem. Sci.* **8**, 974-987 (2017).
- Thureau, P. *et al.* Reducing the computational cost of NMR crystallography of organic powders at natural isotopic abundance with the help of ^{13}C - ^{13}C dipolar couplings. *Magn. Reson. Chem.* **57**, 256-264 (2019).
- Gu, Z., Ridenour, C. F., Bronnimann, C. E., Iwashita, T. & McDermott, A. Hydrogen Bonding and Distance Studies of Amino Acids and Peptides Using Solid State 2D ^1H - ^{13}C Heteronuclear Correlation Spectra. *J. Am. Chem. Soc.* **118**, 822-829 (1996).
- Case, D. H., Campbell, J. E., Bygrave, P. J. & Day, G. M. Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling. *J. Chem. Theory Comput.* **12**, 910-924 (2016).
- Price, S. L. *et al.* Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. *Phys. Chem. Chem. Phys.* **12**, 8478-8490 (2010).
- Nyman, J. & Day, G. M. Static and lattice vibrational energy differences between polymorphs. *Crystengcomm* **17**, 5154-5165 (2015).
- Boles, M. O. & Girven, R. J. The structures of ampicillin: a comparison of the anhydrate and trihydrate forms. *Acta Crystallogr. B Structural Crystallography and Crystal Chemistry* **32**, 2279-2284 (1976).
- Kolossvary, I. & Guida, W. C. Low mode search. An efficient, automated computational method for conformational analysis: Application to cyclic and acyclic alkanes and cyclic peptides. *J. Am. Chem. Soc.* **118**, 5011-5019 (1996).
- Kolossváry, I. & Guida, W. C. Low-mode conformational search elucidated: Application to $\text{C}_{39}\text{H}_{80}$ and flexible docking of 9-deazaguanine inhibitors into PNP. *J. Comput. Chem.* **20**, 1671-1684 (1999).
- MacroModel, V9.0, Schrödinger LLC: New York, NY (2011).
- Harder, E. *et al.* OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* **12**, 281-296 (2015).
- Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456-1465 (2011).

- 343 Stone, A. J. Distributed multipole analysis: Stability for large basis sets. *J. Chem. Theory Comput.* **1**,
1128-1132 (2005).
- 344 Ceriotti, M., More, J. & Manolopoulos, D. E. i-PI: A Python interface for ab initio path integral molecular
dynamics simulations. *Comput. Phys. Commun.* **185**, 1019-1026 (2014).
- 345 Ceriotti, M., Bussi, G. & Parrinello, M. Colored-noise thermostats à la carte. *J. Chem. Theory Comput.* **6**,
1170-1180 (2010).
- 346 Willatt, M. J., Musil, F. & Ceriotti, M. Atom-density representations for machine learning. *J. Chem. Phys.*
150 (2019).
- 347 Chisholm, J. A. & Motherwell, S. COMPACK: a program for identifying crystal structure similarity using
distances. *J. Appl. Crystallogr.* **38**, 228-231 (2005).
- 348 Stievano, L. *et al.* Density Functional Theory modeling and calculation of NMR parameters: An ab
initio study of the polymorphs of bulk glycine. *Cryst. Growth Des.* **10**, 3657-3667 (2010).
- 349 Carignani, E., Borsacchi, S., Bradley, J. P., Brown, S. P. & Geppi, M. Strong intermolecular ring current
influence on ^1H chemical shifts in two crystalline forms of naproxen: a combined solid-state NMR and
DFT study. *J. Phys. Chem. C* **117**, 17731-17740 (2013).
- 350 Uldry, A.-C. *et al.* Quantifying weak hydrogen bonding in uracil and 4-Cyano-4'-ethynylbiphenyl: a
combined computational and experimental investigation of NMR chemical shifts in the solid state. *J.*
Am. Chem. Soc. **130**, 945-954 (2008).
- 351 Harris, R. K. & Jackson, P. High-resolution ^1H and ^{13}C NMR of solid 2-aminobenzoic acid. *J. Phys. Chem.*
Solids **48**, 813-818 (1987).
- 352 Tatton, A. S. *et al.* Probing intermolecular interactions and nitrogen protonation in pharmaceuticals
by novel 15 N-edited and 2D ^{14}N - ^1H solid-state NMR. *Crystengcomm* **14**, 2654-2659 (2012).
- 353 Abraham, A., Apperley, D. C., Gelbrich, T., Harris, R. K. & Griesser, U. J. NMR crystallography—three
polymorphs of phenobarbital. *Can. J. Chem.* **89**, 770-778 (2011).

