Advanced Computational Methods for NMR Crystallography

Thèse N° 7463

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

Albert HOFSTETTER

Acceptée sur proposition du jury

Prof. B. Smit, président du jury
Prof. D. L. Emsley, directeur de thèse
Prof. L. Mueller, rapporteur
Prof. D. Massiot, rapporteur
Prof. A.-C. Corminboeuf, rapporteuse

2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Acknowledgements

To start, I want to thank all the members of my jury, **Dominique Massiot** (external expert), **Leonard Mueller** (external expert), **Clémence Corminboeuf** (internal expert) and **Berend Smit** (president), for taking the time to read my thesis and judge my work. I also want to thank them for coming all the way to the EPFL in Lausanne to attend my private defense and for the stimulating discussions we had during and after the defense.

I want to thank immensely my supervisor **Lyndon Emsley** for giving me the opportunity to work in his lab. This thesis would not have been possible without all the ideas, support and discussion he provided during the four years of my PhD. During this time, he not only thought me all I know about scientific research and integrity but also how to plan, perform and present my research. In that sense, I also want to thank him for all the great group-meetings we had, even for the painfully long ones, during all of which I learned countless valuable lessons.

Also, a big thanks to **Nadia Gauljaux**, who took care of the whole administrative side of my PhD and made my life so much easier. Her help allowed me to completely focus on my research and I'm sure that it would have taken me at least one to two years longer without her.

Further, I would also like to thank all of the members of my group at EPFL. A special thanks to **Federico, Dominik, Brennan, Subba, Baptiste** and **Martins** for the great collaborations and all the fruitful discussion we had. I also want to thank all my other fantastic colleagues: **Claudia, Snædis, Jasmine, Arthur, Gabriele, Pierrick, Andrea, Manuel, Francois, Nicolas, Aditya, Pinelopi, Georges, Mattia, Yu** and **Bruno.** To mention all they did for me during the past four years would fill to many pages. Thus, I just want to generally thank them for some of the best years of my live.

During my time at EPFL, I also had the pleasure to meet several visitors to our lab: **Bradley Chmelka, Michel Bardet** and **Philip Grandinetti.** A big thanks to all of you for the interesting and fruitful discussions and the great times we had. A special thanks to **Brad** for the delicious paella.

In my collaborations I had the pleasure to work with brilliant people from many different research areas. I would like to thank **Michele Ceriotti** and his students and postdocs, **Felix, Sandip, Andrea** and **Edgar** who have been relentless in explaining the concepts of machine learning and scientific computing to me. I also want to thank the groups of **Martin Blackledge, Józef Lewandowski, Michael Grätzel, Paul Bowen and Karen Scriver** for the great collaborations. A special thanks goes to **Abhishek** and **Aslam** for the first real collaboration I had at EPFL and for the great success they turned this work into. Further, I want to thank **Graeme Day, Aaron Rossini** and **Cory Widdifield** for all the insightful discussions and collaborations.

Finally, my biggest thanks goes to **Astrid, Valentin** and **Gabi** as well as my **Mom** and **Dad** who were always there for me. Without your continuous support and believe, I would not be where I am today. A special thanks goes to **Valentin**, for being such a cute and peaceful baby during the last few month of my thesis.

Abstract

Knowledge of the atomic-level structure is key to understanding and predicting properties of materials. X-ray diffraction (XRD) is the methods of choice for structures containing well-defined long-range order. However, many materials contain various degrees of disorder and are thus not characterizable by diffraction methods. In contrast, NMR directly probes local atomic environments and thus allows for structural characterization. In solid-state NMR several types of observables (such as quadrupolar coupling constants, dipole coupling constants, $^1H/^1H$ spin diffusion and chemical shifts) can be used to extract structural information.

In chemical shift driven NMR Crystallography (NMRX) comparisons between experimental and calculated chemical shifts are used to identify the experimental structure from an ensemble of trial structures. The candidate structures are generated either by a comprehensive crystal structure prediction (CSP) search or through searches using different degrees of chemical intuition in combination with constraints extracted from experimental data.

In the present thesis we use chemical shift driven NMRX to investigate materials containing different types of structural disorder, ranging from microcrystalline solids over doped structures up to amorphous materials.

A perfect application for NMRX is the structural determination of drug polymorphs, where the samples are often only available as microcrystalline powders. Here, we investigate a combined CSP-NMRX approach for structure determination of microcrystalline molecular solids. To this end, we first evaluate the positional accuracy of the combined approach. Then, we develop empirical-based methods as well as machine learning algorithms to extend the scope of the CSP-NMRX approach. Finally, we combine the developed methods to determine the crystal structure of powdered ampicillin, for which the traditional approach to CSP-NMRX would have failed.

Another interesting class of structures to investigate with NMRX are amorphous compounds, which are an important component in many industrial devices and materials. Amorphous structures cannot be described by a single crystalline unit-cell, and therefore, the CSP-NMRX approach is no longer applicable. Here, we determine the atomic-level structure of amorphous calcium silicate hydrate by generating a constrained ensemble of local structural motifs using chemical intuition and experimental data. We then evaluate the local structural motifs by comparing calculated and experimental chemical shifts. Finally, we combine the selected local motifs to generate an extended amorphous structural model.

The last applications for NMRX which we investigate are doped structures. Doping is a key technology to design new functional materials with desired properties and has been successfully used in various industrial materials. However, the presence of dopants inevitably leads to disorder within the material. In general, the same approach we investigated for amorphous materials should be applicable. However, the systems analyzed here contain heavy atoms and thus a higher level of theory is required in order to accurately calculate chemical shifts. We investigate different hypothesis for doping mechanisms in a set of photovoltaic lead halide perovskite materials. For these materials, we show that chemical shift based NMRX is able to differentiate between interstitial dopants, surface passivation layers and the formation of segregated phases.

Keywords

solid-state NMR, NMR crystallography, machine learning, pharmaceutical compounds, polymorphism, perovskites, amorphous calcium silicate hydrate, microcrystalline solids, density-functional theory (DFT), crystal structure prediction (CSP)

Abstrakt

Die Kenntnis der atomaren Struktur eines Materials ist fundamental zum Verständnis der Materialeigenschaften. Röntgenstrahlendiffraktion (XRD) ist der Standard zur Analyse von geordneten Strukturen. Viele Materialien beinhalten jedoch diverse Arten von Unordnung und können daher nicht mit diffraktionsbasierten Methoden charakterisiert werden. Als Alternative bietet sich hier die Kernspinresonanzspektroskopie (NMR) an, da diese direkt die lokalen atomaren Umgebungen analysiert. In Festkörper-NMR sind verschiedene Arten von Informationen, z.B. Quadrupole Interaktionen, Dipol-Dipol Interaktionen, Protonen Spin-Diffusion und die chemische Verschiebung, direkt von der atomaren Struktur abhängig und können zur Strukturbestimmung genutzt werden.

Auf chemischer Verschiebung basierte NMR-Kristallographie (NMRX) nutzt den Vergleich zwischen experimentellen und berechneten chemischen Verschiebungen, um aus Teststrukturen die Experimentalstruktur zu bestimmen. Die Teststrukturen werden entweder durch eine vollständige Krystallstrukturvorhersage (CSP), oder durch eine auf experimentellen Daten und Intuition basierten Suche generiert.

In dieser Doktorarbeit nutzen wir auf chemischer Verschiebung basierte NMRX, um Materialien mit verschiedenen Arten von Unordnung zu untersuchen. Die untersuchten Materialien reichen von mikrokristallinen Pulvern über dotierte Strukturen bis hin zu amorphen Materialien.

Die Strukturbestimmung von Arzneistoffpolymorphen, welche oft nur als mikrokristalline Pulver verfügbar sind, ist ein optimales Anwendungsgebiet von NMRX. Hier untersuchen wir eine kombinierte CSP-NMRX Methode. Dazu evaluieren wir zuerst die Positionsgenauigkeit der Methode. Danach entwickeln wir empirische und auf Maschinellem Lernen basierte Algorithmen, um den Umfang der Methode zu erweitern. Zum Schluss kombinieren wir alle entwickelten Ansätze zur Bestimmung der Kristallstruktur von mikrokristallinem Ampicillin.

Amorphe Materialien sind ein wichtiger Bestandteil vieler industrieller Komponenten und Apparaturen. Sie bilden eine weitere interessante Materialklasse, welche mit NMRX untersucht werden kann. Im Gegensatz zu mikrokristallinen Pulvern kann ihre Struktur aber nicht durch ein Kristallgitter definiert werden. Daher ist der CSP-NMRX Ansatz nicht mehr anwendbar. Hier bestimmen wir durch NMRX die atomare Struktur von amorphen Kalzium-Silikat-Hydrat. Dazu nutzen wir chemische Intuition und experimentelle Daten, um ein Ensemble an möglichen lokalen Strukturmotiven zu generieren. Danach bestimmen wir die wahrscheinlichsten Motive durch einen Vergleich der berechneten und experimentellen chemische Verschiebungen. Schliesslich erhalten wir wiederum eine vollständige amorphe Struktur zu erhalten durch die Kombination der ausgewählten Motive.

Zum Schluss untersuchen wir mittels NMRX dotierte Materialien. Dotierung ist eine Schlüsseltechnologie, bei der Spuren von Fremdatomen zu funktionellen Materialien beigemischt werden. Die Anwesenheit von Dotierungen führt aber zwangsläufig zu Unordnung im Material. Im Allgemeinen sollte der gleiche NMRX-Ansatz wie für amorphe Materialien anwendbar sein. Hier untersuchen wir jedoch diverse photovoltaisch aktive Blei-Halogenide-Perowskite; diese beinhalten schwere Atome und müssen daher durch ein höheres Niveau an Theorie beschrieben werden. Für diese Materialien zeigen wir, dass es möglich ist zwischen interstitieller Dotierung, Oberflächen-Passivierung und der Bildung getrennter Phasen zu unterscheiden.

Schlüsselwörter

Festkörper-Kernspinresonanzspektroskopie (NMR), NMR Kristallographie, Maschinelles Lernen, Arzneistoffe, Polymorphie, Perowskit, amorphes Kalzium-Silikat-Hydrat, mikrokristalline Pulver, Dichtefunktionaltheorie (DFT), Krystallstrukturvorhersage (CSP)

Contents

Acknowledgements .... **i**

Abstract .... **ii**

Abstrakt .... **iii**

List of publications .... **vii**

Chapter 1 Introduction **9**

1.1 Structure elucidation of solids 9

1.2 Solid state NMR as probe of local environments 9

1.3 Computational methods for NMR crystallography 11

1.4 Outline of the present thesis 15

Chapter 2 Microcrystalline solids **17**

2.1 Introduction 17

2.2 NMR crystallography directed by unbiased prior constraints 19

2.2.1 Introduction 19

2.2.2 Methods 19

2.2.3 Results and Discussion 22

2.2.4 Conclusion 25

2.2.5 Appendix I 25

2.3 Chemical shifts by machine learning 37

2.3.1 Introduction 37

2.3.2 Results 38

2.3.3 Discussion 44

2.3.4 Methods 45

2.4 Positional variance and uncertainty 63

2.4.1 Introduction 63

2.4.2 Methods 63

2.4.3 Results 66

2.4.4 Conclusion 67

2.4.5 Appendix II 68

2.5 A Bayesian approach to NMRX ... 85
2.5.1 Introduction ... 85
2.5.2 Theory ... 85
2.5.3 Computational Methods ... 88
2.5.4 Results and discussion ... 88
2.5.5 Conclusion ... 93
2.5.6 Appendix III ... 93
2.6 Structure determination of Ampicillin ... 103
2.6.1 Introduction ... 103
2.6.2 Results and Discussion ... 103
2.6.3 Conclusion ... 107
2.6.4 Methods ... 108
2.7 Conclusion and Outlook ... 123
Chapter 3 Amorphous solids ... 125
3.1 Introduction ... 125
3.2 The atomic-level structure of cementitious calcium silicate hydrate ... 127
3.2.1 Introduction ... 127
3.2.2 Methods ... 128
3.2.3 Results and Discussion ... 129
3.2.4 Conclusion ... 135
3.2.5 Appendix IV ... 135
3.3 Conclusion and Outlook ... 161
Chapter 4 Defective and doped solids ... 163
4.1 Introduction ... 163
4.2 Methods ... 165
4.3 Phase Segregation in Cs-, Rb- and K-Doped Mixed-Cation $(MA)_x(FA)_{1-x}PbI_3$ Hybrid Perovskites ... 173
4.3.1 Introduction ... 173
4.3.2 Methods ... 174
4.3.3 Results and Discussion ... 175
4.3.4 Conclusion ... 181
4.3.5 Appendix V ... 181
4.4 Phase Segregation in Potassium-Doped Lead Halide Perovskites ... 189
4.4.1 Introduction ... 189
4.4.2 Methods ... 191
4.4.3 Results and discussion ... 191
4.4.4 Conclusion ... 196
4.4.5 Appendix VI ... 196

4.5 Conclusion and Outlook 213

Chapter 5 Conclusion 215

5.1 Achieved results 215

5.2 Future development. 217

Bibliography 219

Curriculum Vitae 237

List of publications

The present thesis is based on the following publications:

1. Engel, E.A.; Anelli, A.; Hofstetter, A.; Paruzzo, F.; Emsley, L.; Ceriotti, M., "A Bayesian approach to NMR crystal structure determination". *Submitted* **2019**. ***(pre-print)***

2. Hofstetter, A.; Balodis, M.; Paruzzo, F.M.; Widdifield, C..M.; Stevanato, G.; Pinon, A.C.; Bygrave, P.; Day, G.M.; Emsley, L., "Rapid Structure Determination of Molecular Solids Using Chemical Shifts Directed by Unambiguous Prior Constraints". *Journal of the American Chemical Society* **2019**, XXXX, XXX. ***(pre-print)***

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

4. Kubicki, D. J.; Prochowicz, D.; Hofstetter, A.; Zakeeruddin, S. M.; Grätzel, M.; Emsley, L., "Phase Segregation in Potassium-Doped Lead Halide Perovskites from ^{39}K solid-state NMR at 21.1 T". *Journal of the American Chemical Society* **2018**, 140 (23), 7232-7238. ***(post-print)***

5. Kumar, A.; Walder, B. J.; Kunhi Mohamed, A.; Hofstetter, A.; Srinivasan, B.; Rossini, A. J.; Scrivener, K.; Emsley, L.; Bowen, P., "The atomic-level structure of cementitious calcium silicate hydrate". *The Journal of Physical Chemistry C* **2017,** *121* (32), 17188-17196. ***(post-print)***

6. Kubicki, D. J.; Prochowicz, D.; Hofstetter, A.; Zakeeruddin, S. M.; Grätzel, M.; Emsley, L., "Phase Segregation in Cs-, Rb-and K-Doped Mixed-Cation $(MA)_x(FA)_{1-x}PbI_3$ Hybrid Perovskites from Solid-State NMR". *Journal of the American Chemical Society* **2017,** *139* (40), 14173-14180. ***(post-print)***

7. Hofstetter, A.; Emsley, L., "Positional variance in NMR crystallography". *Journal of the American Chemical Society* **2017,** *139* (7), 2573-2576. ***(post-print)***

During my PhD I also have been working on other computational aspects of solid-state nuclear magnetic resonance. Most notably this includes dynamical and structural investigation of Perovskite systems and Proteins at various temperatures. These results are not reported in the present thesis. However, they can be found in the following publications.

1. Ruiz-Preciado, M.A.; Kubicki, D.J.; Hofstetter, A.; Ummadisingu, A.; Gershoni-Poranne, R.; Zakeeruddin, S.M.; Emsley, L.; Milic, J.V.; Grätzel, M., "Supramolecular Modulation of Hybrid Perovskite Solar Cells via Bifunctional Halogen Bonding Revealed by Two-Dimensional ^{19}F Solid-State NMR Spectroscopy". *Submitted* **2019**.

2. Xiang, W.; Wang, Z.; Kubicki, D.J.; Tress, W.; Luo, J.; Wang, X.; Zhang, J.; Hofstetter, A.; Zhang, L.; Emsley, L.; Grätzel, M.; Hagfeldt, A., "Ba-induced phase segregation and band gap reduction in mixed-halide $CsPbI_2Br$ for inorganic perovskite solar cells". *Nature Communications* **2019**.

3. Kubicki, D.J.; Prochowicz, D.; Pinon, A.; Stevanato, G.; Hofstetter, A.; Zakeeruddin, S.M.; Grätzel, M.; Emsley, L., "Doping and phase segregation in Mn^{2+}-and Co^{2+}-doped lead halide perovskites from ^{133}Cs and ^{1}H NMR relaxation enhancement". *Journal of Materials Chemistry A* **2019**, 7, 2326**.**

4. Kubicki, D. J.; Prochowicz, D.; Hofstetter, A.; Saski, M.; Yadav, P.; Bi, D.; Pellet, N.; Lewiński, J.; Zakeeruddin, S. M.; Grätzel, M., "Formation of Stable Mixed Guanidinium–Methylammonium Phases with Exceptionally Long Carrier Lifetimes for High-Efficiency Lead Iodide-Based Perovskite Photovoltaics". *Journal of the American Chemical Society* **2018,** *140* (9), 3345-3351.

5. Busi, B.; Yarava, J. R.; Hofstetter, A.; Salvi, N.; Cala-De Paepe, D.; Lewandowski, J. R.; Blackledge, M.; Emsley, L., "Probing Protein Dynamics Using Multifield Variable Temperature NMR Relaxation and Molecular Dynamics Simulation". *J Phys Chem B* **2018,** *122* (42), 9697-9702.

6. Kubicki, D. J.; Prochowicz, D.; Hofstetter, A.; Pechy, P.; Zakeeruddin, S. M.; Grätzel, M.; Emsley, L., "Cation Dynamics in Mixed-Cation $(MA)_x(FA)_{1-x}PbI_3$ Hybrid Perovskites from Solid-State NMR". *Journal of the American Chemical Society* **2017,** *139* (29), 10055-10061

.

Chapter 1 Introduction

1.1 Structure elucidation of solids

Structure-activity relationships play a central role in chemistry and material science. Determining three-dimensional atomic-level structures is key to understanding and predicting properties and for ultimately designing new functional materials. Many molecules and materials have been characterized by single crystal X-ray (XRD)[1] diffraction and to a lesser extent by neutron[2-3] or electron diffraction.[4-7] However, a wide range of materials are unavailable as single crystals (e.g., composites, amorphous and glassy materials, disordered and doped materials, fine powders, formulated samples and fibrils) and are thus not characterizable by diffraction methods. Examples include active pharmaceutical ingredients (APIs), which are often only available as microcrystalline powders. The determination of their structures and crystal packings is essential to extract structure-property relations for formulations in the drug development process. Also, the optical, electronic, magnetic and energetic properties in amorphous materials and doped solids crucially depends on their intrinsic disorder. For all these materials, the characterization of the atomic level structure using diffraction is extremely challenging, due to the lack of long-range order. In contrast, solid-state nuclear magnetic resonance spectroscopy (NMR) directly probes the local atomic environments and thus allows for characterization without the need for long-range order. This has led to its broad use today in many fields, including materials and pharmaceutical chemistry.[8-59]

1.2 Solid state NMR as probe of local environments

NMR spectroscopy detects the motion of nuclear magnetic moments ($\vec{\mu}$) which are associated to the nuclear spin (I), an intrinsic nuclear property through,

$$\vec{\mu} = \gamma \hbar I, \tag{1-1}$$

where γ is the gyromagnetic ratio of the nuclei and $\hbar$ is the reduced Planck's constant. The energy of the nuclear spin (I) is in turn described by the NMR Hamiltonian, which can contain up to 13 interactions. However, most of them are usually zero or are not observable. The relevant terms for solid-state NMR are usually given by **Equation 1-2**.

$$\mathcal{H}_{NMR} = -\hbar \sum_I \gamma_I B_{ext} \left(\bar{\bar{1}} - \bar{\bar{\sigma}}\right) I_I + \frac{1}{2} \hbar^2 \sum_I \sum_{J \neq I} \gamma_I \gamma_J I_I \left(\bar{\bar{D}}_{IJ} + \bar{\bar{J}}_{IJ}\right) I_J + \sum_{I, |I_I| \geq 1} I_I \bar{\bar{Q}}_I I_I. \tag{1-2}$$

The first term describes the interaction of the nuclear spin operator (I_I) with an external magnetic field (B_{ext}) (Zeeman and chemical shift / magnetic shielding ($\bar{\bar{\sigma}}$) interaction). The second term describes the interaction between two different nuclear spins (I_I and I_J) either through space (through the nuclear magnetic dipolar coupling, $\bar{\bar{D}}_{IJ}$) or through chemical bonds (through the indirect nuclear spin-spin coupling, $\bar{\bar{J}}_{IJ}$). The third term describes the interaction between a spin and an electric field gradient (through the nuclear quadrupolar coupling, $\bar{\bar{Q}}_I$).

In principle structural information can be extracted from all three of the terms in the NMR Hamiltonian. However, here we are going to focus mainly on the structural information contained in the chemical shift interaction. (In **Chapter 2.1** we present a method to extract complementary structural information from dipolar coupling interactions.) The interaction of an isolated nuclear spin with an external magnetic field (B_{ext}) is described through the Zeeman interaction, given as,

$$\mathcal{H}_{Zeeman} = -\hbar \gamma_I B_{ext} I_I.$$

(1-3)

However, the chemical systems investigated here do not consist of isolated spins. Instead the nuclear spins are embedded in an electronic charge density which is determined to 1st order by the investigated nuclei, to 2nd order by the adjacent and bonded nuclei (or by the molecule the nuclei belong to) and to 3rd order by the crystal packing surrounding the investigated nuclei. Additionally, the electronic charge density is also influenced through charges and currents applied over the structure as well as excitations within the crystal. Following Lenz's law, the external magnetic field will induce a current in the electronic density around the nuclei, which according to Biot-Savart's law (**Equation 1-4**) will lead to an induced magnetic field opposing the external magnetic field.

$$B_{\mathrm{ind}}(\mathrm{r}) = \frac{\mu_0}{4\pi} \int_{\mathrm{C}} j(r') \times \frac{\mathrm{r} - \mathrm{r}'}{|r - r'|^3} \mathrm{dr}',$$

(1-4)

where $j(r')$ describes the induced current density at point r' on the closed curve C around the reference point r.

The direct relation between the external magnetic field (B_{ext}) and the locally induced magnetic field (B_{ind}) is given by the magnetic shielding tensor ($\bar{\bar{\sigma}}$), as:

$$B_{\mathrm{ind}} = -\bar{\bar{\sigma}} B_{\mathrm{ext}}.$$

(1-5)

Additionally, for a given nucleus *A* the chemical shielding tensor can be expressed as the 2nd derivative of the electronic energy with respect to the *i*-th component of the external magnetic field (B_i) and the *j*-th component of the nuclear magnetic moment of nucleus *A* (μ_j^A) (**Equation 1-6**). Note, that from this expression the gauge problem becomes apparent, as the external magnetic field appears as a vector potential without fixed origin. This problem is overcome by using so-called gauge invariant or gauge including calculation formalisms (GIPAW and GIAO). [60-63]

$$\sigma_{ij}^A = \frac{\partial^2 E}{\partial B_i \partial \mu_j^A}$$

(1-6)

However, in NMR experiments the magnetic shielding is not measured directly. Instead the relative shielding (or respectively the deshielding) of a nucleus with respect to a fixed reference value ($\bar{\bar{\sigma}}_{ref}$) is measured. This referenced shielding is the chemical shift ($\bar{\bar{\delta}}$) and is given as:

$$\bar{\bar{\delta}} = \bar{\bar{\sigma}}_{ref} - b\,\bar{\bar{\sigma}},$$

(1-7)

where the slope (b) should be fixed at unity for an ideal case, but is typically used to account for systematic errors within calculations, including incomplete basis sets and nuclear quantum effects.[64] From the description above it becomes clear how the motion of nuclear magnetic moments, which is measured in an NMR experiment, depends on the effective magnetic field. The effective magnetic field is given by the response of the local electronic density to an external magnetic field through the chemical shift tensor, which is in turn determined by the local atomic environment. In conclusion, the chemical shifts of a structure are uniquely determined by its electronic structure and thus by the crystal structure. **Therefore, the full structural information of the local environments is contained within the chemical shifts and should be accessible by NMR.**

1.3 Computational methods for NMR crystallography

Initially, structural constraints from solid-state NMR were used to refine diffraction structures and to obtain information for a few ill-defined atomic environments within a given diffraction structure.[41, 55, 65] Additionally, solid-state NMR has also been used as a complementary tool in crystallographic studies to determine the tautomeric form present,[66] to locate regions of disorder[67] or to provide key distance measurements.[51, 68] More recent developments in solid-state NMR and complementary computational methods have led to the point where full atomic level structures can be determined using only NMR without any type of diffraction data.[26, 54, 58] This recently emerged field is now often referred to as NMR crystallography (NMRX).

Most commonly, in NMRX structures have been characterized and / or determined using selected distance constraints extracted via dipolar couplings,[26, 28, 47, 69] quadrupolar couplings[59, 70] or ^{1}H/^{1}H spin diffusion.[46] However, already 1993 Facelli and Grant[71] have shown that the chemical shift tensor contains sufficient information to determine molecular symmetry in crystalline solids. Additionally, de Dios *et al.*[48] and Harper *et al.*[9] and have shown in 1993 and 2001 that the information contained in the chemical shifts is sufficient to characterize the secondary and tertiary structure of proteins as well as the stereochemistry and conformation of molecular solids. Further, in 2006, 2009 and 2011 Harris *et al.*[72], Salager *et al.*[73] and Abraham *et al.*[74] demonstrated that calculated chemical shifts are sufficiently accurate to assign chemical shifts from experimental solid-state NMR and to differentiate between different polymorphs. These developments have given rise to the field of chemical shift based NMRX. Due to the relative simplicity in extracting experimental chemical shifts together with the strong dependence of chemical shifts on the local atomic environment (see **Chapter 1.2**) as well as the progress being made both in measuring as well as in calculating accurate chemical shifts, the scope of chemical shift driven NMRX has steadily increased in the past years.

A large step towards extracting the structural information contained in the chemical shift space has been taken with the development of accurate computational methods to predict chemical shifts[48, 60-63, 75-84] of single molecules as well as extended structures (**Figure 1-1**), as described further below. However, there is currently no method to directly transform the chemical shift information into atomic-level structures, as a direct and simple analytical expression linking the chemical shifts back to the atomic structure does not exist. Instead chemical shift driven NMRX is based on the generation of reasonable structural hypotheses and / or structural models, either by a comprehensive structure search[13, 26, 35, 56, 58] or through searches using different degrees of intuition in combination with experimental constraints.[12, 22, 25, 51, 54, 57, 71, 85] The structural models are then validated by comparing calculated and experimental chemical shifts.[86]

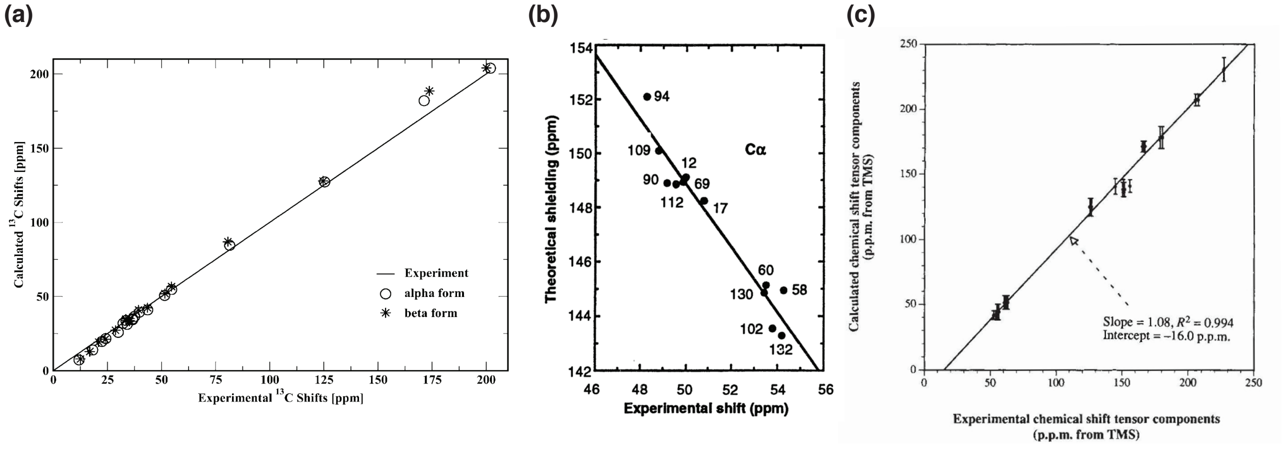


Figure 1-1. Correlation between DFT calculated and experimental ^{13}C chemical shifts of the α and β forms of testosterone adapted with authorization from Harris *et al.*[72] (copyright 2006 Royal Society of Chemistry) **(a),** of the Cα carbons of the 12 Ala sites in SNase as adapted with authorization from de Dios *et al.*[48] (copyright 1993 AAAS) **(b)** and of naphthalene as adapted with authorization from by Facelli and Grant[71] (copyright 1993 Nature Publishing Group).

The result of this indirect structure determination approach means that the power of NMRX critically depends on the methods used to generate the structural models as well as on the number of structural models for which chemical shifts can be calculated with reasonable computational cost. Further, both of these points strongly depend on the type of investigated system. For example, the ^{1}H, ^{13}C and ^{15}N chemical shifts of proteins are typically calculated using statistical[87-94] or machine learning[95-97] approaches based on large experimental databases. This allows for the screening of thousands of structural models and has met with considerable success in predicting local sequences and structural motifs.[15, 98-100]

However, for most materials such database approaches do not exist or are currently being developed[101-103] (as discussed in **Chapter 2.2**). For these systems, the development of accurate *ab-initio* methods to calculate chemical shifts,[104] in particular using plane wave density functional theory (DFT) methods based on the gauge including projected augmented wave (PAW/GIPAW) approach[62-63, 81] as well as fragment based DFT methods[83-84] in combination with the Gauge-Independent Atomic Orbital (GIAO) method,[60-61, 75, 77, 82] has greatly contributed to the success of NMRX. For a wide range of organic and inorganic materials, such as molecular solids, graphite, silicates, zeolites and oxides, the DFT errors on the isotropic chemical shift values are around 1-2% of the chemical shift range of the investigated nucleus. Example nuclei include, but are not limited to, ^{1}H, ^{13}C, ^{15}N, ^{17}O, ^{19}F, ^{27}Al, ^{29}Si and ^{43}Ca.[18, 36, 70, 83, 105-112] **Figure 1-1** shows the correlation between the DFT calculated and experimental ^{13}C chemical shifts for a set of reference structures as given by Harris *et al.*,[72] de Dios *et al.*[48] and Facelli and Grant.[71]

Figures 1-2 and **1-3** show the ^{1}H, ^{13}C and ^{19}F DFT isotropic chemical shift accuracy that can be obtained today for a set of example crystal structures. The investigated crystal structures were obtained as described in **Table 1-1** and optimized using plane-wave DFT as described in **Chapter 2**. The ^{1}H, ^{13}C and ^{19}F chemical shifts were calculated using plane-wave DFT as described in **Chapter 2**. The experimental ^{13}C and ^{1}H chemical shifts were acquired as described in **Chapter 2**. The ^{19}F chemical shifts were obtained as described in **Table 1-1**. Using the shielding (σ) to shift (δ) conversion, given in **Equation 1-7,** we obtain a chemical shift root-mean-square error (RMSE) of 0.42 ppm for ^{1}H, 2.49 ppm for ^{13}C and 2.96 ppm for ^{19}F.

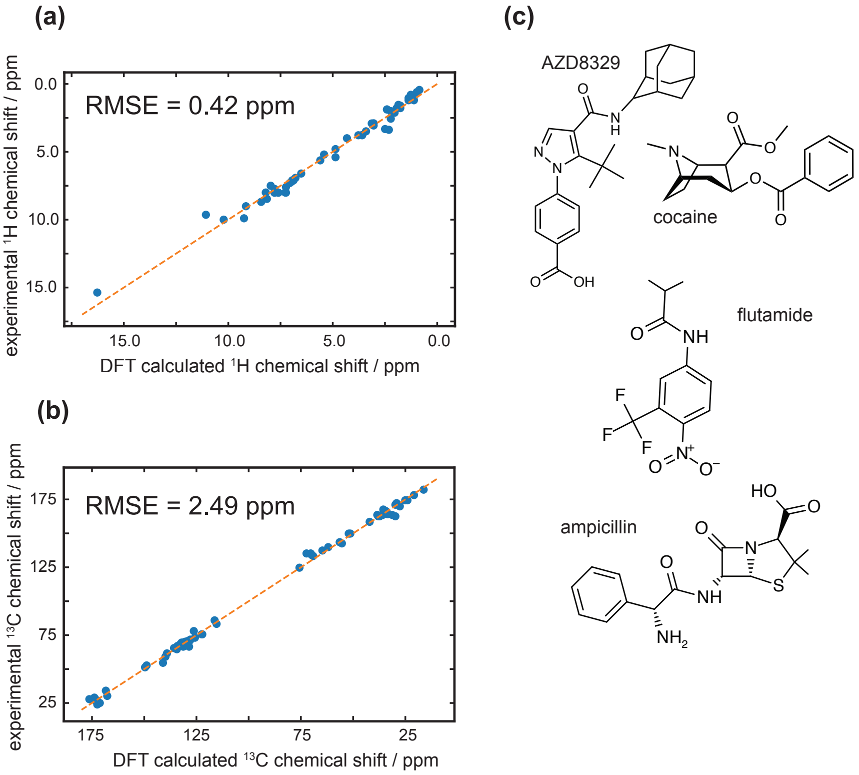


Figure 1-2. Correlation between DFT calculated and experimental ^{1}H **(a)** and ^{13}C **(b)** chemical shifts for a set of example crystal structures **(c)**, which are given in **Table 1-1**. For ^{1}H we obtain a chemical shift RMSE of 0.42 ppm and a slope (b) of 0.912. For ^{13}C we obtain a chemical shift RMSE of 2.49 ppm and a slope (b) of 0.962. In **(a)** and **(b)** the dotted orange line indicates a perfect linear correlation.

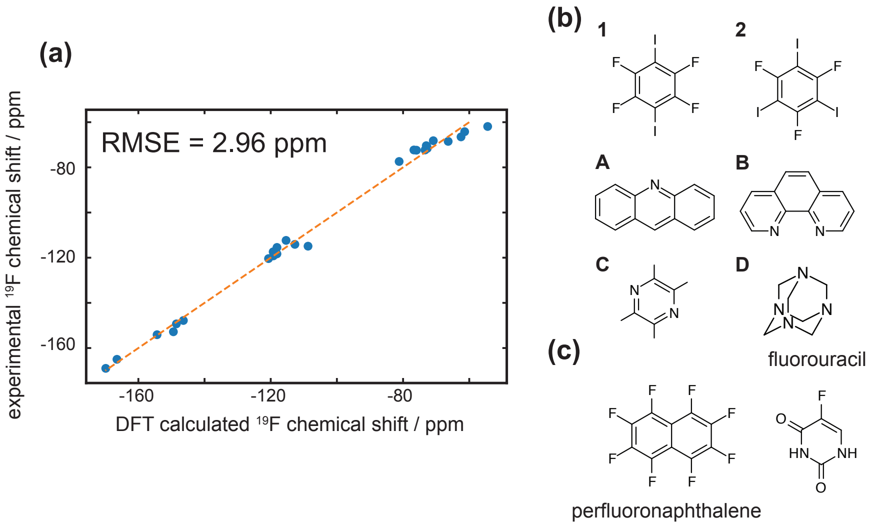


Figure 1-3. Correlation between DFT calculated and experimental ^{19}F chemical shifts **(a)** for a set of example crystal structures **(b, c)**, which are given in **Table 1-1**. Panel **(b)** shows the donors: 1,4-diiodotetrafluorobenzene (p-DITFB, 1) and 1,3,5-trifluoro-2,4,6-triiodobenzene (sym-TFTIB, 2) and acceptors: acridine (ACD, A), 1,10-phenanthroline (PHN, B), 2,3,5,6-tetramethylpyrazine (TMP, C), and hexamethylenetetramine (HMT, D) for the investigated cocrystals. For ^{19}F we obtain a chemical shift RMSE of 2.96 ppm and a slope (*b*) of 0.913. In **(a)** the dotted orange line indicates a perfect linear correlation.

For microcrystalline powders and amorphous materials, the power of chemical shift based NMRX arises from the fact that DFT is today accurate enough to reproduce the exquisite sensitivity of chemical shifts to changes in local atomic environments. However, compared to database approaches for chemical shift predictions, the computational cost of DFT chemical shift calculations prevents the extensive screening of the structural landscape. Therefore, NMRX is often combined with structure selection algorithms to identify relevant structural motifs and regions. Here, the choice of the selection algorithm critically depends on the type of structure present.

Microcrystalline powders of 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. Here, the relatively small size of the crystals in powders limits the diffraction approach, whereas the atomic level structure is still uniquely determined through the single-crystal parameters. Thus, NMRX has been combined with crystal structure prediction[113] (CSP) protocols to generate a set of trial crystal structures, which are then evaluated by comparing DFT calculated and experimental chemical shifts, to determine de novo crystal structures from powders.[34-35, 38, 44, 58, 114] However, CSP and accurate DFT chemical shift calculations still require considerable computational resources thus limiting the combined approach to relatively small systems. Additionally, errors and uncertainties, both for the full structural model and for individual atomic positions, are not determined by CSP-NMRX and in that sense the structures remain just models.

For amorphous and doped materials, disorder is present on a more local level and the atomic-level structure is not uniquely determined by the single crystal parameters. Additionally, the disorder leads to a distribution in the observable chemical shifts, which generally makes it more challenging to extract the structural information contained in the experimental NMR spectra. Thus, the CSP-NMRX approach described above is not applicable. Here, NMRX can be combined with large-scale molecular-dynamics (MD) simulations to generate an ensemble of trial structural motifs, which can then be evaluated by comparing DFT calculated and experimental chemical shifts.[40, 115-121] However, the large-scale structures generated in the MD simulations are too large for DFT chemical shift calculations. Thus, an approach has to be developed to generate representative structural fragments and motives, which are amendable for accurate chemical shift calculations.

NMRX is further complicated for materials containing heavy atoms, e.g., nuclei heavier than the 5th row of the periodic table. For these systems it has been shown that a full relativistic treatment of the electronic density has to be considered.[122-127] This prohibits the use of periodic DFT calculations and leads to a drastic increase in the required computational resources, which in turn strongly limits the number of computationally investigable trial structures. Therefore, NMRX for these systems is often limited to the evaluation of relatively general structural hypotheses. However, in many cases the local structural information extracted from the NMRX evaluations complements the information which can be extracted from other characterization methods and thus can lead to novel structural insights.

Note that, open-shell and paramagnetic systems as well as metallic materials further complicate DFT chemical shift calculations.[128-130] These systems are not investigated here. However, similar considerations as in **Chapter 4** must be considered.[16, 20, 33]

Table 1-1. List of references for the crystal structure coordinates and ^{1}H, ^{13}C and ^{19}F experimental chemical shifts of the investigated crystal structures.

Structure	CSD Refcodes / structure reference	Experimental chemical shift reference
Cocaine	COCAIN10	**Chapter 2.**
AZD8329	CCDC 957764	**Chapter 2.**
Flutamide	WEZCOT	**Chapter 2.**
Ampicillin	AMCILL	**Chapter 2.**
Fluorouracil	FURACL	Viger-Gravel *et al.*[131]
Perfluoronaphtalene	OFNAPH01	Robbins *et al.*[14]
1 (*p*-DITFB)	CCDC 819337	Szell *et al.*[36]
2 (*sym*-TFTIB)	CCDC 293751	Szell *et al.*[36]
A1 (ACD-DITFB)	CCDC 712048	Szell *et al.*[36]
A2 (ACD-TFTIB)	Szell *et al.*[36]	Szell *et al.*[36]
B1 (PHN-DITFB)	CCDC 259705	Szell *et al.*[36]
B2 (PHN-TFTIB)	Szell *et al.*[36]	Szell *et al.*[36]
C1 (TMP-DITFB)	CCDC 259702	Szell *et al.*[36]
C2 (TMP-TFTIB)	Szell *et al.*[36]	Szell *et al.*[36]
D1 (HMT-DITFB)	CCDC 161327	Szell *et al.*[36]
D2 (HMT-TFTIB)	CCDC 1018109	Szell *et al.*[36]

1.4 Outline of the present thesis

In this chapter I have presented an overview of NMR as probe of local environments and how solid-state NMR can be used for structure elucidation of materials which are not amendable by diffraction-based methods. Additionally, I have briefly discussed the applicability and the current limitations of NMRX to materials containing different degrees of disorder. The focus of my PhD has been the application and development of computational methods for NMRX. In the following chapters I will present selected results on method development and applications of NMRX for microcrystalline molecular solids, amorphous materials and doped systems containing heavy atoms.

Chapter 2 describes chemical shift based NMRX in combination with CSP for the atomic-level structure determination of microcrystalline molecular solids. We investigate the current limitations of CSP-NMRX with respect to the computational cost, the structure selection confidence and the structural uncertainty. For this, we investigate the positional accuracy of the combined CSP-NMRX approach and we develop machine learning and empirically based methods to extend the scope of NMRX for molecular crystals. Furthermore, we combine the presented methods to correctly determine the atomic-level structure, including positional uncertainties, of microcrystalline ampicillin with up to 95% confidence.

Chapter 3 discusses the application of NMRX to amorphous materials. We determine the atomic-level structure of amorphous calcium silicate hydrate using NMRX. In contrast to the comprehensive CSP based approach for molecular crystals, we use experimental data and chemical intuition to generate a constrained ensemble of local structural motifs. The individual motifs are then evaluated by comparing their calculated ^{1}H and ^{29}Si chemical shifts to experiment. Further we use MD simulations to verify the stability of the proposed structures.

Chapter 4 investigates the atomic-level nature of doped materials containing heavy atoms. Here, we propose and evaluate a set of possible doping mechanism using NMRX in combination with other characterization methods. We investigate the doping mechanism for different cation dopants (^{39}K, ^{133}Cs and ^{87}Rb) in hybrid organic-inorganic multi-cation lead halide perovskites.

Chapter 5 summarizes the achieved results and presents a general outlook on advanced computational methods for NMRX.

Chapter 2 Microcrystalline solids

2.1 Introduction

The 40,000-60,000 crystal structures published every year[132-135] perfectly illustrate the importance of the knowledge of atomic level structures of solids. In pharmaceutical compounds, crystal structures guide the understanding of physicochemical and pharmacokinetic properties such as bioavailability or solubility.[136] However, many active pharmaceutical ingredients (APIs) are only available as powders that are not amenable to resolution with X-ray diffraction methods if, for example, they are sub-micron in size, or they contain elements of disorder.

For microcrystalline powders of molecular solids NMRX often involves CSP[113] protocols to generate reliable trial crystal structures (see **Chapter 1.3**) and has already been used to determine *de novo* crystal structures from powders[34-35, 38, 44, 58, 114] as well as to determine elements of structure such as hydrogen bonding, proton positions and stereochemistry,[9-10, 13, 137-139] to validate and refine crystal structures of molecular solids, or to identify known polymorphs.[8, 12-13, 17-18, 24, 29-32, 37, 39, 47, 56, 73, 83-84, 138, 140-149]

However, CSP requires considerable computational resources, which increases rapidly with the structural degrees of freedom. Thus, CSP based NMRX (CSP-NMRX) for *de novo* determination is currently limited to systems with up to about 10 degrees of torsional freedom within the molecule,[150] and going beyond this requires some prior knowledge or intuition.[35, 114] Indeed, in order to circumvent these limitations CSP methods often make assumptions based on space groups or predicted conformational energies for example to help limit the search space of possible structures. However, this can lead to failure of the CSP-NMRX method to determine the crystal structures when the correct structure is excluded from the search space.

A common feature of CSP-NMRX methods developed to date is that they exploit structural constraints from solid-state NMR only in the final step, 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.

In **Chapter 2.2**, we introduce a CSP-NMRX method to determine crystal structures in which we use unambiguous constraints from solid-state NMR on microcrystalline samples to restrict the CSP search space to the 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.

The power of the CSP-NMRX method for molecular solids 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 (see **Figures 1-1** to **1-3**). However, this approach also has severe limitations such as 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.[151-155] Notably, prediction of chemical shifts for the specific case of proteins in solution using methods based on large experimental databases, with traditional[87-94] or machine learning approaches,[95-97] have met with considerable success in predicting shifts based on local sequence and structural motifs, and are widely used today (see **Chapter 1.3**). While there are some examples of machine learned experimental and ab-initio chemical shifts of liquid and gas phase molecules,[156-160] at the start of this work 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.[101] 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 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 **Chapter 2.3**, we develop a machine learning framework to predict chemical shifts in solids which is based on capturing the local environments of individual atoms, and thus suitable for the prediction of local properties such as chemical shifts. Most significantly, even though no experimental shifts were used in training, we show that the model has sufficient accuracy to be used in a chemical shift driven CSP-NMRX protocol to correctly determine, based on the match between experimentally-measured and ML-predicted shifts, the correct structure of cocaine, and the drug 4-[4-(2-adamantylcarbamoyl)-5-tert-butylpyrazol-1-yl]benzoic acid (AZD8329). We also show that this method allows to calculate the NMR spectrum of very large molecular crystals, which cannot be calculated using DFT.