Acknowledgements

There are no words to describe the past four years. They have completely changed my life, both personally and professionally, and I am extremely grateful to all the people I have crossed in my path and that have helped me reach this milestone, directly or indirectly. Thank you, to all of you !

This thesis would have not been possible without the support and the direction of my supervisor, **Lyndon Emsley**. Lyndon, to join your group four years ago was by far the best decision I have ever taken. I cannot explain, or perhaps even realize, everything you have done for me. Thank you for guiding me toward the realization of a thesis I am proud of, for your invaluable ideas and the always interesting discussions. But thank you also for making me understand that it is not all about science, for teaching me how to present yourself and your results, for always supporting my professional growth and, especially, for showing me the importance of leading by example and the impact that a great leader can have. Thank you, Lyndon, I will always be grateful to you.

Besides Lyndon, I would like to thank the other members of my thesis committee: **Maria Baias**, **Luís Mafra**, **Jean-Philippe Ansermet** and **Clémence Corminboeuf**. Thank you for taking the time to read my work, for the very interesting discussions and for the detailed revisions which have improved and strengthened this thesis.

A special thanks to **Nadia Gauljaux**, for perfectly taking in care of all the administrative works of course, but also for being a wonderful colleague I love to chat with (in italiano, ovviamente). Getting to know you and **Pascal** more was a lot of fun !

Over the years, many people came to visit the lab and some of them left a mark on me. Thanks to **Phil Grandinetti** for introducing me to theoretical aspects of the dipolar coupling. Dealing with all these Hamiltonians and trying to find the analytical solutions of the dipolar coupling Hamiltonian for a 3 spin systems made me clearly realize what I do not want to do in my future, but ended up being very helpful for my PhD. Thank you, Phil, also for teaching me phase cycling and for making the moka break after lunch always remarkable. Thanks to **Brad Chmelka** for always being super enthusiastic and full of ideas for every project I have told him about, but also for being extremely proactive in organizing activities with the lab, for taking his time to take a beer together and for making the best paella in town. Thank you, Brad, it was great to have you here in Lausanne. Many thanks also to **Michel Bardet** and **Anne Lesage** for the very interesting discussions on NMR and much more. Anne, in a couple of hours at the spectrometer with you I have learned more than in months alone, thank you.

I am grateful to all the EPFL technical staff that have helped me with the most diverse tasks, from buying pens at the magasin to run XRD measurements. In particular, thanks to the NMR platform team: **Anto Barisic**, **Aurélien Bornet**, **Emilie Baudat**, and **Pascal Mieville**. Your help was gold to make everything run smoothly, and working along with you I could learn a lot about troubleshooting.

Thanks to all my colleagues, especially to the ones that are more friends than colleagues. You have made these years fun and unforgettable, in and outside the office. Thanks to **Albert Hofstetter** *in primis*, the best office mate ever, with who I have shared most of the projects but also many amazing moments. Thank you, Albi, for always finding incredible solutions to all the problems we have faced and for being always there when needed. Spending time with you was so much fun, and I am so happy we will finish this path together. Thanks to **Arthur Pinon**, for