An additional key difference between NMR and XRD crystallographic methods is that there exists no protocol to quantify the positional errors on individual atoms for structures determined by chemical shift-based NMRX.

In **Chapter 2.4**, we introduce a method, based on MD, DFT and machine learning methods, to estimate the correlation between the root mean squared deviation (RMSD) of the experimental and calculated chemical shifts, and the variances of atomic positions of individual atoms in structures determined by CSP-NMRX, thereby making them directly comparable to structures determined by other methods. The approach is demonstrated on multiple crystal structures recently characterized by CSP-NMRX.[56, 58, 142]

While usually sufficiently accurate, DFT chemical 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.[18] Conventionally candidates are therefore considered to be consistent with experiment if the RMSE of their shifts from the experimentally measured values falls within these uncertainties. However, this approach is severely limited. It neither allows determination of the experimental structure when multiple candidates exhibit similar RMSEs 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.

In **Chapter 2.5**, we propose a probabilistic approach to overcome these limitations in the evaluation of candidate structures in chemical shift based NMRX. Whereas previously, structures were considered either in agreement or not with the data, this method allows one 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. We demonstrate the method on structures determined with different levels of confidence. As a demonstration of the capabilities of the method, we combine experimental NMR data with DFT and ML predictions of the shifts of a set of CSP candidates to determine the confidence in the structure determination of five different molecular crystals.

In **Chapter 2.6**, we combine the unambiguous prior constraints for CSP-NMRX together with chemical shifts calculated with both DFT and ML[161] as well as the Bayesian approach to correctly determine the full crystal structure, including positional uncertainties, of powdered ampicillin with up to 95% confidence, for which the usual approach to CSP-NMRX would have failed.

2.2 NMR crystallography directed by unbiased prior constraints

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.; Day, G.M.; Emsley, L., "Rapid Structure Determination of Molecular Solids Using Chemical Shifts Directed by Unambiguous Prior Constraints". *Journal of the American Chemical Society* **2019**, XXXX, XXX. ***(pre-print)***

2.2.1 Introduction

The CSP-NMRX approach (see **Chapter 1.3**) involves the combination of crystal structure prediction methods, ab-initio calculated chemical shifts and solid-state NMR experiments and is a powerful tool for crystal structure determination of microcrystalline powders.[34-35, 38, 44, 58, 114] However, currently structural information obtained from solid state NMR is usually included only after a set of candidate crystal structures has already been independently generated, starting from a set of single molecule conformations. Here, we show that this can lead to failure of the structure determination. We thus propose a crystal structure determination method that includes experimental constraints already during conformer selection. To overcome the problem that experimental measurements on the crystalline samples are not obviously translatable to restrict the single molecule conformational space, we propose constraints based on the analysis of absent cross-peaks in solid-state NMR correlation experiments. We show that these absences provide unambiguous structural constraints on both the crystal structure and the gas phase conformations, and therefore can be used for unambiguous selection. The approach is parameterized on the crystal structure determination of flutamide, flufenamic acid, and cocaine, where we reduce the computational cost by around 50%.

2.2.2 Methods

Figure 2-1a schematically illustrates the workflow in a successful case for the current CSP-NMRX approaches.[56-58] In the first step, the torsional degrees of freedom are explored to generate a comprehensive ensemble of energetically stable single molecule conformers. The 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 an empirical cut-off energy. Although flexible molecules often do not assume their lowest energy molecular conformation in their observed crystal structures,[162] the assumption here is that low energy crystal structures, including the correct (observed) polymorph, will generally result from low energy molecular conformers. However, this is not always the case, as will be demonstrated in **Chapter 2.6** below.

The selected conformations are then each subjected to a crystal structure search, during which trial structures are generated by varying the unit cell dimensions, molecular positions, packing symmetry, and the number of molecules per asymmetric unit, leading to hundreds or thousands of possible crystal structures from each single molecular conformer. The energy of each structure is then minimized, typically using atom-atom force fields and DFT.[113]

Next, this ensemble is ranked by calculated lattice energy and again only the structures below a given cut-off energy are retained. In the final step, these structures are further optimized, typically using periodic boundary DFT calculations, and then the chemical shifts (or other experimental data such as dipolar couplings or chemical shift anisotropies)[13, 35, 85, 149, 163-164] for this sub-ensemble of crystal structures are calculated and compared to experimental chemical shifts measured on a powder sample. The error between the calculated and the experimental chemical shift data is then used to determine the unique crystal structure present in the powder. Note, that the computational cost rises sharply when moving from the energy calculations of a single molecule to lattice energy calculations to DFT GIPAW chemical shift calculations, thus requiring the use of successive selection steps to reduce the number of candidate structures at each stage.

From the description of the NMRX procedure above, it is evident that a gas phase conformer similar to the one present in the correct crystal structure must be among those initially selected.

Figure 2-1b illustrates a case where the current CSP-NMRX method fails. Analogously to the previous case, 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 too energetically unfavorable in the gas phase, thus failing to pass the selection criteria 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 structure search, and consequently 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 in two ways. Options are to loosen the initial selection criteria, thus allowing more conformers to proceed to the following steps, increasing the computational cost, usually prohibitively, or to use a different initial selection criterion including information from experiment.

Figure 2-1c 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 guarantees that the conformational sub-ensemble selection is guided towards the correct crystal conformer, and thus that the structure determination is not limited by possibly erroneous assumptions.

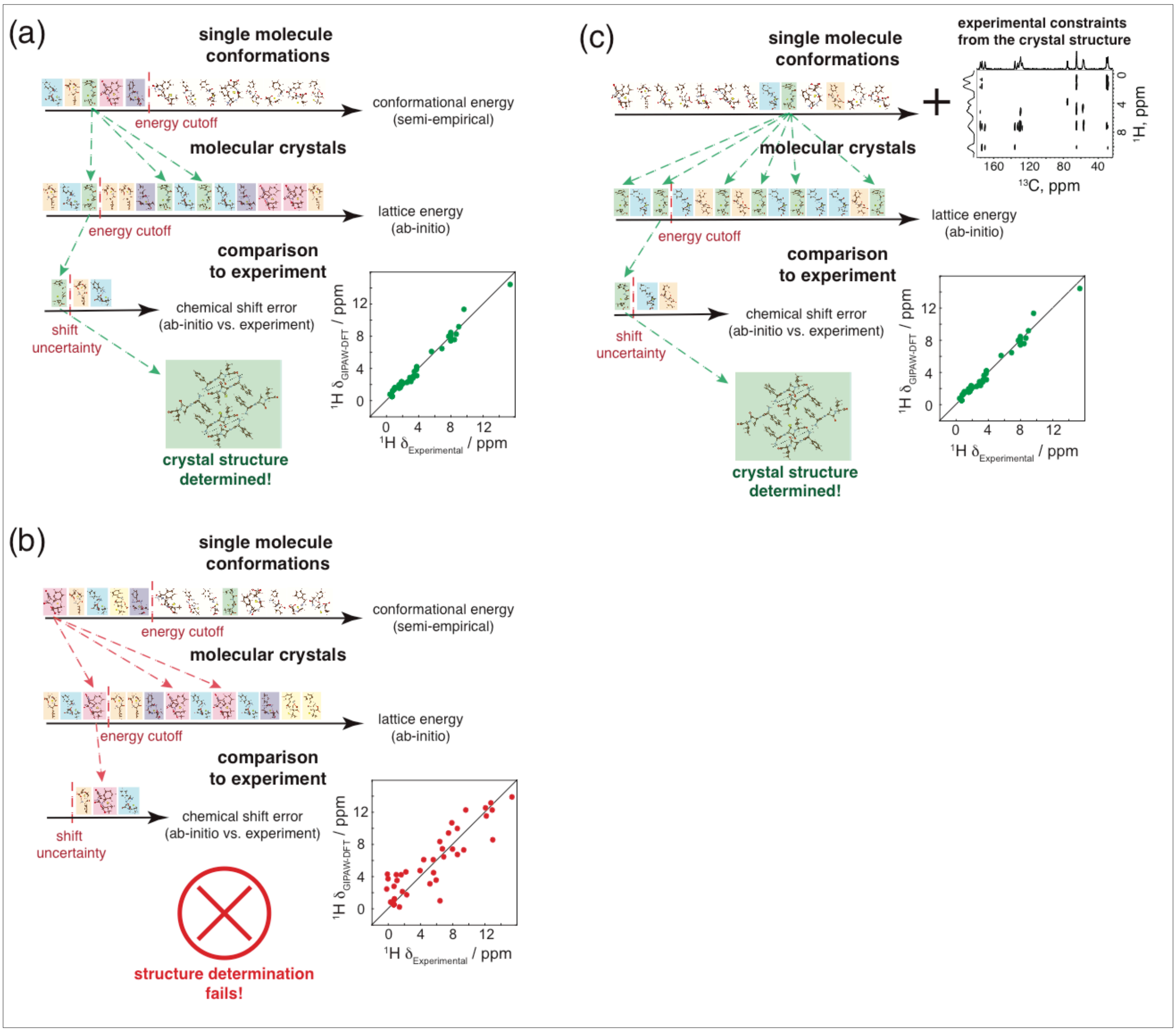

(a)
single molecule conformations
energy cutoff
conformational energy (semi-empirical)
molecular crystals
energy cutoff
lattice energy (ab-initio)
comparison to experiment
chemical shift error (ab-initio vs. experiment)
shift uncertainty
crystal structure determined!
^{1}H $\delta_{GIPAW-DFT}$ / ppm
0
4
8
12
^{1}H $\delta_{Experimental}$ / ppm
(b)
single molecule conformations
energy cutoff
conformational energy (semi-empirical)
molecular crystals
energy cutoff
lattice energy (ab-initio)
comparison to experiment
chemical shift error (ab-initio vs. experiment)
shift uncertainty
structure determination fails!
(c)
single molecule conformations
experimental constraints from the crystal structure
+
^{13}C, ppm
160
120
80
40
^{1}H, ppm
molecular crystals
energy cutoff
lattice energy (ab-initio)
comparison to experiment
chemical shift error (ab-initio vs. experiment)
shift uncertainty
crystal structure determined!

Figure 2-1. Schematic of the current and proposed CSP-NMRX methods. **(a)** an example of a successful structure prediction using the current CSP-NMRX method. **(b)** an example of a failed structure prediction using the current CSP-NMRX method. **(c)** an example of the proposed experimentally constrained CSP-NMRX method, which successfully overcomes the failure of the current CSP-NMRX method shown in panel **(b)**. 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 criteria a reduced conformer set is used to generate an ensemble of possible crystal structures (represented by the 2nd line in each panel). The colored boxes are indented 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. 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 ^{1}H chemical shift plotted against the DFT calculated ^{1}H chemical shift for the trial structure with the lowest error between DFT and experimental chemical shifts.

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[28, 34, 46, 50, 165-172] 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.

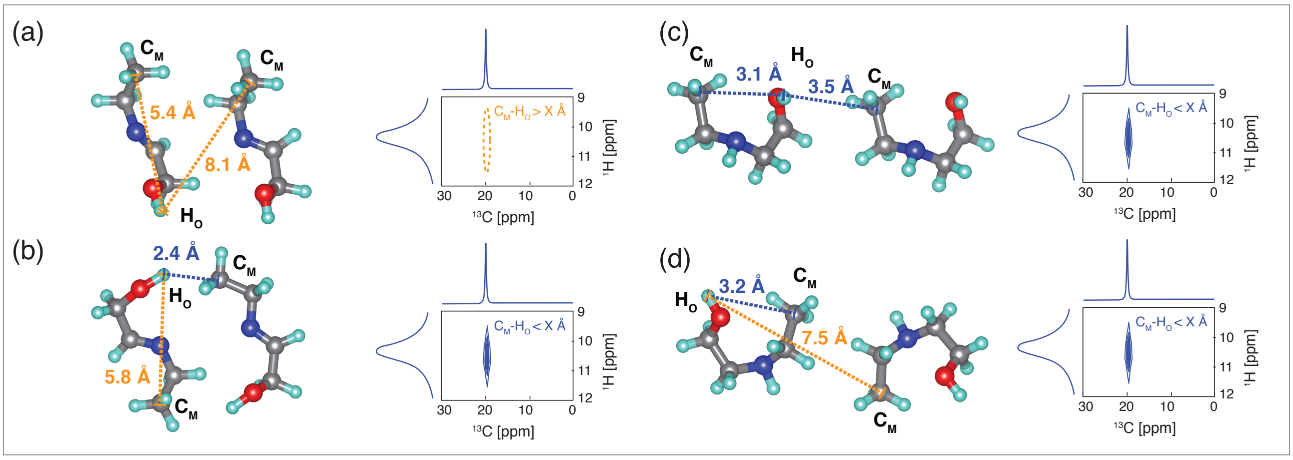


Figure 2-2. Schematic illustrations of ^{1}H-^{13}C HETCOR spectra **(right)** for four different structural fragments **(left)** and the derived constraints. Structures **(a)** and **(b)** contain an “open” conformer. Structures **(c)** and **(d)** contain a “closed” conformer. Blue dotted lines are 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.

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 2-2**, where we differentiate between two conformers (“open” and “closed”) by analyzing a ^{1}H-^{13}C HETCOR spectrum.

The ^{1}H-^{13}C HETCOR spectrum contains two different types of information. First, cross-peaks which 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 2-2** 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 best demonstrated with a thought experiment. If the heteroatoms C_M and H_O are close in space, the cross-peak at C_M-H_O will be present in the HETCOR spectra. However, the cross-peak can result either from a short intra-molecular C_M-H_O distance (i.e. the “closed” conformer) (**Figure 2-2c**-**d**) or from a short inter-molecular interatomic distance (which can be from the “closed” or the “open” conformer) (**Figure 2-2b**-**c**). Thus, the presence of a cross peak does not contain unambiguous information about the single molecule conformer, as the fragments in **Figure 2-2b**-**d** contain both possible conformations.

An absent cross-peak for C_M-H_O however indicates that C_M and H_O are at least “X” angstroms apart, for both intra- and inter-molecular C_M-H_O distances (**Figure 2-2a**). This can only happen for the “open” conformer. Thus, the 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 2-2b** also contains the “open” conformation, but does 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 ^{1}H-^{13}C HETCOR cross-peaks at least up to 3.0 Å.[173] In order to establish a reliable value for the threshold distance “X”, accessible in the ^{1}H-^{13}C HETCOR experiments used here, we investigate the correlation between interatomic ^{1}H-^{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, spinning-rates and on different spectrometers. **Figure 2-5a** 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 ^{1}H-^{13}C HETCOR experiment, where the signal is transferred from the ^{1}H 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.

To make different spectra comparable, we first estimate the number of active protons for a given cross-peak in a spectrum to be proportional to the maximum signal intensity at a given frequency in ω_1. The signal intensity of each cross-peak is then re-normalized by this number of active protons. Then, we consider the difference in overall SNR between spectra by re-normalizing each cross-peak with respect to the maximum proton-normalized SNR per spectra. This leads to a normalized SNR per ^{1}H, which is comparable across all experiments, and which is shown in **Figure 2-5b.**

Once we have selected a reliable threshold distance X Å 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_x must exceed X Å.*"

For each single molecule conformer all the generated constraints are checked and the conformers are sorted according to the number of constraints violated. This procedure allows 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 amount of 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 amount of 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.

2.2.3 Results and Discussion

In a first step, we 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. Then, we investigate the application of the parametrized constraints to CSP-NMRX structure determination of these three compounds.

Parametrization using known structures.

For cocaine, flufenamic acid and flutamide, ^{1}H-^{13}C HETCOR experiments were performed with ^{1}H-^{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. We re-normalized the spectra as described above, (see **Appendix I** for details). The resulting normalized SNR per ^{1}H is then comparable between compounds, see **Figure 2-5b.**

However, **Figure 2-5b** shows that although there is a correlation between the normalized SNR and the corresponding inter-atomic distance, there are significant fluctuations. This is expected since the HETCOR experiment is quite simple (and robust) but is subject to spin relayed transfer and dipolar truncation effects, among others. We find that the effect of these fluctuations can be minimized by only considering correlations/distances from protons which are situated towards the extremities of the molecules. 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 shown in **Table 2-4**, and marked with a green circle in **Figure 2-6a**. This results in a much clearer correlation between normalized SNR and the corresponding inter-atomic distances, as shown in **Figure 2-6b.**

From **Figure 2-6b** it is clear that only a very limited number of inter-atomic distances below 3 Å result in a *SNR* above 0.2. We then test a range of S_{norm} cut-off values from 0.08 to 0.22 with threshold distances "*X*" ranging from 2.0 to 5.0 Å. For this we use the single molecule conformer ensembles previously generated for the successful CSP-NMRX structure determination protocol described by Baias *et al.*[56] Our goal is to verify that the proposed parameterization can select the gas-phase conformer that leads to the correct crystal structure while at the same time significantly reducing the total amount of conformers which have to be considered.

Figure 2-3a shows the set of parameters for which the selection procedure is successful for all three molecules simultaneously. **Figure 2-10** shows the set of successful parameters for each molecule individually. The dashed orange line in **Figure 2-3a** shows the limit at which the selection process starts to fail. To obtain maximal selection power, the parameters should be chosen as close as possible to this limit. For cocaine, flufenamic acid and flutamide the highest selection power within the investigated conformer ensembles explored here was obtained using S_{norm} =0.14 and "*X*" = 3.5 Å.

To aid our interpretation of the selection procedure we apply a sketch-map[174-177] 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 the **Appendix I** in **Figures 2-7** to **2-9**.

Flutamide. The initial gas-phase ensemble of flutamide generated in the first step of CSP contains 15 conformers,[56] of which 7 are in the *trans* and 8 are in the *cis* conformation with respect to the amide group (**Figure 2-8**). The absent cross-peaks in a series of ^{1}H-^{13}C HETCOR spectra (**Figure 2-4a**) are used to generate the conformational constraints shown in **Figure 2-4a**. **Figure 2-3b** shows the selected sub-ensemble of 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,[56] while being able to reduce the gas-phase 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 select all 7 structures in the trans amide conformation (see **Figure 2-11**).

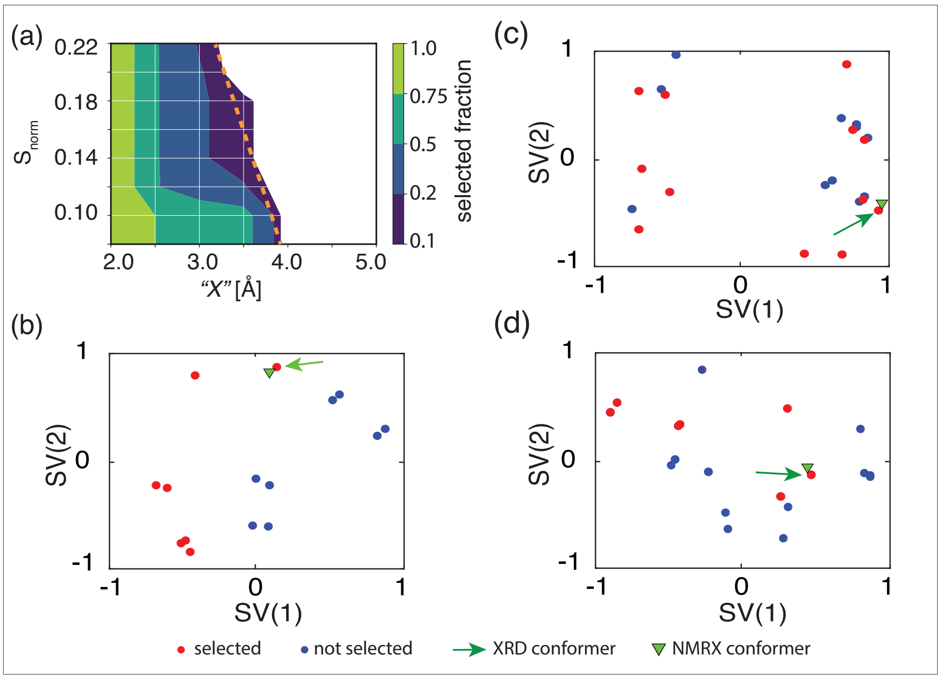


Figure 2-3. (a) Grid search results of the threshold distance "X" and S_{norm} cut-off values for flutamide, cocaine and flufenamic acid. The color-map shows the fraction of selected structures from within the conformer ensemble. The white area indicates the region where the correct conformer is not selected. Optimal selection parameters should select the smallest conformer ensemble, while still containing the correct structure. This corresponds to the dark blue regions within the different panels. The dashed orange line shows the limit, 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 which are selected for a threshold distance of 3.5 Å and a S_{norm} cut-off value of 0.14. The green triangle shows the gas-phase conformer of the XRD crystal structure. The green arrow points to the gas-phase conformer which results in the correct crystal structure after the CSP procedure.

Cocaine. The initial CSP ensemble for cocaine contains 27 single molecule conformers.[56] Figure 3d shows the sub-ensembles with the lowest number of violated constraints (2 out of 10 total constraints) extracted from the ^{1}H-^{13}C HETCOR spectra (**Figure 2-4b**). As for flutamide, these two constraints are 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 2-12** shows that the HETCOR constraints can distinguish between the folding and stretching of the cocaine molecule with respect to the aromatic group as well as a flip in the methylamine group. Here, the relevant ensemble is reduced by around 55% (from 27 to 12 conformers), while retaining the conformer that leads to the correct crystal structure.

Flufenamic acid. The gas-phase ensemble for flufenamic acid contains 26 molecular conformations.[56] **Figure 2-3c** shows the sub-ensembles with the lowest number of violations (0 of 2 total constraints) selected from ^{1}H- ^{13}C HETCOR. The extracted constraints are shown in **Figure 2-4c**. 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 ^{1}H dimension. However, the distance constraints extracted solely from the carboxyl proton (see **Figure 2-4c** and **Figure 2-13**) are 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.

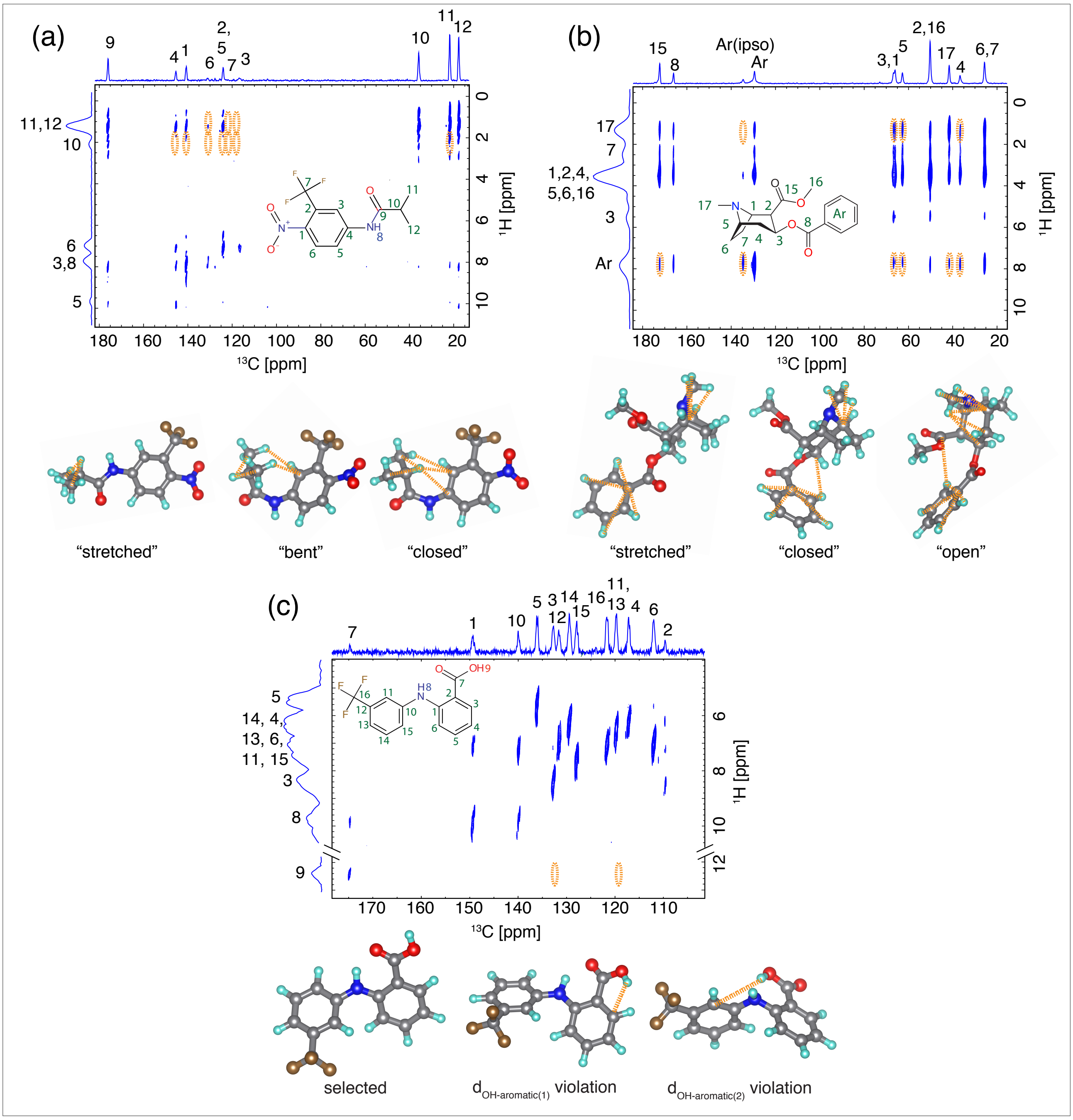


Figure 2-4. The top part in each panel shows the 1H-13C HETCOR spectrum of: flutamide with 1.25 ms contact time **(a)**, cocaine with 1.0 ms contact time **(b)** and flufenamic acid with 1.5 ms contact time **(c)** (further details and raw data in **Appendix I**). 13C peaks are assigned based on the literature[178] and 1H peaks are assigned from HETCOR spectra and DFT chemical shift calculations (see **Appendix I**). The cross-peaks from the terminal protons (**Figure 2-6**) below a Snorm 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 the 1H-13C HETCOR cross-peaks for different example conformers within the ensembles.

2.2.4 Conclusion

The most severe limitations of CSP-NMRX 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 crystal packing.

However, removing any conformer selection and including all possible conformers during crystal structure generation can lead to prohibitively high computational costs. To overcome this, we propose a modified CSP-NMRX method which includes unambiguous prior NMR constraints, in this case ^{1}H-^{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 parametrize the proposed method on the crystal structure determination of three flexible molecules that we previously studied using CSP-NMRX: cocaine, flutamide and flufenamic acid. For all these compounds we found that the method reproduces CSP-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.

The compounds studied here were not subjected to any modification prior to the experiments, and they were investigated using powder samples at natural isotopic abundance.

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 to reduce its computational cost

We believe that the method is robust and we have chosen the experimental constraints, based on ^{1}H-^{13}C NMR correlation experiments, for their relative simplicity and ease of access. However, we note that ^{1}H-^{13}C correlation-based experiments are not the only ones that can give conformational constraints. Future work could incorporate other types of experiments such as ^{13}C-^{13}C correlations, or more accurate ^{1}H-^{13}C correlation experiments, which could be simpler to parameterize. 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 ^{1}H-^{1}H spin diffusion or dipolar truncation, the selection criteria can be made stronger, further reducing the conformational space and improving the computational efficiency and reliability of the methodology.

2.2.5 Appendix I

Samples

The powdered samples of free base cocaine (Methyl (1R,2R,3S,5S)-3-(benzoyloxy)-8-methyl-8-azabicyclo[3.2.1]octane-2-carboxylate, purity > 98.0%) was purchased from Toronto Research Chemicals, while the powdered samples of flutamide (2-Methyl-N-[4-nitro-3-(trifluoromethyl)phenyl]propenamide, purity > 98.0%) and flufenamic acid (2-((3-(Trifluoromethyl)phenyl)amino)benzoic acid, purity > 98.0%) were purchased from Tokyo Chemical Industry. All samples were used without further purification. For all compounds, the reference crystal structures were previously determined by single-crystal XRD.[179-181]

The reference structure of flutamide, (CSD entry: WEZCOT) contains 4 molecules in the unit cell, and it is orthorhombic, space group $Pna2_1$, with unit cell parameters a = 11.856(2) Å, b = 20.477(3) Å, c = 4.9590(9) Å.

The crystal structure of cocaine, (CSD entry: COCAIN10) contains 2 molecules in the unit cell, it is monoclinic, space group $P2_1$, with unit cell parameters a = 10.130(1) Å, b = 9.866(2) Å, c = 8.445(1) Å.

The flufenamic acid structure (CSD entry: FPAMCA11) is monoclinic, space group P21/c, with unit cell parameters a = 12.523(4) Å, b = 7.868(6) Å, c = 12.874(3) Å and 4 molecules in the unit cell.

Solid-state NMR experimental setup

Experiments were performed at room temperature on a Bruker 500 wide-bore Avance III and a Bruker 900 US^2 wide-bore Avance Neo NMR spectrometers operating at Larmor frequencies of 500.43 and 900.13 MHz, equipped with H/X/Y 3.2 mm and H/C/N/D 1.3 mm probes.

The 2D ^{1}H-^{13}C dipolar heteronuclear correlation (HETCOR) experiments were performed at 12.5 kHz MAS rate for flutamide and cocaine and at 24.0 kHz MAS rate for flufenamic acid. In all experiments, we used SPINAL-64 for heteronuclear decoupling during t1 and eDUMBO-1$_{22}$ for homonuclear decoupling in the indirect dimension. 64 transients and 256 increments for flutamide, 4 transients with 64 increments for flufenamic acid and 16 transients with 256 increments for cocaine.

All chemical shifts were referenced indirectly to tetramethylsilane using the methyl signals of l-alanine at 1.3 ppm (^{1}H) and 20.5 ppm (^{13}C).[182] ^{1}H chemical shifts were corrected for the scaling factor due to homonuclear decoupling, which was determined using ^{1}H 1D spectra acquired under fast spinning on Bruker 900 spectrometer. Post-processing was done using Topspin 3.5.

Assignment of experimental NMR spectra

The assignment of ^{13}C and ^{1}H chemical shifts for flutamide, flufenamic acid and cocaine was taken from the paper by M. Baias *et al.*[56]

Experimental chemical shifts

Table 2-1. Cocaine 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 2-2. Flufenamic acid 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 2-3. Flutamide 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

Signal to Noise analysis

The signal to noise ratio (SNR) extraction and analysis was 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:

$$SNR = maxval(S)/(2 * noise), \tag{2-1}$$

where maxval(S) is the maximum intensity at a given ^{1}H and ^{13}C coordinate $\pm 0.2ppm$. Note, that after a first extraction of maxval(S) the ^{1}H 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 ($0.4 \times 0.4\ ppm$) 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-or-equal to two times the maximum signal intensity in the already chosen areas. **Figure 2-5a** shows the extracted SNR of all ^{1}H-^{13}C HETCOR spectra for cocaine, flufenamic acid and flutamide against the corresponding inter-atomic distance.

First, we normalize each cross-peak by the number of active protons. For this we estimate the number of active 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 ^{1}H coordinate. In a next step, we consider the difference in sensitivity between the spectra, due to the specific experimental setups, by normalizing each cross-peak with respect to the maximal proton-normalized SNR per spectrum. This leads to a normalized SNR per ^{1}H, which is comparable across all experiments and is shown in **Figure 2-5b**.

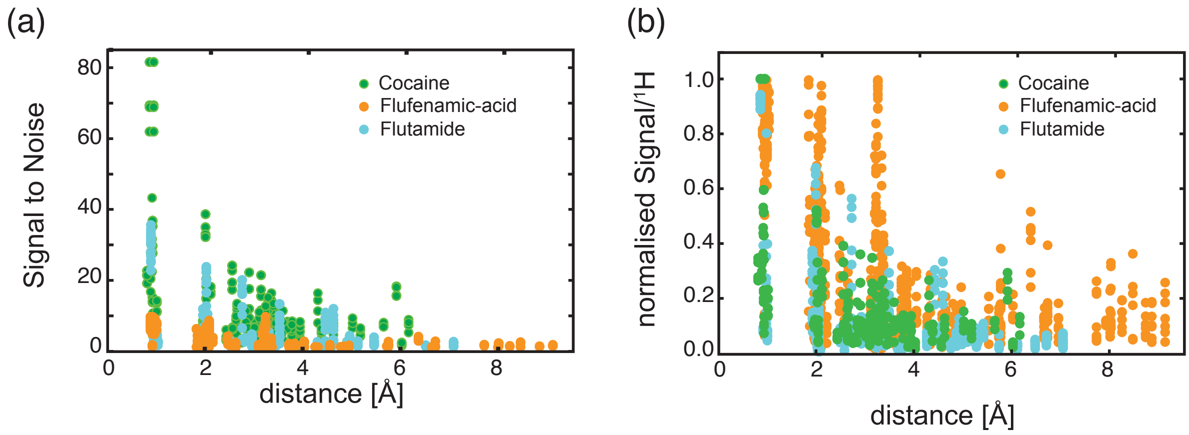


Figure 2-5. Signal intensity of ^{1}H-^{13}C HETCOR cross-peaks plotted against the corresponding interatomic distance for cocaine (green), flufenamic-acid (orange) and flutamide (cyan). **(a)** The SNR is extracted directly for all ^{1}H-^{13}C HETCOR at different contact-times and different experimental setups. **(b)** The normalized SNR per ^{1}H allows a direct comparison across different experimental setups and for cross-peaks with a different number of active protons.

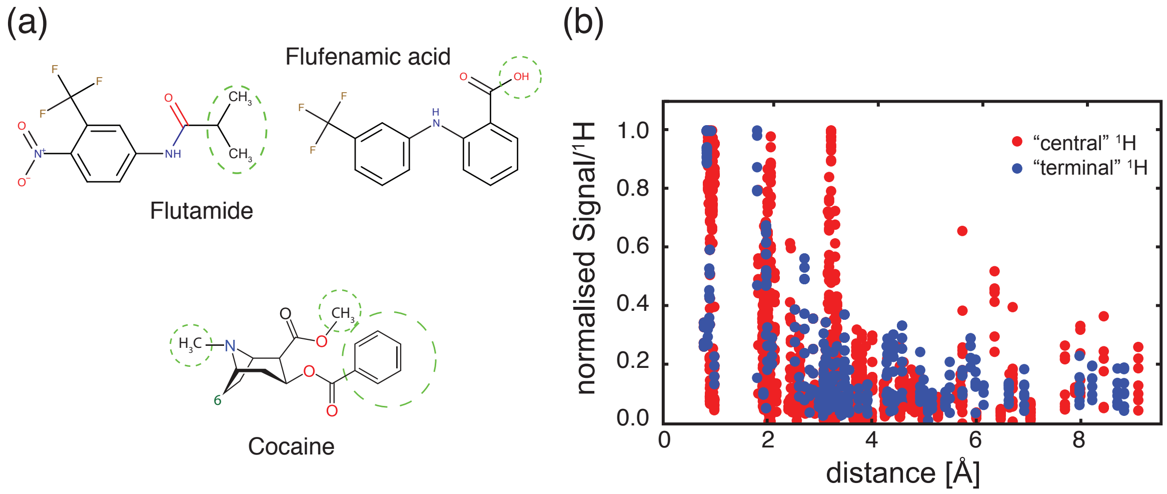


Figure 2-6. (a) Illustration of terminal protons, for which cross-contribute to conformational constraints. **(b)** normalized SNR of ^{1}H-^{13}C HETCOR cross-peaks plotted against the corresponding interatomic distance for center protons (red) and terminal protons (blue), which are used to generate conformational constraints.

Table 2-4. Terminal protons contributing to conformational constraints for cocaine, flufenamic acid, flutamide and ampicillin

Molecule	**terminal ^{1}H**
Cocaine	Ar
	16
	17
Flufenamic acid	9
Flutamide	10
	11
	12

Gas-phase conformer generation

For cocaine, flutamide and flufenamic acid the CSP conformers and crystal structures were generated as described in the paper by M. Baias *et al.*[56]

Sketch-map analysis

The cluster generation and analysis were performed with home-written Python and MATLAB codes and using the sketch-map package.[174-177] The sketch-map parameters are given **Table 2-5**. They were chosen following the procedure described in Ceriotti *et al.*[175] and the tutorial on sketchmap.org. The sketch-map analysis was not sensitive to small variations in the chosen parameters, as was already noted in the references.[175-177] As starting point for the sketch-map analysis we used all dihedral angles, not containing protons, over the full 2π range. This gives 47, 31 and 35 dihedral angles for cocaine, flutamide and flufenamic acid, within a range of $-\pi$ to π.

Table 2-5. Sketch-map parameters for all compounds.

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

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. The 2nd conformer resulted in the correct crystal structure after the remaining CSP procedure.[56] **Figure 2-7** 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)).

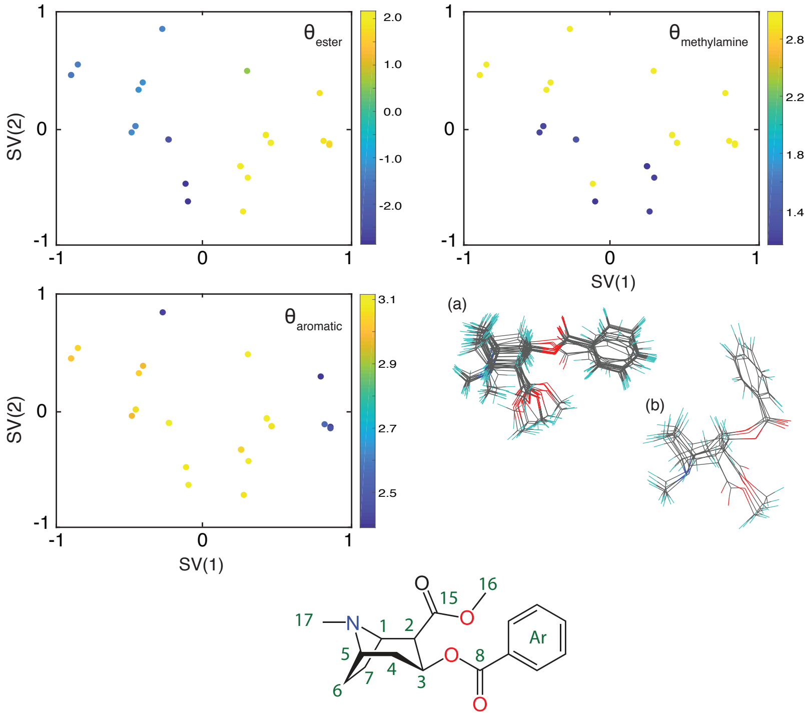


Figure 2-7. Top) Sketch-map representation of the locally stable cocaine conformations. To show the extent of the sub-clustering the panels are colored according to different torsion angles reporting on different rotations in the molecule. Θ_{ester} is defined as the torsion angle between C1-C2-C15-O4 and reports on rotations of the ester group. $\Theta_{methylamine}$ is defined as the torsion angle between C2-C1-N-C17 and reports on rotations of the methyl group attached to the nitrogen. $\Theta_{aromatic}$ is defined as the torsion angle between C(ortho)-C(ipso)-C8-O2 and reports on flips of the aromatic group. The lower right panel shows the overlapped conformation without **(a)** and with **(b)** a flipped aromatic ring. **Bottom)** 2D structure of cocaine with the used labelling scheme.

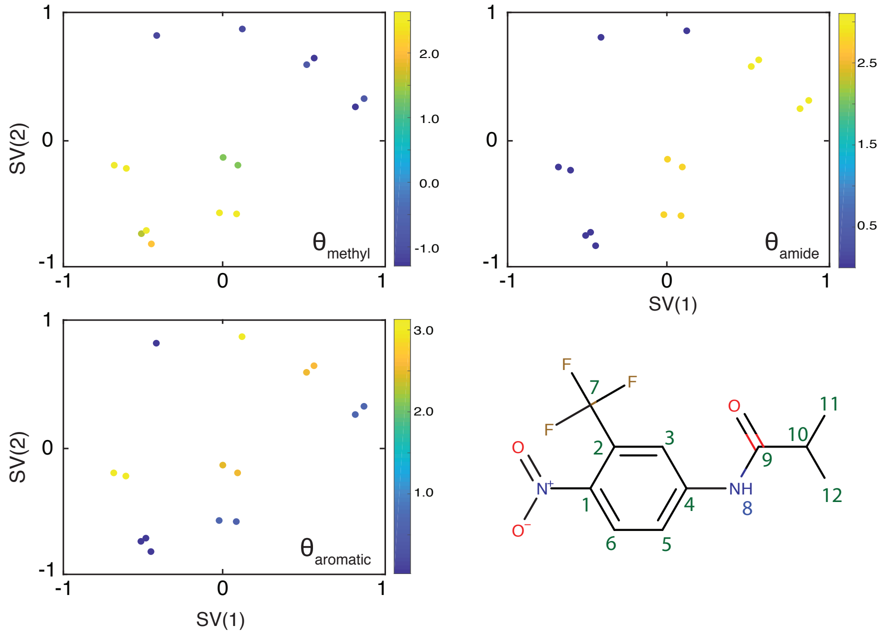


Figure 2-8. Sketch-map representation of the gas-phase flutamide ensemble. To show the extent of the sub-clustering the panels are colored according to different torsion angles reporting on different rotations in the molecule. Θ_{methyl} is defined as the torsion angle between C11-C10-C9-N(H) and reports on rotations of the methyl groups. Θ_{amide} is defined as the torsion angle between C4-N(H)-C9-O1 and reports on the amide conformation. $\Theta_{aromatic}$ is defined as the torsion angle between C3-C4-N(H)-C9 and reports on rotations of the aromatic group. The lower right panel shows the 2D structure of flutamide with the used labelling scheme.

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. The 1st conformer resulted in the correct crystal structure after the remaining CSP procedure.[56] **Figure 2-8** shows the sketch-map representation of the locally stable flutamide conformers. The sketch-map representation shows a relative distinctive 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 correspond 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 3rd conformer resulted in the correct crystal structure after the remaining CSP procedure.[56] **Figure 2-9** shows the sketch-map representation of the flutamide gas-phase 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)).

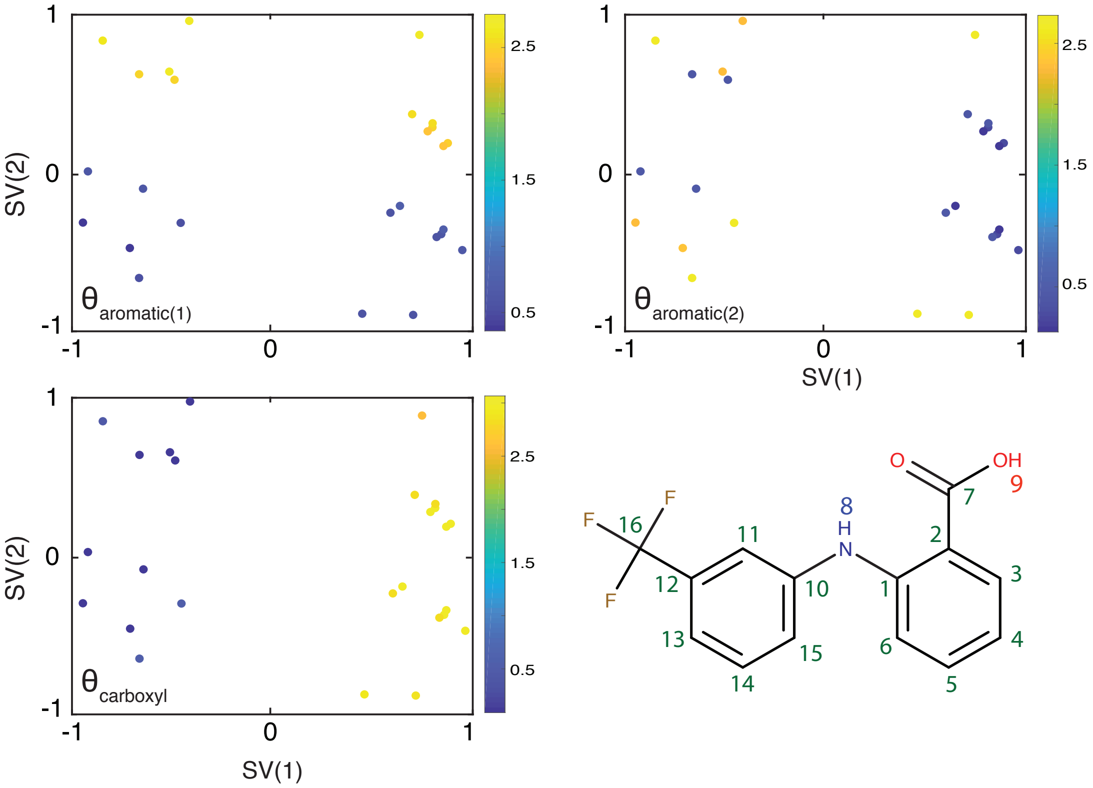


Figure 2-9. Sketch-map representation of the gas-phase flufenamic acid conformations. To show the extent of the sub-clustering the panels are colored according to different torsion angles reporting on different rotations in the molecule. $\Theta_{aromatic(1)}$ is defined as the torsion angle between C15-C10-N(H)-H(N) and reports on rotations of aromatic ring with the attached trifluoromethyl. $\Theta_{aromatic(2)}$ is defined as the torsion angle between C2-C2-N(H)-H(N) and reports on rotations of aromatic ring with the attached carboxyl. $\Theta_{carboxyl}$ is defined as the torsion angle between C1-C2-C7-O(H) and reports on rotations of the carboxyl group. The lower right panel shows the 2D structure of flufenamic acid with the used labelling scheme.

Parametrization of the constraints

It is not a priori clear as to what the threshold distance "X" should be but in general we expect ^{1}H-^{13}C HETCOR cross-peaks in solid-state NMR for 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 2-10** shows the set of successful parameters for each molecule individually.

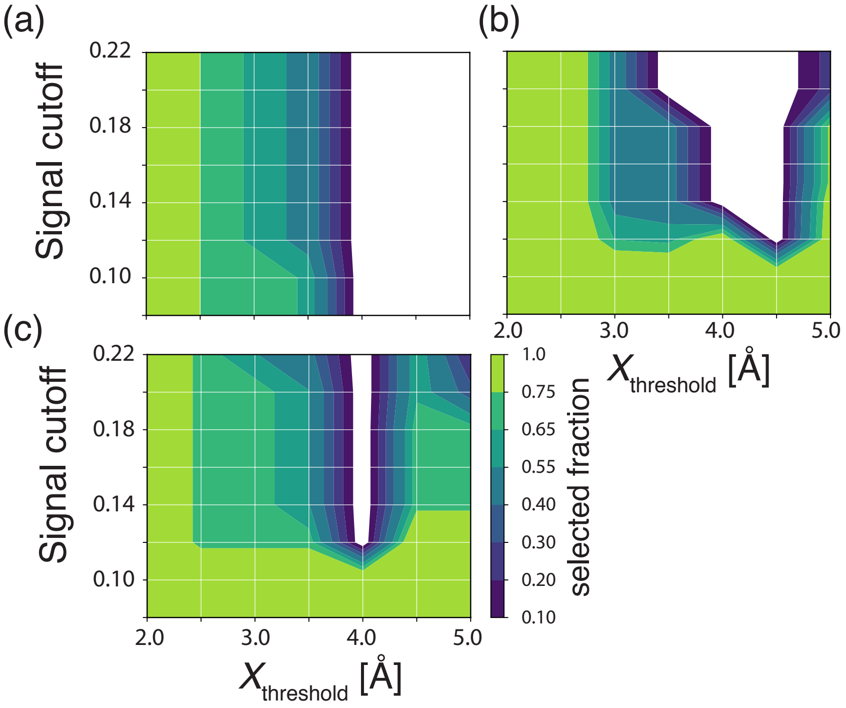


Figure 2-10. Grid search results of the threshold distance "X" and S_{norm} cut-off values for **(a)** flutamide, **(b)** cocaine and **(c)** flufenamic acid. The color-map shows the percentage of selected structures from within the conformer ensemble. The white area indicates the region where the correct conformer is not selected. Optimal selection parameters should select the smallest conformer ensemble, 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. For the constraints the peaks below a S_{norm} cut-off value of 0.14 were interpreted as proton-carbon distances greater than a threshold distance “X” of 3.5 Å. For each conformation the number of fulfilled constraints was counted and the conformations were sorted in decreasing order.

Flutamide. The sub-ensemble selection for flutamide is done based on constraints from multiple HETCOR contact times 0.1, 0.3, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75 and 2.0ms. The ^{1}H and ^{13}C cross peaks from the two methyl groups were not distinguished. Also, the ^{1}H cross peaks from H3 and H8 as well as the ^{13}C cross peaks from C5 and C2 are too close and not distinguishable. Therefore, if a cross-peak was seen it was attributed to all the atoms within the given group.

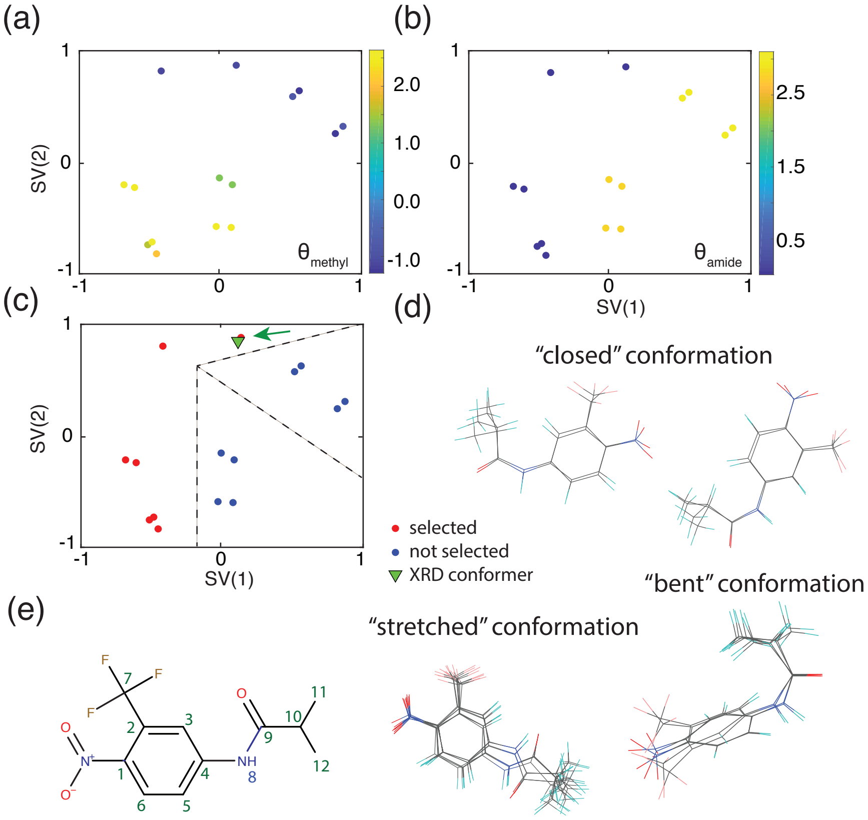


Figure 2-11. Sketch-map representation of the gas-phase flutamide ensemble. To show the extent of the sub-clustering the panels are colored according to different torsion angles reporting on different rotations in the molecule. Θ_{methyl} is defined as the torsion angle between C11-C10-C9-N(H) and reports on rotations of the methyl groups. Θ_{amide} is defined as the torsion angle between C4-N(H)-C9-O1 and reports on the amide conformation. **(c)** Sketch-map projection of the gas-phase flutamide ensemble. Red dots represent the structures with the lowest violations that are selected. The green triangle shows the gas-phase conformer of the XRD 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)** Overlap 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 used labelling scheme.

Cocaine. The cocaine HETCOR were performed at the contact times 0.5, 0.75, 1.0 and 1.5 ms. The ^{1}H and ^{13}C cross peaks from the aromatic group were not distinguished. Also, the ^{13}C cross peaks from C6 and C7, the ^{13}C cross peaks from C2 and C16 as well as the ^{1}H cross peaks from H1, H2, H4, H5, H6 and are too close and not distinguishable. Therefore, if a cross-peak was seen it was attributed to all the atoms within the given group.

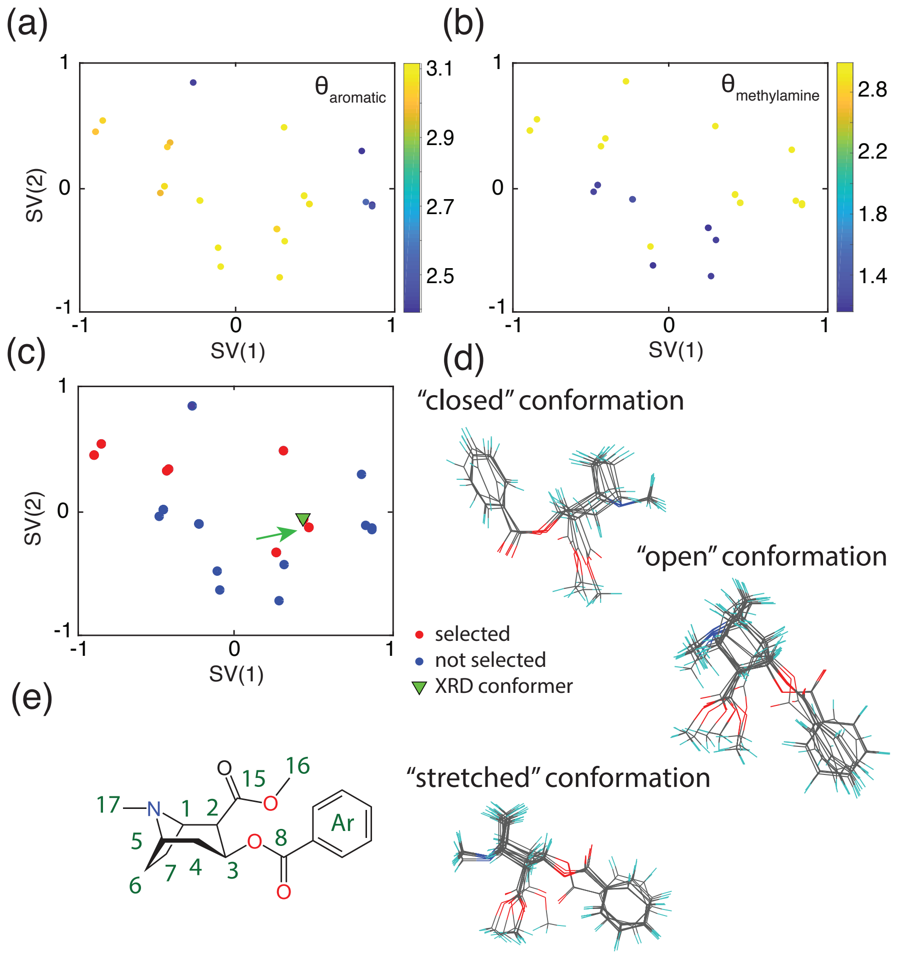


Figure 2-12. (a-b) Sketch-map representation of the locally stable cocaine conformations. To show the extent of the sub-clustering the panels are colored according to different torsion angles reporting on different rotations in the molecule. $\Theta_{\text{methylamine}}$ is defined as the torsion angle between C2-C1-N-C17 and reports on rotations of the methyl group attached to the nitrogen. Θ_{aromatic} is defined as the torsion angle between C(ortho)-C(ipso)-C8-O2 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 violations that are selected. The greed triangle shows the gas-phase conformer of the XRD crystal structure. The green arrow points to the gas-phase conformer, which resulted in the correct crystal structure after the CSP procedure. **(d)** Overlap of the structures within the different sketch-map clusters. The "stretched" conformations correspond to the selected conformers. The "closed" conformation contain a 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 used labelling scheme.

Flufenamic Acid. The sub-ensemble selection for flufenamic acid is done based on constraints from multiple HETCOR contact times 0.1, 0.5, 1.0, 1.5, 3.0 and 3.5 ms. The ^{1}H cross peaks from H4, H13 and H14 as well as the ^{1}H cross peaks from H6, H11 and H15 are too close and not distinguishable. Therefore, if a cross-peak was seen it was attributed to all the atoms within the given group.

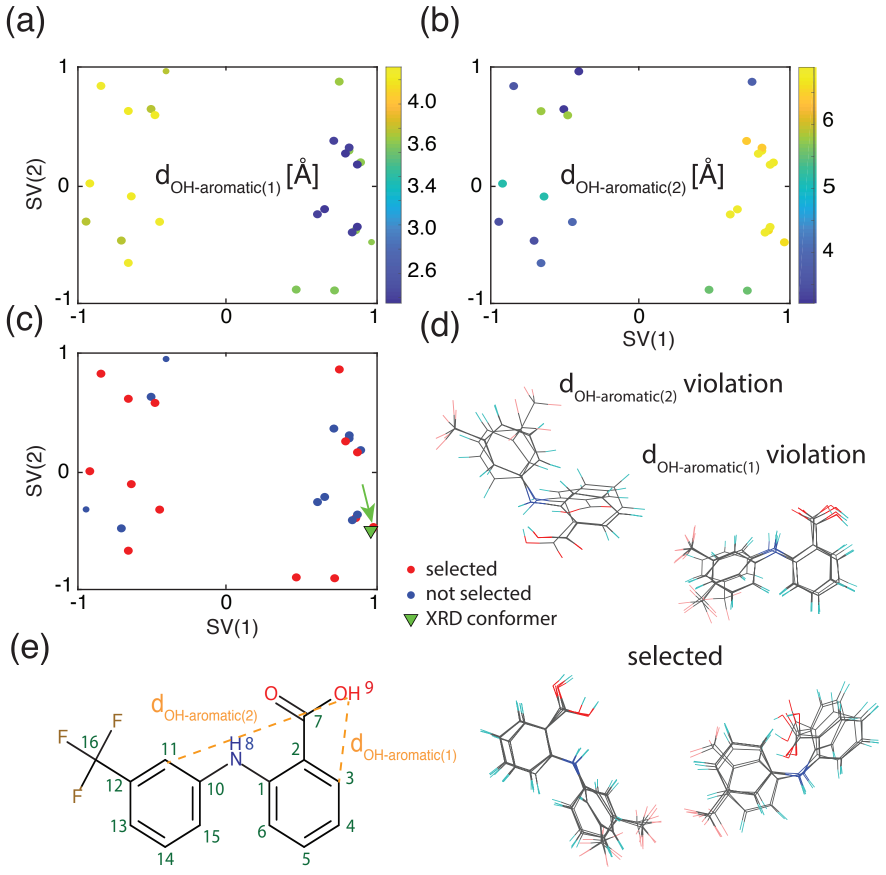


Figure 2-13. (a-b) Sketch-map representation of the gas-phase flufenamic acid conformations. To show the extent of the sub-clustering the panels are colored according to the distance [Å] between the OH group and the two aromatic rings. The distance is expressed as the distance between the carboxyl proton and C3/C11 (as shown in **e**). **(c)** Sketch-map projection of the gas-phase flufenamic acid ensemble. Red dots represent the structures with the lowest violations that are selected. The green triangle shows the gas-phase conformer of the XRD crystal structure. The green arrow points to the gas-phase conformer, which resulted in the correct crystal structure after the CSP procedure. **(d)** Overlap of the structures within the different sketch-map clusters. **(e)** 2D structure of flufenamic acid with the used labelling scheme.

2.3 Chemical shifts by machine learning

This chapter has been adapted with permission from: Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L., "Chemical shifts in molecular solids by machine learning". *Nature Communications* **2018**, 9 (1), 4501. **(post-print)**

2.3.1 Introduction

For microcrystalline powders of molecular solids, the scope of CSP-NMRX is mainly limited by the considerable computational resources required by both the CSP search and the calculation of accurate DFT chemical shifts (see **Chapters 1.3** and **2.1**). In **Chapter 2.2** we have introduced an approach to reduce the computational cost of the CSP search. In this chapter we investigate an approach to reduce to computational cost of the chemical shift calculations, while maintaining sufficiently high accuracy needed in chemical shift driven NMRX.

ML has recently emerged as a way to overcome the need for quantum chemical calculations but for chemical shifts in solids it is hindered by the chemical and combinatorial space spanned by molecular solids, the strong dependency of chemical shifts on their environment, and the lack of an experimental database of shifts.

Here, we propose a ML method based on local environments to accurately predict chemical shifts of molecular solids and their polymorphs to within DFT accuracy. The protocol is schematically illustrated in **Figure 2-14.** 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),[135] chosen to be as diverse as possible, and then show that the method can predict chemical shifts in a test set with a R^2 coefficients between the chemical shifts calculated with DFT and with ML of 0.97 for ^{1}H, 0.99 for ^{13}C, 0.99 for ^{15}N, and 0.99 for ^{17}O, corresponding to RMSEs of 0.49 ppm for ^{1}H, 4.3 ppm for ^{13}C, 13.3 ppm for ^{15}N, and 17.7 ppm for ^{17}O. Predicting the chemical shifts for a polymorph of cocaine, with 86 atoms in the unit-cell, using the ML method takes less than a minute of CPU time, thus reducing the computational time by a factor of between 5 to 10 thousand, without any significant loss in accuracy as compared to DFT.

We also demonstrate that the trained model is able to determine, based on the match between experimentally-measured and ML-predicted shifts, the structures of cocaine and the drug 4-[4-(2-adamantylcarbamoyl)-5-tert-butylpyrazol-1-yl]benzoic acid, even though no experimental shifts were used in training. We also show that it is possible to calculate the NMR spectrum of very large molecular crystals. For this we calculate the chemical shifts of six structures from the CSD with between 768 and 1,584 atoms in the unit-cells.

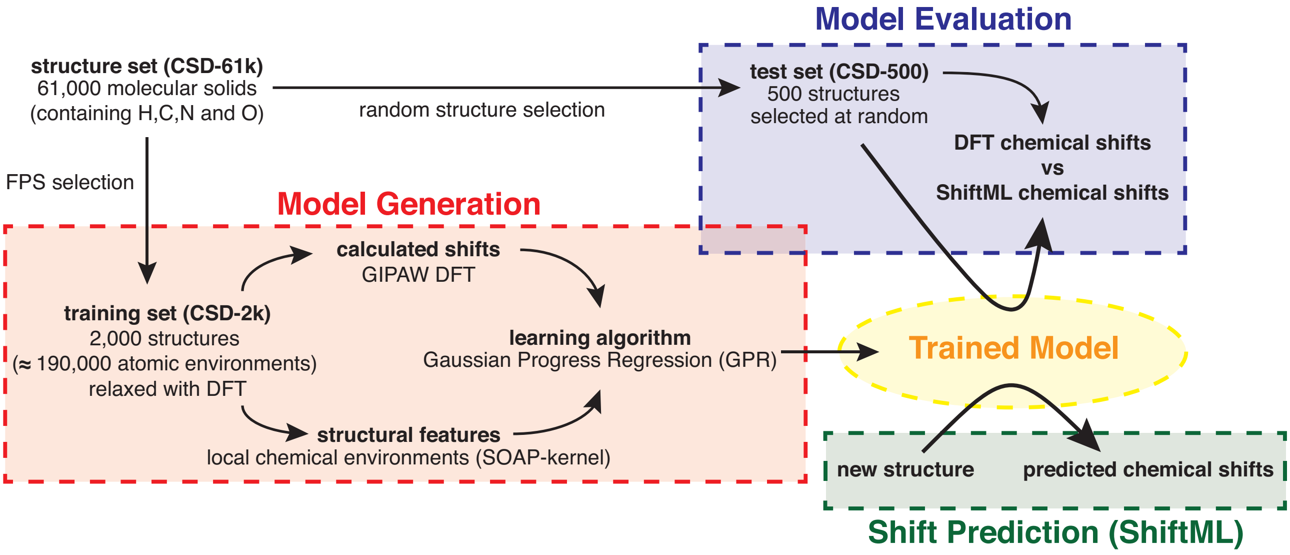


Figure 2-14. Scheme of the machine learning model used for the chemical shift predictions.

2.3.2 Results

Training and validation using DFT calculated shifts of known crystal structures.

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 ^{1}H 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.[18, 83] 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 hypothesize 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.

The approach we take to predicting chemical shifts in molecular solids is illustrated in **Figure 2-14**. We use the Gaussian Process Regression (GPR) framework[183] 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,

$$\sigma(X) = \sum_{i} \alpha_i k(X, X_i), \tag{2-2}$$

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 2-2** 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.

Our model relies on the Smooth Overlap of Atomic Positions (SOAP) kernel,[176, 184] in which any atomic environment is represented as a three dimensional neighborhood density given by a superposition of Gaussians, one centered at each of the atom positions in a spherical neighborhood within a cut-off radius r_c from the core atom. This framework, combined with GPR, has been used to model the stability and properties of a number of different systems,[155, 176, 184] and has been extended to the prediction of tensorial properties.[185] 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 chemical shift-based NMR crystallography.

As shown in **Figure 2-14**, in the absence of an experimental database of shifts the model is developed by using a reference training set of structures for which chemical shifts are calculated with GIPAW DFT. To obtain a model which is robust and general, the training set should be as large, as reliable, and as diverse as possible. 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 Methods 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, see Methods) 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 algorithm (FPS)[186-187] (CSD-2k, see Methods). 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.

To avoid including spurious environments in the model, e.g. environments which 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 in the Methods. 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 ^{1}H, or 0.3%). This is discussed in detail in the Methods. 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 ^{1}H, this reduced the size of the training set from 70,000 to about 35,000 different atomic environments. (Details of the selection method and the members of the different sets used are given in the Methods section).

All the atomic positions of the structures in the training and testing sets were relaxed with DFT, using the Quantum Espresso suite,[188-190] prior to calculation of the chemical shieldings using the GIPAW DFT method.[62-63] Note that the DFT relaxation ensures "reasonable" geometries will be used even for crystal structures containing errors (e.g. improbable ^{1}H positions). Parameters for the DFT calculations are given in the Methods section. The calculated chemical shieldings σ are converted to the corresponding chemical shifts δ through the relationship $\delta = \sigma_{\text{ref}} - \sigma$. Here, we used a σ_{ref} of 30.8 ppm (for ^{1}H) and 169.5 ppm (for ^{13}C), found through linear regression between the calculated and experimental chemical shifts for cocaine.

Figure 2-15 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 ^{1}H chemical shifts as RMSE in ppm between the shifts calculated with DFT and with the protocol described above, which we refer to in the following as 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=2Å 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=7Å 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 behavior is observed for the prediction errors of the ^{13}C, ^{15}N and ^{17}O chemical shifts (see **Figures 2-25** to **2-28**).

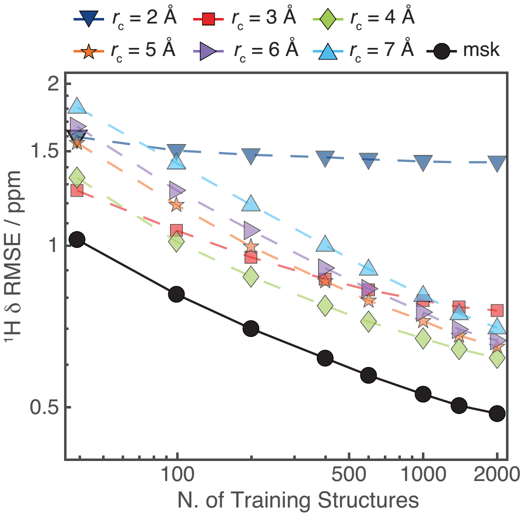


Figure 2-15. ^{1}H chemical shift prediction error of the trained model for the CSD-500 set. The RMSE prediction error between chemical shifts calculated with ShiftML and GIPAW DFT is shown for different local environment cut-off radii, and for the multi-kernel (labelled as msk), as a function of the training set size.

The observed differences in the behavior 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 neighbors 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,[155] with weightings of 256 (r_c=2Å), 128 (r_c=3Å), 32(r_c=4Å), 8 (r_c=5Å and r_c=6Å) and 1 (r_c=7Å). This weighting was determined by rough optimization around values inspired by previous experience,[155] and by cross-validation on the CSD-2k training set (as described in the Methods section). 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 **Figures 2-25** to **2-28**).

Figure 2-16a-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,[18, 83] 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 2-24**). 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),[18, 83] which is not surprising as there are (currently) significantly less training environments for the heteronuclei than for ^1H.

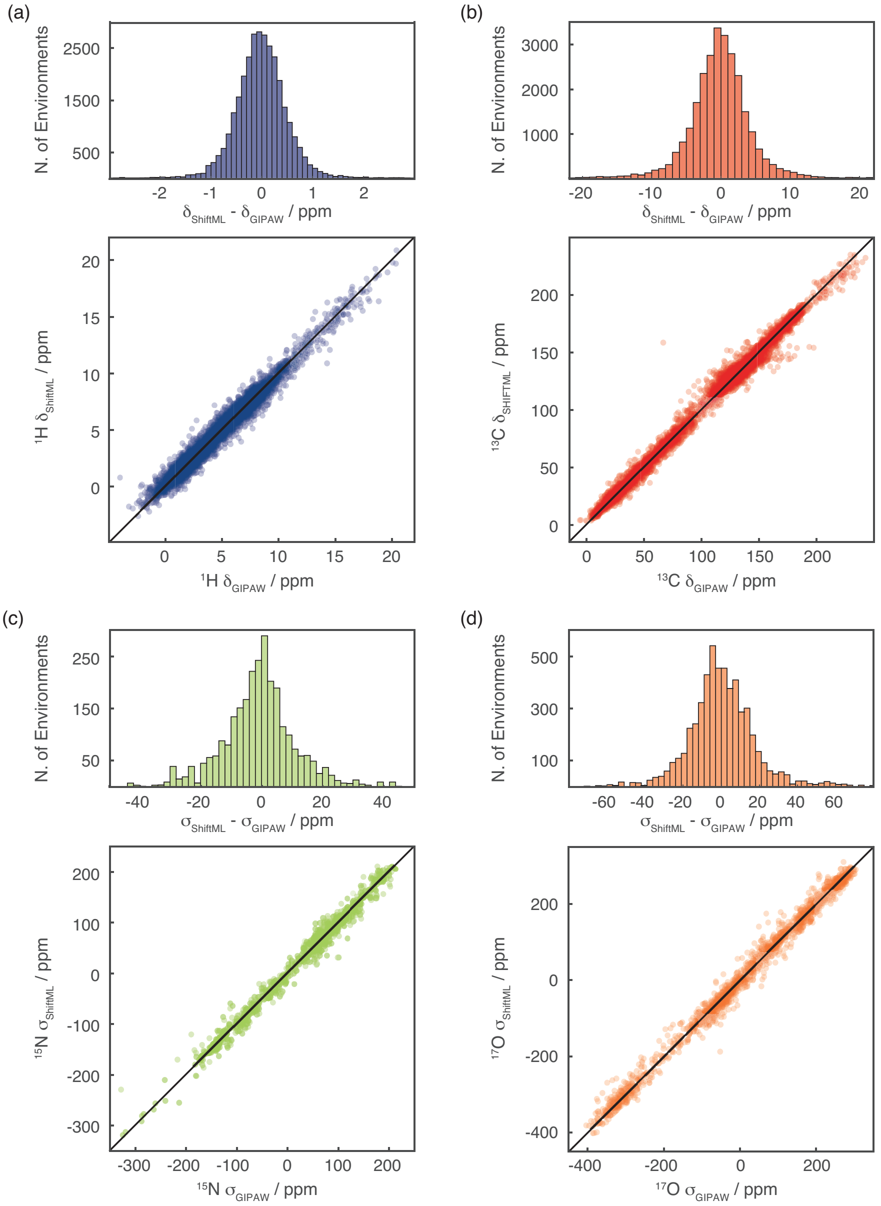


Figure 2-16. Comparison of predictions from ShiftML and GIPAW DFT. Histograms and scatterplots showing the correlation between ^1H **(a)**, ^{13}C **(b)**, ^{15}N **(c)** and ^{17}O **(d)** chemical shifts (shieldings) calculated with GIPAW and ShiftML. The black lines indicate a perfect correlation.

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 delocalized 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 2-17a** 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 crystal structure prediction (CSP) procedure.[56, 58] 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)[191], which suggests that ShiftML is robust with respect to small deviations from the fully optimized structures. (As shown in the **Figure 22**, performing the prediction using Quantum Espresso consistently leads to comparable prediction accuracy.)

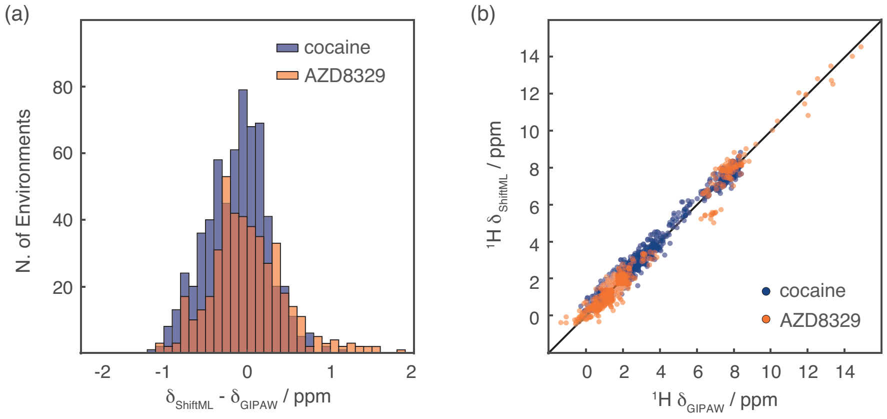


Figure 2-17. Comparison of predictions from ShiftML and GIPAW DFT for polymorphs of cocaine and AZD8329. **(a)** Histogram showing the distribution of the differences between ^1H chemical shifts calculated with GIPAW and with ShiftML for the polymorphs of cocaine (blue), and the polymorphs of AZD8329 (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

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 relaxed either by the CASTEP protocol used in ref. 30, or the Quantum Espresso protocol used here (the RMSE between ML and DFT for the Quantum Espresso relaxed 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.

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 in a chemical shift driven NMR crystallography approach.

Figure 2-18a 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 2-19** (numerical values of the experimental chemical shifts, the crystal structures, and the shifts calculated with ShiftML are given in the Methods section). 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 favorably to the equivalent agreement found between GIPAW DFT and experiment which for this set of structures is an RMSE of 0.38 ppm.

Figure 2-18a 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,[18] 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[18, 83] 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.

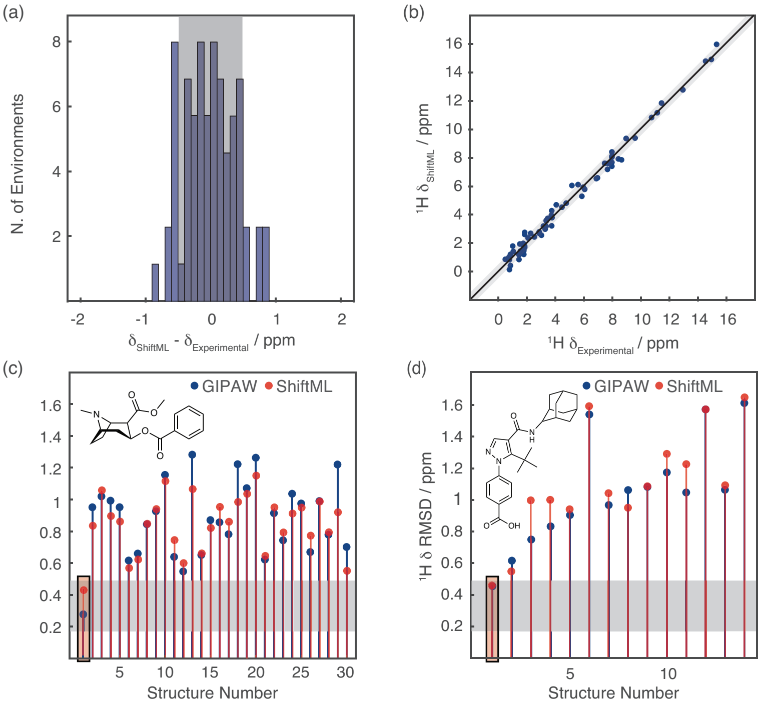


Figure 2-18. Comparison of ShiftML to experimentally measured shifts. **(a)** Histogram showing the distribution of differences between experimentally measured ^1H chemical shifts and ^1H chemical shifts calculated with ShiftML for six different crystal structures (see Methods section for the structures 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)**. 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 RMSD, as described in the text,[18] 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.

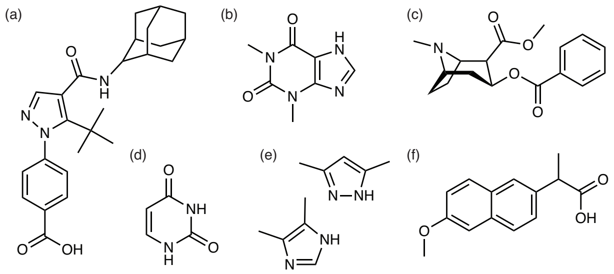


Figure 2-19. Chemical structures of the six molecules used to evaluate the correlation between experimentally measured ^{1}H chemical shifts and the shifts calculated by ShiftML. The structures are given as AZD8329 **(a)**, theophylline **(b)**, cocaine **(c)**, uracil **(d)**, 3,5-dimethylimidazole and 4,5-dimethylimidazole **(e)** and naproxen **(f)**.

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 appears nearly constant, because it is dominated by the loading time of the reference SOAP vector (see **Figure 2-20a**). We have used this method to calculate the NMR spectra (shown in **Figure 2-20b-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 2-30** for the chemical formula). The values of the predicted chemical shifts are given as CSD-6 in the Methods section. **Figure 2-20a** 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.

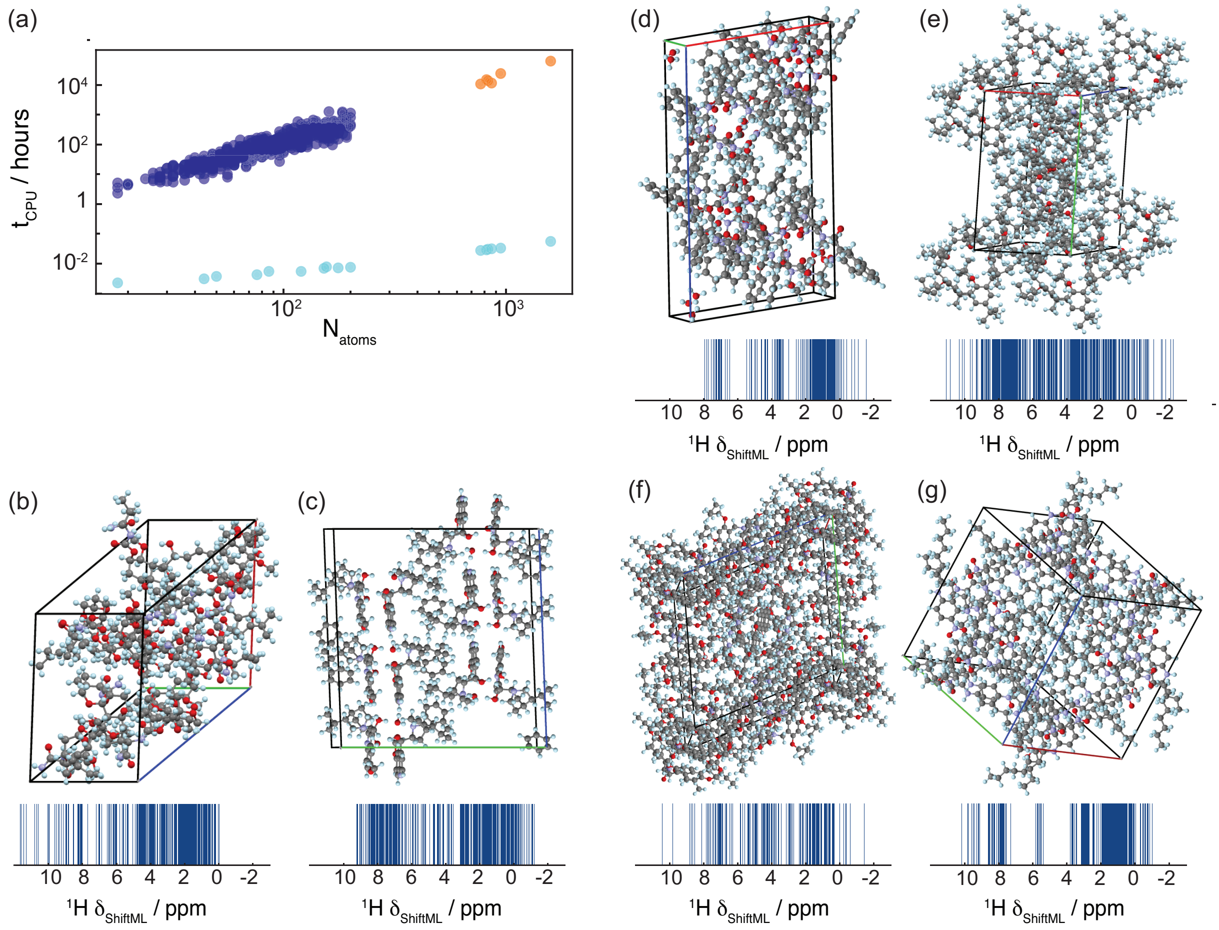


Figure 2-20. Chemical shift calculation times and large structures. **(a)** DFT GIPAW calculation time (blue) and ShiftML prediction time (turquoise) for different system sizes. The GIPAW DFT calculation time for the six large structures (orange) is estimated from a cubic dependence on the number of valence electrons in the structure (see Methods section). **(b-g)** 3D-shemes and ^{1}H NMR spectra predicted with ShiftML, of the six large molecular crystals with CSD Refcodes: **(b)** CAJVUH,[192] N_{atoms} = 828, **(c)** RUKTOI,[193] N_{atoms} = 768, **(d)** EMEMUE,[194] N_{atoms} = 860, **(e)** GOKXOV,[195] N_{atoms} = 945, **(f)** HEJBUW,[196] N_{atoms} = 816, **(g)** RAYFEF,[197] N_{atoms} = 1,584.

2.3.3 Discussion

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. 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 NMR crystallography 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 on the protocol described here 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 has already been updated (see **Chapters 2.5** and **2.6**). The training set of the current version has been extended to include structures containing H, C, N, O and S atoms (see **Chapter 2.6**). Additionally, the prediction procedure has been changed to reduce the memory requirements, while maintaining comparable chemical shift accuracy. Most notably, the new ShiftML version contains one radially SOAP kernels[176, 184, 198] as opposed to the seven multi-scale SOAP kernels (see **Chapter 2.6**). Further, a projected process[183, 199-200] (PP) strategy is used for the prediction, in which the full $(N \times N)$ kernel matrix is approximated by a lower rank $(M \times M)$ kernel matrix corresponding to an "active set" of M training data containing the most relevant information (see **Chapter 2.5**). Note, that the PP strategy allows for the rapid calculation of uncertainties associated with the individual chemical shift predictions.[201] To further accelerate the ML predictions, we also sparsified the SOAP fingerprints using an FPS strategy[202] (see **Chapter 2.5**).

We also note, that after the publication of the initial ShiftML paper in 2018, already a number of further ML models being applied to calculations of chemical shifts in solids has been published.[101-103]

2.3.4 Methods

Crystal Structures

All the crystal structures of CSD-61k and CSD-500 were obtained from the Cambridge Structural Database (CSD).[135] 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 a maximum of 200 atoms, containing either (i) only H and C or (ii) H, C and one heteroatom between N and O or both. 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 that 60,000 structures, the whole procedure was automatized and we didn't 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, given as Supplementary Dataset 1) and the testing set (CSD-500, given as Supplementary Dataset 2) for the ^{1}H, ^{13}C, ^{15}N and ^{17}O chemical shift prediction as described in the main text. The test set (CSD-500) was created by randomly picking 500 structures from the CSD-61k excluding the structures already selected for the training set. The Refcodes of all CSD sets are given in Paruzzo *et al.*[161]

Crystal Structure Prediction

Here we use a set of possible polymorphs predicted by CSP for cocaine and the drug 4-[4-(2-adamantylcarbamoyl)-5-tert -butylpyrazol-1-yl]-benzoic acid (also referred as AZD8329). General details on the CSP protocol can be found in ref. [203]. In chemical shift based NMR crystallography, the CSP trial polymorphs are tested against experimental parameters (^{1}H chemical shifts) to determine the experimental crystal structure.

In this work we used 30 possible polymorph structures of cocaine and 14 trial structures of AZD8329 generated with CSP. The 30 structures of cocaine were obtained from the Electronic Supporting Information (ESI) of ref. [56], and correspond to the most stable polymorphs obtained with CSP. Crystal structures of AZD8329 were obtained from the ESI of ref. [58], and correspond to the 14 most stable predicted polymorphs with the *cis* conformation of the amide bond. From the same sources we obtained chemical shifts for each structure calculated with GIPAW[62-63] using the DFT program CASTEP[191] and the experimental chemical shifts. Labels for the different polymorphs of each structure are based on their DFT calculated energy, with 1 being the most stable trial polymorph of a given molecule.

DFT Calculations

All the DFT calculations were carried out using the DFT program Quantum ESPRESSO.[188, 190] 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[62-63] 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].[204] The optimizations were done with the generalized-gradient-approximation (GGA) density functional PBE,[205] using 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 correction[206] was included in order to account for van der Waals interactions. The geometry optimization was done relaxing all atomic positions while keeping the lattice parameters fixed.

A single point energy (scf) was then computed for the relaxed geometry, using higher wave-function and charge density energy cut-offs which were set to 100 Ry and 400 Ry respectively. For this calculation we also used a Monkhorst-Pack grid of *k*-points[207] corresponding to a maximum spacing of 0.06 Å^{-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 shielding σ_{DFT} using the GIPAW method, with the same parameters as used in the scf calculation.

Note that using a convergence threshold of in the scf calculation of 1e^{-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 1e^{-12}-1e^{-14} Ry.

Machine Learning

For the SOAP kernels,[176, 184] each atomic environment is represented as a three dimensional neighborhood density given by a superposition of Gaussians, one centered at each of the atom positions in a spherical neighborhood within a cut-off radius r_c from the core atom. The Gaussians have a variance ς^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'.

SOAP-based structural kernels contain several adjustable hyper-parameters, which are discussed in refs.[176] 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 experience[155] and with some optimization by cross-validation on the CSD-2k training set (see Methods for details).We also combine kernels computed for different cutoff radii to capture the contributions to shifts from different length scales,[155] as is described in detail above. The calculations of the local environment, the similarity kernel and the weighted correlations were done using the glosim2 package.[208]

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

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

(2-3)

where the function f is a Gaussian Process[183] and ε represents the error of the prediction, which is modeled as independent identically distributed Gaussian variates, with variance σ_n^2. Following the Gaussian Process Regression framework, the isotropic chemical shielding function becomes:

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

(2-4)

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 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, due to errors in the training set, the limitations of the model or the reduced number of training configurations

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

(2-5)

To assess the correlation between local atomic environments *A* and *B*, we use the SOAP kernel[184] defined by the rotationally invariant overlap between smooth representations of their atomic density:

$$k(A, B) = \int_{SO(3)} \left| \int_{\mathrm{R}^3} \rho_{\mathrm{A}}(\vec{\mathbf{r}}) \rho_{\mathrm{B}}(\hat{R}\vec{\mathbf{r}}) \mathrm{d}\vec{\mathbf{r}} \right|^2 d\hat{R},$$

(2-6)

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

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

(2-7)

The details of the construction, and the extension to the case with many atomic species, are given in refs. [176] and [155].

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 farthest point selection algorithm[186-187] to sort the CSD-61k in descending order of "diversity". Essentially, we select a first structure at random, and then pick the others in the sequence such that

$$k = \underset{k \in CSD-61k}{argmax} \underset{j \in selection}{min} \left| X_k - X_j \right|, \tag{2-8}$$

where the distance is the kernel-induced distance associated with an average SOAP kernel for the entire structure.[176] 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:

$$\varepsilon(X) = \frac{1}{M} \sum_{i=1}^{M} (y_i(X) - y(X)), \tag{2-9}$$

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 2-21**, while further details for ^1H and the other nuclei are listed in the below.

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.

Of the 76,214 ^1H environments of the CSD-2k, 211 environments were detected as unusual. Of the 58,148 ^{13}C environments of the CSD-2k, 1,419 environments were detected as unusual. Of the 27,814 ^{13}C environments of the CSD-2k, 514 environments were detected as unusual. Of the 25,924 ^{13}C environments of the CSD-2k, 441 environments were detected as unusual. The unusual environments are detailed in Paruzzo *et al.*[161]

Most of the environments detected as "unusual" are part of zwitterionic structures or charged structures (such as VIWYEH, ZACSOO or EKUJIF). Others are metallic structures ($E_{LUMO} - E_{HOMO} = 0$), such as HAZQUV, QUICNA02, DMEBQU01 or AYUKIP, or have a partially empty unit cell (QAHVUQ). An intrinsic limit of this procedure is the fact that it might detect structures with uncommon functional groups as “anomalies” (e.g. TIMCHX, which is an aziridine – a three membered heterocycle with one amine group, or FIGMAJ which has a cubane group), due to the fact that these structures are not well represented by the used training set. However, with increasing training size, we expect these structures to be better represented and they will not be detected as anomalies anymore.