being the first person in the lab I have often hang out with, and for all the crazy evenings, conferences and trips we have done together. You have left some unforgettable memories, and some memories I will never be able to remember, but I am sure were a lot of fun. Thanks to **Baptiste Busi** for great moments such as our trip to Amsterdam, and for making me want to have an ant colony as well. Thanks to **Claudia Avalos** for being super fun, hosting great parties and cooking the best Mexican food ever, and thanks to **John** for always creating amazing cocktails. Thanks to **Snædis Björgvinsdóttir** for sometimes being the only one that can get my accent and for teaching me how to say bjorgonaveste. Grazie a **Gabriele Stevanato** per esserci sempre quando c'è bisogno di smadonnare in italiano, sparare stronzate o fare una bella chiacchierata. Thanks to **Pinelopi Moutzouri** for taking over the duty to provide the world with high-resolution ^1H spectra in the solid-state. Good luck with that, it was a lot of fun to spend time at the spectrometer together. Thanks to **Pierrick Berruyer** for having been a great coloc for a month. Also Sigma is grateful for that. Grazie ad **Andrea Bertarello** perché un giorno andremo a cantare tutta la discografia di Raffaella Carrà al car-au-gay. Thanks to **Michael Hope**. Even if it was not much, the time we have spent together was fun and you have been a great office mate. Take care of my workstation. Thanks to **Manuel Cordova** and **Bruno Simões de Almeida**, the future of NMR crystallography. It was great to overlap a few months with you, and best luck with your PhD projects. Thanks to **Brennan Walder** for always providing brilliant technical solutions, no matter what the problem was. Thanks to **Subba** for helping me with my first steps in NMR. Thanks also to all the other past and present colleagues for the interesting discussions and the time spent together: **Dominik Kubicki**, **Aditya Mishra**, **Martins Balodis**, **Georges Menzildjian**, **François Pichard**, **Fanny Racine**, **Nicolas Matthey**, and **Yu Rao**.

Many thanks to **Michele Ceriotti** for the extremely interesting collaborations we had over the years, and for welcoming Albert and me in his group where we could learn about machine learning. Thank you very much Michele, my thesis and my future opportunities would have not been the same without you. And thank you very much also to Michele's boys: **Andrea Anelli**, **Félix Musil**, and **Edgar Engel**. It has been a lot of fun working and spending time with you !

I am very grateful also to the many new friends I have made during my time in Lausanne. Thanks to **Natalia**, for being a very good friend, for always bringing good vibes and for taking care of my beloved Sigma. Thanks to **Mohamed**, my first coloc, for all the time and the evenings spent together, starting from the ones in Denges. Grazie a **Daniele**, per esserci sempre per uno svago o una chiacchierata, per i mille consigli che mi ha dato e per essere ancora un grande amico, nonostante la distanza. Thanks to **Estel** and **Arnau** for being great friends, for all the fun moments we have spent together, and for all the trip and adventures we have done and the ones that still have to come. Thanks also to **Marina**, who has joined us for many trips and contributed to making them unique and incredibly fun. Thanks to **Jade**, for being an amazing person I love to talk and spend time with, for making me laugh and for perfectly drunk-cutting my hair in a hotel room. I will never forget that.

Un grazie a tutti i gli amici che ho portato con me dall'Italia, che arrivino da Carmagnola o dall'Università, per aver reso indimenticabile ogni ritorno in patria e per esserci sempre quando c'è da far festa. Un grazie speciale a **Nico**, **Andre** e **Drupi**, che han trovato il tempo per venire a trovarmi a Losanna, e son diventati leggenda da queste parti. Specialmente Drupi. Grazie a **Panta** per i fantastici e sempre sobri weekend passati insieme. E un mega grazie va naturalmente a **Andy** e **Cate**, per tutto, non saprei neanche da dove iniziare. Passare del tempo con voi é sempre fantastico, grazie di esserci.

Thanks to all the people I have met and spent time with in Lyon during this last year, and especially to **Adrian**, **Ribal** and **Monu**.

Un grazie speciale va a tutta la mia famiglia. A cominciare da **Nonna Ninni**, a cui questa tesi è dedicata, e che sarebbe stata super orgogliosa di essere qua oggi. Mi manchi nonna. Grazie a **mamma** e **papà**, per esserci sempre, per tutto quello che avete fatto per me e per darmi sempre il vostro supporto e incoraggiamento in tutto quello che faccio. Siete i migliori, e senza di voi non sarei mai arrivato fin qua. Grazie a **Matte** e **Carol**, i migliori amici e fratelli che avessi mai potuto immaginare, e che nonostante la distanza e gli impegni ci sono sempre per me. Non

posso neanche descrivere quanto bene vi voglio. Un mega grazie a **Nonno Mario**, per essere il nonno più cool di sempre, e per essere sempre stato un grande esempio e un'ispirazione. Grazie a **Nonno Gianni** e **Nonna Maria**, per aver sempre creduto in me e per esserci sempre stati qualsiasi cosa di cui abbia avuto bisogno. Grazie a **Zia Manu**, con cui son sempre contento di passare del tempo e fare una bella chiacchierata, a **Zio Mauri** e ai miei fantastici cugini **Franci**, **Ale** e **Umbi**. Grazie a **Zio Simone** e **Zia Angela**, anche se non succede spesso ,mi fa sempre piacere vederli. Il grazie piu speciale di tutti va a **Leonardo**, l'ultimo arrivo in famiglia, che mi rende uno zio super felice. Spero che tra un po' di anni sia tu a scrivere i ringraziamenti per la tua tesi. E grazie mille anche ad **Alice** e **Paolo** per i sempre bei momenti passati insieme. A special thanks goes also to the canadian side of my family, **Jean** and **Joanne**, for giving me such a warm welcome in the family and for always making me feel at home. I wish you could be here too, but I know we will celebrate soon together.