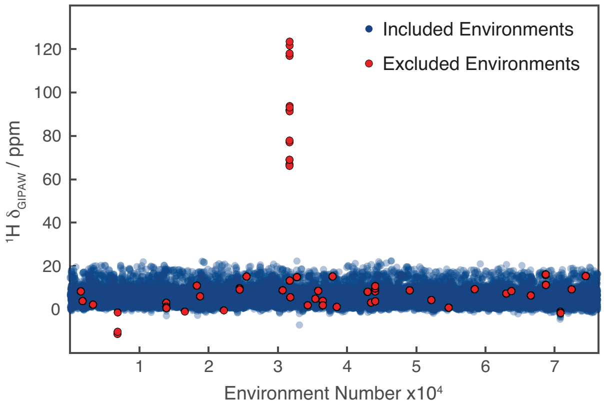


Figure 2-21. ^{1}H chemical shifts of the 76,214 environments in the CSD-2k set. The environments excluded using the unusual structures detection procedure described in the text are shown in red.

NMR Crystallography

To validate the accuracy of the chemical shifts calculated with ShiftML, we replicated the last step of the protocol for the *ab initio* crystal structure determination of powdered solids[18, 56, 58] using predicted shifts. This step consists in the comparison between experimental and predicted ^{1}H chemical shifts for the candidate crystal structures selected from a crystal structure prediction method. We perform this analysis for cocaine and form 4 of AZD8329.[56, 58] 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 ^{1}H 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. [56] and [58].

Remarkably, the high accuracy shown in **Figure 2-17** was obtained using crystal structures with only ^{1}H positions relaxed 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 2-22** 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 2-17**, 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 2-17**, although the all-atom optimization leads to some significant structural differences compared to the only ^{1}H relaxed structures, especially for AZD8329. We find a chemical shift prediction error (RMSE) for ^{1}H 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 AZD 8329 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.

Experimental chemical shifts were referenced to the ^{1}H 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 ^{1}H 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 2-23** and **Table 2-6**.

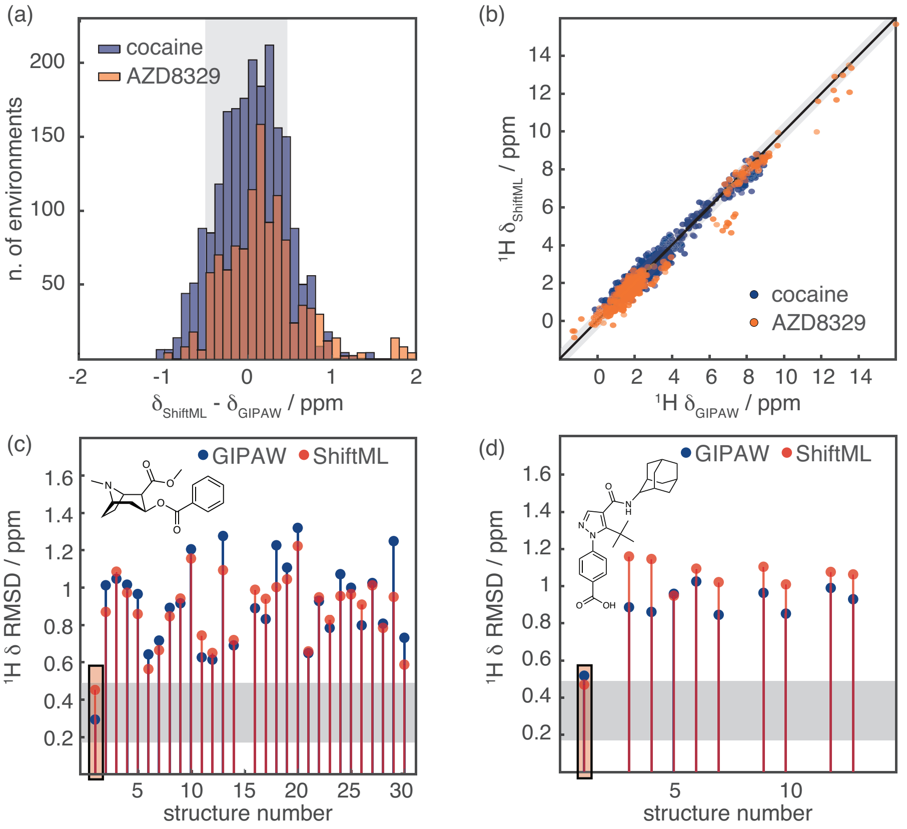


Figure 2-22. NMR crystallography of cocaine and the form 4 of AZD8329. **(a)** Histogram showing the distribution of the differences between chemical shifts calculated with GIPAW and ShiftML. The blue bars were calculated for the polymorphs of cocaine, and the orange ones for the polymorphs of AZD8329. **(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 ^{1}H 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 ^{1}H chemical shift RMSD, as described in the text.[18]

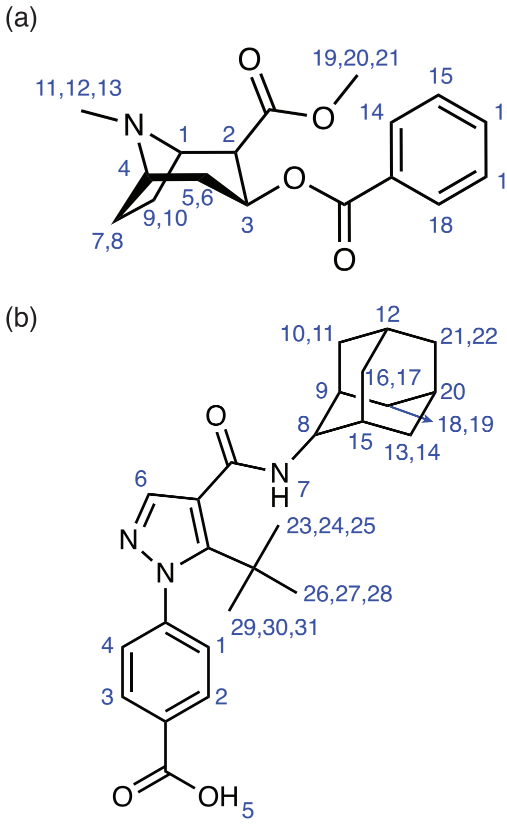


Figure 2-23. Chemical structure of cocaine (a) and AZD8329 (b) and the labelling scheme used here.

Table 2-6. Experimental chemical shifts of cocaine and the form 4 of AZD8329. The labelling scheme is given in **Figure 2-23**. When more than one atom corresponds to a single chemical shift value, their values were averaged.

Cocaine		**AZD8329**	
Atom Label	^{1}H δ (ppm)	Atom Label	^{1}H 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
15	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

DFT Calculation Times

Figure 2-24 shows the CPU time needed for part of the GIPAW DFT calculations done for this work. The calculations shown in Supplementary **Figure 2-24a were** done on polymorph 1 of the cocaine dataset, which contains 86 atoms per unit-cell, while the one in Supplementary **Figure 2-24b** were done on 500 structures of the CSD-2k set. In **Figure 2-24a** the calculation time is plotted as a function of the number of Monkhorst-pack k-points 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 2-24b** 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 0.06 Å^{-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.[209] For this reason, the predictive accuracy of ShiftML for non-equilibrium structures has not yet been quantified. This will be the subject of further work. However, **Figure 2-24b** 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.

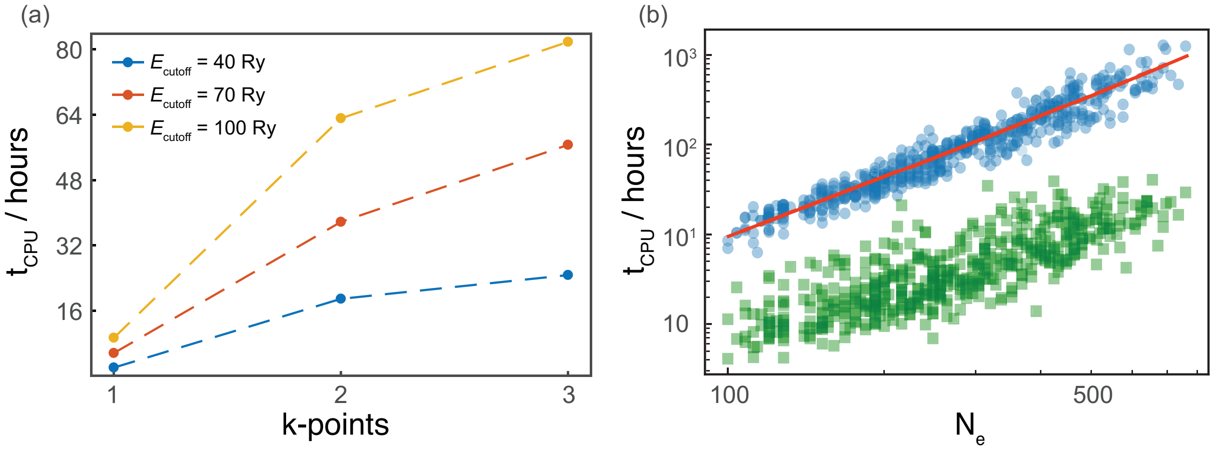


Figure 2-24. CPU time for NMR chemical shift calculations using the GIPAW method. **(a)** The CPU time is shown as function of the DFT accuracy, determined by the plane-wave cutoff energy E_{cutoff} and the number of k-points in each dimension for polymorph 1 of cocaine. The charge density energy cut-offs 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 $b = 5.91e-06$.

ShiftML Prediction Times

The ShiftML run-times are shown in **Figure 2-18**. They 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.

Prediction Parameters, Leaning and Evaluation Curves

Tables 2-7 and **2-8** and show the parameters used for the single and the multi-scale kernel predictions respectively. Using these parameters, we obtained the curves shown in **Figure 2-15** and the ones shown in **Figures 2-25**-to **2-28**. **Figures 2-25** and show the RMSE and MAE learning curves for ^{1}H, ^{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)

Figures 2-27 and **2-28** show 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 and are given in Paruzzo *et al.*[161]

Table 2-7. Kernel weights and GPR parameters used for multi-scale kernel prediction.

Atom	**Multi-Scale Kernel Weights**						σ_n	ζ
	r_c = 2 Å	r_c = 3 Å	r_c = 4 Å	r_c =5 Å	r_c = 6 Å	r_c = 7 Å		
^{1}H	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 2-8. Kernel and GPR parameters. The GPR parameters (σ_n and ζ) are the ones used in single kernel predictions.

Atom	Cut-off (r_c)	Gaussian width (ς)	l_{max}	n_{max}	σ_n	ζ
^{1}H	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

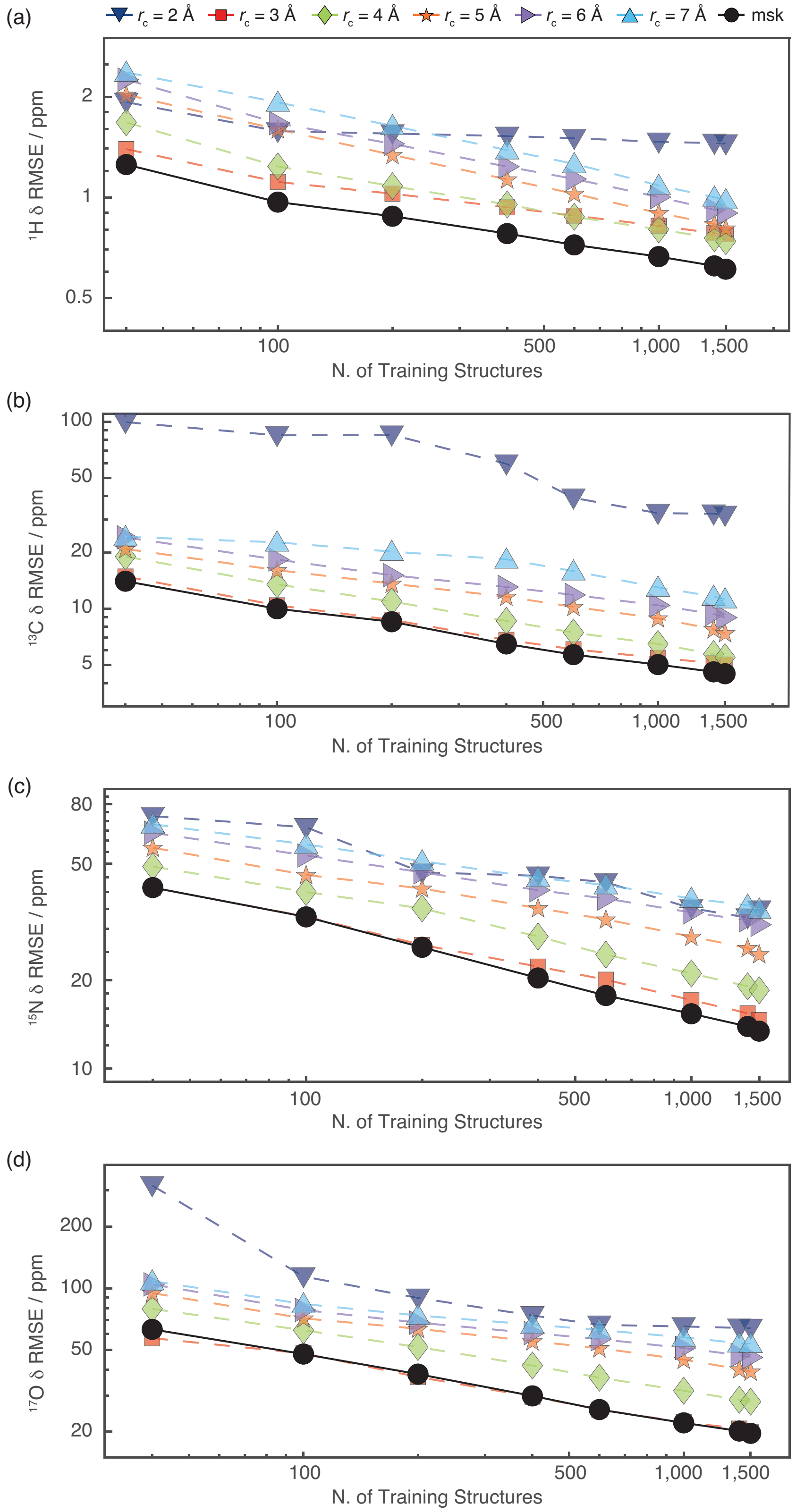


Figure 2-25. RMSE learning curves showing the error between chemical shifts calculated with DFT and ShiftML. The curves are for ^{1}H **(a)**, ^{13}C **(b)**, ^{15}N **(c)** and ^{17}O **(d)** chemical shieldings. The multi-kernel learning-curve is labelled as msk.

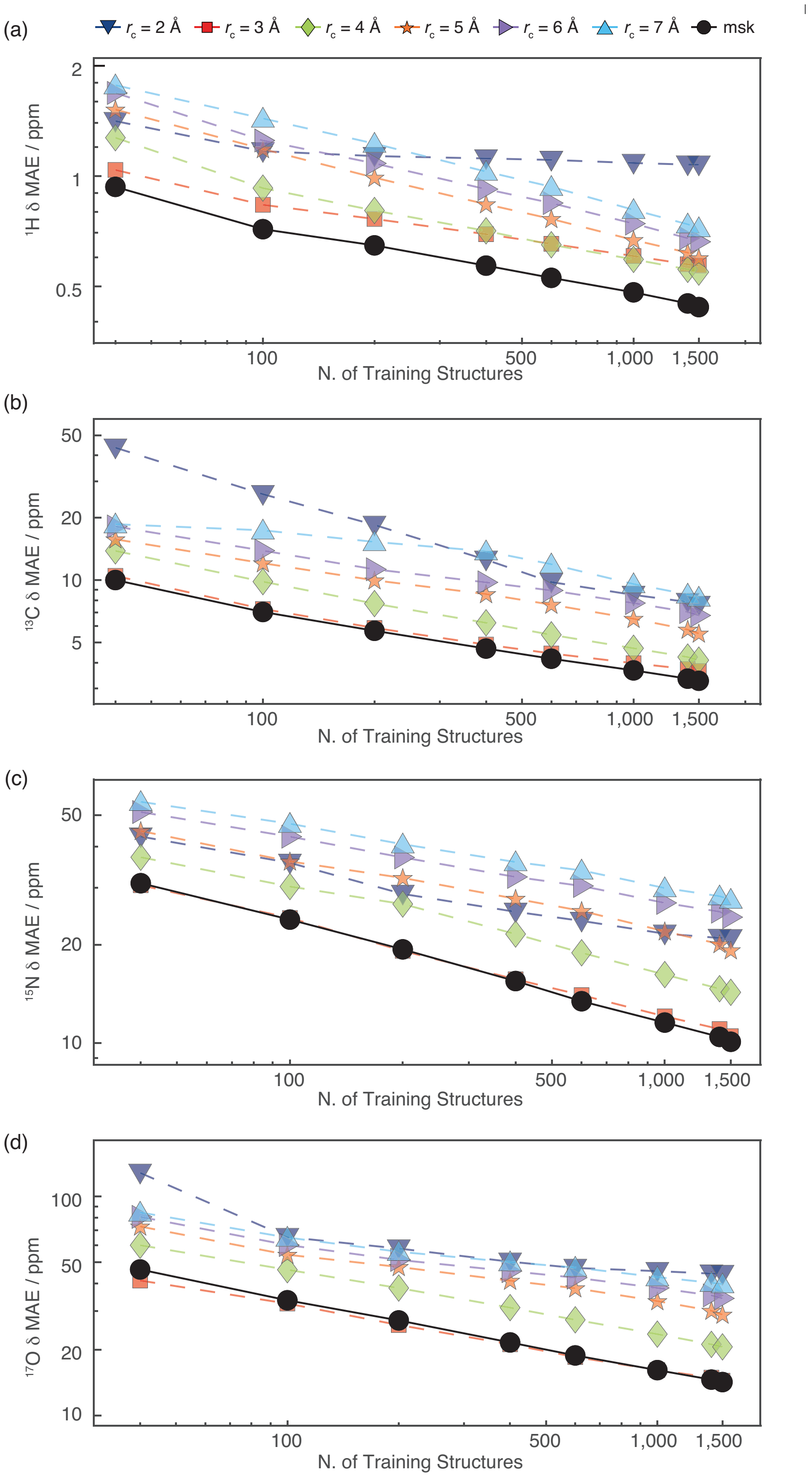


Figure 2-26. MAE learning curves showing the error between chemical shifts calculated with DFT and ShiftML. The curves are relative to ^{1}H **(a)**, ^{13}C **(b)**, ^{15}N **(c)** and ^{17}O **(d)** chemical shieldings. The multi-kernel learning-curve is labelled as msk.

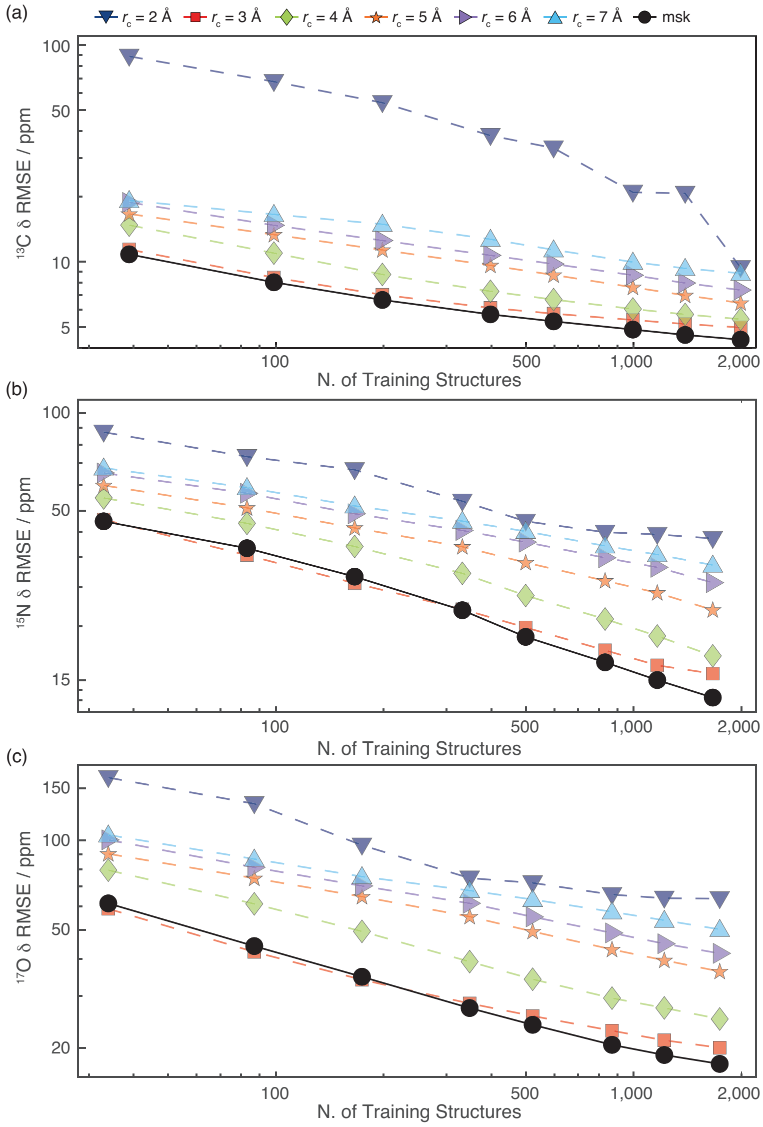


Figure 2-27. RMSE evaluation curves showing the error between chemical shifts calculated with DFT and ShiftML. The curves are relative to ^{13}C **(a)**, ^{15}N **(b)** and ^{17}O **(c)** chemical shieldings. The errors were measured for different training set sizes, and evaluated on the CSD-500 test set. The multi-kernel learning-curve is labelled as msk.

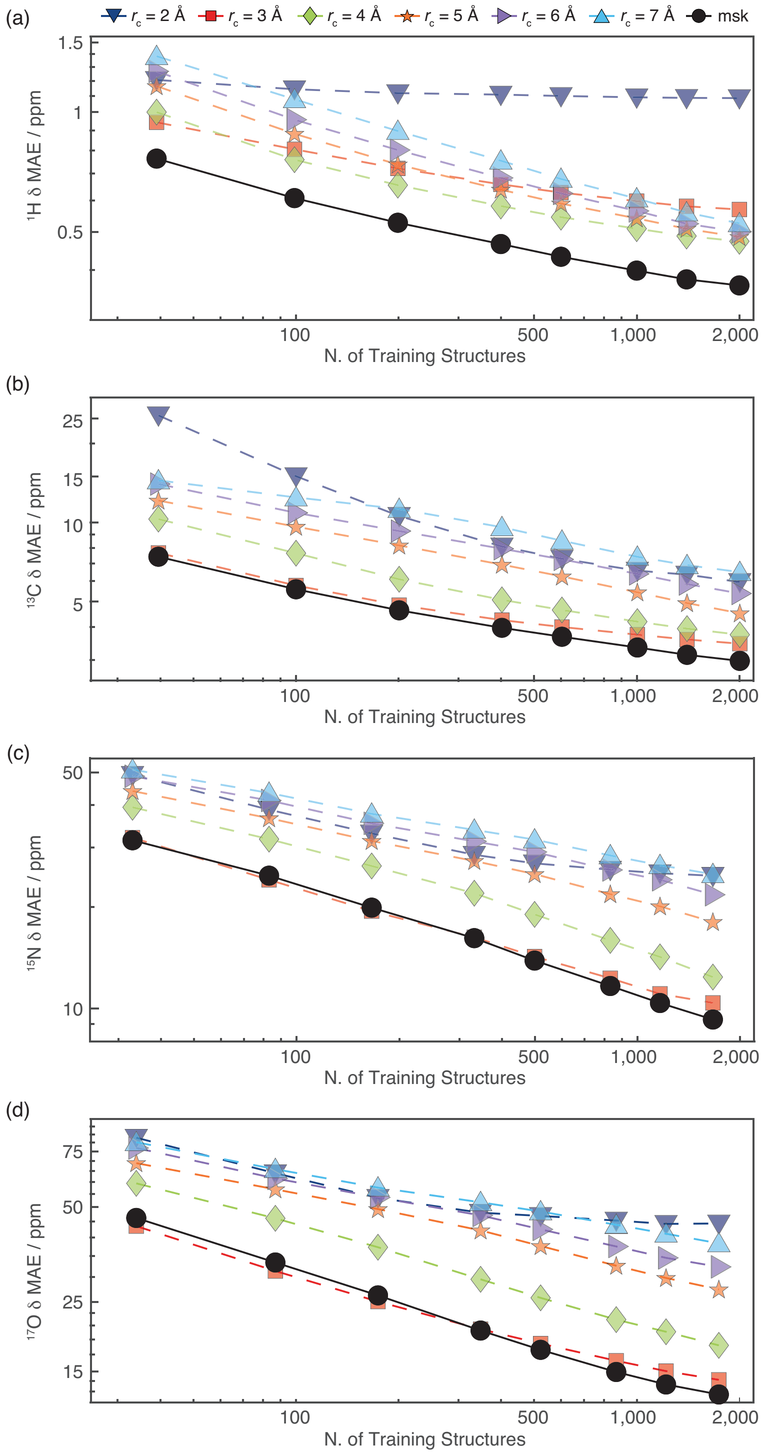


Figure 2-28. MAE evaluation curves showing the error between chemical shifts calculated with DFT and ShiftML. The curves are relative to ^{1}H **(a)**, ^{13}C **(b)**, ^{15}N **(c)** and ^{17}O **(d)** chemical shielding. The errors were measured for different training set sizes, and evaluated on the CSD-500 test set. The multi-kernel learning-curve is labelled as msk.

Comparison to Experiments

Comparison between ^{1}H experimental chemical shifts and ^{1}H chemical shifts calculated with ShiftML were carried out analyzing 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 analyzed crystal structures are the following:

(i) Naproxen, (2S)-2-(6-Methoxy-2-naphthyl)propanoic acid, COYRUD11, Ref. [210]

(ii) Uracil, Pyrimidine-2,4(1H,3H)-dione, URACIL, Ref. [211]

(iii) Co-crystal of 3,5-dimethylimidazole and 4,5-dimethylimidazole, Ref. [212]

(iv) Theophylline, 1,3-Dimethyl-3,7-dihydro-1H-purine-2,6-dione, BAPLOT01, Ref. [56]

(v) Cocaine, methyl (1R,2R,3S,5S)-3- (benzoyloxy)-8-methyl-8-azabicyclo[3.2.1] octane-2-carboxylate, COCAIN10, Ref. [56]

(vi) AZD8329, 4-[4-(2-adamantylcarbamoyl)-5-tert-butylpyrazol-1-yl]benzoic acid, Ref. [58]

The crystal structures (i-iv) were obtained from Ref. [83], 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. [56] and [58] 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 ^{1}H 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 $\delta = \sigma_{\text{ref}} - \beta\sigma$. 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 2-29** and **Table 2-9**.

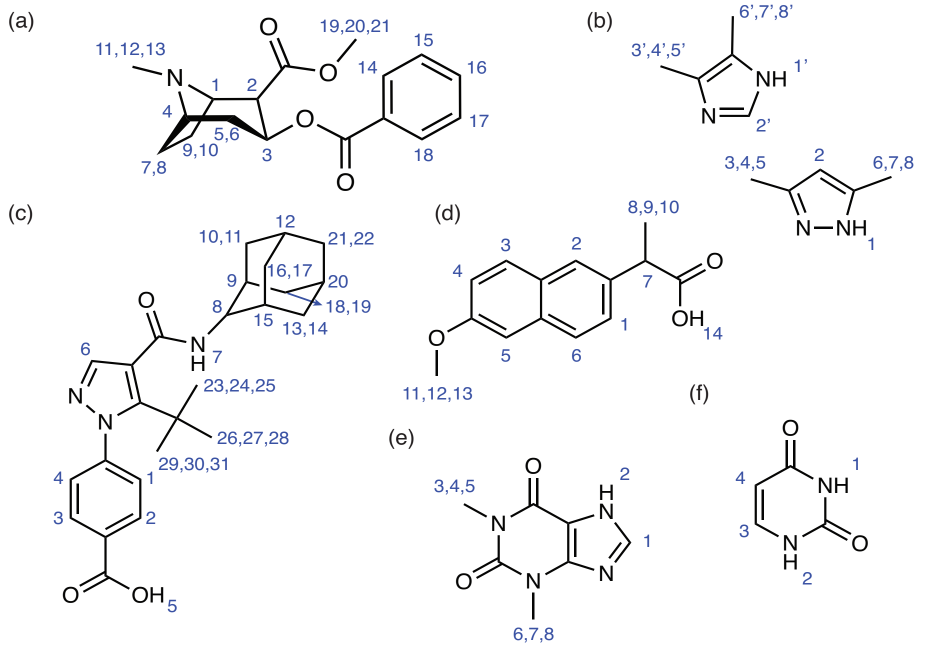


Figure 2-29. Chemical structures of the compounds used for experimental comparison. In order, cocaine **(a)**, 3,5-dimethylimidazole and 4,5-dimethylimidazole **(b)**, AZD8329 **(c)**, naproxen **(d)**, theophylline **(e)** and uracil **(f)**, and the labelling scheme used here.

Table 2-9. 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 2-29**. When more than one atom corresponds to a single chemical shift value, their values were averaged.

Naproxen			**Uracil**		
Atom Label	Experimental ^{1}H δ (ppm)	ShiftML ^{1}H δ (ppm)	Atom Label	Experimental ^{1}H δ (ppm)	ShiftML ^{1}H 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 ^{1}H δ (ppm)	ShiftML ^{1}H δ (ppm)	Atom Label	Experimental ^{1}H δ (ppm)	ShiftML ^{1}H 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
15	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

Structures and Chemical Shifts of the CSD-6 Set

For all 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.

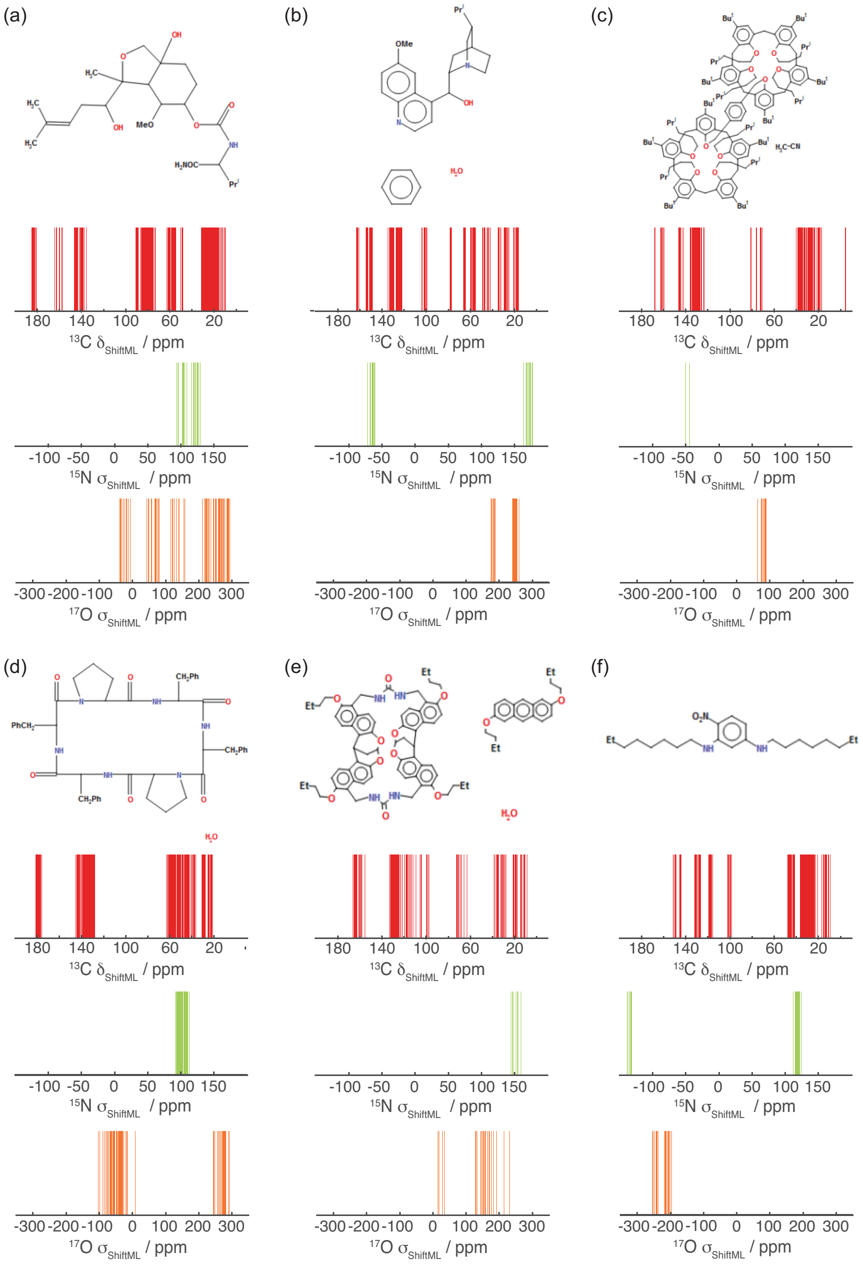


Figure 2-30. Chemical formula and corresponding 13C, 15N and 17O NMR spectra predicted using ShiftML of the six large molecular crystals with CSD Refcodes. **(a)** CAJVUH,[192] Natoms = 828, **(b)** RUKTOI,[193] Natoms = 768, **(c)** EMEMUE,[194] Natoms = 860, **(d)** GOKXOV,[195] Natoms = 945, **(e)** HEJBUW,[196] Natoms = 816, **(f)** RAYFEF,[197] Natoms = 1,584.

2.4 Positional variance and uncertainty

This chapter has been adapted with permission from: Hofstetter, A.; Emsley, L., “Positional variance in NMR crystallography”. *Journal of the American Chemical Society* **2017**, 139 (7), 2573-2576. **(post-print)**

2.4.1 Introduction

The scope of the combined CSP-NMRX approach has rapidly increased and today there are many examples of structure validation and determination by chemical shift measurements combined with DFT[8, 18, 37, 56-58, 72-73, 141-142, 148] In **Chapters 2.2** and **2.3** we have introduced approaches to reduce the computational cost of CSP-NMRX and thus to further extend the scope of this combined approach. In this chapter we investigate a further aspect of CSP-NMRX, namely the positional accuracy of the determined structures. In contrast to diffraction based methods, there exists no protocol to quantify the positional errors on individual atoms for structures determined by chemical shift based NMRX.

We propose a method to quantify positional uncertainties in crystal structures determined by chemical shift based NMR crystallography. The method combines MD simulations and DFT calculations with experimental and computational chemical shift uncertainties. In this manner we determine the average positional accuracy as well as the isotropic and anisotropic positional accuracy associated with each atom in a crystal structure determined by NMRX. The approach is demonstrated on the crystal structures of cocaine, flutamide, flufenamic acid, the K salt of penicillin G, and form 4 of the drug 4-[4- (2-adamantylcarbamoyl)-5-tert-butyl-pyrazol-1-yl]benzoic acid (AZD8329), which have been recently characterized by NMRX.[56, 58, 142] We find that, for the crystal structure of cocaine, the uncertainty corresponds to a positional root mean squared deviation (RMSD) of 0.17 Å. This is a factor of 2.5 less than for single crystal X-ray diffraction based structure determination.

2.4.2 Methods

First, we generate an ensemble of slightly perturbed crystal structures with 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 temperature ranges used and the associated computational costs are detailed in the **Appendix II**. Predicted ^{1}H and ^{13}C chemical shifts are then calculated for the members of the ensemble using plane wave DFT and the GIPAW[63] method. Given the estimated errors in the measured and predicted chemical shifts, we then correlate this directly with the atomic positions that are compatible with the measured chemical shifts to within the error, yielding a distribution of positions for each atom. The positional distributions are then converted into anisotropic displacement parameters (ADPs)[213], which can be represented by ellipsoids on the determined structure. The results of this process are given in **Figure 2-31** for cocaine, flutamide, flufenamic acid, AZD8329[58] and the K salt of Penicillin G.

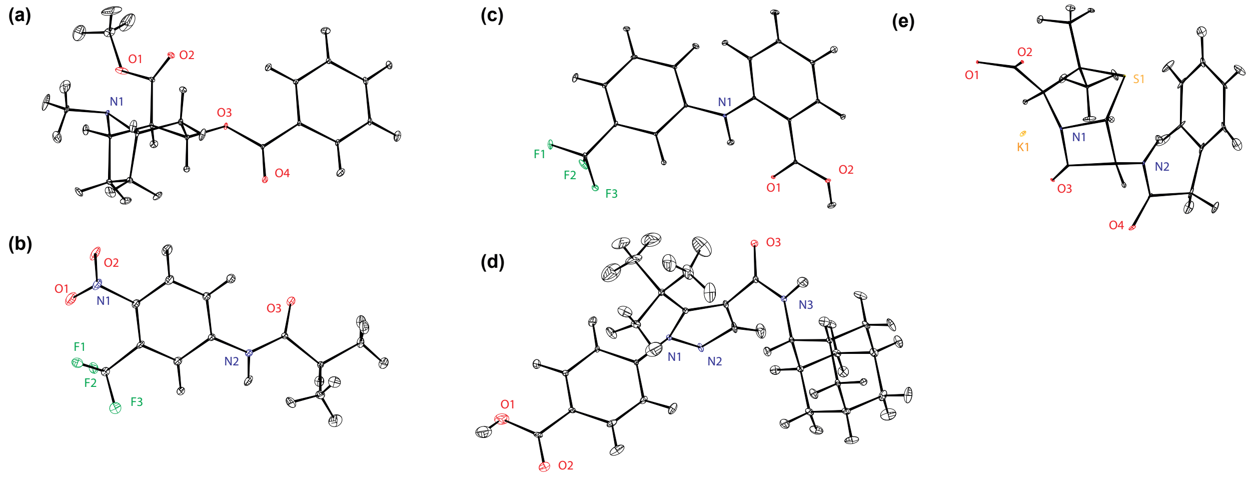


Figure 2-31. ORTEP plots drawn at the 90% probability level for the NMR determined crystal structures of **(a)** cocaine **(b)** flutamide, **(c)** flufenamic acid, **(d)** AZD8329 and **(e)** the K salt of Penicillin G. The ellipsoids correspond to positions within a ^{1}H chemical shift RMSD of 0.49 ppm.

To obtain the correlation between the chemical shift uncertainty and the ADPs, first the chemical shift RMSD between each structure in an ensemble and a reference structure from the ensemble is calculated. Next the positional deviations between each structure and the reference structure are calculated. For each individual atom the principle axis system (PAS) of the ensemble of positional deviations is determined using principle component analysis (PCA) as detailed in the **Appendix II**. This results in a scatter plot of the type shown in **Figure 2-32a**.

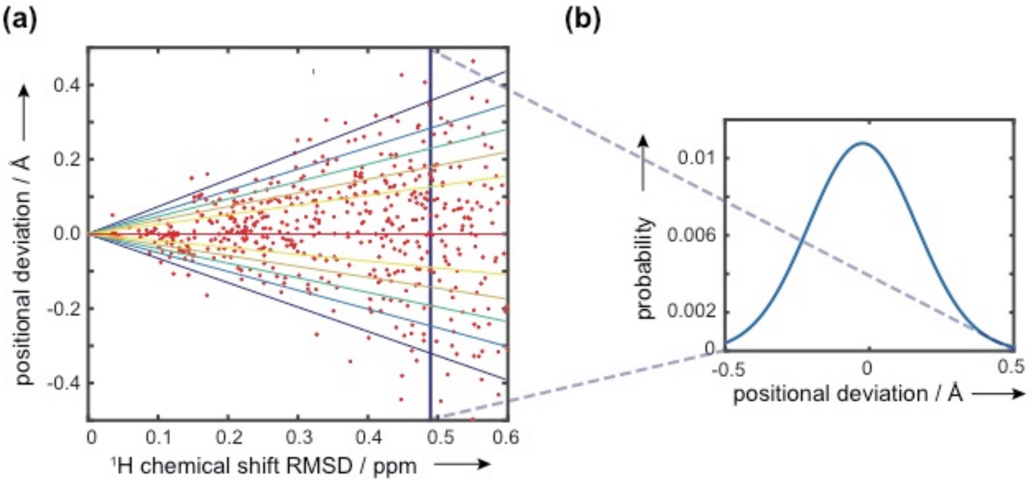


Figure 2-32. (a) Contour plot of the Gaussian fit of the correlation between the positional displacement (Å) and the ^{1}H chemical shift RMSD (ppm) along one principal axis of the anisotropic displacement tensor for the O1 atom for the cocaine crystal structure. **(b)** Probability distribution of the positional displacement (Å) for to a ^{1}H chemical shift RMSD of 0.49 ppm.

A continuous correlation function is obtained by maximizing the log-likelihood between the correlation points and a Gaussian distribution:

$$G(\langle r_{i,l}\rangle,\langle\delta\rangle)=\frac{1}{\sqrt{2\pi\Sigma_{i,l}^2\langle\delta\rangle^2}}\exp\left\{-\frac{(\langle r_{i,l}\rangle-\mu_{i,l}\langle\delta\rangle)^2}{2\Sigma_{i,l}^2\langle\delta\rangle^2}\right\},$$

(2-11)

where $< r >$ denotes the positional deviation, $< \delta >$ the chemical shift RMSD, Σ the scaling of the variance and μ the scaling of the mean. The indices *l* and *i* denote the atom and the principle axis respectively. The fit parameters are Σ and μ. The detailed procedure is given in **Appendix II**. The result of this procedure for the O1 atom of cocaine is shown in **Figure 2-32**. Please note, that the uncertainty prediction method described here is not limited to the use of a Gaussian distribution function (details in the **Appendix II**).

The principal values of the ADPs in the PAS are calculated as the mean-square displacements, which for Gaussian distributions is given as the variance, as a function of the chemical shift RMSD,

$$U_{ii,l}^{PAS}=\Sigma_{i,l}^2\langle\delta\rangle^2.$$

(2-12)

The amplitudes of the second rank tensors describing the ellipsoids at a given probability (*W*) are calculated in the PAS, where they are diagonal, as,

$$T_{ii,l}^{PAS}=p_{i,l}(W,\langle\delta\rangle)^2,$$

(2-13)

where $p_{i,l}(W, < \delta >)$ denotes the *W*th percentile of the fitted Gaussian for a chemical shift RMSD $< \delta >$. These are the quantities that are usually plotted in so-called ORTEP plots as anisotropic displacement ellipsoids, and this is what is shown in **Figure 2-1**.

Note that, for simplicity, or for cases with insignificant anisotropy in the displacements, the second rank ADP can be replaced by the equivalent isotropic displacement parameter.[214-215]

$$U_{eq}^{l} = \frac{1}{3}\left(U_{11,l}^{PAS} + U_{22,l}^{PAS} + U_{33,l}^{PAS}\right).$$

(2-14)

Note also that, from the equivalent isotropic displacement parameters, we can derive a global measurement of the positional uncertainty (U_{eq} and T_{eq}) for the whole structure, which is given as the average of the equivalent isotropic displacement parameters over all the *N* atoms in the structure,

$$U_{eq} = \frac{1}{N}\sum_{l=1}^{N} U_{eq}^{l}.$$

(2-15)

The radii of the isotropic spheres and of the average isotropic spheres at a certain probability (*W*) are calculated analog to the axes of the anisotropic displacement ellipsoids (**Equation 2-13**), the formula is detailed in the **Appendix II**. The isotropic spheres and the average isotropic spheres are shown for cocaine in **Figure 2-33b** and **Figure 2-33c** respectively. The average positional RMSD <r_{av}> for a given chemical shift RMSD <δ> is then calculated as ,

$$< r_{av} > = \sqrt{3U_{eq}}.$$

(2-16)

The factor $\sqrt{3}$ results from the fact that the isotropic displacement parameter is given as in **Equation 2-14**, while the RMSD is calculated as $< \mathrm{r} > = \sqrt{\Delta \mathrm{x}^2 + \Delta \mathrm{y}^2 + \Delta \mathrm{z}^2}$. (2-17)

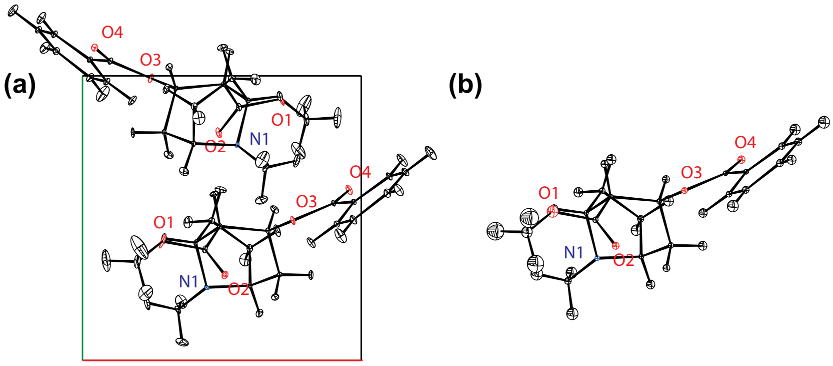


Figure 2-33. ORTEP plot of the cocaine structure drawn at the 90% probability level. **(a)** Anisotropic ellipsoids, corresponding to a ^{1}H chemical shift RMSD of 0.49 ppm. **(b)** Equivalent isotropic spheres, corresponding to a ^{1}H chemical shift RMSD of 0.49 ppm. **(c)** Average thermal spheres for a chemical shift RMSD <δ> of 0.49 ppm , corresponding to an average positional RMSD <r_{av}> of 0.169 Å.

2.4.3 Results

As indicated in **Figure 2-34**, we find that the positional RMSD <r$_{av}$> shows an approximately linear correlation with the average chemical shift RMSD $< \delta >$ for each of the five structures, but that the slope of the correlation is different for each structure. For example, for a given chemical shift RMSD the structure determined for penicillin has more than a factor two less uncertainty than that for flutamide. This is not surprising. The sources of this variation depend on the rigidity of the molecule, its hybridization, and the electron density gradients in the crystal structure. A detailed investigation of these factors will be the subject of future studies. Also, the positional uncertainty depends on how internal dynamics (such as methyl rotation) is accounted for (detailed in the **Appendix II**), as one of the main contributors to the positional RMSD. The positional uncertainty presented here should therefore be viewed as an upper limit.

From **Equations 2-12** and **2-14** to **2-16** the correlation between the chemical shift RMSD <δ> and the average positional RMSD <r$_{av}$> is,

$$< r_{av} > = \sqrt{\frac{1}{N}\sum_{i,l} \Sigma_{i,l}^{2}} < \delta > = \bar{\Sigma} < \delta >.$$

(2-18)

For the crystal structure of cocaine, we find a direct correlation ($\bar{\Sigma}$) of 0.345. Given an average chemical shift RMSD <δ> of 0.49 ppm, which is the current estimated upper limit for the accuracy in ^{1}H chemical shift based crystallography methods[18], this leads to an average positional RMSD <r$_{av}$> of around 0.169 Å, corresponding to an average equivalent displacement parameter (U_{eq}) of 0.0095 Å^{2}. Compared to other structure determination methods, for example XRD which yielded an average positional RMSD of 0.458 Å for the crystal structure of cocaine,[180 C 17 H 21 NO 4] we find an increase in positional accuracy by a factor 2.5.

It is interesting to note that for XRD the positional uncertainty mainly results from the thermal motion of the atoms and is a direct result of the decrease in scattering amplitude due to vibrations. In contrast, in NMR spectroscopy thermal motion and fast lattice vibrations lead to motional narrowing of the measured signal, and if anything, are likely to increase accuracy; thus, we see that the different techniques naturally have different limits on the positional accuracy.

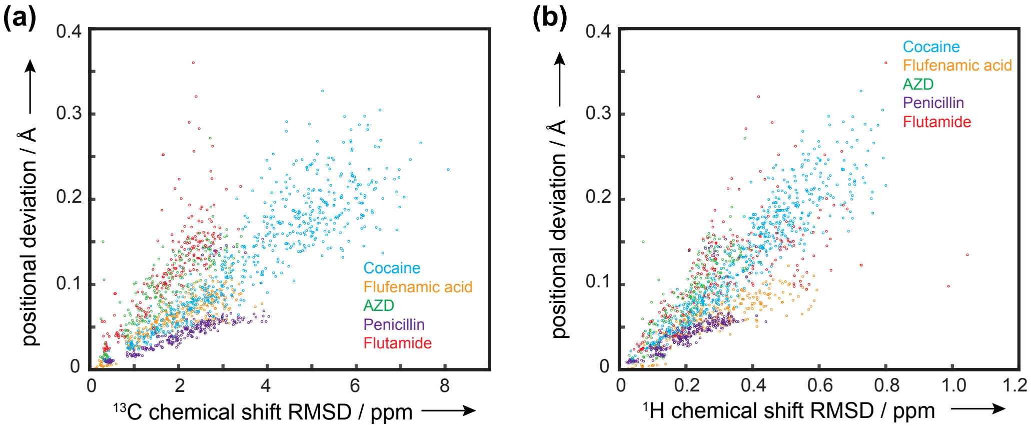


Figure 2-34. (a) Correlation between positional RMSD (Å) and ^{1}H chemical shift RMSD (ppm) for five ensembles of perturbed crystal structures generated by MD. **(b)** Correlation between positional RMSD (Å) and ^{13}C chemical shift RMSD (ppm) for five ensembles of slightly perturbed crystal structures.

We remark that the methods used to create the ensemble of structures and to calculate the chemical shifts are important in determining the positional errors. We have evaluated the use of different force-fields in the MD simulation, as well as a fixed versus a variable unit cell as discussed in the **Appendix II**, and we find they have no significant effect on the uncertainty quantification.

Comparable calculations were also done for an ensemble of perturbed cocaine crystal structures generated by random uncorrelated displacement of the atoms (i.e. this correspond to systematic uncorrelated bond stretching). For this ensemble the correlation predicts much larger deviations in chemical shift for a given average displacement (see **Figure 2-47**), and would lead to much higher apparent positional accuracy. This is expected, due to the generation of physically improbable structures resulting in an unreasonable electronic density. A possibility to overcome this would be to weight the random structures with a Boltzmann factor based on their calculated energy, but this should provide no direct advantage compared to the MD method. The MD method on the other hand searches the conformational space more efficiently and implicitly weights the generated structures with a Boltzmann factor. The random displacement method thus severely underestimates the positional errors. The MD ensemble allows for a significantly larger uncertainty in position than the random displacement method for a given chemical shift RMSD, and it is thus a better representation of the uncertainty in positions in the experimentally determined structures. We are currently exploring other methods to generate physically reasonable ensembles, for example through the exploitation of vibrational modes of the crystal structures.

Finally, it is possible, that the choice of the DFT functional might have an influence on the calculated uncertainties. The PBE[205] functional used here is the current standard for the computation of chemical shifts in molecular crystals,[216] and we remark that the systematic error in chemical shift calculations has shown to be similar for different functionals.[217] This systematic error likely results from the difficulty for DFT to correctly describe polar groups and long range dispersion forces, e.g. Hydrogen bonds. However, here we would not be sensitive to this systematic error, but only to any systematic variation within the error, which is likely to be small.

2.4.4 Conclusion

In conclusion, we have introduced a method to quantify positional uncertainties in crystal structures derived from NMR chemical shifts. The structures quantified here were determined by chemical shift based NMR crystallography, but in principle structures determined by other methods, e.g. XRD, could be refined with this method. An ensemble of structures around the experimentally determined structure is generated *in silico*, and the predicted chemical shift deviations for this ensemble are compared to the positional deviations. In this way we determine the average positional error of the experimentally determined structure for each atom in the crystal structure. We find that the average positional uncertainty in the five structures studied here yield an RMSD of 0.17 Å, or an average value of the equivalent displacement parameter of 0.0095 Å^2. We find that chemical shift based NMR crystallography methods provide a gain in positional accuracy of around a factor 2 compared to XRD structure determination. This is mainly because thermal vibrations are not limiting for chemical shift based NMR methods.

2.4.5 Appendix II

Experimental and Computational Details

Crystal Structures. The initial NMR determined crystal structures were obtained from the supplementary information of M. Baias et al[56], for cocaine, flutamide and flufenamic acid, and from the Cambridge Crystallographic Database for AZD8329 and the K salt of penicillin G . The CSD Refcodes for the structures are: BZPENK01 for the K salt of penicillin G and the CCDC number 957764 for AZD8329.

Molecular Dynamics. The all-atom optimized potential for liquid simulations (OPLS-aa) force-field[218] within the GROMACS suite[219] flexible, and free was used for the MD simulations. The force-field was chosen after performance tests with multiple force-fields. It is the same force-field as in the crystal prediction method used for the structure elucidation of cocaine, flutamide and flufenamic acid.[56] Prior to the MD simulations the crystal structures were relaxed in the force-field. During the relaxation the structures changed on average by a RMSD of 0.74 ± 0.18 Å. The relaxation was done to avoid any significant structural changes during the MD simulation. During the MD simulations the crystal structures were kept in a constant heat bath, at multiple values between 1° and 250° K, for 300ps. For the different compounds the simulation temperatures were set as given in **Table 2-10**. For each temperature the simulation was run for around 20min on 1 node with 2 Ivy Bridge processors running at 2.6 GHz, each with 8 cores and 64 GB of DDR3 RAM.

Table 2-10. Temperatures used during the MD simulations of the crystal structures.

cocaine	flutamide	flufenamic acid	AZD 8329	K salt of penicillin G
1° K, 5° to 50° K in steps of 5° K, 60° to 250° K in steps of 10° K.	1° K, 5° to 50° K in steps of 5° K,	1° K, 5° to 60° K in steps of 5° K,	1° K, 5° to 55° K in steps of 5° K,	1° K, 5° to 55° K in steps of 5° K,

DFT Chemical Shift Calculations. The DFT calculations were performed using the generalized gradient approximation (GGA) functional PBE[205] within the Quantum Espresso code[188]. The plane-wave cutoff energy and the reciprocal grid spacing were optimized for each crystal structure, and found to be: cocaine, Ecutoff = 70 Ry with a 2 x 2 x 2 Monkhorst-Pack grid of k-points[220]; flutamide, Ecutoff = 80 Ry and a 2 x 1 x 3 Monkhorst-Pack grid of k-points; Penicillin G, Ecutoff = 50 Ry and a 2 x 2 x 1 Monkhorst-Pack grid of k-points; flufenamic acid, Ecutoff = 110 Ry and a 1 x 2 x 1 Monkhorst-Pack grid of k-points; AZD8329, Ecutoff = 60 Ry and a 2 x 2 x 2 Monkhorst-Pack grid of k-points. The chemical shifts δ_{calc} were calculated using the GIPAW method[63] with the parametrization described above. For each compound chemical shifts were calculated for an ensemble of 620 (cocaine), 260 (flufenamic acid), 240 (penicillin and AZD8329) and 220 (flutamide) structures extracted uniformly from the 1.92 x107 (cocaine), 7.8 x106 (flufenamic acid), 7.2 x106 (penicillin and AZD8329) and 6.6 x106 (flutamide) in the complete MD set. For each structure the DFT calculation was run for around 120min on 2 nodes each with 2 Ivy Bridge processors running at 2.6 GHz, each with 8 cores and 64 GB of DDR3 RAM.

Continuous Positional and Chemical Shift Correlation

Chemical Shift RMSD. The chemical shift RMSD is determined by a linear regression between the reference and calculated chemical shifts ($\delta_{ref} = a - b\,\delta_{calc}$), illustrated in **Figure 2-35**. It is used as a measurement of the goodness of the fit for a given trial structure.

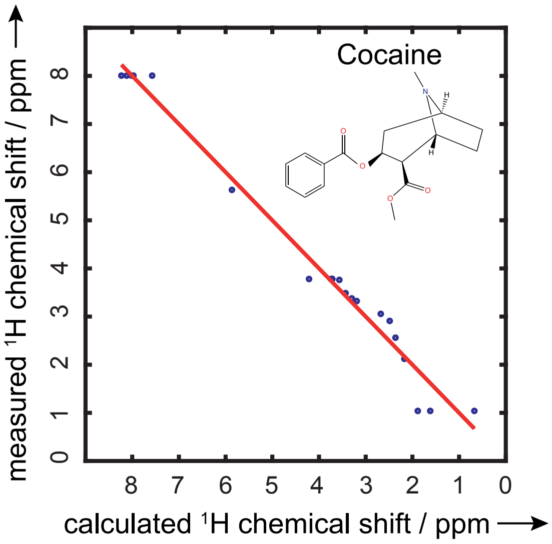


Figure 2-35. Linear regression between the reference and calculated ^{1}H chemical shifts for a predicted crystal structure of cocaine.

Principle Component Analysis (PCA)

For each atom (*l*) the mean atomic position ($\vec{m}_l$) and the covariance matrix of displacements (Σ_l) over the whole ensemble of slightly perturbed structures is calculated:

$$\vec{m}_l = \frac{1}{M}\sum_{k=1}^{M} \vec{r}_l(k), \tag{2-19}$$

$$\Sigma_l = \frac{1}{M}\sum_{k=1}^{M} (\vec{r}_l(k) - \vec{m}_l)\,(\vec{r}_l(k) - \vec{m}_l)^{\mathrm{T}}, \tag{2-20}$$

where *k* indexes the structure and *M* denotes the total number of structures in the ensemble.

The eigenvectors $\hat{u}_l = [\vec{u}_1^l, \vec{u}_2^l, \vec{u}_3^l]$ of the covariance matrix of displacements (Σ_l) are used to transform the displacement vectors in the reference frame ($\Delta\vec{A}_l\,(k)$) into the displacements in the principle axis system (PAS) ($\Delta\vec{B}_l\,(k)$).

$$\Delta\vec{A}_l\,(k) = \vec{r}_l(k) - \vec{r}_l(k_{ref}), \tag{2-21}$$

$$\Delta\vec{B}_l\,(k) = \hat{u}_l^{\mathrm{T}} \Delta\vec{A}_l\,(k). \tag{2-22}$$

Maximum Likelihood Estimation (MLE)

The correlation function ($G\left(\langle r_{i,l}(k)\rangle, \langle\delta(k)\rangle\right)$), which is detailed in the main text, is fitted to the discrete correlation data in order to find the parameters $\Sigma_{i,l}$ and $\mu_{i,l}$. Where *i* denotes the principle axis, *l* the atom and *k* the structure in the ensemble.

$$G\left(\langle r_{i,l}(k)\rangle, \langle\delta(k)\rangle\right) = \frac{1}{\sqrt{2\pi\Sigma_{i,l}^2\langle\delta(k)\rangle^2}} \exp\left\{-\frac{(\langle r_{i,l}(k)\rangle - \mu_{i,l}\langle\delta(k)\rangle)^2}{2\Sigma_{i,l}^2\langle\delta(k)\rangle^2}\right\}.$$

(2-23)

For this subchapter the indices *i* and *l* are omitted, but the procedure is done for each principle axis (*i*) of each atom (*l*) individually. The fit is done by maximizing the logarithm of the Likelihood functional ($L[G|\mu,\Sigma]$).

$$L[G|\mu,\Sigma] = \prod_{k=1}^{M} G(\langle r(k)\rangle, \langle\delta(k)\rangle) = (\prod_{K=1}^{M} \langle\delta(k)\rangle^{-1}) \left(\frac{1}{2\pi\Sigma^2}\right)^{M/2} \exp\left\{-\sum_{k=1}^{M} \frac{(\langle r(k)\rangle - \mu\langle\delta(k)\rangle)^2}{2\Sigma^2\langle\delta(k)\rangle^2}\right\},$$

(2-24)

$$\log L[G|\mu,\Sigma] = -\left(\sum_{k=1}^{M} \log\langle\delta(k)\rangle\right) - \frac{M}{2}\log 2\pi\Sigma^2 - \sum_{k=1}^{M} \frac{(\langle r(k)\rangle - \mu\langle\delta(k)\rangle)^2}{2\Sigma^2\langle\delta(k)\rangle^2}.$$

(2-25)

Where *k* indexes the structure and *M* denotes the total number of structures in the ensemble. The maximum of the log-Likelihood is found by differentiating with respect to the parameters Σ and μ.

$$\frac{\partial \log L[G|\mu,\Sigma]}{\partial\mu} = \frac{\partial \log L[G|\mu,\Sigma]}{\partial\Sigma} = 0,$$

(2-26)

$$\mu = \frac{1}{M}\sum_{k=1}^{M} \frac{\langle r(k)\rangle}{\langle\delta(k)\rangle},$$

(2-27)

$$\Sigma^2 = \frac{1}{M}\sum_{k=1}^{M} \frac{(\langle r(k)\rangle - \mu\langle\delta(k)\rangle)^2}{\langle\delta(k)\rangle^2}.$$

(2-28)

Isotropic and average isotropic spheres

The isotropic spheres are calculated as,

$$T_{eq}^{l} = \frac{1}{3}\left(T_{11,l}^{PAS} + T_{22,l}^{PAS} + T_{33,l}^{PAS}\right),$$

(2-29)

where *l* denotes the index of the atom. Analog, the average isotropic spheres are calculated as,

$$T_{eq} = \frac{1}{N}\sum_{l=1}^{N} T_{eq}^{l},$$

(2-30)

where N denotes the total number of atoms in the structure.

Cauchy-Lorentz distribution as an alternative kernel

As mentioned above, the uncertainty quantification method described above is not limited to the use of a Gaussian distribution function as the kernel for the MLE. For the observed correlation a Gaussian kernel seems to be a pertinent choice, but in other cases a different kernel might be more appropriate. One of the main advantages of a Gaussian kernel is the existence of an analytical solution for the MLE. For other kernels, where no analytical solution exists, a more general procedure, as illustrated with a Cauchy-Lorentz distribution, can be used.

A Cauchy-Lorentz distribution ($C\big(\langle r_{i,l}(k)\rangle, \langle\delta(k)\rangle\big)$), following the same principles as the Gaussian distribution above, is fitted to the discrete correlation data in order to determine the parameters $\gamma_{i,l}$ and $\mu_{i,l}$.

$$C\big(\langle r_{i,l}(k)\rangle, \langle\delta(k)\rangle\big) = \frac{1}{\pi\gamma_{i,l}\langle\delta\rangle}\left(1 + \left(\frac{\langle r_{i,l}(k)\rangle - \mu_{i,l}\langle\delta(k)\rangle}{\gamma_{i,l}\langle\delta(k)\rangle}\right)^{2}\right)^{-1}.$$

(2-31)

For this subchapter the indices *i* and *l* are omitted, but the procedure is done for each principle axis (*i*) of each atom (*l*) individually. The fit is done by numerically minimizing the logarithm of the inverse Likelihood functional ($L[C|\mu,\Sigma]$).

$$L[C|\mu,\Sigma] = \prod_{k=1}^{M} C(\langle r(k)\rangle, \langle\delta(k)\rangle) = \prod_{k=1}^{M} \frac{\gamma\langle\delta(k)\rangle}{\pi}\left(\gamma_{i,l}^{2}\langle\delta(k)\rangle^{2} + \big(\langle r_{i,l}(k)\rangle - \mu_{i,l}\langle\delta(k)\rangle\big)^{2}\right)^{-1},$$

(2-32)

$$-\log L[C|\mu,\Sigma] = -M\log[\gamma] + M\log[\pi] - \left(\sum_{k=1}^{M}\log\left[\langle\delta(k)\rangle\right]\right) + \left(\sum_{k=1}^{M}\log\left[\gamma_{i,l}^{2}\langle\delta(k)\rangle^{2} + \big(\langle r_{i,l}(k)\rangle - \mu_{i,l}\langle\delta(k)\rangle\big)^{2}\right]\right).$$

(2-33)

Due to the fact that the Cauchy-Lorentz distribution does not possess any moments of finite order, it is impossible to calculate the principal values of the ADPs in the PAS as the mean-square displacements of the Cauchy-Lorentz distribution. A possible estimation of the mean-square displacement can be to use the value of displacement at the 68th percentile, in accordance with the Gaussian mean-square displacement. By applying this estimation, we calculate an average positional RMSD $< r_{av} >$ of 0.061 Å for a ^{1}H chemical shift RMSD of 0.49 ppm for Cocaine, compared to an average positional RMSD of 0.169 Å for a Gaussian kernel. Please note, that the Cauchy-Lorentz kernel is just used here to illustrate the potential applicability of the uncertainty quantification described here for different MLE kernels. ***Also note that the Cauchy-Lorentz kernel is not an appropriate choice for the correlation we observe here and thus severely underestimates the average positional RMSD*** $<\boldsymbol{r_{av}}>$.

Displacement Parameters from ^{1}H Chemical Shifts

The ORTEP plots were made with the programs CRYSTALS[221] and CAMERON[222 1996, CAMERON].

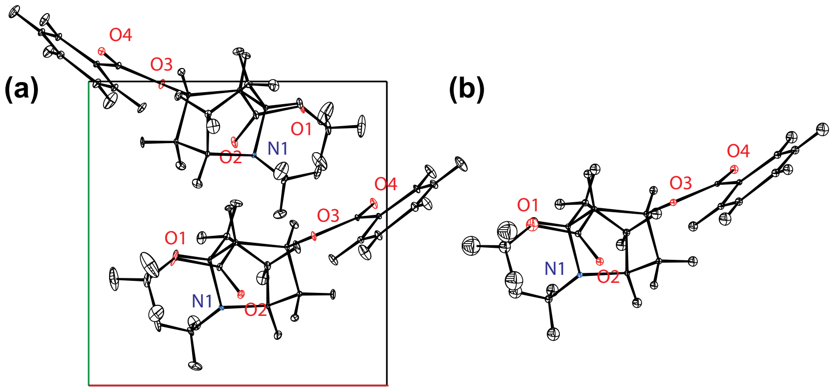


Figure 2-36. ORTEP plot of the unperturbed cocaine structure drawn at the 90 % probability level**. (a)** Anisotropic ellipsoids, corresponding to a ^{1}H chemical shift RMSD of 0.49 ppm. **(b)** Equivalent isotropic spheres, corresponding to a ^{1}H chemical shift RMSD of 0.49 ppm. **(c)** Average thermal spheres. The ^{1}H chemical shift RMSD of 0.49 ppm leads to a structural positional RMSD with a 90% confidence interval of 0.169 Å.

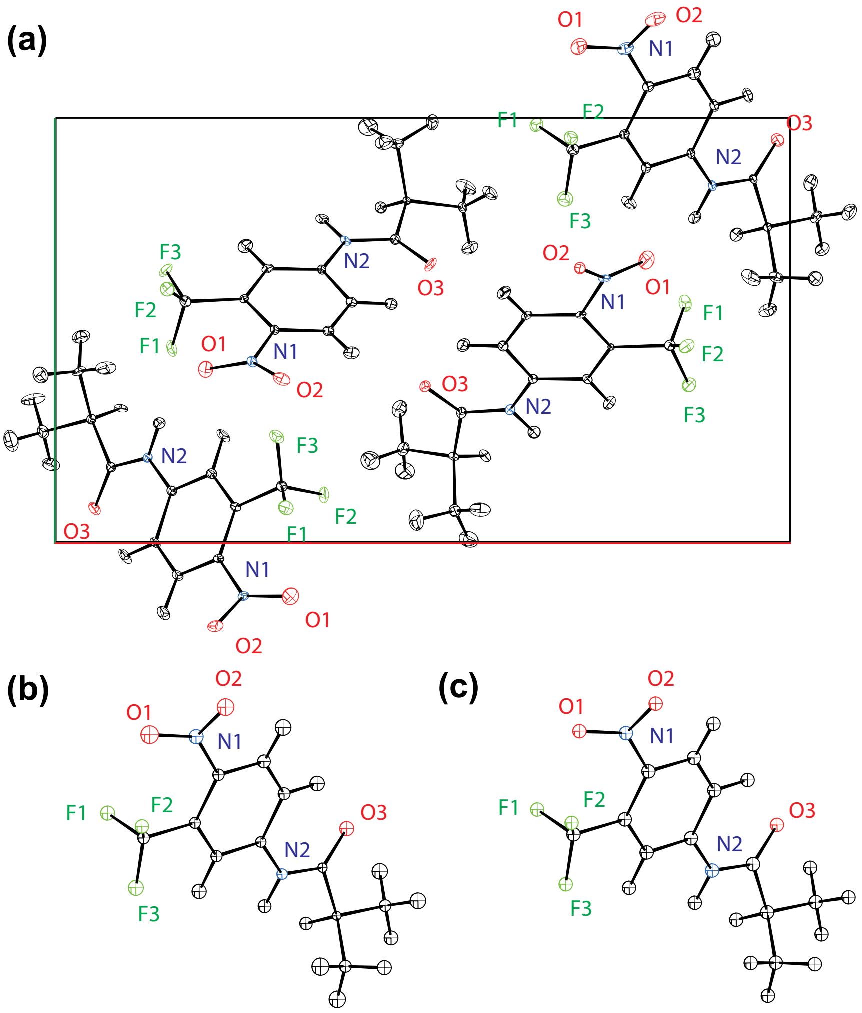


Figure 2-37. ORTEP plot of the unperturbed flutamide structure drawn at the 90 % probability level**.** **(a)** Anisotropic ellipsoids, corresponding to a ^{1}H chemical shift RMSD of 0.49 ppm. **(b)** Equivalent isotropic spheres, corresponding to a ^{1}H chemical shift RMSD of 0.49 ppm. **(c)** Average thermal spheres. The ^{1}H chemical shift RMSD of 0.49 ppm leads to a structural positional RMSD with a 90% confidence interval of 0.202 Å.

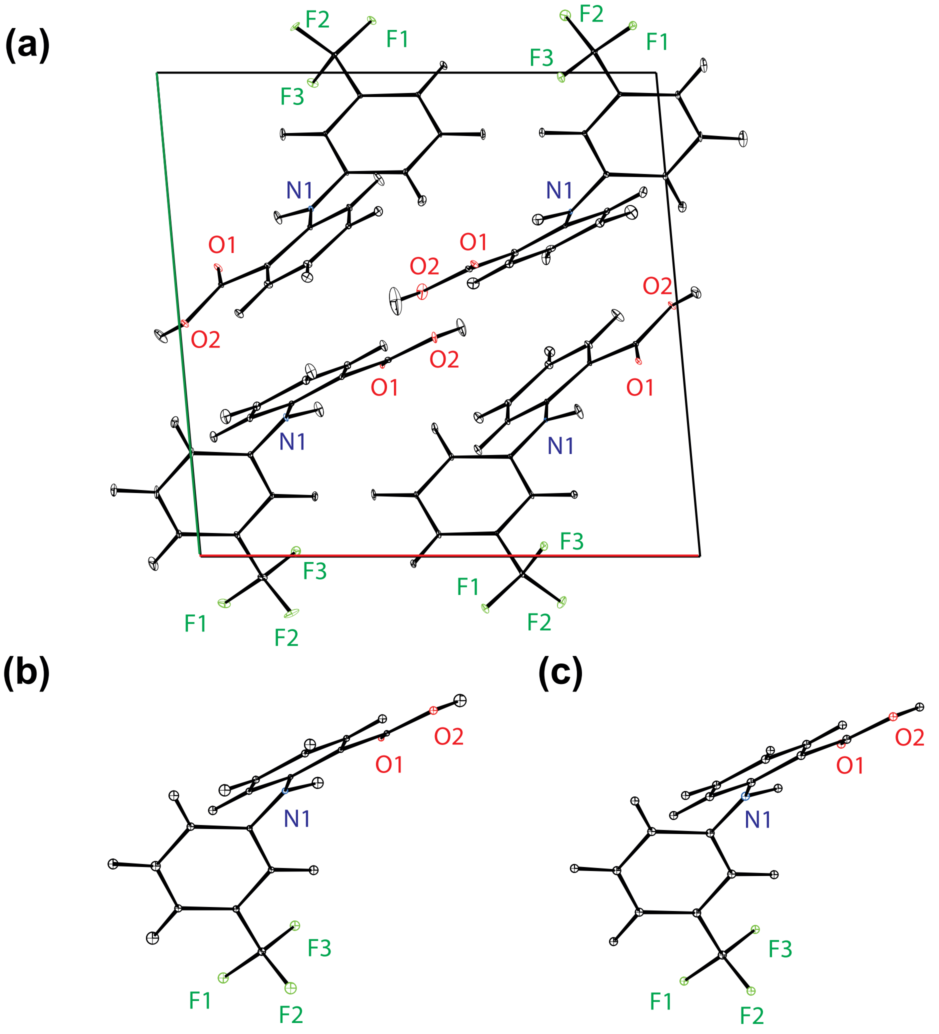


Figure 2-38. ORTEP plot of the unperturbed flufenamic acid structure drawn at the 90 % probability level**. a)** Anisotropic ellipsoids, corresponding to a ^{1}H chemical shift RMSD of 0.49 ppm. **b)** Equivalent isotropic spheres, corresponding to a ^{1}H chemical shift RMSD of 0.49 ppm. **c)** Average thermal spheres. The ^{1}H chemical shift RMSD of 0.49 ppm leads to a structural positional RMSD with a 90% confidence interval of 0.111 Å.

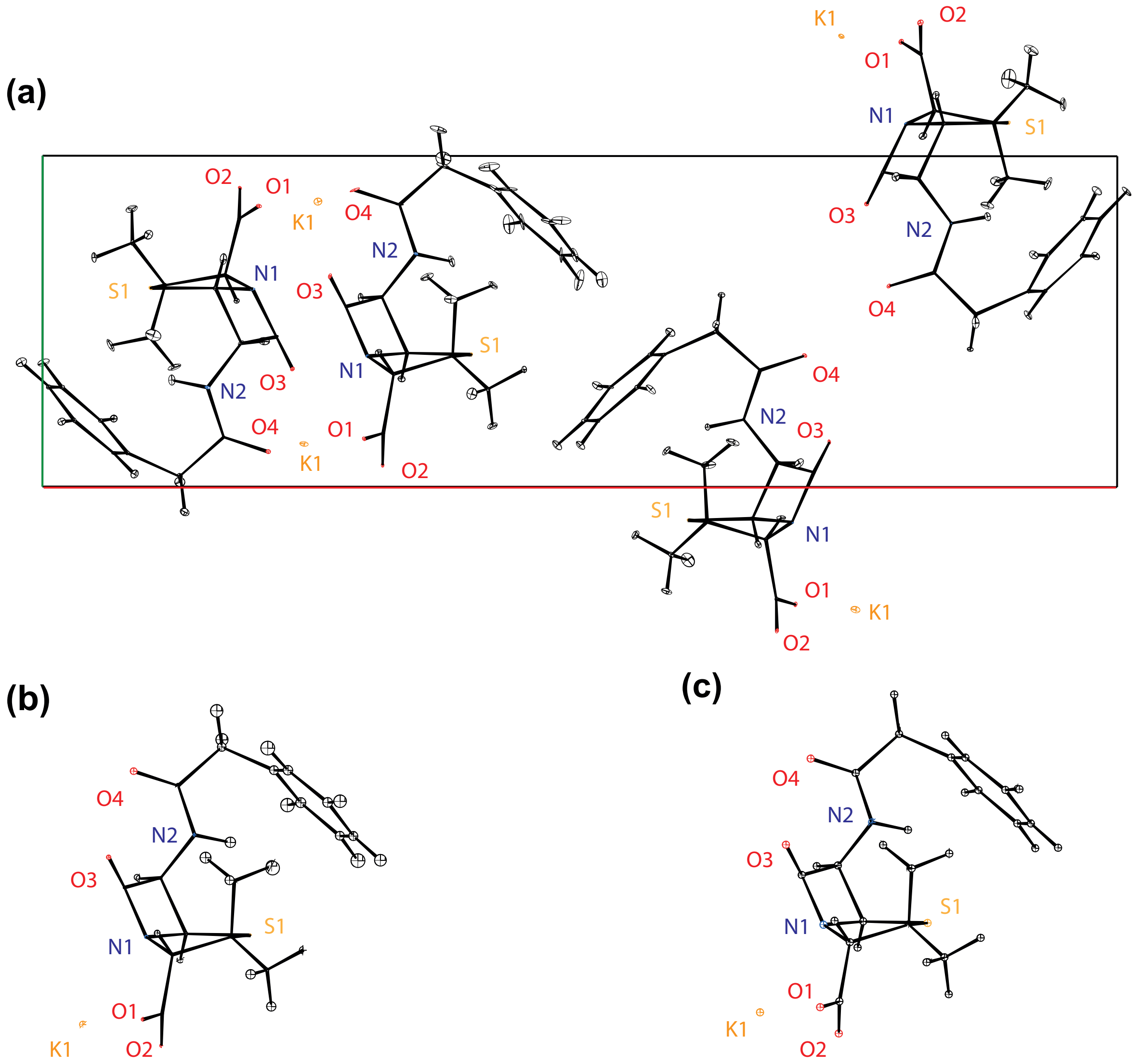


Figure 2-39. ORTEP plot of the unperturbed penicillin structure drawn at the 90 % probability level. **(a)** Anisotropic ellipsoids, corresponding to a ^{1}H chemical shift RMSD of 0.49 ppm. **(b)** Equivalent isotropic spheres, corresponding to a ^{1}H chemical shift RMSD of 0.49 ppm. **(c)** Average thermal spheres. The ^{1}H chemical shift RMSD of 0.49 ppm leads to a structural positional RMSD with a 90% confidence interval of 0.109 Å.

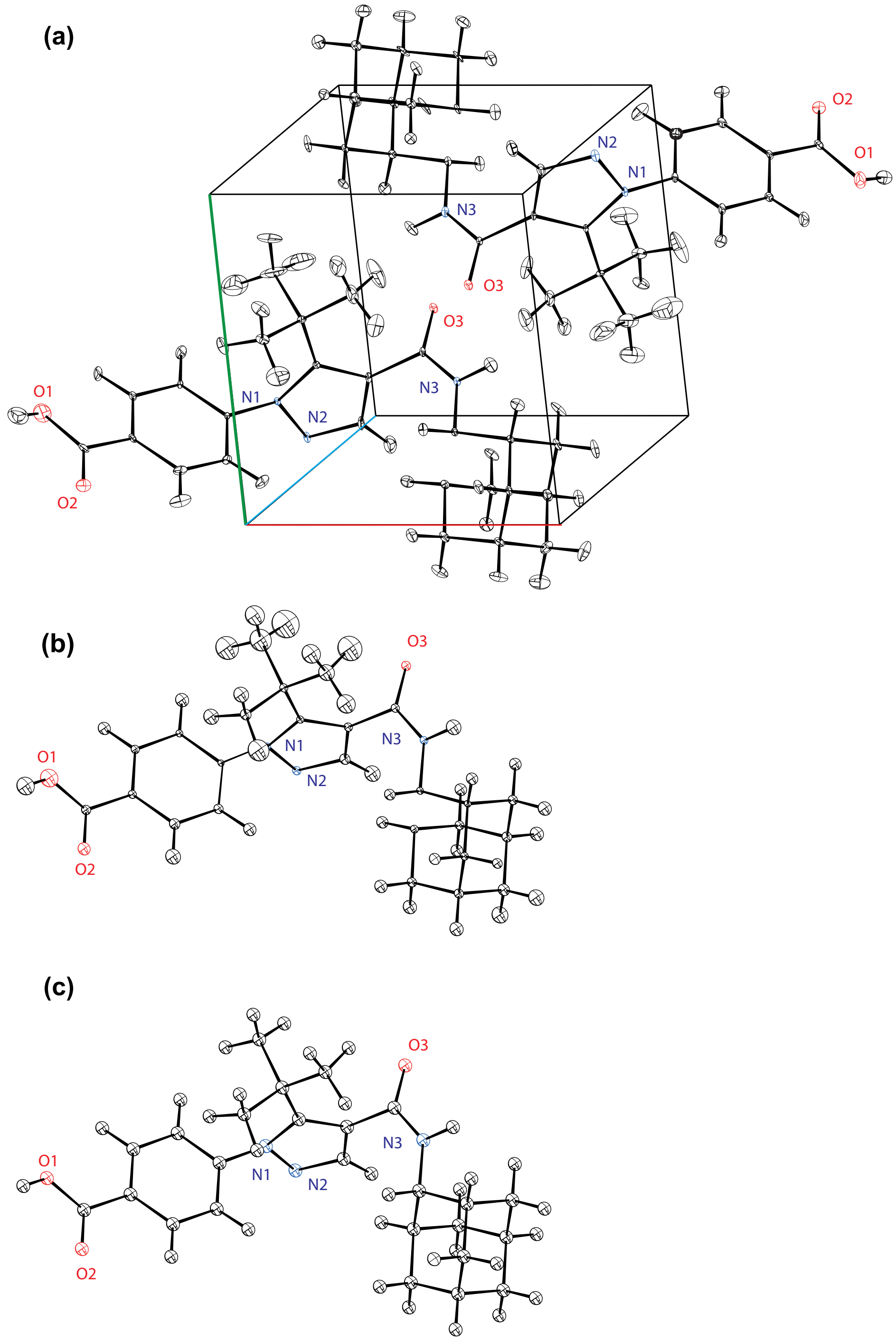


Figure 2-40. ORTEP plot of the unperturbed AZD8329 structure drawn at the 90 % probability level**. (a)** Anisotropic ellipsoids, corresponding to a ^{1}H chemical shift RMSD of 0.49 ppm. **(b)** Equivalent isotropic spheres, corresponding to a ^{1}H chemical shift RMSD of 0.49 ppm. **(c)** Average thermal spheres. The ^{1}H chemical shift RMSD of 0.49 ppm leads to a structural positional RMSD with a 90% confidence interval of 0.215 Å.

Displacement Parameters from 13C Chemical Shifts

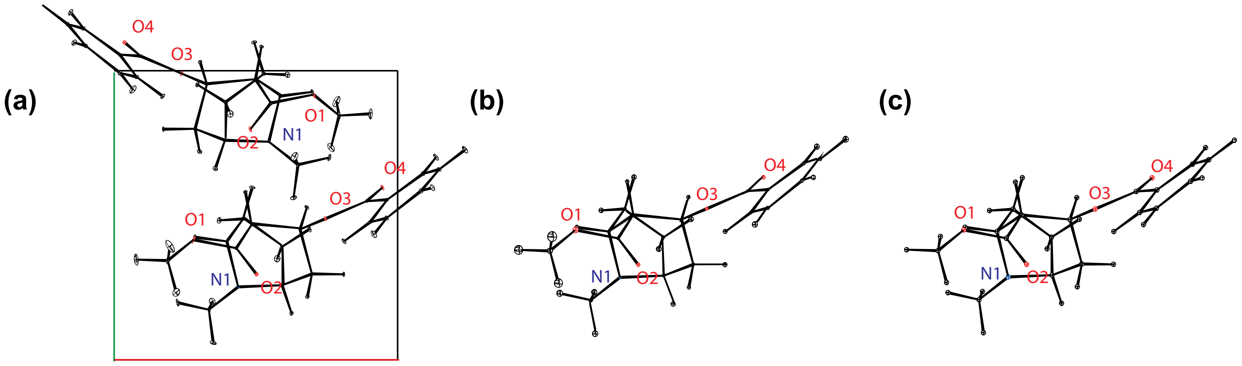


Figure 2-41. ORTEP plot of the unperturbed cocaine structure drawn at the 90 % probability level**. (a)** Anisotropic ellipsoids, corresponding to a ^{13}C chemical shift RMSD of 2.3 ppm. **(b)** Equivalent isotropic spheres, corresponding to a ^{13}C chemical shift RMSD of 2.3 ppm. **(c)** Average thermal spheres. The ^{13}C chemical shift RMSD of 2.3 ppm leads to a structural positional RMSD with a 90% confidence interval of 0.083 Å.

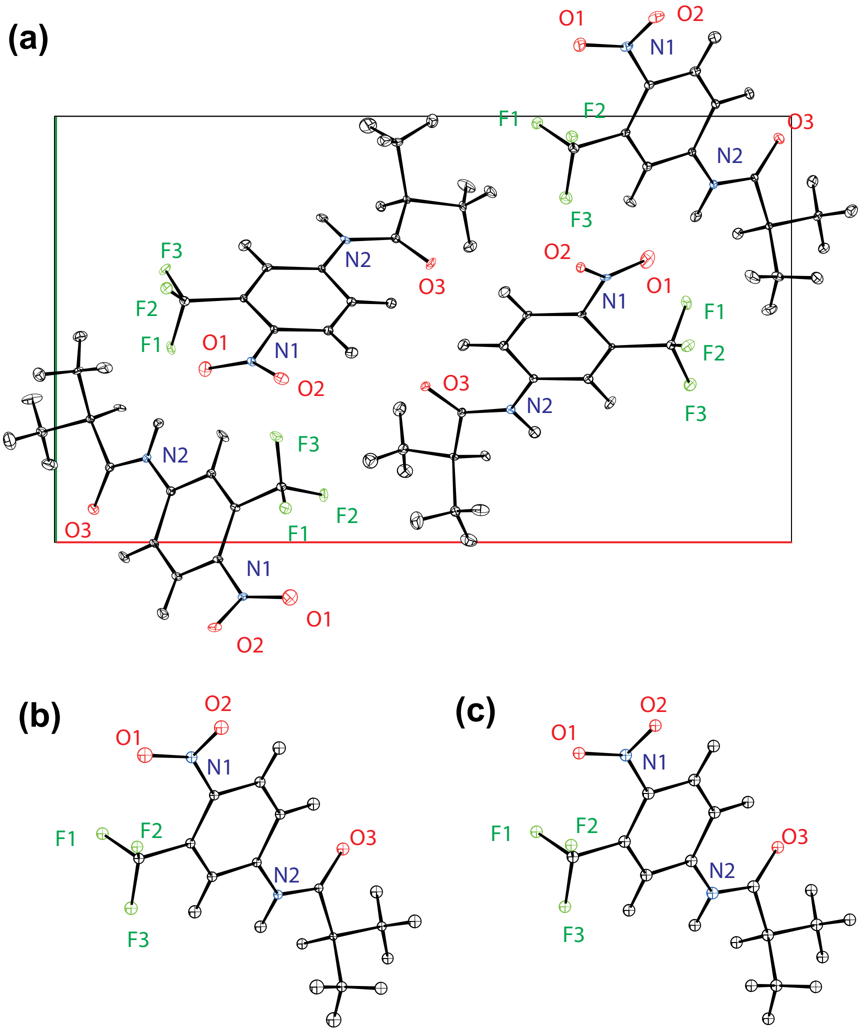


Figure 2-42. ORTEP plot of the unperturbed flutamide structure drawn at the 90 % probability level**. (a)** Anisotropic ellipsoids, corresponding to a ^{13}C chemical shift RMSD of 2.3 ppm. **(b)** Equivalent isotropic spheres, corresponding to a ^{13}C chemical shift RMSD of 2.3 ppm. **(c)** Average thermal spheres. The ^{13}C chemical shift RMSD of 2.3 ppm leads to a structural positional RMSD with a 90% confidence interval of 0.171 Å.

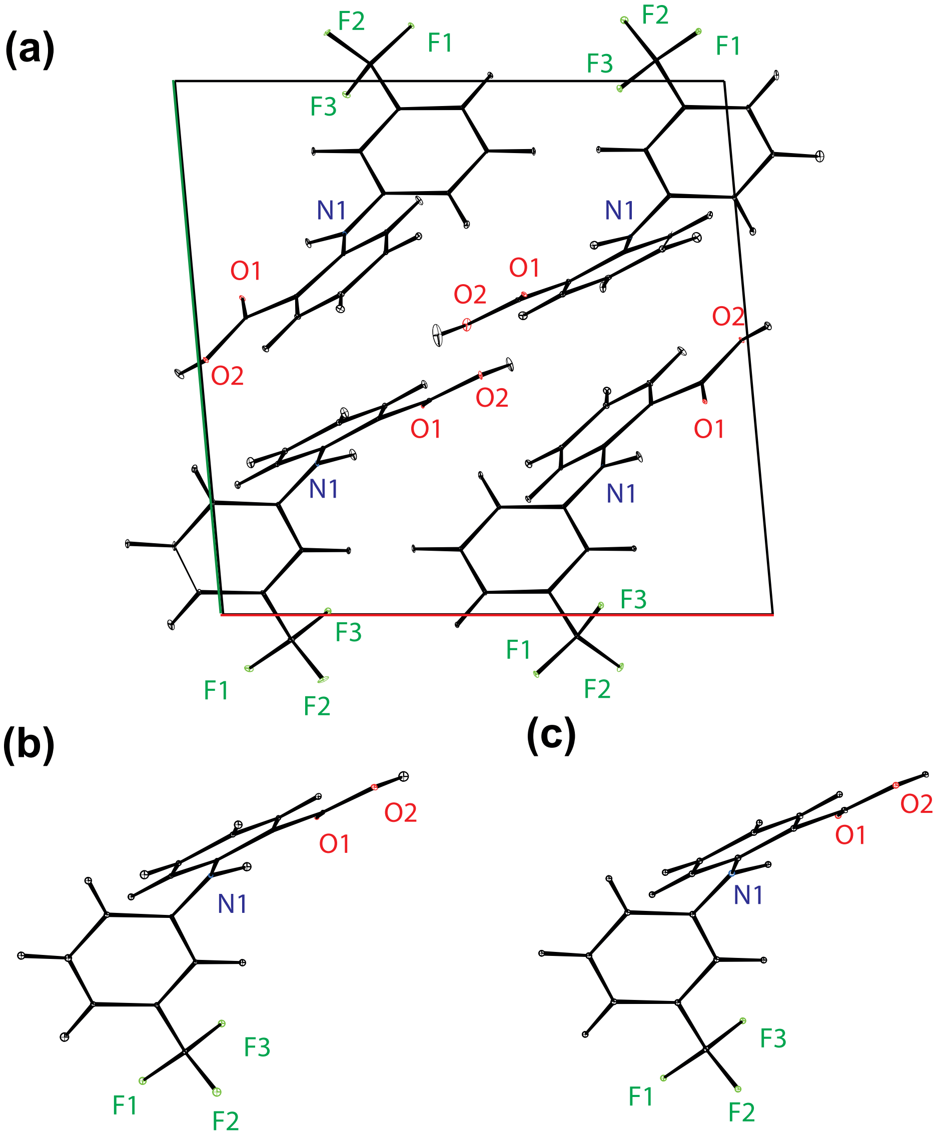


Figure 2-43. ORTEP plot of the unperturbed flufenamic acid structure drawn at the 90 % probability level**.** **(a)** Anisotropic ellipsoids, corresponding to a ^{13}C chemical shift RMSD of 2.3 ppm. **(b)** Equivalent isotropic spheres, corresponding to a ^{13}C chemical shift RMSD of 2.3 ppm. **(c)** Average thermal spheres. The ^{13}C chemical shift RMSD of 2.3 ppm leads to a structural positional RMSD with a 90% confidence interval of 0.072 Å.