Thanks to my best buddy **Sigma**, for always being there waiting for me and ready to chat a little bit.

For last, I want to thank the most important person for me. Thanks to my love and girlfriend **Jasmine**. Thank you, Jasmine, for all the support you have given me during these years, for making this time wonderful and unforgettable, for all the adventures and fun moments we have spent together. Thank you for always being there when I need, for believing in me and for being an amazing partner every day. If these years were amazing, it is mostly thanks to you. I love you.

PS: If you were not mentioned above, thank you ! For reading this thesis, or at least for opening it.

Curriculum Vitae

First Name Federico Maria
Last Name Paruzzo
Date of birth October 2, 1991
Place of birth Turin, Italy
Nationality Italian
Address Chemin des Paleyres 16, 1006 Lausanne (Switzerland)
Email federico.paruzzo@epfl.ch
fedeparuzzo@gmail.com
Phone number +41 78 774 72 52



Professional Experience

- 2015-2019 **Doctoral Assistant, Laboratory of Magnetic Resonance**
Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
Method development for solid-state nuclear magnetic resonance (NMR), with focus on experimental and computational aspects of NMR crystallography. 20% of the time spent as teaching assistant (supervision of master students and teaching of undergraduate chemistry courses).
- 2014-2015 **Undergraduate Research Assistant, Laboratory of Metallurgy and Materials**
University of Turin, Turin, Italy & University of Birmingham, Birmingham, United Kingdom
Experimental study of the diagram phase of the binary system $\text{LiBH}_4\text{-LiNH}_2$.
- 2013-2013 **Undergraduate Research Assistant, Laboratory of Computational Organic Chemistry**
University of Turin, Turin, Italy
Computational study of the reaction pathway between singlet oxygen and graphene model systems.

Education

- 2015-2019 **Ph.D. in Chemistry**
Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
Thesis title: "New Approached to NMR Crystallography", under the supervision of Prof. L. Emsely.
- 2013-2015 **M.Sc. in Chemistry, summa cum laude**
University of Turin, Turin, Italy
Thesis title: "Hydrogen Storage Materials: Phase Diagram Study of the Binary System $\text{LiBH}_4\text{-LiNH}_2$ " under the supervision of Prof. R. Gobetto and Prof. M. Baricco.
- 2010-2013 **B.Sc. in Chemistry, summa cum laude**
University of Turin, Turin, Italy
Thesis title: "Theoretical Study of the reaction pathway between singlet oxygen and graphene models", under the supervision of Dr. G. Ghigo

Industrial Collaborations

2018-2019 **AstraZeneca**

Languages

Italian (First language) English (Proficient) French (Upper Intermediate) German (Elementary)

Certifications

- PRINCE2® Foundation Certificate in Project Management (2018), PeopleCert, No. GR656027344FP.

Trainings

- Advanced Technical Presentation 1: Core Techniques by Damian Conway (2018), Université de Lausanne, Lausanne, Switzerland
- Advanced Technical Presentation 2: Design and Delivery Skills by Damian Conway (2018), Université de Lausanne, Lausanne, Switzerland
- Persuasive Speaking (2018), Université de Lausanne, Lausanne, Switzerland
- Entrepreneurial Opportunity Identification and Exploitation (2018), École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland.
- Body Language (2018), École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland.
- Developing Leadership Skills (2019), École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland.
- Transversal Leadership (2019), École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland.
- Resolving Conflicts (2019), École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland.