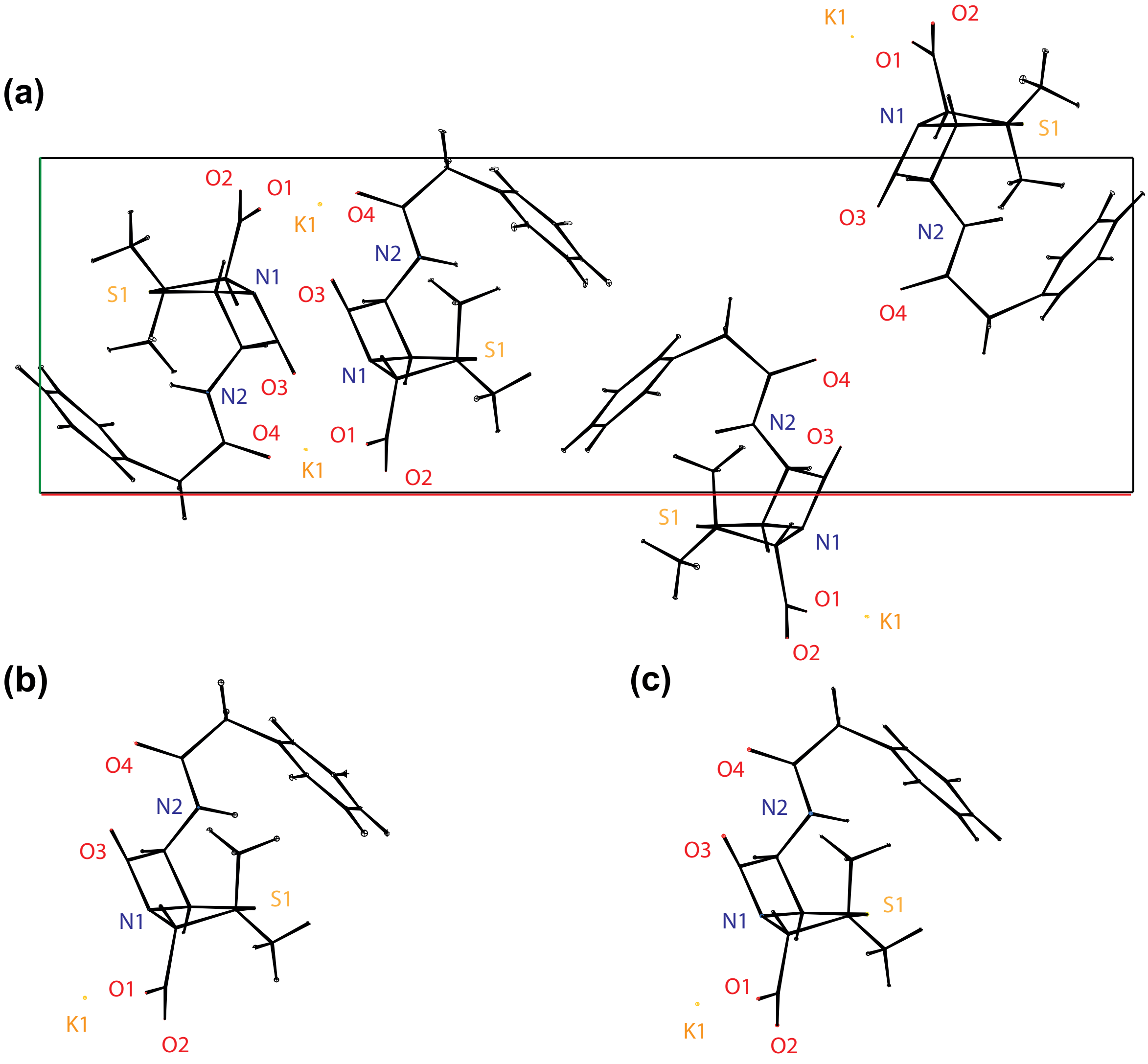


Figure 2-44. ORTEP plot of the unperturbed penicillin structure drawn at the 90 % probability level. **(a)** Anisotropic ellipsoids, corresponding to a ^{13}C chemical shift RMSD of 2.3 ppm. **(b)** Equivalent isotropic spheres, corresponding to a ^{13}C chemical shift RMSD of 2.3 ppm. **(c)** Average thermal spheres. The ^{13}C chemical shift RMSD of 2.3 ppm leads to a structural positional RMSD with a 90% confidence interval of 0.050 Å.

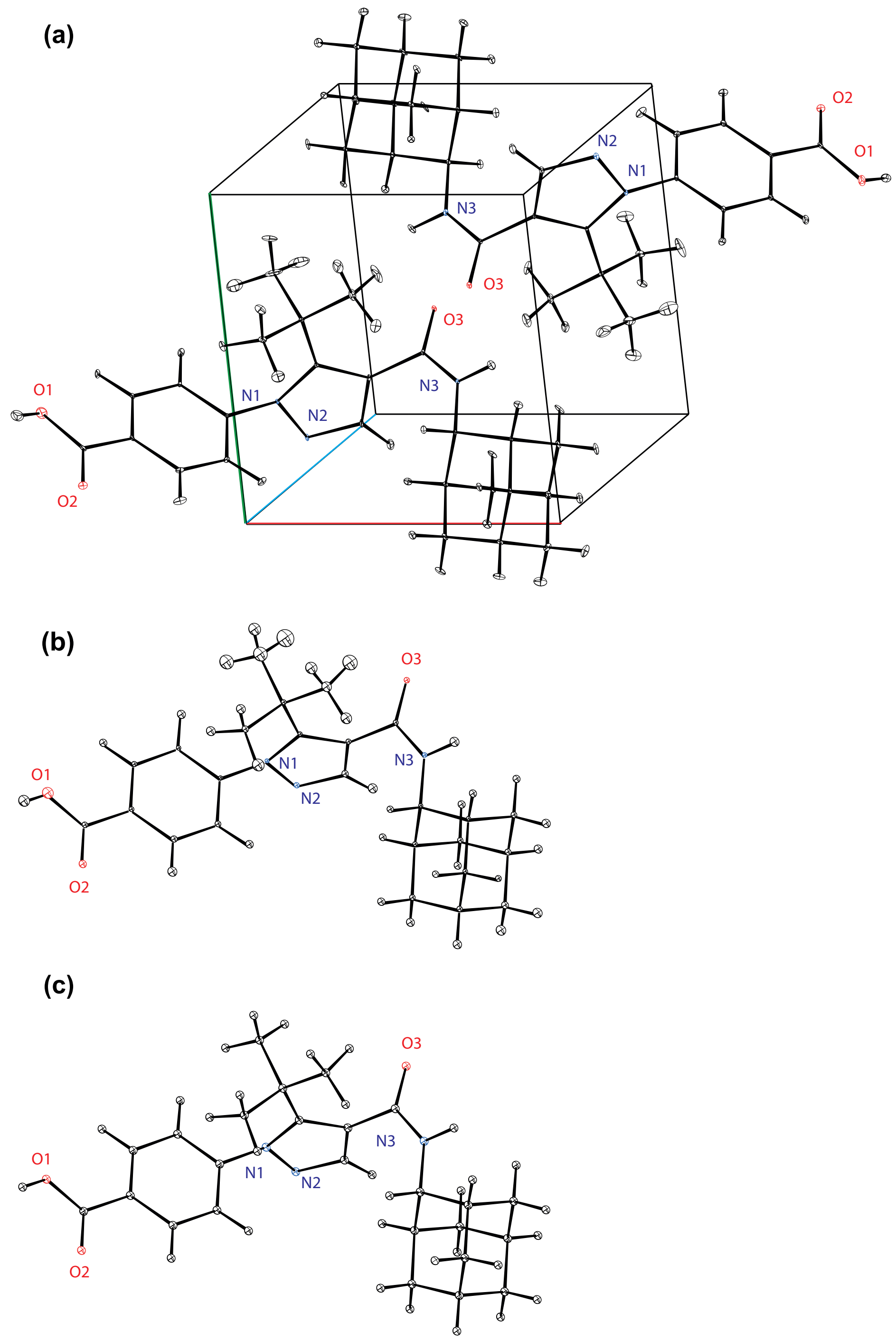


Figure 2-45. ORTEP plot of the unperturbed AZD8329 structure drawn at the 90 % probability level**.** **(a)** Anisotropic ellipsoids, corresponding to a ^{13}C chemical shift RMSD of 2.3 ppm. **(b)** Equivalent isotropic spheres, corresponding to a ^{13}C chemical shift RMSD of 2.3 ppm. **(c)** Average thermal spheres. The ^{13}C chemical shift RMSD of 2.3 ppm leads to a structural positional RMSD with a 90% confidence interval of 0.130 Å.

Temperature Ranges used for Ensemble Generation

The temperatures used for the ensemble generation are detailed in **Table 2-10**. **Note, that the MD simulation is not required to replicate any potential actual dynamical behavior of the molecule in the crystal structure.** Instead the MD simulation is simply used as a method to generate the potentially most physically reasonable ensemble of distorted structures. (As mentioned in the above, there are potentially other approaches to generate such an ensemble including weighting structures having small random displacements with a Boltzmann factor, or exploiting vibrational modes of the molecular crystal.) For these reasons, the ensemble can be generated using one or more MD runs over temperature ranges that are sufficient to generate an ensemble that covers a space of distortions that is larger than that needed to explain the uncertainty in the chemical shifts. The temperatures used in the MD run(s) have no relation to the temperatures used to determine the crystal structures (whether they were determined by NMR or XRD). In this context we consider slightly perturbed structures as crystal structures with small atomic displacements about the local minima, but which do not undergo significant conformational changes or jumps to other local minima. The temperature range used in the MD runs is also chosen so as to avoid such larger structural changes.

We are aware that there might be an explicit temperature dependence of calculated or measured NMR chemical shifts.[223-225] Although it is out of the scope of this work to investigate this effect in detail, we see here that the temperature dependence can be neglected for the uncertainty quantification here (see **Figure 2-46**). The calculated average positional RMSD $< \boldsymbol{r_{av}} >$ only changes by ± 0.015 Å for changes in the temperature ranges of around ± 100°K (see **Table 2-11**).

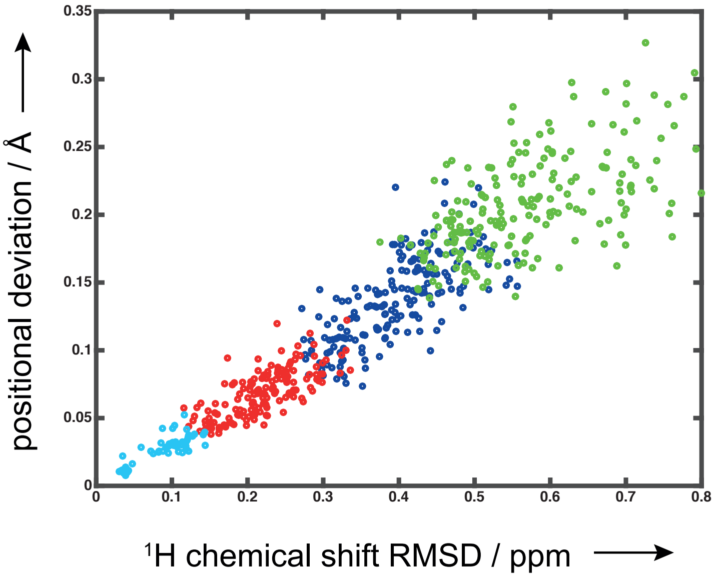


Figure 2-46. Correlation between the overall positional RMSD (Å) and the ^{1}H chemical shift RMSD (ppm) for an ensemble of cocaine crystal structures generated for different temperature ranges (1°-10°K in cyan, 15°-50°K in red, 60°-150°K in blue and 160°-250°K in green). The figure illustrates that different temperature ranges display a nearly identical correlation between the positional deviation and the chemical shift RMSD. The figure also shows that higher temperature ranges can be used to access higher chemical shift RMSD ranges if required.

Table 2-11. Average positional RMSD $< \boldsymbol{r_{av}} >$ calculated for different temperature ranges.

temperature range	**1°-10°K**	**1°-50°K**	**1°-150°K**	**1°-250°K**	**60°-150°K**	**60°-250°K**	**160°-250°K**	**mean value**
average positional RMSD $< \boldsymbol{r_{av}} >$	0.146 Å	0.153 Å	0.163 Å	0.169 Å	0.168 Å	0.174 Å	0.176 Å	0.164 ± 0.015 Å

Comparison to Randomly Generated Ensembles

For cocaine the correlation between the overall positional RMSD and the chemical shift RMSD was also calculated for an ensemble generated by a random displacement method, where all atoms are randomly placed such that a certain total positional RMSD for each structure, with respect to the initial structure, is achieved. Neither the individual mobility of the atoms nor their vibrational properties are considered. **Figure 2-47** shows clearly that for a comparable positional RMSD, an unreasonably high chemical shift RMSD is generated. **Figure 2-47** also indicates, that for the same positional RMSD a huge variance (up to 2 ppm) in the ^{1}H chemical shift is obtained. This is likely due to the generation of physically improbable structures resulting in an unreasonable electronic density. The MD method on the other hand creates an ensemble of more physically reasonable structures for a given average displacement. The random displacement method would thus severely underestimate the positional errors. The MD ensemble allows for a significantly larger uncertainty in position than the random displacement method for a given chemical shift RMSD, and it is thus a better representation of the uncertainty in positions in the experimentally determined structures.

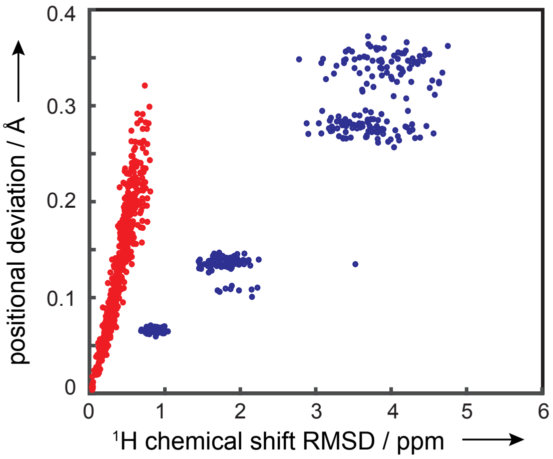


Figure 2-47. Correlation between the overall positional RMSD (Å) and the ^{1}H chemical shift RMSD (ppm) for an ensemble of cocaine crystal structures genera

Comparison to MD with a Variable Unit Cell

The effects of a variable unit cell on the correlation between the positional displacement and the chemical shift RMSD are studied by performing a MD simulation with identical temperature protocol but with a variable unit cell. **Figure 2-48** clearly indicates that for small positional deviations the correlation between the positional displacement and the chemical shift RMSD is not strongly influenced by a variable unit cell.

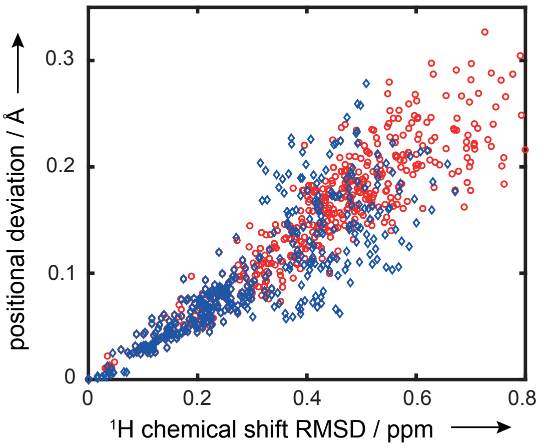


Figure 2-48. Correlation between the average positional RMSD (Å) and the ^{1}H chemical shift RMSD (ppm) for an ensemble of cocaine crystal structures generated by a MD simulation with fixed unit cell (red circles) and a MD simulation, where the unit cell was allowed to vary (blue diamonds).

Force Field used for Ensemble Generation

By employing different force-fields in the MD simulation we generate a set of different ensembles in order to evaluate the effect of the force field used in the MD simulation. The tested force fields are OPLS-aa, Amber-03[226 Nucleic Acids and Organic Molecules] and the Charmm-27[227] force-field, all within the GROMACS suite. The bond distance, bond angle and dihedral angle distributions found for a given positional RMSD (0.05 ± 0.01 Å) are evaluated. The distributions generated by the random displacement method, described above, are also compared to the other distributions. **Figures 2-49** to **2-51** illustrate that the different force-fields generate ensembles with similar molecular properties. In contrast, the distributions generated by the random displacement method display a much larger spread in the molecular property distributions.

We are aware that for large scale dynamics, including conformational changes, distinct differences between different force-fields can be observed,[228] but since we do not need to correctly model the dynamic behavior of the molecular crystal and do not need to explore the whole energy landscape or model conformational changes of the molecule, this does not matter here. We only use the MD to create a physically reasonable ensemble of slightly perturbed crystal structures within a local energy minimum. Our calculations summarized in **Figures 2-49** to **2-51** suggest that all the tested force-fields generate similar ensembles within the desired boundaries. We therefore conclude, that the force-field does not significantly impact the evaluation of the positional uncertainty.

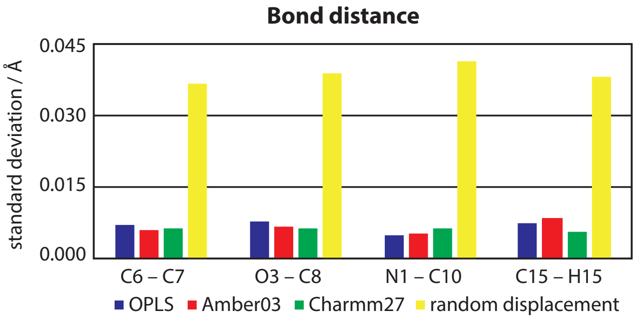


Figure 2-49. Standard deviations of the bond distance distribution at a given positional RMSD of 0.05 ± 0.01 Å for different force-fields (OPLS-aa blue, Amber03 red and Charmm27 green) and the random displacement method (yellow) for four selected bond distances labeled according to the scheme in **Figure 2-52**.

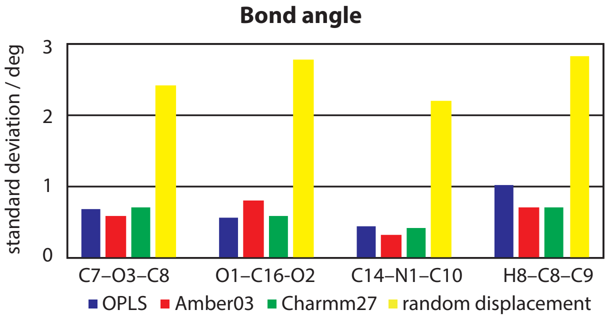


Figure 2-50. Standard deviations of the bond angle distribution at a given positional RMSD of 0.05 ± 0.01 Å for different force-fields (OPLS-aa blue, Amber03 red and Charmm27 green) and the random displacement method (yellow) for four different bond angles labeled according to the scheme in **Figure 2-52**.

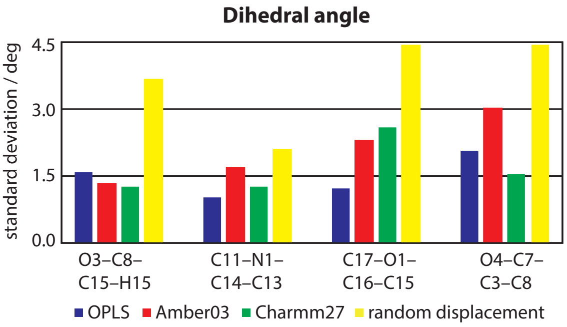


Figure 2-51. Standard deviations of the dihedral angle distribution at a given positional RMSD of 0.05 ± 0.01 Å for different force-fields (OPLS-aa blue, Amber03 red and Charmm27 green) and the random displacement method (yellow) for four different dihedral angles. The dihedrals are labeled according to the scheme in **Figure 2-52**.

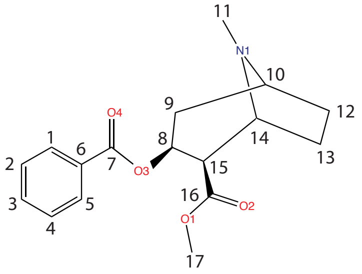


Figure 2-52. Scheme of the cocaine molecule with the atom labels used in the bond, angle and dihedral labeling.

Treatment of Internal Dynamics and Rotating Groups

Here we neglect any potential internal dynamics, except for the case of methyl group rotation. In solid-state NMR chemical shift measurements at room temperatures methyl group rotation usually leads to a single average line shape[229] for the ^{1}H chemical shifts of the methyl protons. On the other hand, methyl group rotation leads to high positional deviations for the individual atoms involved, leading to an overestimation of the correlation between positional deviations and the chemical shift RMSD, if methyl group rotations are present. This overestimation can be corrected in two ways. Either by calculating the positional deviation for crystallographic sites instead of individual atoms. This still leads to a slight overestimation, but it allows a direct positional uncertainty quantification for each proton site in the methyl group. This method was applied in the calculations here. Another method is to consider only a single average proton position for the whole methyl group, thus significantly lowering the positional deviation but disabling a direct quantification of uncertainty for each individual proton site.

2.5 A Bayesian approach to NMRX

This chapter has been adapted with permission from: Engel E.A.; Anelli, A.; Hofstetter, A.; Paruzzo, F.; Emsley, L.; Ceriotti, M., "A Bayesian approach to NMR crystal structure determination", *submitted* **2019**, **(pre-print)**

2.5.1 Introduction

In **Chapter 1** we discussed how the scope of chemical shift driven NMRX has been greatly extended by the development of accurate computational methods to calculate chemical shifts.[48, 60-63, 75-84] However, th error contained within DFT (and ML) chemical shifts leads to uncertainties in the predicted NMR shifts (see **Chapter 1**). [18, 36, 70, 83, 105-112] In **Chapter 2.4** we investigated how the prediction uncertainties in ^{1}H and ^{13}C chemical shifts translate to variances of individual atomic positions in the determined structures. In this chapter we investigate how the chemical shift prediction uncertainty within a set of candidate structures, generated by a structure search (see **Chapter 1**),[12-13, 22, 25-26, 35, 51, 54, 56-58, 71, 85] can be translated into a quantitative probability that one of the candidate structures corresponds to the experimental structure.

In the CSP-NMRX approach presented in **Chapters 2.2** to **2.4**, structures were considered indistinguishable from experiment if the RMSE of their shifts falls within the currently expected chemical shift accuracy.[18, 83, 112] However, this approach fails when multiple candidates exhibit similar RMSEs within the "confidence interval". In this chapter we propose a Bayesian framework to determine the confidence, on a continuous scale from 0 to 100%, in the identification of the experimental crystal structure from a set of candidate structures. As a demonstration of the capabilities of the method, we combine experimental NMR data with GIPAW-DFT and ML predictions of the shifts of a set of CSP candidates to determine the confidence in the structure determination of five different molecular crystals. 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 confirm the low confidence (13%) in the capability to determine the structure of theophylline.[58] We further introduce a method to visualize 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. 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 ^{1}H shifts) leads to more reliable structure determinations.

2.5.2 Theory

In our probabilistic approach to chemical shift driven NMRX each candidate structure constitutes a "model", M, for which we determine the posterior probability, $p(M|\boldsymbol{y}^*)$, of corresponding to the experimental structure, given experimentally determined shifts, $\boldsymbol{y}^*$. 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 $\boldsymbol{y}$ for a given model M as $p(\boldsymbol{y}|M)$ and the probability of observing a shift $\boldsymbol{y}$ before we run the experiment as $p(\boldsymbol{y}) = \sum_M p(\boldsymbol{y}|M)p(M)$. Bayes theorem dictates that

$$p(M|\boldsymbol{y}^*) = \frac{p(\boldsymbol{y}^*|M)p(M)}{p(\boldsymbol{y}^*)} = \frac{p(\boldsymbol{y}^*|M)p(M)}{\sum_{M'} p(\boldsymbol{y}^*|M')p(M')}. \tag{2-34}$$

Clearly, in order to evaluate the posterior $p(M|\boldsymbol{y}^*)$, the conditional probability distribution $p(\boldsymbol{y}|M)$ must be defined. Given GIPAW or ML estimates of the shifts $\boldsymbol{y}^M$ 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.

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

(2-35)

The width σ_j represents an estimate of the typical error in the calculated shift with respect to experiment. We will discuss different approaches to determining σ_j later, and will start by discussing how to translate **Equation 2-35** into a posterior $p(M|\boldsymbol{y}^*)$, which quanties the confidence in designating the model M as the experimental structure.

With full assignments of shifts

In order to evaluate $p(M|\boldsymbol{y}^*)$, one needs to combine information from all experimental shifts $\boldsymbol{y}^* = \{y_j^*\}$, determining the conditional probability $p(\boldsymbol{y}^*|M)$ based on the probabilities for individual shifts in **Equation 2-35**. 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 in Baias *et al.*[56] Assuming independent errors on shifts from distinct nuclei, $p(\boldsymbol{y}^*|M)$ becomes,

$$p(\boldsymbol{y}^*|M) = \prod_j p_j(\boldsymbol{y}_j^*|M).$$

(2-36)

Without assignments of shifts

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 $\boldsymbol{a}$, the conditional probability may be written as a sum over assignments,

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

(2-37)

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

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

(2-38)

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

Partial assignments of shifts

Cases in which none of the experimental shifts can be assigned are rare. In most cases the sum in **Equation 2-37** only needs to be evaluated over a subset of all the possible permutations of indices $\boldsymbol{a}$. In practice this means that $\boldsymbol{P}$ can be made block-diagonal, each block $\boldsymbol{P}_k$ 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,

$$p(\boldsymbol{y}^*|M) = \prod_k \mathrm{perm}\,\boldsymbol{P}_k/n_k!$$

(2-39)

Where n_k indicates the size of the *k*-th block. While evaluating the permanent has a cost that grows combinatorically with the size of $\boldsymbol{y}^*$, algorithms with a low pre-factor make its evaluation an ordable 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 2-37** and **2-39** to be dominated by the contribution from the assignment producing the best-match between $\boldsymbol{y}^M$ and $\boldsymbol{y}^*$.

Examples and a discussion of chemical shift driven NMRX with partial assignments or without assignments of shifts are given in the original publication: Engel E.A.; Anelli, A.; Hofstetter, A.; Paruzzo, F.; Emsley, L.; Ceriotti, M., "A Bayesian approach to NMR crystal structure determination", *submitted* **2019**, **(pre-print).**

Estimate of the reference errors

Clearly, the evaluation of $p(M|\boldsymbol{y}^*)$ 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 (see **Appendix III**), but estimating the error of the underlying GIPAW shifts with respect to experiment σ_j^{DFT}, usually requires extensive benchmarks. Existing datasets[18, 83, 112] 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 $p(\boldsymbol{y}^*)$ 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 for all environments in the crystal, or use the benchmark values to define a prior distribution for σ_j. In the following we discuss results obtained using a single, global value of 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 driven NMRX relies strongly on CSP to generate candidate structures. Although CSP is constantly improving in thoroughness and energetic accuracy,[113] 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 $\widetilde{M}$ 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 $\widetilde{M}$ 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 principal component analysis (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 $k(M; M')$, which can be found in **Appendix III** 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 $p(\boldsymbol{y}|M)$ 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 $p(\boldsymbol{y}|M)$ is defined by **Equations 2-35** and **2-36**, 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 uniquely be 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.[176, 184] Loosely speaking, atomic configurations are represented in terms of an atom-density, which distinguishes the different involved chemical species.[230] It is constructed as the sum of Gaussian distributions centered on the atomic positions and symmetrized 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. A more detailed description can be found in **Chapter 2.4.2** and **Appendix III**. 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 (**Chapter 2.3**), 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 but one particular choice. In general, any metric of structural (dis-)similarity for example the single molecule RMSE[231] – can be used as a basis for a KPCA projection of structural similarity.

2.5.3 Computational Methods

In **Chapter 2.5.4** we discuss chemical shifts predicted using a ML model, which extends the GPR model built around the SOAP framework[176, 184] presented in **Chapter 2.3** by (i) training set sparsification via a projected process (PP) strategy,[183, 199-200] (ii) the efficient estimation of the uncertainty in predictions using a resampling approach,[201] and (iii) the radial scaling approach introduced in Ref.[230] and **Chapter 2.6**, 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 construction of the ML model is described in detail in the **Appendix III**. The new model extends the original model presented in **Chapter 2.3** by incorporating sulfur-containing compounds thereby increasing the training set from 2000 to 2500 structures, and (slightly) outperforming it (see **Chapter 2.6**). 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 $0.33 \pm 0.16\, ppm.$[18, 83, 112]

It is worth noting that Liu *et al.*[103] have recently demonstrated that, despite replacing the SOAP description of atomic densities with a non-symmetry-adapted real-space discretized 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 generalized to the prediction of tensorial properties such as (anisotropic) chemical shielding tensors.[185] Furthermore, it provides accurate estimates of prediction uncertainty.[201] 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 (see **Appendix III**).

2.5.4 Results and discussion

In order to demonstrate the Bayesian approach to NMRX, we use it to quantify the confidence in the structure determination of five molecular crystals (see **Figure 2-53**). We also demonstrate the use of two-dimensional visualizations of the similarity between candidate structures, both in terms of their structural features and in terms of their predicted chemical shifts, following the recipe of **Chapter 2.5.2**.

Benchmark systems

Cocaine, 4-[4-(2-adamantylcarbamoyl)-5-tert-butyl-pyrazol-1-Yl] benzoic acid (referred to as AZD8329), theophylline, flufenamic acid, and flutamide (see **Figure 2-53**) have all previously been studied using NMRX.[56, 58, 148] 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 five 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 III.**

Figure 2-54 shows examples of the analysis that is traditionally performed in chemical shift driven NMRX. The RMSE between the experimental shifts and those predicted for multiple CSP candidates is computed using fully assigned ^{1}H 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 of cocaine and AZD8329, only one structure is consistent with experiment, making the structure determination conclusive. In the case of flufenamic acid, 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 favor the correct candidate and facilitate determining the correct crystal structure.

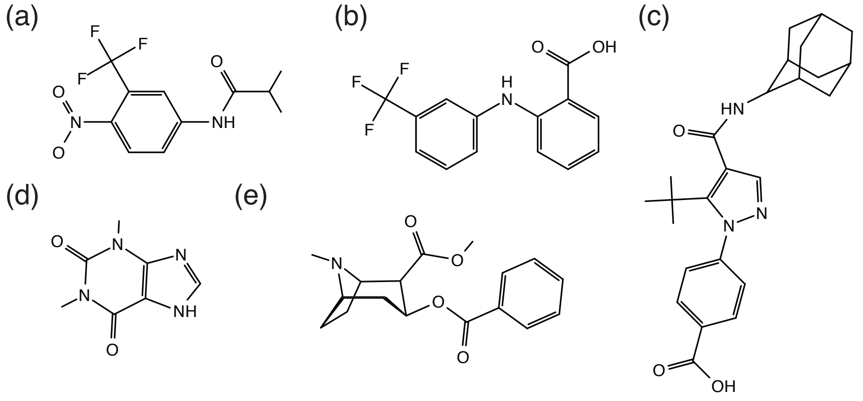


Figure 2-53. Chemical structures of flutamide **(a)**, flufenamic acid **(b)**, AZD8329 **(c)**, theophylline **(d)** and cocaine **(e).**

Quantitative structure determination and visualization

Cases such as flufenamic acid, in which chemical shift driven 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. From **Figure 2-55** we see that on the basis of the same ^{1}H shifts from GIPAW calculations, we estimate that the correct structure is identified with high confidence in 4 out of the 5 benchmark cases (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 (see Baias *et al.*[56]).

In order to elucidate why the level of confidence in the structural determination varies among the benchmark problems, we generate a two dimensional visualization 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. **Figure 2-56** shows the representations of the similarity of the CSP candidates for each of the five compounds. For each compound we show the similarity in terms of ^{1}H 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 2-56e**) shows that structures 8 and 16, which are identified as the most likely candidates, exhibit very distinct ^{1}H chemical shifts from structure 13, which is the correct candidate. Hence, even though there are only four ^{1}H shifts, this analysis suggests that more accurate predictions of the ^{1}H shifts would probably suffice to correctly determine the structure. In contrast, in the case of flufenamic acid (**Figure 2-56c**) the three structures with non-zero probability are all similarly close to experiment as they are to each other. (Actually, **Figure 2-56c** seems to indicate that structure 41 is closer to experiment than structure 14, whose chemical shifts agree most closely with the experimentally measured ones and which happens to be the correct candidate. This distortion is an artifact of the projection of the NMR (and geometric) similarities, which correspond to a distance in a high-dimensional space, onto a two-dimensional representation.) 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.

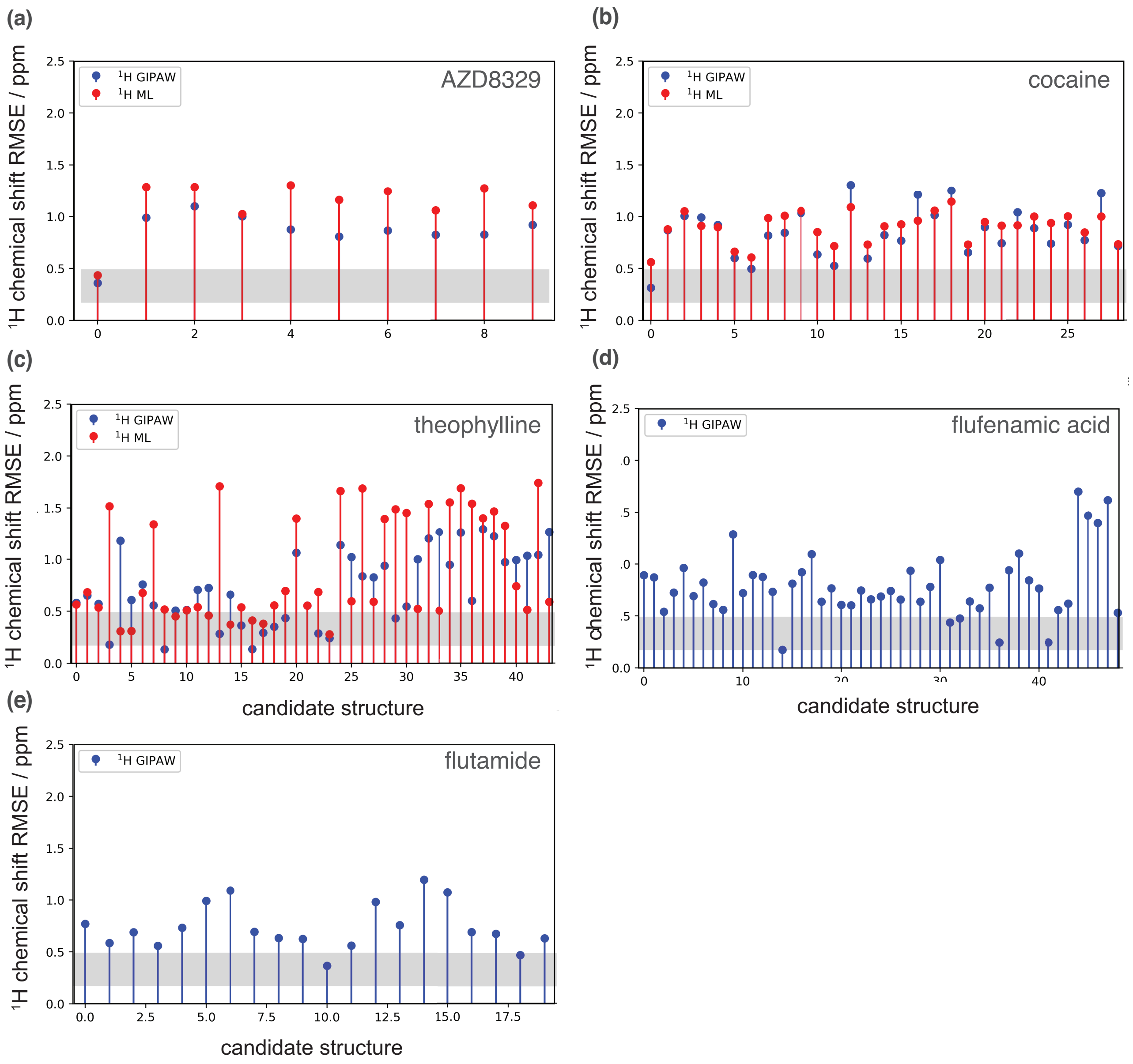


Figure 2-54. RMSEs of the GIPAW (blue) and ML (red) ^{1}H chemical shifts of the most stable cocaine **(a)**, AZD8329 **(b)** and flufenamic acid **(c)** CSP candidates with respect to experiment. The gray area indicates the one sigma confidence interval for the GIPAW-DFT ^{1}H chemical shifts as determined 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.

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 2-56e**. Meanwhile, the geometry-based representation, which is also shown in **Figure 2-56e** 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.

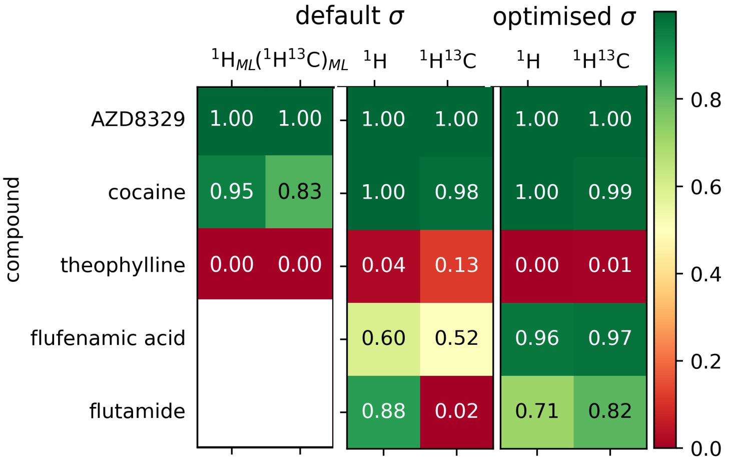


Figure 2-55. Overview of the results of NMR crystal structure determinations for the five benchmark compounds using ^{1}H and ^{13}C shifts calculated with ML or GIPAW, respectively. Each cell is colored and labeled according to the Bayesian probability of matching experiment assigned to the representative 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 $\sigma_H^{DFT} = 0.33 \pm 0.16\, ppm$, and $\sigma_C^{DFT} = 1.9 \pm 0.4\, ppm$. The **right panel** shows the Bayesian probabilities based on uncertainties estimated for each individual compound under consideration by maximizing $p(\boldsymbol{y}^*)$ with respect to $\{\sigma_j\}$ as described in **Chapter 2.5.2.** The estimated uncertainties are given as, $\sigma_H^{DFT} = 0.28 \pm 0.09\, ppm$, and $\sigma_C^{DFT} = 2.7 \pm 0.9\, ppm$.

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 permits 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 2-55** and demonstrate that ML-based NMRX almost matches the resolving power achieved with GIPAW predictions of NMR shifts.

^{13}C NMRX

Irrespective of whether NMR chemical shifts are predicted using GIPAW-DFT calculations or ML methods, ^{1}H shifts do not always suffice to pin down the experimental structure. The cases of flufenamic acid and theophylline highlight the limits of ^{1}H 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,[56] **Figure 2-55** 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[18, 83, 112] substantially underestimates the actual error for the compounds considered here.

Following the strategy of maximizing $p(\boldsymbol{y}^*)$ with respect to $\{\sigma_j\}$ proposed in **Chapter 2.5.2**, 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 $2.6 \pm 1.4\, ppm$. For comparison, the corresponding RMSD of the ^{1}H GIPAW shifts is $0.28 \pm 0.09\, ppm$ and thus entirely consistent with the global estimate of $\sigma_H^{DFT} = 0.33 \pm 0.26\, ppm$.

Figure 2-55 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 ^{1}H (and indeed ^{13}C) shifts, but rather by the accuracy of the estimates of the typical errors in those shift. ***Accordingly, its structure can be determined with almost complete confidence (96%) provided accurate estimates of the typical errors in ^{1}H (and ^{13}C) shifts*** *(see **Figure 2-55**).*

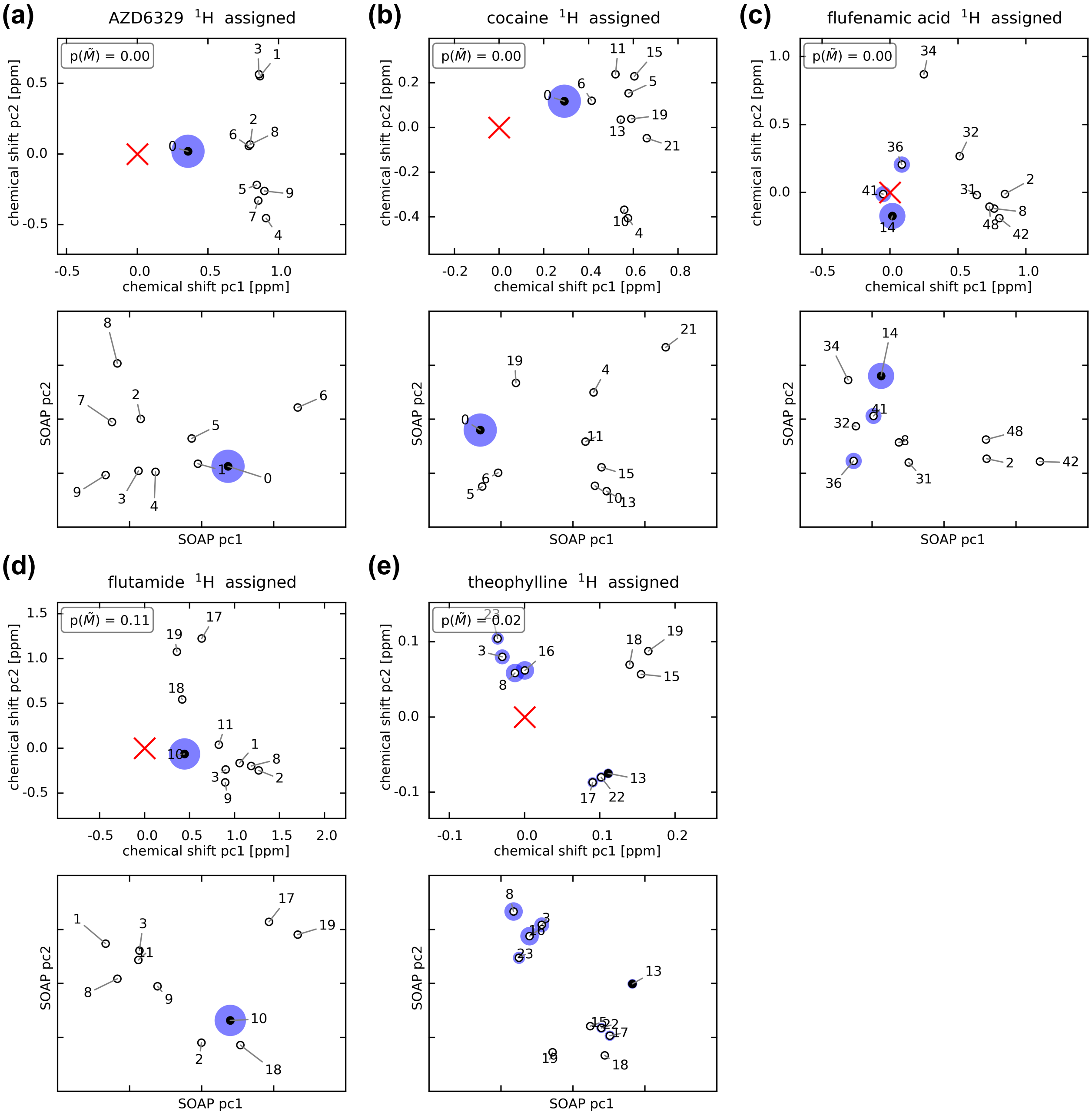


Figure 2-56. Evaluation of the top 10 AZD8329 **(a)**, cocaine **(b)**, flufenamic acid **(c)**, flutamide **(d)** and theophylline **(e)** CSP candidates. The correct candidates are shown as filled circles and the others as empty circles. For each candidate the probability of matching experiment $p(M|\boldsymbol{y}^{*})$ 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 favorable 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 ^{1}H GIPAW-DFT shifts. $p(\widetilde{M})$ 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 **Chapter 2.5.2** has no intuitive physical meaning and is therefore not shown.

2.5.5 Conclusion

We have introduced an analysis framework for chemical shift driven NMRX, which is suited to a variety of experimental (and computational) setups. By quantifying the confidence in identifications of experimental structures our analysis framework demonstrates 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 $2.7 \pm 0.9\ ppm$ as opposed to previous, estimates of $1.9 \pm 0.3\ ppm$. If we use our corrected error estimates, incorporating ^{13}C shifts into the analysis improves the reliability of structure determination. In one of the cases we considered, 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 visualizations 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

(i) identifying among a pool of candidate structures which most closely approximates the experimental one,
(ii) performing sanity checks of the comprehensiveness of the pool, the associated predicted NMR shifts, and the initial identification,
(iii) quantifying the confidence in the identification assuming the sanity checks have provided satisfactory results,
(iv) 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.

2.5.6 Appendix III

Applications

Crystal structure prediction. Detailed descriptions of the generation and refinement of the candidate crystal structures for all compounds discussed in this work can be found in the original publications. [56, 58, 148] In summary, the theophylline, flutamide, flufenamic acid, cocaine, and AZD8329 candidates were generated starting from their chemical formulae using CrystalPredictor[232] to perform a quasi-random sampling of unit cells and molecular predictions within the most commonly observed Söhnke space groups, all with one molecule (geometry optimized using DFT with the hybrid B3LYP functional [233-234] in the asymmetric unit cell. For cocaine this was prefaced by an automated conformer search using the low-mode search method[235] leading to 16 starting conformations, while for the other compounds a search of their torsional energy profiles[203] provided eight (flutamide) and six (flufenamic acid and AZD8329) starting conformations, respectively.

Subsequently, the **theophylline** candidates were geometry optimized at fixed molecular geometry using the DMACRYS code[236] with the FIT potential of Coombes *et al.*[237] and electrostatics based on atomic multipoles from a distributed multipole analysis[238] of the electron density at the B3LYP/6-31G(d,p) DFT level of theory. For **flufenamic acid** and **flutamide** the candidates were geometry optimized using a molecular mechanics description of inter- and intra-molecular interactions using an atom-atom model with exp-6 + atomic multipoles electrostatics and B3LYP/6-31G(d,p) DFT, respectively. The influence of polarization effects was approximated by performing the molecular calculations in a continuum dielectric ($\varepsilon = 3$). For **cocaine** the lowest energy structures were geometry optimized using CrystalOptimizer[239] using the same description of the intra- and inter-molecular interactions as for flufenamic acid and flutamide. 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 CIF format) in the supplementary information of Ref.[56]

The **AZD8329** structures were geometry optimized using the molecular mechanics description outlined in Ref.,[203] using the Open Force Field module of the Cerius2 v4.6 package, and refined using DMACRYS[236] with DFT calculations in the Gaussian03 software[240] for the intra-molecular contribution and an atom-atom model of inter-molecular interactions with atomic multipole electrostatics. 11 AZD8329 candidates within 30 *kJ/mol* of the most stable predicted crystal structure for a given conformation were further geometry optimized using CASTEP[191] at the PBE-DFT level of theory and can be found in the supplementary information of Ref.[58]

DFT chemical shift calculations. The GIPAW DFT calculations for the different compounds were performed as follows:

- **Flutamide and theophylline :** the NMR calculations were performed using CASTEP v5.0 with the PBE exchange-correlation functional[205] without dispersion correction, an equivalent plane-wave energy cut-off of 550 *eV* and a Monkhorst-Pack k-point grid[207] with a maximum spacing of $2\pi \times 0.05\ \text{Å}^{-1}$. The calculations used on-the-fly generated GIPAW pseudopotentials.[62]

- **Flufenamic acid :** the NMR calculations were performed using CASTEP v5.5 with the PBE exchange-correlation functional[205] with a Tkatchenko-Scheffler semi-empirical dispersion correction,[241] an equivalent plane-wave energy cut-off of 700 *eV* and a Monkhorst-Pack k-point grid with a maximum spacing of $2\pi \times 0.05\ \text{Å}^{-1}$. The calculations used on-the-fly generated GIPAW pseudopotentials.

- **AZD8320 and cocaine :** the NMR calculations were performed using Quantum Espresso v6.3. with the PBE exchange-correlation functional[205] with a Grimme D2 semi-empirical dispersion correction[206] 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 PS library database.[242]

Experimental chemical shifts. For flufenamic acid, flutamide and cocaine the fully assigned experimental ^{1}H and ^{13}C chemical shifts with the corresponding labels are given in **Appendix I.** For AZD8329 and theophylline the fully assigned experimental shifts were taken from Refs.[58, 148] and are given in **Tables 2-12** and **2-13**. The corresponding labels are given in **Figure 2-57**.

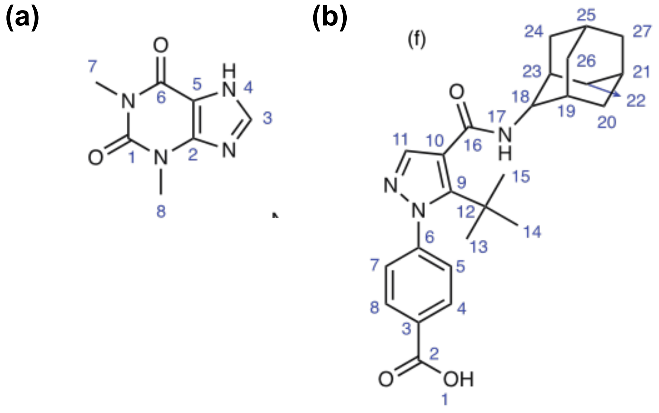


Figure 2-57. Chemical structures of theophylline **(a)** and AZD8329 **(b)**. The distinct ^{1}H and ^{13}C sites are labeled.

Table 2-12. AZD8329 experimental chemical shifts.

Label	^{1}H, 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 2-13. Theophylline 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

Machine-learning with uncertainty estimation

Above, we discuss chemical shifts predicted using a ML model which extends that of **Chapter 2.3** by training set sparsification and the efficient estimation of the uncertainty in predictions. It is built on the same framework that combines physically-motivated structural representations with a GPR framework. Properties *y* are predicted from inputs *X* via an interpolating function *f(X)* assuming normally distributed noise $\varepsilon \sim \mathcal{N}(0,\sigma)$:

$$y(X) = f(X) + \varepsilon.$$

(2-40)

Given a training set of *N* input-property pairs $(\boldsymbol{X},\boldsymbol{Y}) = \{(X_i,Y_i)\}$ one can model *f* as a Gaussian process *GP(0,K)*, where *K* is the covariance function between the inputs. The prediction for an input *X* can then be written as a linear combination :[183]

$$y(X) = \sum_{i=1}^{N} w_i k(X_i,X) = K_{XN} K_{NN}^{-1} \boldsymbol{y}.$$

(2-41)

where $k(X_i,X) = (K_{XN})_i$ and $w_i = \sum_j (K_{NN}^{-1})_{ij} y_j$. While predictions can in principle be converged to any desired level of accuracy by including more training data, this rapidly produces kernel matrices K_{NN} of considerable dimensions, slowing down training and predictions. We thus follow a projected process (PP) strategy,[183, 199-200] in which the full $(N \times N)$ kernel matrix K_{NN} is approximated by a lower rank $(M \times M)$ K_{MM} corresponding to an "active set" composed of the *M* training data which retain the most relevant information. The correlations between all the training points and the active set are encoded in an $(M \times N)$ kernel matrix K_{MN}, and predictions for new points *X* are calculated as,

$$y(X) = K_{XM}(K_{MM} + \varsigma^{-2} K_{MN} K_{MN}^{T})^{-1} K_{MN} \boldsymbol{y}.$$

(2-42)

Here ς is a regularisation parameter. During training, the size of the matrix to be inverted is thereby reduced to \$$(M \times M)$, at the cost of computing, once, the Gram matrix of the active-passive kernel. Conversely, when predicting, only similarities between the new structures and the active set have to be considered.

In principle the uncertainty associated with a PP prediction can be calculated directly as,

$$\sigma(X)^2 = \varsigma^2 + K_{XX} - K_{XM}K_{MM}^{-1}K_{XM} + K_{XM}\left(K_{MM} + \varsigma^{-2}K_{NM}^{T}K_{NM}\right)^{-1}K_{XM}^{T}. \tag{2-43}$$

This estimate, however, is considerably more demanding than that of y. We therefore instead employ the scheme for accurate and efficient uncertainty estimation proposed in Ref.[201] which is based on a committee of models. An ensemble of N_m models is trained on subsamples of the full training set of size $N_s < N$. Crucially, the different structural variance covered by the subsamples affects the spread of predictions $\{y^{(m)}(X)\}$ obtained from the different models m. This is corrected for by rescaling,

$$y^{(m)}(X) \to \bar{y}^{(m)}(X) + \alpha\left(y^{(m)}(X) - \bar{y}^{(m)}(X)\right)^{\gamma/2+1}, \tag{2-44}$$

where $\bar{y}^{(m)}(X) \equiv 1/N_m \sum_m y^{(m)}(X)$, using the constants α and γ which maximise the log-likelihood of the rescaled ensemble predictions for a validation set of choice,

$$P\left(\boldsymbol{y}\middle|\{X_n\}_{n=0,1,..}\right) = \prod_{n=0}^{N_v} \frac{1}{\sqrt{2\pi\sigma^2(X_n)}} \exp\frac{\left(y_n - y(X_n)\right)^2}{2\sigma^2(X_n)}, \tag{2-45}$$

where $\sigma^2(X) \equiv 1/N_m \mathrm{Var}(\{y^{(m)}(X)\})$ and N_v is the size of the validation set. In practice we apply a linear rescaling ($\gamma = 0$), for which the log-likelihood is maximised by,

$$\alpha^2 = \frac{1}{N_v}\sum_n \frac{\left(y_n - \bar{y}(X_n)\right)^2}{\sigma^2(X_n)}. \tag{2-46}$$

Uncertainties in predictions can then simply be estimated as the standard deviation over the ensemble of models,

$$\sigma^{ML}(X) \approx \sqrt{\frac{\sum_m \left(y^{(m)}(X) - \bar{y}(X_n)\right)^2}{N_m - 1}}. \tag{2-47}$$

It is worth noting that the resultant uncertainties are environment- and model-dependent. Further they are statistical uncertainties which are uncorrelated with the inherent errors of the underlying reference (GIPAW-DFT) data relative to experiment. In consequence they must be added to the GIPAW-DFT error(s) in quadrature.

In practice our GPR model is built around SOPA kernels,[176, 184] in which atomic environments are represented as species-dependent atomic densities constructed by associating a Gaussian density with each atomic position within a cut-off radius of the central atom. Using the radially-scaled variant of the SOAP framework[198] drastically improves the computational performance compared to the multi-scale approach described in **Chapter 2.3**, which effectively requires the construction and evaluation of multiple GPR models per chemical species. The associated hyperparameters were determined using a cross-validation scheme and are detailed in **Table 2-14**. The SOAP-GPR framework has proven successful in the context of regressions for different systems[155, 243-244]and (scalar as well as tensorial) properties.[185] Most importantly, SOAP-GPR has previously proven suitable for GIPAW-DFT accurate predictions of NMR chemical shifts (see **Chapter 2.3**).

Table 2-14. SOAP hyperparameters and sparsification parameters for all species.

	H	**C**	**N**	**O**
cut-off radius r_c [Å]	4.5	4.0	4.5	4.5
Gaussian width σ [Å]	0.3	0.3	0.3	0.3
radial basis set size *n*	12	12	12	12
angular basis set size *l*	9	9	9	9
kernel exponent ζ	3	3	3	3
scaling onset r_s [Å]	2.0	2.0	2.0	2.0
scaling exponent e_s	3	3	3	3
training set size *N*	50k	50k	40k	40k
active set size *M*	20k	20k	20k	20k
number of FPS features	8000	8000	8000	8000
regularization ς	1800	3200	5300	3000
test set RMSE [ppm]	0.48	4.13	13.70	17.05

A critical element of the ML model are the underlying training and test sets, which are detailed in **Chapter 2.6**. Shifts are calculated for atomic centers, i.e. for local “environments”, rather than structures. Crystal structures often contain redundant environments, for example due to crystal symmetries. Hence, the training set was reduced in size by FPS ordering the individual environments and retaining only the 100,000 (^{1}H and ^{13}C) and 40,000 (^{15}N and ^{17}O) most structurally diverse and therefore informative ones. At this point environments exhibiting GIPAW-DFT shifts far outside the physical ranges of around 5 to -50 *ppm* for ^{1}H (64 unphysical environments), around -100 to -200 *ppm* for ^{13}C (149 unphysical environments), around -700 to -400 *ppm* for ^{15}N (12 unphysical environments), and around -1250 to -350 *ppm* for ^{17}O (13 unphysical environments) were eliminated. Their presence highlights that GIPAW-DFT shifts are not always reliable. Initial ML models were therefore trained in a cross-validation scheme to assess (i) the residual error with respect to the GIPAW-DFT reference and (ii) the estimated ML uncertainty for all training environments. These were then used to identify anomalous environments with residual errors outside the 3σ confidence interval associated with the estimated ML uncertainty, suggesting a possible failure of the GIPAW-DFT shift calculation. For each anomalous environment the entire associated structure was purged from the training set. We found this procedure to improve the accuracy of the model when applied to the validation set, which suggests that indeed “outliers” in the train set affect adversely the accuracy of the model. All in all, 373 ^{1}H, 347 ^{13}C, 44 ^{15}N, and 113 ^{17}O environments were eliminated.

Active sets were then extracted on the basis of the FPS order, so as to incorporate the largest amount of information for a given size.[186-187, 202, 245] The ”learning curves” with respect to the size of the active set in **Figure 2-58** suggests that for all species active sets of $M = 20{,}000$ environments suffice to match the accuracy of the non-sparsified models to within less than 1% of the RMSE of the full model. It is worth noting that within the PP framework the underlying training set can be arbitrarily large since $K_{MN}K_{MN}^T$ in **Equation 2-42** can be calculated in chunks, so that the only limiting factor in constructing and applying the PP model is the size of the active set. In practice underlying training sets of $N = 50{,}000$ for ^{1}H and ^{13}C and $N = 40{,}000$ for ^{15}N and ^{17}O were found to be sufficient to saturate the accuracy of the respective models.

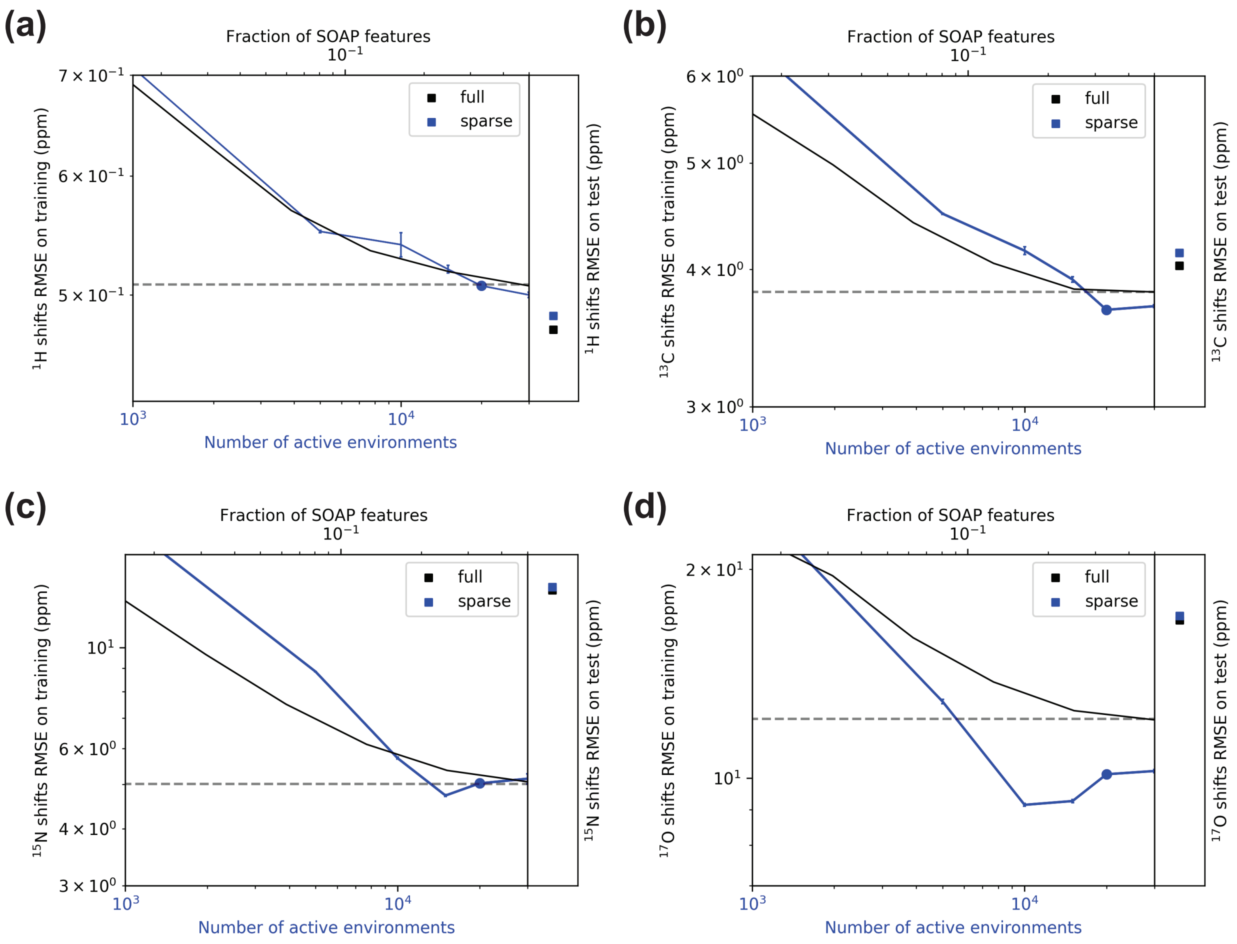


Figure 2-58. Convergence of training set RMSE from a CV scheme with the number of environments in the active set (blue) and the fraction of retained SOAP features (black) for ^{1}H **(a)**, ^{13}C **(b)**, ^{15}N **(c)** and ^{17}O **(d)**. Results from models using the full training set explicitly are shown using square symbols.

To further accelerate the ML predictions, we also sparsified the SOAP fingerprints, using an FPS strategy, [202] performing a separate selection for each element. Cross-validation (CV) demonstrates that the first 20,000 training environments for any given chemical species suffice to guide the FPS of the SOAP features. The FPS-based choice of SOAP features is guided by structural variance and consequently leads to sparsified fingerprints which should be suitable for regressions of general observables. The RMSE of models built with an increasing number of SOAP features (see **Figure 2-58**) shows that sparsifying from an initial 18,301 components to 8000 leads to a negligible decrease in model accuracy for all species (less than 1% increase in the RMSE).

The full sets of hyperparameters defining the specific ML models constructed in this work are collected in **Table 2-14**.The final accuracy of this sparse model is (slightly) better than that of the original ShiftML model presented in **Chapter 2.3**. The expected errors of 0.48 *ppm* for out-of-sample predictions of ^{1}H shifts are comparable to the inherent error of the underlying GIPAW-DFT predictions with respect to experiment of around $0.33 \pm 0.26\ ppm.$ [18, 83, 112] Further reductions in ML errors would reap insignificant improvements to the resolving power of ML-based NMRX without accompanying reductions in the underlying GIPAW-DFT errors with respect to experiment. For ^{13}C the expected ML errors of 4.13 ppm are about twice as large as the typical error in GIPAW-DFT predictions of $1.9 \pm 0.4\ ppm.$ [18, 83, 112] Even though, as demonstrated in **Chapter 2.5.4**, GIPAW-DFT ^{13}C errors are often much larger than this value, an improvement in the accuracy of ShiftML for carbon, oxygen and nitrogen would be desirable, and will be the subject of future improvements of ShiftML.

Finally, **Figure 2-59** demonstrates the agreement between the distributions of ML errors with respect to GIPAW-DFT, $|\bar{y}(X_i) - y_i|$, and that predicted in terms of the distribution around the mean of the ensemble of subsampling models, $\left|\sum_m y^{(m)}(X_i) - \bar{y}(X_i)\right|$. The qualitative agreement between the distributions confirms that the standard deviation over the ensemble of models provides a good estimate of the uncertainty in the ML predictions.

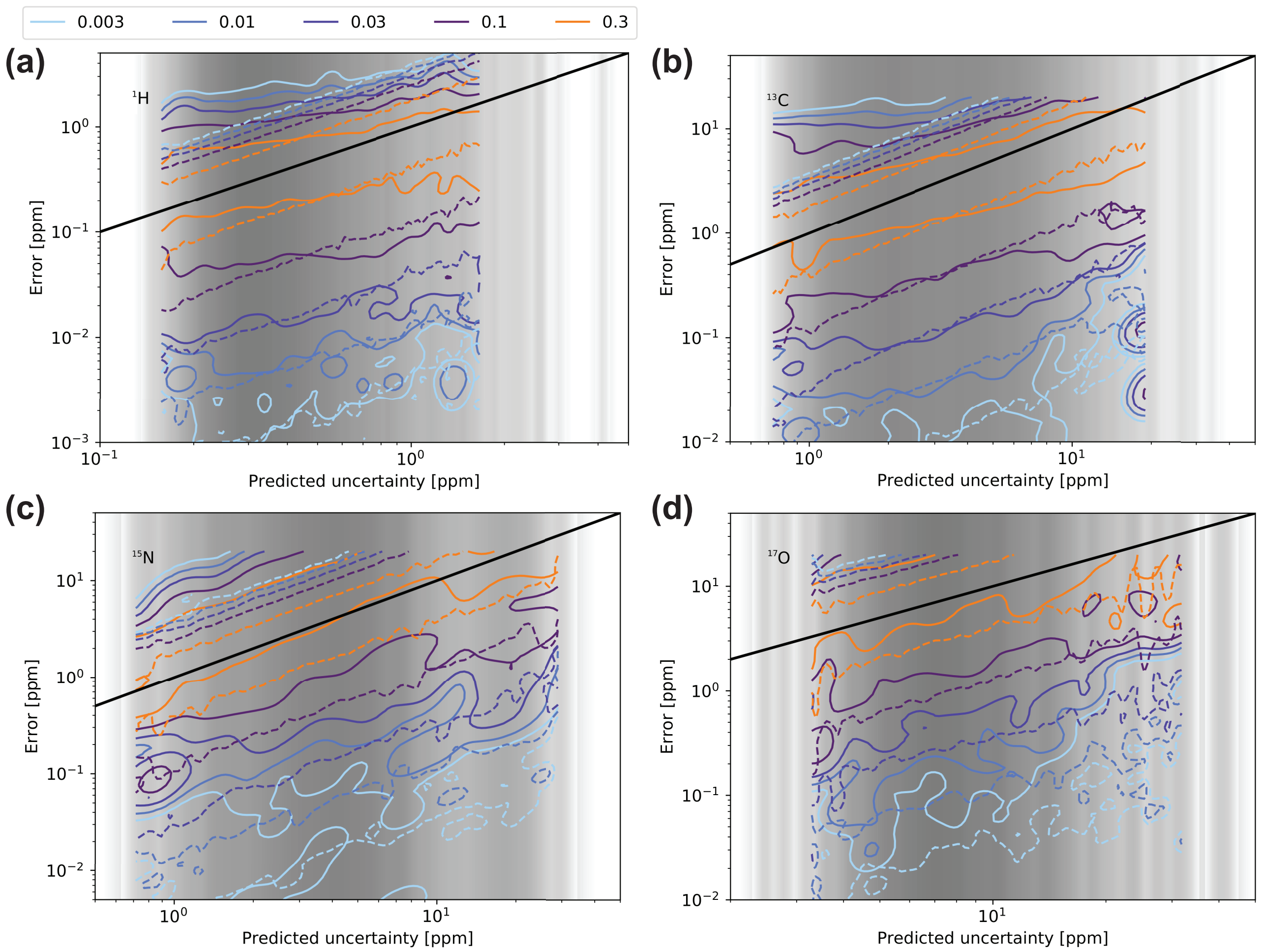


Figure 2-59. Distribution of ^{1}H **(a)**, ^{13}C **(b)**, ^{15}N **(c)** and ^{17}O **(d)** chemical shielding predictions. The colored solid lines show contours of the distribution of actual errors relative to the reference, $P(\ln|\bar{y}(X_i) - y_i|\,\ln\sigma^{ML}(X))$, while the colored dashed lines show contours of distribution of the predictions of the subsampling models around their mean, $P\left(\ln\left|y^{(m)}(X_i) - \bar{y}(X_i)\right|\,\ln\sigma^{ML}(X)\right)$. The gray scale density plot corresponds to the marginal distribution of the predicted uncertainty $P(\ln\sigma^{ML}(X))$. The solid black line shows $y = x$ to guide the eye.

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 normalization, $p(M, M')$ can be calculated as,

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

(2-48)

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

$$\lim_{\sigma^{M'} \to \varepsilon} p(M, M') \propto \varepsilon p(\boldsymbol{y}^{\boldsymbol{M}'}|M),$$

(2-49)

which is then symmetrized and normalized, giving

$$p(M, M') = \frac{p\left(\boldsymbol{y}^{\boldsymbol{M}'}|M\right) + p(\boldsymbol{y}^{\boldsymbol{M}}|M')}{2\sqrt{p(\boldsymbol{y}^{\boldsymbol{M}}|M) p(\boldsymbol{y}^{\boldsymbol{M}'}|M')}}.$$

(2-50)

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,*

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

(2-51)

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

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

2.6 Structure determination of Ampicillin

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.; Day, G.M.; Emsley, L., “Rapid Structure Determination of Molecular Solids Using Chemical Shifts Directed by Unambiguous Prior Constraints”. *Journal of the American Chemical Society* **2019**, XXXX, XXX. ***(pre-print)*** and Engel E.A.; Anelli, A.; Hofstetter, A.; Paruzzo, F.; Emsley, L.; Ceriotti, M., “A Bayesian approach to NMR crystal structure determination”, *submitted* **2019**, **(pre-print)**

2.6.1 Introduction

In the current CSP-NMRX approaches, structural information obtained from solid state NMR is usually included only in the final step, to select the correct crystal structure from an ensemble of predicted structures. Here, we show with the case of ampicillin that this can lead to failure of structure determination, as the correct structure is excluded from the search space during the preceding conformer selection in the CSP approach. In **Chapter 2.2** we proposed a crystal structure determination method, based on the analysis of absent cross-peaks in solid-state NMR correlation experiments, that includes experimental constraints already during conformer selection. In **Chapter 2.2** we also showed that these absences provide unambiguous structural constraints on both the crystal structure and the gas phase conformations, and therefore can be used for unambiguous selection. The approach was also parameterized on the crystal structure determination of flutamide, flufenamic acid, and cocaine.

Here, we apply the approach presented in **Chapter 2.2** to correctly determine the crystal structure of ampicillin, which would have failed using current methods because ampicillin adopts a high energy conformer in its crystal structure. Additionally, we adapt the machine learning approach (ShiftML) presented in **Chapter 2.3,** to predict the chemical shifts of molecular solids containing H,C,N,O and S atoms with an ^{1}H RMSE compared to experiment of 0.346 ± 0.195 ppm. Thus, making ShiftML applicable to the NMRX crystal structure determination of ampicillin. Further, we apply the Bayesian approach to NMRX, described in **Chapter 2.5**, to determine the crystal structure of powdered ampicillin with up to 95% confidence. Finally, we apply the positional uncertainty approximation presented in **Chapter 2.4** to determine that 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 $U_{eq} = 0.0103$ Å^2.

2.6.2 Results and Discussion

In contrast to the three cases discussed in **Chapter 2.2** the crystal structure determination of ampicillin would have failed using the usual CSP-NMRX protocol. In the first step, an ensemble of 16 locally stable gas-phase conformers is generated (for details, see Methods) and the ensemble is then sorted according to the isolated molecule conformational energy. **Figure 2-61b** and 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 2-61b 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[56, 162] 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 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.[162]

To solve this problem, we apply experimental constraints extracted from ^{1}H-^{13}C HETCOR spectra at different contact times 0.1, 0.3, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0 and 2.25 ms, detailed in the Methods section. **Figure 2-60** shows the assigned HETCOR spectrum of ampicillin at 1.5 ms contact time together with the labelled 2D structure. Following the protocol established for cocaine, flutamide and flufenamic acid, the SNR is then normalized over all experimental setups and for the amount of active ^{1}H. As we did for the other three molecules, we only consider cross-peaks resulting from *terminal*-protons, see **Figure 2-67**. Using the S_{norm} of 0.14, and X of 3.5 Å, that were parametrized on the reference compounds **in Chapter 2.2**, the extracted constraints are circled in orange and are shown on three example conformers below the spectra. **Figure 2-61a** shows the sub-ensembles with no violations (0 out of 1 total constraint). **Figures 2-60** and **2-69** show that only conformers without an intra-molecular hydrogen bond are selected. Also, from **Figure 2-61b** it is clear 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 we now continue with only 7 out of the original 16 structures. This reduces the computational cost by approximately 55%

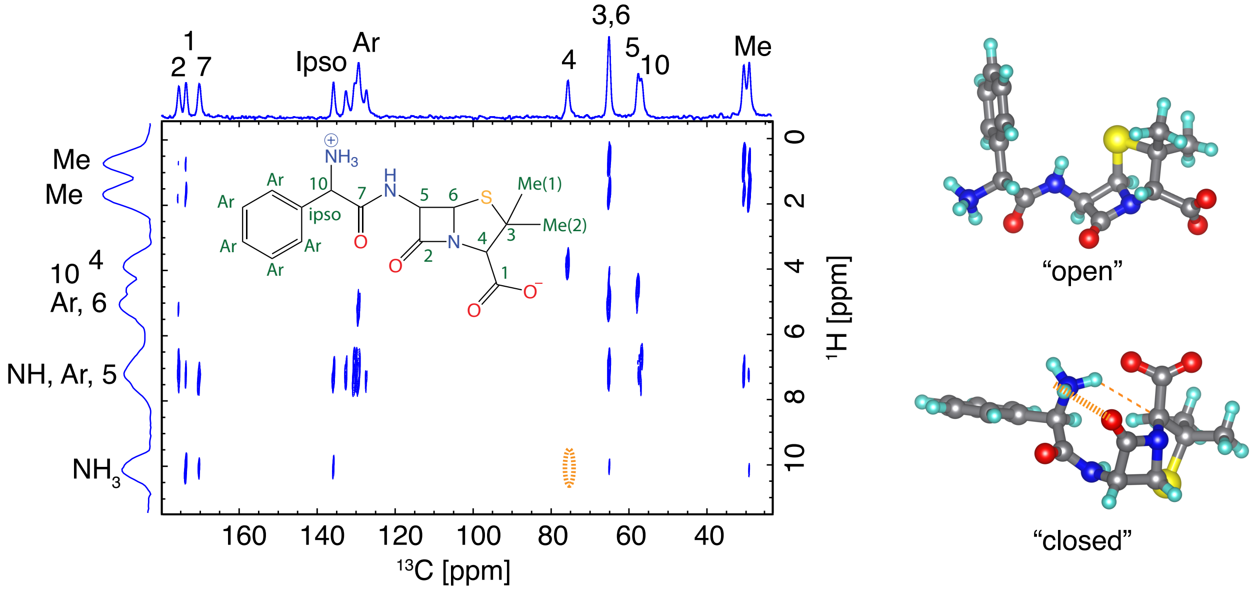


Figure 2-60. The left part shows the ^{1}H-^{13}C HETCOR spectrum of ampicillin with 1.5 ms contact time (further details in Methods). ^{13}C peaks are assigned based on the literature[178] and ^{1}H peaks are assigned from HETCOR spectra and DFT chemical shift calculations (see Methods). The cross-peaks from the terminal protons (**Figure 2-67**) below a S_{norm} of 0.14 were used as constraints on the conformer ensembles, and are indicated as orange ellipsoids. The right part shows the violated constraints extracted from all of the ^{1}H-^{13}C HETCOR cross-peaks for different example conformers within the ensemble.

For each conformer within this reduced gas-phase ensemble, we generated a crystal structure ensemble using a quasi-random sampling[246] 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,[236] followed by DFT re-optimization of the lowest energy crystal structures, producing a final set of 75 candidate crystal structures. The full procedure is detailed in the Methods.

^{1}H chemical shieldings were then calculated with GIPAW DFT and a machine learned method (ShiftML)[161] for each candidate structure and compared to the experimental chemical shifts (details are given in the Methods). **Figure 2-62** shows the RMSE between DFT calculated and measured ^{1}H chemical shifts together with the calculated relative lattice energies for the candidate set. With current accuracy we expect a correct structure to have a ^{1}H RMSE of 0.33 ppm (±0.16 ppm) or lower.[18] This is indicated as the grey zone in **Figure 2-62**. Predicted structures with ^{1}H chemical shift errors within this zone are thus considered to be indistinguishable from experiment.

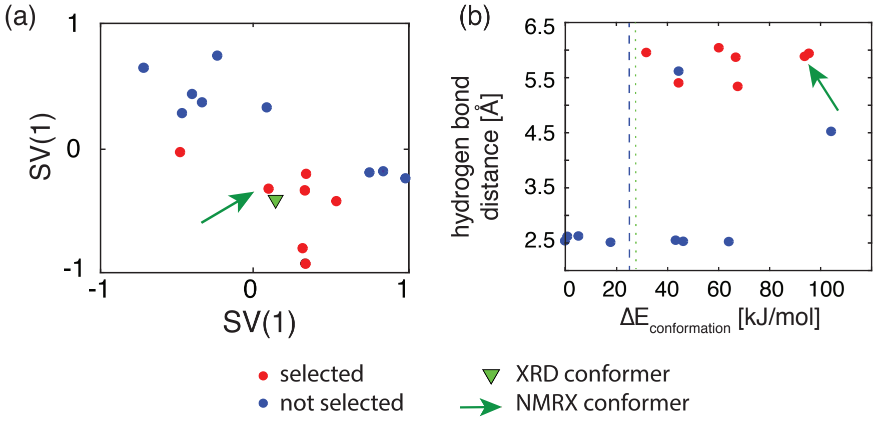


Figure 2-61. Conformer selection for ampicillin. **(a)** The panel shows the sketch-map projections of the gas-phase ensemble. Red dots represent the structures which are selected for a threshold distance of 3.5 Å and a S_{norm} cut-off value of 0.14. The green triangle shows the gas-phase conformer of the XRD 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 intramolecular hydrogen-bond distance (N-O distance). The blue dashed line is the typical cut off energy (25 kJ/mol) used for 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.

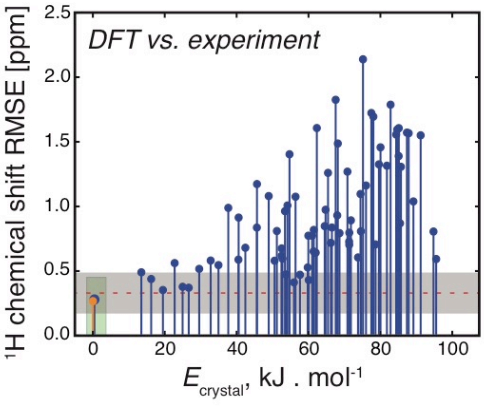


Figure 2-62. Comparison of crystal structure candidates. The structures are sorted according to their relative lattice energy, horizontal axis. The vertical axis shows ^{1}H chemical shift RMSE between DFT calculated and experimental chemical shifts. The orange marker shows the ^{1}H chemical shift RMSE for the single-crystal XRD structure. The red line shows the mean of the current error between experimental and DFT calculated ^{1}H chemical shifts with the limits indicated as grey shaded zone, as described in the main text.

Figure 2-71 shows the RMSE between ShiftML calculated and measured ^{1}H 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 ^{1}H chemical shifts (as described in the Methods, **Table 2-19**) we expect a correct structure to have a ^{1}H RMSE of 0.346 ppm (±0.195 ppm) or lower. Note that the RMSE between experiment and the predicted chemical shifts follows the same trends as for the DFT calculated shifts (**Figure 2-62**).

Based on the agreement between experimental and calculated ^{1}H 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 (**Figures 2-62** and **2-69**). 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 ^{1}H 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)[247] above the best candidate structure. Thus, our final structure selection relies on both the chemical shifts and calculated lattice energies.

Further, we apply the Bayesian approach to NMRX, described in **Chapter 2.5**. **Figure 2-63a** shows the prediction confidence with which we identify the correct crystal structure – 75% confidence using DFT calculated 1H chemical shifts and 94% confidence using ShiftML calculated chemical shifts. If we include the information obtained from ^{13}C chemical shifts the prediction confidence increases to 90% and 95% confidence. Note that, contrary to the RMSE based structure determination, calculated lattice energies do not have to be considered too clearly determine the correct crystal structure.

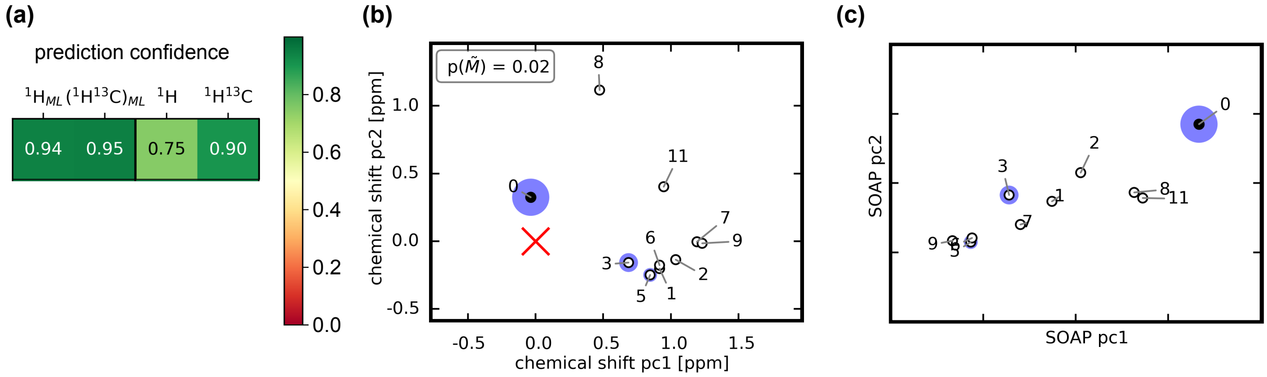


Figure 2-63. Bayesian approach to NMRX for powdered ampicillin. Prediction confidence of the determined ampicillin crystal structure using ^{1}H and ^{13}C chemical shifts calculated with DFT and ShiftML **(a).** Evaluation of the top 10 ampicillin CSP candidates. The correct candidates are shown as filled circles and the others as empty circles. For each candidate the probability of matching experiment $p(M|\boldsymbol{y}^*)$ 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 favorable candidate **(b-c)**. Panel **(b)** shows 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 ^{1}H DFT chemical shifts. $p(\widetilde{M})$ denotes the probability that the virtual candidate, which represents structures potentially missing from the CSP candidate pool, matches experiment. Panel **(c)** shows 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 **Chapter 2.5.2** has no intuitive physical meaning and is therefore not shown.

Figures 2-63b-c show the representations of the similarity of the CSP candidates for ampicillin. We show the similarity in terms of DFT calculated ^{1}H chemical shifts **(b)** and in terms of structure **(c)**. 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. Both in term of ^{1}H chemical shifts and in terms of structure the determined crystal structure is clearly distinguishable from the remaining CSP candidates.

The structure determined here agrees very well with the known reference structure determined by single-crystal XRD,[248] as illustrated in **Figure 2-64a**. 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, demonstrating an excellent determination of the molecular conformation in the crystal structure. 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 **Table 2-20**). 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 2-64b-c**. The positional error analysis is performed using the DFT calculated ^{1}H chemical shifts following the procedure outlined by in **Chapter 2.4** and is detailed in the Methods (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 $U_{eq} = 0.0103$ Å^2. This compares with $\langle r_{av}\rangle = 0.149$ Å and $U_{eq} = 0.0074$ Å^2 for the single-crystal XRD structure.[248] 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 ^{1}H atoms.

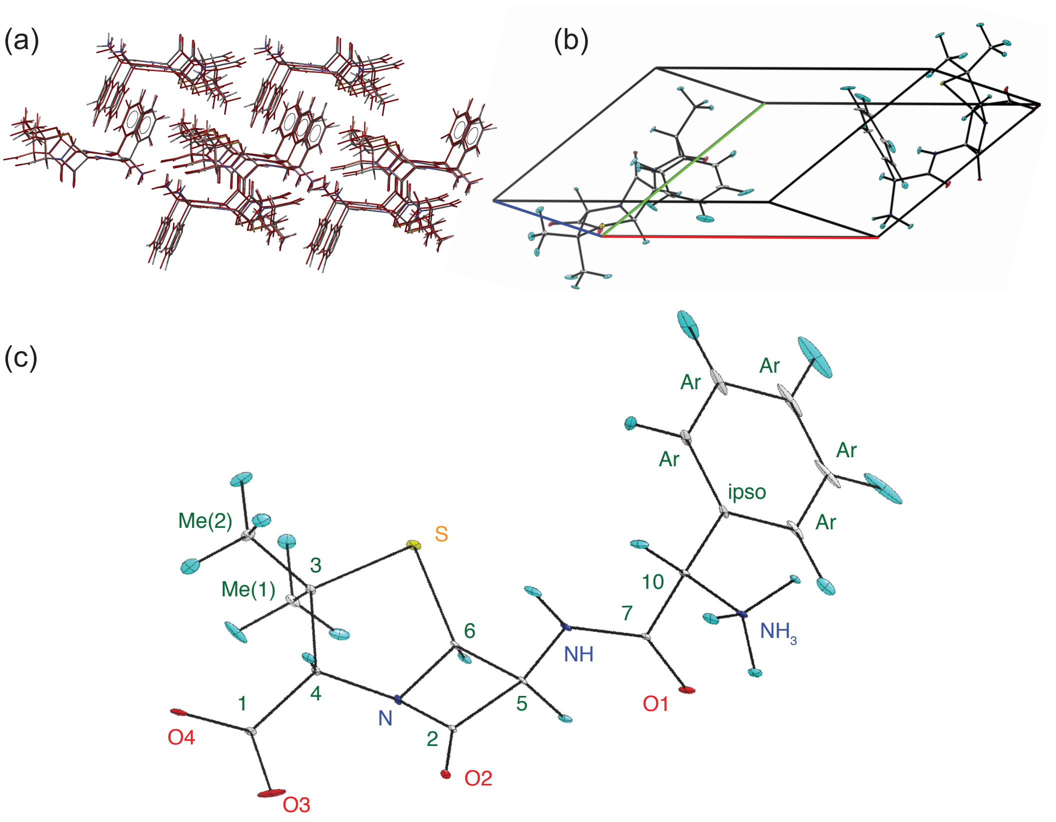


Figure 2-64. (a) Comparison between the structure of ampicillin as determined by the constrained powder ^{1}H CSP-NMRX and the single crystal XRD determined structure.[248] **(b-c)** ORTEP plot of the ampicillin crystal **(b)** and single molecule **(c)** structure drawn at the 90% probability level. The anisotropic ellipsoids correspond to a ^{1}H chemical shift RMSE of 0.49 ppm and to an average positional RMSE of $\langle r_{av} \rangle = 0.144$ Å. (d)

2.6.3 Conclusion

Here we demonstrated the capability of the novel constrained CSP-NMRX method and the Bayesian approach to NMRX by successfully determining the crystal structure of powdered ampicillin with up to 95% confidence, which would have been very challenging for previous methods and either requiring that no energetic limit was applied to the conformational energy, or likely missing the correct crystal structure. Here, a rough estimation shows that to run the CSP-NMRX calculations, including CSP search, DFT optimization and chemical shift calculations, for all 16 conformers would take approximately 54 days on 200 dedicated CPUs. By constraining the structural search space, we were able to more than halve this 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 CSP-NMRX based determination of larger and more flexible molecules, which would previously have been out of the scope of the CSP-NMRX approach.

Note that the compounds studied here were not subjected to any modification prior to the experiments, and they were investigated using powder samples at natural isotopic abundance. The resulting structures have a positional accuracy that is comparable to structures from, for example, single crystal XRD, while including the positions of the light atoms.

2.6.4 Methods

Samples

The powdered sample of anhydrous ampicillin ((2S,5R,6R)-6-([(2R)-2-amino-2-phenylacetyl]amino)-3,3-dimethyl-7-oxo-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylic acid, purity > 98.0%) was purchased from Sigma-Aldrich The reference crystal structure (CSD entry: AMCILL) was previously determined by single-crystal XRD[179-181, 248] and is monoclinic, space group $P2_1$, with unit cell parameters a = 12.40 Å, b = 6.20 Å, c = 12 Å, and 2 molecules in the unit cell.

Solid-state NMR experimental setup

Experiments were performed at room temperature on a Bruker 500 wide-bore Avance III and a Bruker 900 US^2 wide-bore Avance Neo NMR spectrometers operating at Larmor frequencies of 500.43 and 900.13 MHz, equipped with H/X/Y 3.2 mm and H/C/N/D 1.3 mm probes.

The 2D ^{1}H-^{13}C dipolar heteronuclear correlation (HETCOR) experiments were performed at 12.5 kHz MAS. In all experiments, we used SPINAL-64 for heteronuclear decoupling during t1 and eDUMBO-1_{22} for homonuclear decoupling in the indirect dimension. 16 and 128 transients with 256 increments were acquired for ampicillin.