Awards and distinctions

- Featured in the EPFL news (2018). <https://actu.epfl.ch/news/ai-and-nmr-spectroscopy-determine-atoms-configure-5/>
- Best Poster Presentation Award at the Swiss Chemical Society Fall Meeting (2018).
- Scholarship “Fondazione Cecilia Gilardi O.N.L.U.S.” (2015). Merit-based scholarship awarded to one student in Chemical Science at the University of Turin to finance a research period abroad.
- Scholarship “Fondazione Saracco” (2013). Merit-based scholarship awarded to 7 students of the University of Turin.

Scientific Publications

- Engel E.A., Anelli, A., Hofstetter, A., Paruzzo, F.M., Emsley, L. and Ceriotti, M., (2019) A Bayesian approach to NMR crystal structure determination, *Phys. Chem. Chem. Phys.*, 21, 23385-23400.
- Paruzzo, F.M. and Emsley, L. (2019) High-Resolution ¹H NMR of Powdered Solids by Homonuclear Dipolar Decoupling, *J. Mag. Res.*, 309, 106598.
- Walder, B.J., Prisco, N.A., Paruzzo, F.M., Yarava, J.R., Chmelka, B.F., Emsley, L. (2019) Measurement of proton spin polarization diffusivity in hydrated cementitious solids, *J. Phys. Chem. Lett.*, 10, 17, 5064-5069.
- Hofstetter, A., Balodis, M., Paruzzo, F.M., Widdifield, C., Stevanato, G., Pinon, A.C., Bygrave, P.J., Day, G.M., Emsley, L. (2019) Rapid Structure Determination of Molecular Solids Using Chemical Shifts Directed by Unambiguous Prior Constraints, *J. Am. Chem. Soc.*, 141, 42, 16624-16634.
- Paruzzo, F.M., Walder B., Emsley, L. (2019) Line Narrowing in ¹H NMR of Powdered Organic Solids with TOP-CT-MAS Experiments at Ultra-Fast MAS, *J. Mag. Res.*, 305, 131.
- Paruzzo, F.M., Hofstetter, A., Musil, F., De, S., Ceriotti, M., Emsley, L. (2018) Chemical Shifts in Molecular Solids by Machine Learning. *Nat. Comm.*, 9(1), 4501.

2. Paruzzo, F.M., Stevanato, G., Halse, M., Schlagnitweit, J., Mammoli, D., Emsley, L. (2018) Refocused Linewidths Less Than 10 Hz in ^1H Solid-State NMR. *J. Mag. Res.*, 293, 41.
1. Leclaire, J., Poisson, G., Ziarelli, F., Pepe, G., Fotiadu, F., Paruzzo, F.M., Rossini, A.J., Dumez, J.N., Elena-Herrmann, B., Emsley, L. (2016). Structure elucidation of a complex CO_2 -based organic framework material by NMR crystallography. *Chem. Sci.*, 7, 4379.

Website Development

1. Paruzzo, F.M., Hofstetter, A., Musil, F., De, S., Ceriotti, M., Emsley, L., (2018, May 3). ShiftML: Chemical Shifts in Molecular Solids by Machine Learning, <https://shiftml.epfl.ch>.

Oral Presentations

4. Paruzzo, F.M., Hofstetter, A., Musil, F., De, S., Ceriotti, M., Emsley, L., "NMR Meets Machine Learning: Chemical Shift Predictions in Solids in Less Than a Minute", *XLVII National Congress on Magnetic Resonance*, Turin, Italy, 19-21 Sept. 2018, (national conference).
3. Paruzzo, F.M., Hofstetter, A., Musil, F., De, S., Ceriotti, M., Emsley, L., "Predicting ^1H and ^{13}C Chemical Shifts of Molecular Crystals with DFT Accuracy Using Machine Learning", *59th Experimental Nuclear Magnetic Resonance Conference (ENC)*, Orlando, Florida (USA), 29 Apr.- 4 May 2018, (international conference).
2. Paruzzo, F.M., Stevanato, G., Halse, M., Schlagnitweit, J., Emsley, " ^1H Refocused Linewidths Less Than 15 Hz in Organic Solids at Natural Abundance", *10th Alpine Conference on Solid State NMR*, Chamonix Mont-Blanc, France, 10-14 Sept. 2017, (international conference).
1. Paruzzo, F.M., Stevanato, G., Halse, M., Schlagnitweit, J., Emsley, "High Resolution ^1H NMR in Organic Solids at Natural Abundance", *SCS Fall Meeting*, Bern, Switzerland, 21-22 Aug. 2017, (national conference)