The ^{1}H and ^{13}C chemical shifts were referenced indirectly to tetramethylsilane using the methyl signals of L-alanine at 1.3 ppm (^{1}H) and 20.5 ppm (^{13}C),[182] while ^{15}N chemical shifts were referenced using glycine at –347.54 ppm. ^{1}H chemical shifts were corrected for the scaling factor due to homonuclear decoupling, which was determined using ^{1}H 1D spectra acquired under fast spinning on a Bruker 900 spectrometer. Post-processing was done using Topspin 3.5 or 3.6.1.

The 11.7 T 2D ^{13}C-^{13}C refocused Incredible Natural Abundance Double Quantum Transfer Experiment (INADEQUATE) was performed using a 13.0 kHz MAS frequency at a temperature of 295 K. Prior to the indirect evolution period, cross-polarization (CP) from the ^{1}H nuclei was carried out (contact time of 2.5 ms). SPINAL-64 heteronuclear decoupling (100 kHz nutation frequency) was used during both evolution dimensions. 1760 transients with 128 t_1 increments were used. Each τ delay during the indirect dimension evolution was set to 3.84 ms, the length of the z-filter was 1.0 ms, and the recycle delay was 1.0 s.

The 16.4 T ^{1}H-^{15}N CP-HETCOR NMR experiment was carried out at T = 265 K using a 15 kHz MAS rotation frequency, while a ^{15}N magic-angle-turning (MAT) experiment was performed at T = 266 K and a 1.90 MAS rotation frequency. For the ^{1}H-^{15}N HETCOR experiment, SPINAL-64 heteronuclear decoupling was used during the t_2 dimension (83 kHz nutation frequency), and eDUMBO-1_{22} was used for homonuclear decoupling in the indirect dimension (the scaling factor was set to 0.564). Prior to the indirect evolution period, CP from the ^{1}H nuclei was done (contact time = 300 μs). 1440 transients with 64 t_1 increments were used. For the ^{15}N MAT experiment, SPINAL-64 heteronuclear decoupling was used during both the t_1 and t_2 dimensions (100 kHz nutation frequency). Prior to the indirect evolution period, CP from the ^{1}H nuclei was done (contact time = 5.5 ms), with 1024 transients being acquired and averaged per t_1 increment, and with 125 t_1 increments being used.

Assignment of experimental NMR spectra

The assignment of the ^{13}C spectra of ampicillin has been done by Clayden *et al.*[249] and then revised by Antzutkin *et al.*[178], but as the above authors mentioned, the assignment remains ambiguous, and so we revised it. To assign the ^{13}C NMR spectra at natural abundance a ^{13}C-^{13}C INADEQUATE experiment was done. To assign the ^{1}H directly attached to ^{13}C, the ^{1}H-^{13}C HETCOR spectra were used. To assign the ^{1}H directly attached to ^{15}N, a ^{1}H-^{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, from which it was possible to tell that the NH_3 resonance corresponds to the peak with negligible chemical shift anisotropy due to the fast exchange of the three attached ^{1}H 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.

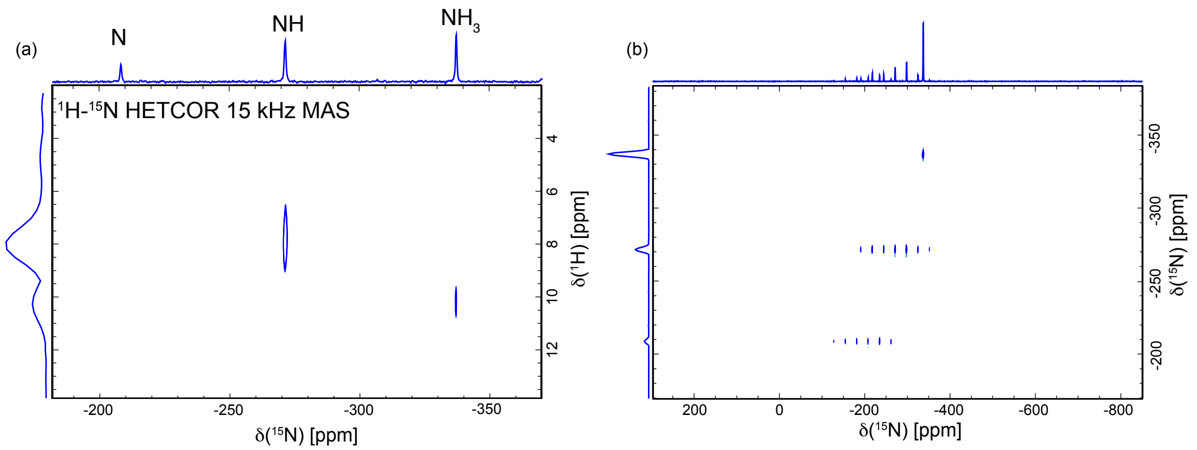


Figure 2-65. ^{15}N spectra of ampicillin used for the ^{1}H and ^{15}N assignments. **(a)** ^{1}H-^{15}N HETCOR spectra of ampicillin measured at 16.4 T and 15 kHz MAS. **(b)** ^{15}N MAT spectra of ampicillin at 16.4 T.

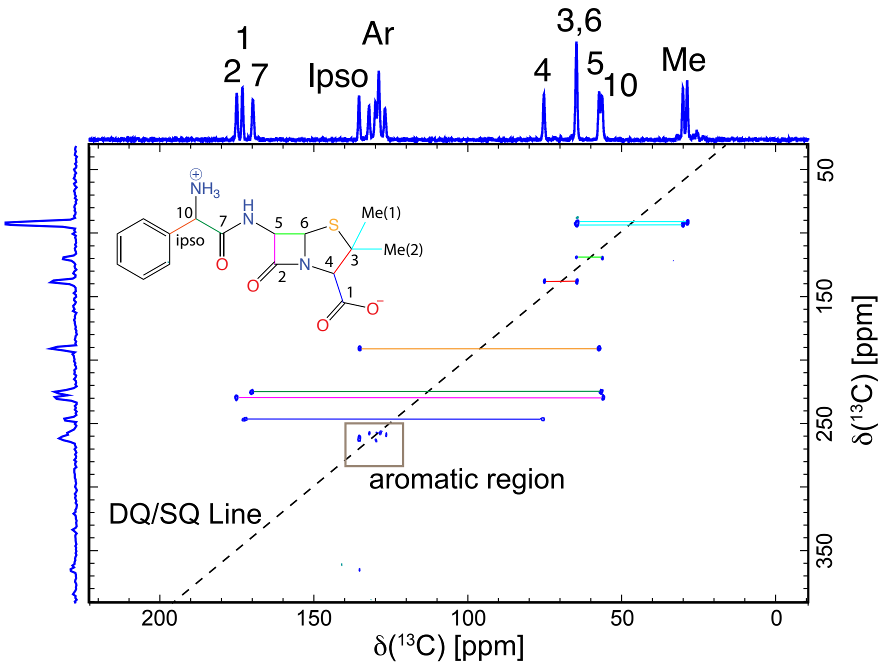


Figure 2-66. ^{13}C-^{13}C INADEQUATE spectra of ampicillin used for the ^{13}HC assignments, measured at 11.7 T and 13 kHz MAS.

Experimental chemical shifts

Table 2-15. Ampicillin experimental chemical shifts.

Label	^{1}H, 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(meta)	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	

Signal to Noise analysis

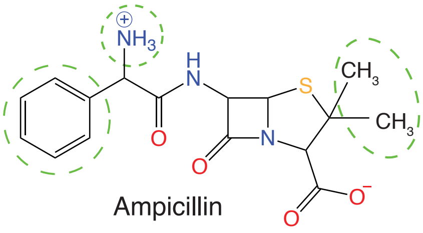


Figure 2-67. Illustration of terminal protons, for which cross-peaks contribute to conformational constraints.

Table 2-16. Protons contributing to conformational constraints for ampicillin

Molecule	terminal ^{1}H
Ampicillin	Ar
	NH_3
	Me(1)
	Me(2)

Gas-phase conformer generation

For ampicillin, we generated as complete set of gas phase conformers as possible using a low-mode conformational search (LCMS) method,[235, 250] as implemented in MacroModel.[251] Energies were calculated during the conformer search using the OPLS3 force field.[252] The only prior knowledge used was that bonding within the molecule was fixed in the zwitterionic configuration throughout the conformer search; this information is readily available from NMR. 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 within Gaussian09 using dispersion corrected density functional theory (DFT-D) at the B3LYP/6-311G** level of theory with the D3BJ dispersion correction. 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 proton transfers from the amino to the carboxyl group during DFT optimization of many of the conformers. However, the resulting non-zwitterionic conformers are not relevant to the crystal structure of ampicillin.

In analyzing the conformers resulting from the search, we found that the configuration around chiral centers could be reversed during the LCMS 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 selection was performed for CSP.

Sketch-map analysis

The cluster generation and analysis were performed as described in **Chapter 2.2** The sketch-map parameters are given **Table 2-17**. They were chosen following the procedure described in Ceriotti *et al.*[175] and the tutorial on sketchmap.org. The sketch-map analysis was not sensitive to small variations in the chosen parameters, as was already noted in the references.[175-177] As starting point for the sketch-map analysis we used all dihedral angles, not containing protons, over the full 2π range. This gives 55 dihedral angles for ampicillin, within a range of $-\pi$ to π.

Table 2-17. Sketch-map parameters for all ampicillin.

Structure	$\Sigma = \sigma$	A	B	a	b
Ampicillin	6	2	2	1	1

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. The 14th conformer is the most similar to the crystal conformer and resulted in the correct crystal structure after the remaining CSP procedure. **Figure 2-68** shows the sketch-map analysis of the ampicillin gas-phase ensemble.

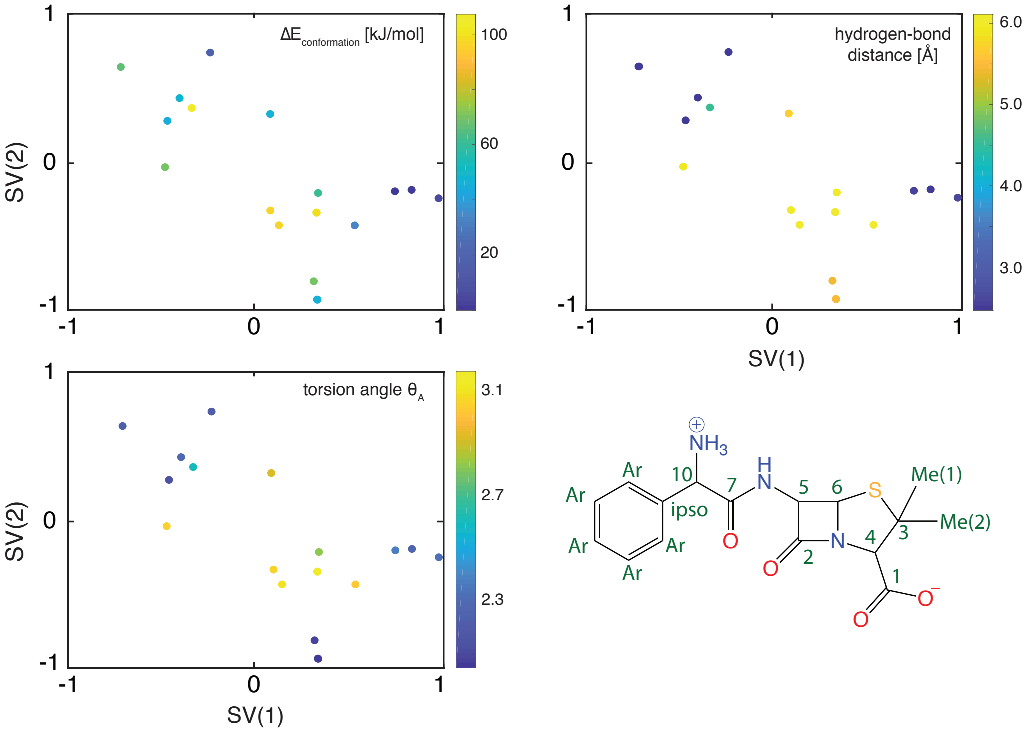


Figure 2-68. Sketch-map representation of the locally stable ampicillin conformers. To show the extent of the sub-clustering the panels are colored according to different molecular properties. **Top left** shows the difference in conformational energy ($\Delta E_{conformation}$). **Top right** shows 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, defined as the torsion angle between C_{10}-C_7-N(H)-(N)H. In general, the clustering seems to correspond to conformational changes along the C_{ipso}-C_{10}-C_7-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 used labelling scheme.

Conformer selection

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

The sub-ensemble selection for ampicillin, is done based on constraints from multiple HETCOR contact times 0.1, 0.3, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75 and 2.25 ms. The ^{1}H and ^{13}C cross peaks from the two methyl groups were not distinguished. Also, the ^{1}H cross peaks from Ar2-6, H5 and NH, the ^{1}H 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 are too close and not distinguishable. Therefore, if a cross-peak was seen it was attributed to all of the atoms within the given group.

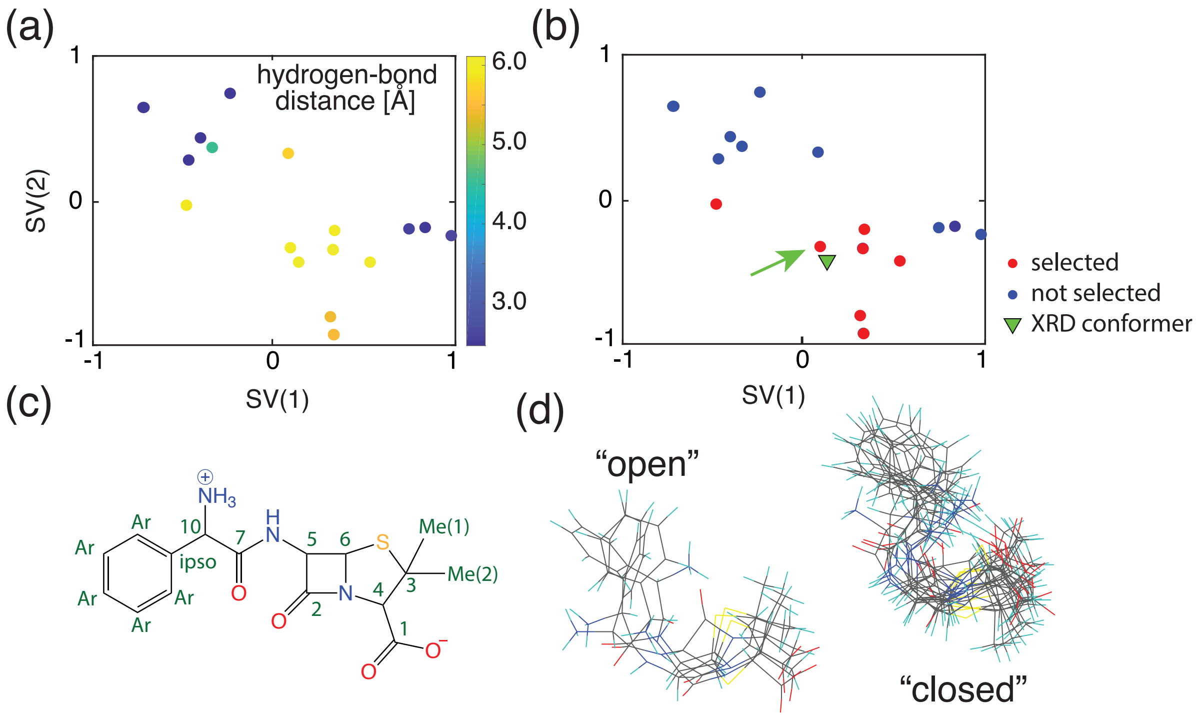


Figure 2-69. (a) Sketch-map representation of the locally stable ampicillin conformations. To show the extent of the sub-clustering the panel is colored according to the shortest intra-molecular hydrogen-bond distance [Å] between either NH_3 or NH and the carboxyl group. **(b)** Sketch-map projection of the gas-phase ampicillin ensemble. Red dots represent the structures with the lowest violations that are selected. The greed triangle shows the gas-phase conformer of the XRD crystal structure. The green arrow points to the gas-phase conformer, which resulted in the correct crystal structure after the CSP procedure. **(c)** 2D structure of ampicillin with the used labelling scheme. **(d)** Overlap of the structures within the different sketch-map clusters. The “open” conformations correspond conformers without an intra-molecular hydrogen bond and are selected. The “closed” conformations mostly contain an intra-molecular hydrogen bond and are 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) 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.[246] 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 Söhnke 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 DMACRYS[236] with the molecular geometry kept fixed at the gas phase geometry. Intermolecular interactions were evaluated using the empirically parameterized FIT force field[237] with electrostatic interactions modelled using atomic multipoles, up to hexadecapolar on each atom, derived using a distributed multipole analysis[253] 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 solid state 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 relaxed with DFT using the Castep[191] suit, using the PBE- functional, the D2 dispersion correction, a 500eV 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 relaxed, including the unit cell and all atomic positions. The resulting structures were used as starting points for the chemical shift modelling (see below).

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 ampicillin[248] were fully relaxed, including the unit cell and all atomic positions, using the same DFT parametrization as for the chemical shift calculations described below.

DFT chemical shielding calculation. For the ampicillin crystal structure selection the magnetic shielding of the 75 trial crystal structures were calculated with plane-wave DFT using the GIPAW formalism[254] and the Quantum ESPRESSO suite.[188] For the GIPAW DFT calculations the generalized-gradient-approximation (GGA) density functional PBE[205] was used. We used the ultrasoft pseudopotentials with GIPAW[62-63] reconstruction, 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 PS library database.[242] 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[207] corresponding to a maximum spacing of 0.04 $Å^{-1}$ in the reciprocal space was used. The electron density self-consistency convergence threshold was set to 10^{-12} Ry.

ShiftML chemical shielding calculation. For the ampicillin crystal structure selection, the magnetic shieldings of the 75 trial crystal structures were calculated as described below.

Shielding to shift conversion. The calculated magnetic shielding was referenced to the experimental chemical shifts, using the linear relationship $\delta_{exp} = a - b\sigma_{DFT}$, where the slope (*b*) and the offset (*a*) were fit for each trial structure individually. For the ^{1}H 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 (ShiftML) chemical shifts, and then compared to each other. This was done for each crystal structure individually. The RMSE was calculated as,

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(\delta_{i,exp} - \delta_{i,calc})^2}{N}}, \tag{2-52}$$

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

ShiftML

The machine-learning model used to predict the ^{1}H chemical shifts follows the basic concepts behind ShiftML,[161] which are detailed in **Chapter 2.2.** However, the original implementation of ShiftML is only able to predict ^{1}H chemical shifts of structures containing H,C,N and O atoms. Thus, we extended the training set in the following manner:

a) Starting from the CSD-61k set and including the CSD-2k training set, described in **Chapter 2.2**, we used a farthest point sampling algorithm (FPS) to include an additional 1,000 training structures.

b) From the Cambridge Structural Database (CSD),[135] we extracted a set of around 22'000 molecular crystal structures containing less than 200 atoms in the unit-cell and containing H,C and S atoms as well as optionally N and O atoms (CSD-S22k). This set was curated analogously to the CSD-61k set and using an FPS algorithm we selected 546 structures from this set.

c) These three structure sets were combined to form the CSD-3k+S546 set.

d) As structures often contain redundant environments, for example due to crystal symmetries, the training set was reduced by FPS ordering the individual environments and retaining only the 65,000 most structurally diverse.

Additionally, the ShiftML model was changed to contain radially scaled smooth overlap of atomic positions (SOAP) kernels[176, 184, 198] as opposed to the seven multi-scale SOAP kernels described in **Chapter 2.2.** This change was implemented to increase the computational efficiency of the model. The parameters of the used ShiftML implementation are given in **Table 2-18**, using the same notation as in **Chapter 2.2** and Willatt *et al.* **[198]**

In order to estimate the prediction accuracy of the updated ShiftML model, we combined the CSD-500 test set from **Chapter 2.2** with 104 random structures extracted from the CSD-S22k set. For this combined CSD-500+S104 test set, we find a RMSE of 0.44 ppm between ^{1}H chemical shifts calculated with DFT and ShiftML. This is directly comparable to the ^{1}H RMSE of 0.49 ppm reported in **Chapter 2.2.** We ascribe the slightly lower ^{1}H chemical shift RMSE to the fact that a larger training set was used.

The Refcodes of all CSD sets are given in the original publication : Hofstetter, A.; Balodis, M.; Paruzzo, F.M.; Widdifield, C..M.; Stevanato, G.; Pinon, A.C.; Bygrave, P.; Day, G.M.; Emsley, L., “Rapid Structure Determination of Molecular Solids Using Chemical Shifts Directed by Unambiguous Prior Constraints”. *Journal of the American Chemical Society* **2019**, XXXX, XXX.

Note, that all the DFT calculations and all of the treatment of the training set, e.g. the detection of unusual environments, was done as described in **Chapter 2.2.**

Table 2-18. Parameters used for the implemented ShiftML version.

Atom	r_c (cutoff)	**c** (cutoff rate)	**m** (cutoff dexp)	**r0** (cutoff scale)	**u0** (central weight)	**gw** (atom sigma)	**nmax**	**lmax**	**cutoff transition width**
^{1}H	5	1	4	2.5	1.0	0.3	9	9	0.5

The RMSE between the ^{1}H chemical shifts calculated with DFT and ShiftML is calculated as 0.464 ppm over all the ampicillin trial structures. This agrees with the overall error reported for this implementation of ShiftML. **Figure 2-70** shows the correlation between ^{1}H chemical shieldings calculated with DFT and ShiftML.

Figure 2-71 shows the RMSE between ShiftML calculated and measured ^{1}H chemical shifts together with the DFT calculated relative lattice energies for the candidate set. Note that the RMSE between experiment and the predicted chemical shifts follows the same trends as for the DFT calculated shifts (**Figure 2-62a**).

Note that, for the Bayesian approach to NMRX for ampicillin (see Figure 2-63) the ML chemical shifts were calculated with the ShiftML version described in Appendix III.

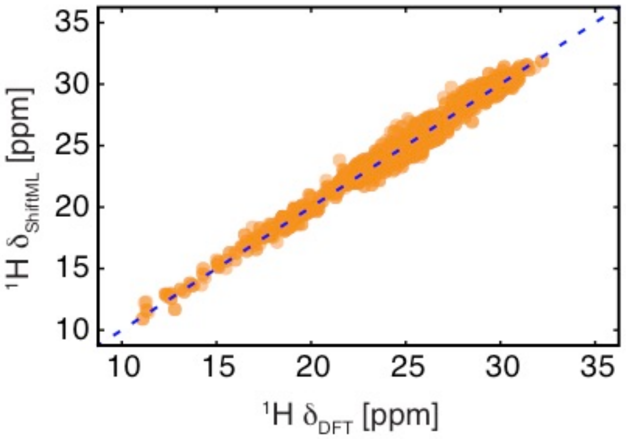


Figure 2-70. Scatterplot showing the correlation between ^{1}H chemical shieldings calculated with DFT and ShiftML, with a RMSE of 0.464 ppm. The blue dotted line indicates a perfect correlation.

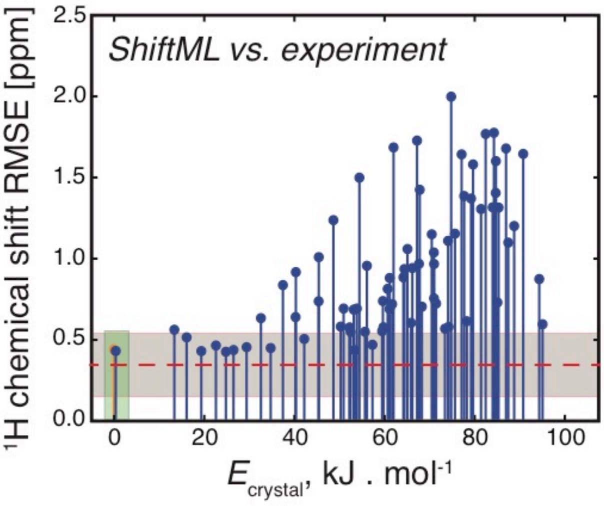


Figure 2-71. Comparison of crystal structure candidates. The structures are sorted according to their relative lattice energy, horizontal axis. The vertical axis shows ^{1}H chemical shift RMSE between ShiftML calculated and experimental chemical shifts. The orange marker shows the ^{1}H chemical shift 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 ^{1}H chemical shifts with the limits at one standard deviation (0.195 ppm) indicated as grey shaded zone, as described below.

ShiftML error estimation.

Comparison between ^{1}H experimental chemical shifts and ^{1}H chemical shifts calculated with ShiftML were carried out analyzing around 150 chemical shifts obtained from 11 crystal structures. The names, IUPAC IDs, CSD reference codes (when available) and references to the experimental NMR data of the analyzed crystal structures are the following:

(i) Naproxen, (2S)-2-(6-Methoxy-2-naphthyl)propanoic acid, COYRUD11, Ref.[210]

(ii) Uracil, Pyrimidine-2,4(1H,3H)-dione, URACIL, Ref. [211]

(iii) Co-crystal of 3,5-dimethylimidazole and 4,5-dimethylimidazole, Ref. [212]

(iv) Theophylline, 1,3-Dimethyl-3,7-dihydro-1H-purine-2,6-dione, BAPLOT01, Ref. [56]

(v) Anthranilic acid, AMBACO05, Refs.[83, 255]

(vi) Cimetidine, CIMETD, Refs.[83, 256]

(vii) Phenobarbital, PHBARB06, Refs.[74, 83]

(viii) Thymol, IPMEPL, Ref.[18]

(ix) Terbutaline hemi-sulfate, ZIVKAQ, Refs.[21, 83]

(x) Cocaine, methyl (1R,2R,3S,5S)-3- (benzoyloxy)-8-methyl-8-azabicyclo[3.2.1] octane-2-carboxylate, COCAIN10, Ref. [56]

(xi) AZD8329, 4-[4-(2-adamantylcarbamoyl)-5-tert-butylpyrazol-1-yl]benzoic acid, Ref.[58]

The crystal structures (i-ix) were obtained from Ref. [83], where the experimentally determined crystal structures were subjected to all-atom geometry optimization with fixed lattice parameters, as described in the reference. Crystal structures (x) and (xi) were obtained from Refs. [56] and [58] respectively. We only used the ^{1}H 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 magnetic shieldings σ are converted to the corresponding chemical shifts δ through the relationship $\delta_{exp} = a - b\sigma_{DFT}$, where the slope ($b$) and the offset ($a$) were fit for each reference structure individually. The chemical structures, 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 2-72** and **Table 2-19**. For the entire reference set we calculate an average RMSE of 0.346 ppm and a standard deviation of 0.195 ppm.

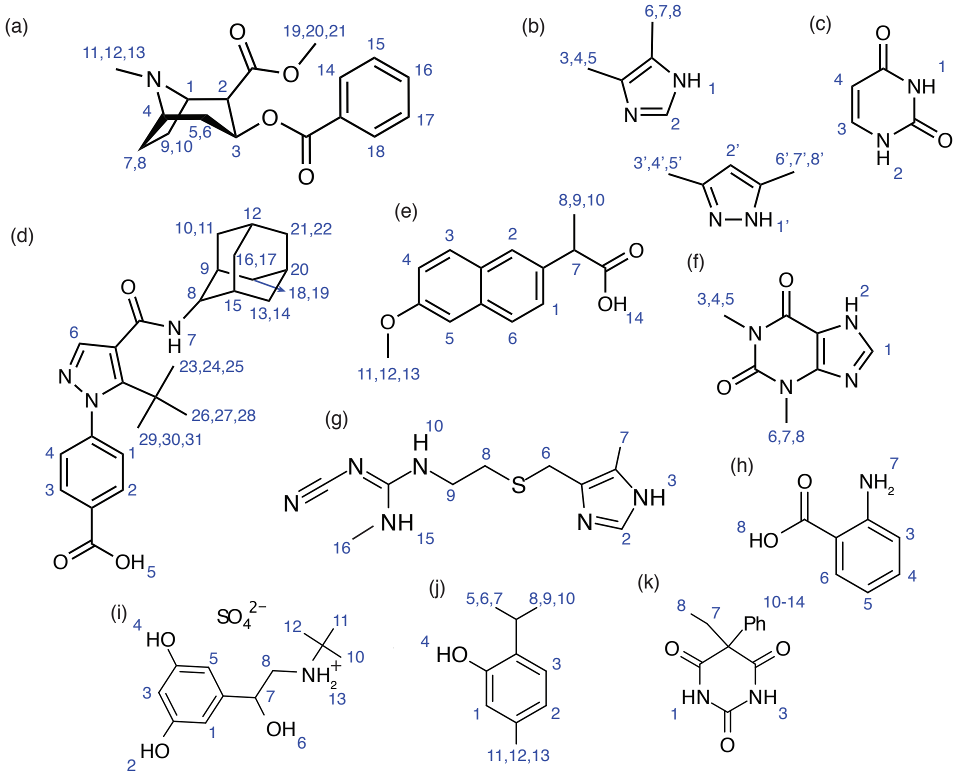


Figure 2-72. Chemical structures of the compounds used for experimental comparison. In order, cocaine (a), 3,5-dimethylimidazole and 4,5-dimethylimidazole (b), uracil (c), AZD8329 (d), naproxen (e), theophylline (f), cimetidine (g), anthranilic acid (h), terbutaline hemi-sulfate (i), thymol (j) and phenobarbital (k) and the labelling scheme used here.

Table 2-19. Experimental and calculated chemical shifts of the structures used in the ShiftML benchmarking, . The labelling scheme is given in **Figure 2-72**. When more than one atom corresponds to a single chemical shift value, their values were averaged.

Naproxen			**Uracil**		
Atom Label	Experimental ^{1}H δ(ppm)	ShiftML ^{1}H δ(ppm)	Atom Label	Experimental ^{1}H δ(ppm)	ShiftML ^{1}H δ(ppm)
1	7	6.44	3	7.5	7.43
2	6.1	5.60	2	10.8	10.79
3	3.8	3.86	1	11.2	11.22
4	4.5	4.65	4	6	6.05
5	4.1	4.65			
6	5.9	5.48			
7	3.2	2.88			
8,9,10	1.8	1.56			
11,12,13	2.3	2.80			
14	11.5	11.75			
a = 4.79 ppm	***b = 0.81***	***RMSE = 0.393 ppm***	***a = 5.15 ppm***	***b = 0.77***	***RMSE = 0.048 ppm***

3,5-dimethylimidazole & 4,5-dimethylimidazole			**Theophylline**		
Atom Label	Experimental ^{1}H δ(ppm)	ShiftML ^{1}H δ(ppm)	Atom Label	Experimental ^{1}H δ(ppm)	ShiftML ^{1}H δ(ppm)
1’	13.0	13.25	2	14.6	14.79
2’	4.8	5.03	1	7.7	7.10
3’,4’,5’	1.4	1.12	3,4,5	3.4	3.54
6’,7’,8’	0.7	1.07	6,7,8	3.4	3.40
1	15	14.62			
2	5.2	5.52			
3,4,5	1.5	1.47			
6,7,8	1.4	1.20			
a = 4.85 ppm	***b = 0.92***	***RMSE = 0.27 ppm***	***a = 5.19 ppm***	***b = 0.84***	***RMSE = 0.24 ppm***

Cocaine			**AZD8329**		
Atom Label	Experimental ^{1}H δ(ppm)	ShiftML ^{1}H δ(ppm)	Atom Label	Experimental ^{1}H δ(ppm)	ShiftML ^{1}H δ(ppm)
1	3.76	4.17	1	6.92	6.54
2	3.78	2.79	2	8.69	8.30
3	5.63	5.78	3	9.01	8.74
4	3.32	3.54	4	8.47	7.64
5	3.06	1.83	5	15.37	14.90
6	3.49	2.56	6	7.73	8.04
7	2.91	2.17	7	9.64	10.70
8	3.38	3.04	8	2.90	2.72
9	2.56	2.19	9	1.78	2.03
10	2.12	2.37	10	1.88	2.28
11,12,13	1.04	1.87	11	1.88	2.28
14	8.01	7.90	12	1.8	1.99
15	8.01	7.90	13	1.6	1.48
15	8.01	7.90	14	0.44	1.21
17	8.01	7.90	15	1.54	1.71
18	8.01	7.90	16	1.88	2.10
19,20,21	3.78	4.27	17	1.88	2.10
			18	0.8	1.39
			19	0.8	1.39
			20	1	1.85
			21	1.74	1.75
			22	1.74	1.75
			23,24,25	0.73	0.39
			26,27,28	0.73	0.83
			29,30,31	0.73	-0.16
a = 5.88 ppm	***b = 1.05***	***RMSE = 0.59 ppm***	***a = 5.40 ppm***	***b = 1.06***	***RMSE = 0.50 ppm***

Cimetidine

Atom Label	Experimental ^{1}H δ(ppm)	ShiftML ^{1}H δ(ppm)
2	7.64	7.55
3	11.84	11.55
7	2.24	2.17
10	8.44	9.00
15	9.94	9.86
16	2.24	2.28
a = 5.12 ppm	***b = 0.86***	***RMSE = 0.21 ppm***

Anthranilic acid

Atom Label	Experimental ^{1}H δ(ppm)	ShiftML ^{1}H δ(ppm)
Aromatic (1)	5.8	5.74
Aromatic (2)	6.8	6.66
NH2	5.4	5.52
COOH	12.3	12.33
a = 4.95 ppm	***b = 0.79***	***RMSE = 0.095 ppm***

Phenobarbital

Atom Label	Experimental ^{1}H δ(ppm)	ShiftML ^{1}H δ(ppm)
1	10.3	10.49
3	8.1	8.34
7a	2.7	2.69
7b	1.7	1.63
8a-c	0.6	0.78
9-14	6.9	6.60
a = 5.08 ppm	***b = 0.78***	***RMSE = 0.33 ppm***

Thymol

Atom Label	Experimental ^{1}H δ(ppm)	ShiftML ^{1}H δ(ppm)
1	5.4	5.80
2	6.19	5.90
3	7.08	6.35
4	3.38	2.91
5-7	1.05	0.44
8-10	1.45	1.14
11-13	0.42	1.68
14	9.99	10.08
a = 4.93 ppm	***b = 0.85***	***RMSE = 0.72 ppm***

Terbutaline hemi-sulfate			
Atom Label	Experimental ^{1}H δ(ppm)	ShiftML ^{1}H 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	
a = 5.25 ppm	***b = 1.00***	***RMSE = 0.44 ppm***	

Ampicillin lattice parameters

A comparison between lattice parameters of the ampicillin crystal structure, as primitive cell, determined with XRD[248] and NMRX are given in **Table 2-20.**

Table 2-20. Comparison between ampicillin lattice parameter of the crystal structures, as primitive cell, determined with XRD[248] and NMRX.

	XRD[248]	**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^{3}]	839.494	777.213	-7.4

Positional error estimation

The positional error estimation, using DFT calculated chemical shifts, is done following the procedure described in **Chapter 2.4**. 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-PI[257] together with the Quantum ESPRESSO suite.[188] During the MD simulations the crystal structures were kept at a constant temperature using the NVT ensemble and a GLE thermostat.[258] 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 ^{1}H chemical shifts are calculated. This leads to a maximal chemical shift RMSD for ^{1}H of 1.99 ppm. **Figure 2-73** shows the correlation between positional deviations and the ^{1}H chemical shift RMSD.

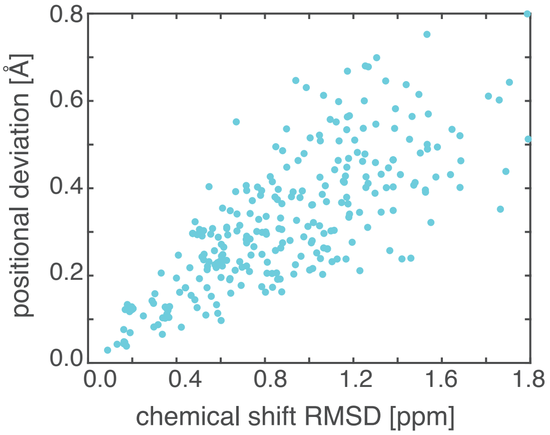


Figure 2-73. Correlation between positional RMSD (Å) and ^{1}H chemical shift RMSD (ppm) for an ensembles of perturbed crystal structures of ampicillin generated by MD. With $< r_{av} > = \sqrt{\frac{1}{N}\sum_{i,l} \Sigma_{i,l}^2} < \delta > = \bar{\Sigma} < \delta >$, we find a slope ($\bar{\Sigma} = 0.36$) for the crystal structure of ampicillin.

For the MD DFT calculations the generalized-gradient-approximation (GGA) density functional PBE[205] was used. We used the ultrasoft pseudopotentials with GIPAW[62-63] reconstruction, 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 PS library database.[242] 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.

2.7 Conclusion and Outlook

In conclusion, we demonstrated how the abundant information on the electronic structure contained in the ^1H and ^{13}C chemical shifts of a molecular solid can be used to extract structural information. Previously it had already been demonstrated how this structural information in combination with CSP protocols can be used to determine *de novo* crystal structures from powders[34-35, 38, 44, 58, 114] as well as to validate and refine crystal structures of molecular solids, or to identify known polymorphs.[8, 13, 17-18, 24, 29-31, 37, 39, 47, 56, 58, 73, 138, 140-149]

Here, we extended this CSP-NMRX approach by including structural information extracted from absent signals in 2D solid-state NMR correlation experiments. As a result, we were able to transfer the structural information extracted from solid-state NMR experiments on the crystal phase directly to the single molecule conformational search.

Additionally, we demonstrated how the ^1H and ^{13}C chemical shifts of molecular solids not only contain enough information to validate and determine *de novo* crystal structures from powders but also to quantify the positional uncertainties of these crystal structures. From this we determined that the average positional errors of crystal structures determined by NMRX are more than comparable to structures determined by single crystal XRD.

Further, we demonstrated a direct mapping between the structural information and the chemical shifts of a molecular solid, without the necessity to calculate the electronic structure. We used this direct mapping to train a ML model based on local environments to predict chemical shifts of molecular solids containing H, C, N, O and S nuclei to within current DFT accuracy. Thus, 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.

Finally, we extended the existing chemical shift based CSP-NMRX approach by including these three approaches to successfully determining the crystal structure of powdered ampicillin, which would have been very challenging for previous methods.

Note that the greatly demonstrated approaches greatly extend the scope of the existing CSP-NMRX methods, which should allow for the routine CSP-NMRX based structure determination of larger and more flexible molecules. However, at the moment, all of the proposed methods are still strongly dependent on the traditional CSP-NMRX approach and do not, on their own, present a novel method of NMRX based structure determination. However, starting from the premise, that the chemical shift information of a crystal structure is uniquely determined by its electronic structure, which is in turn uniquely determined by the crystal structure, it should theoretically be possible to determine a crystal structure using solely the information contained in the chemical shifts without the need for an elaborate CSP protocol.

One possible approach to include information from solid-state NMR experiments more actively into the structure determination would fully discard the need of a CSP structure search. In **Chapter 2.3** we have shown, that it is possible to map the chemical shift information directly onto the structural information of an atomic environment, without the need for electronic structure calculations. If it would be possible to reverse this process, meaning to map the structural information directly onto the chemical shift information, the NMRX structure determination process would be revolutionized. However, this mapping is not straightforward and poses a difficult combinatorial problem. A possible approach would be the use of deep learning methods, e.g. neural networks[101, 259-262] in combination with the SOAP fingerprints, to evaluate the difference between the input and the target output structures during training.

Chapter 3 Amorphous solids

3.1 Introduction

Amorphous and glassy solids are present in a large number of industrial devices and materials, such as in optical fibers,[263] as construction materials[264-265] or for the storage of nuclear waste.[266-267] Note that in XRD, the diffraction pattern of ideal crystals is characterized by well-defined Bragg peaks. However, the chemical and geometrical disorder, which is characteristic for amorphous materials, leads to a large bump in the diffraction pattern and thus renders these materials unsuitable for determination with X-ray or neutron diffraction spectroscopy (**Figure 3-1**). In contrast, solid-state NMR directly probes the local atomic environment (see **Chapter 1.3**) thus making it one of the most powerful tools for the structural characterization of amorphous and glassy solids. [45, 49, 268-270]

However, the structural disorder present in amorphous materials leads to a distribution in both the isotropic and the anisotropic parts of the observable NMR parameters. This generally makes it very challenging to extract the structural information present in the NMR spectra. Therefore, it is crucial to develop additional computational approaches to complement the experimental measurements. **Figure 3-1** shows the difference between the ^{29}Si MAS NMR spectra of an amorphous and a crystalline lithium silicate. Note that in contrast to XRD, detailed structural information can still be extracted from the amorphous NMR spectra.

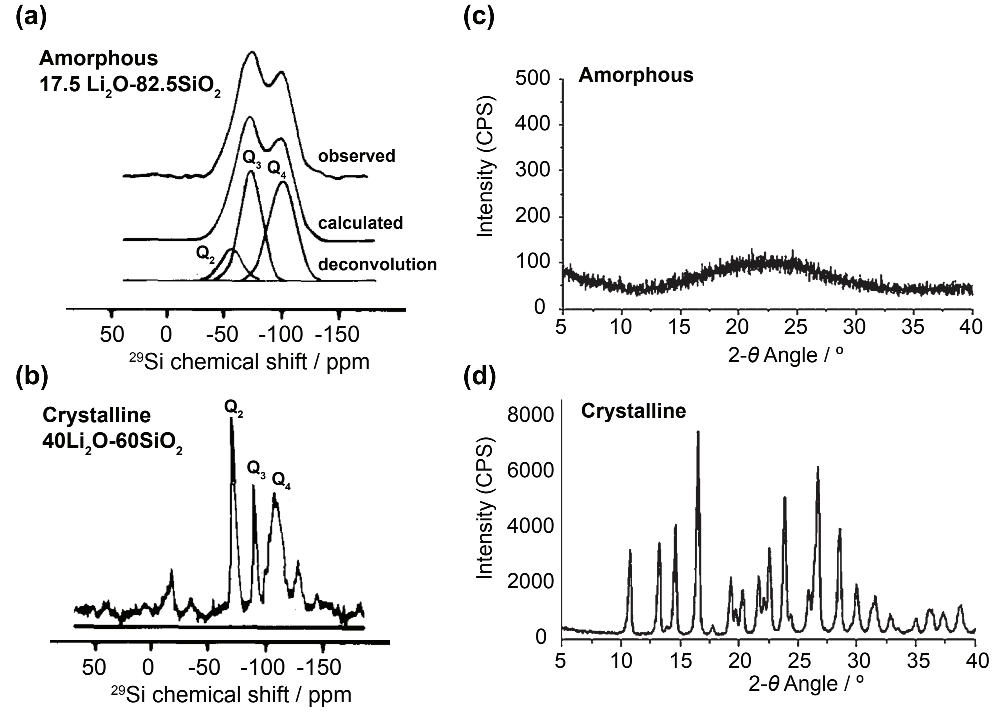


Figure 3-1. Examples of amorphous and crystalline NMR spectra **(a-b)** and powder XRD patterns **(c-d)**. ^{29}Si MAS NMR spectra of an amorphous lithium silicate glass **(a)** and of a crystalline lithium silicate **(b)**. Both spectra were adapted with permission from De Jong *et al.*[271] (copyright 1984 American Chemical Society) Powder XRD patterns of amorphous **(c)** and crystalline **(d)** griseofulvin samples. Both patterns were adapted with permission from Feng *et al.*[272] (copyright 2008 Elsevier)

Analog to the method presented in **Chapter 2** for microcrystalline powders, a general approach to get a better understanding of the local structure of amorphous material is to compare the experimental NMR parameters with calculated NMR parameters generated for a theoretical model of the structure.[254, 273-274] Often this is combined with a MD simulation in order to represent the large amount of structural disorder present in amorphous materials. This has been successfully demonstrated for several amorphous systems, such as phosphate glasses,[42, 275] chalcogenide glasses,[120, 276-277] silicate and aluminosilicate glasses[40, 45, 49, 115-119, 121, 278-279] and proton-conducting polymers.[280-281]

The number of atoms required to quantitatively represent the statistical distribution of disorder and chemical environments present in an amorphous solid is large. For example, for bulk silica it has been shown that a unit-cell of around 3000 atoms is needed to generate a realistic model of the atomic structure.[282] As a consequence, MD simulations modelling amorphous systems have generally been performed using several hundreds or thousands of atoms.[283-285] However, the calculation of sufficiently accurate NMR parameters, as required for structural characterization, currently relies on DFT and can only handle a limited number of atoms. Therefore, the application of such combined methods to amorphous and glassy solids is limited to small systems containing only a few hundred atoms [40, 115-121]. Thus, structural characterization using NMR crystallography, as described in **Chapter 2** for microcrystalline solids, is not directly applicable to amorphous systems.

In **Chapter 3.2** we present a combined approach for the determination of the atomic-level structure of amorphous calcium silicate hydrate (C-S-H) based on local structural motifs. First, we characterize the composition and the uniformity of the material using various spectroscopic methods. In a second step, we constrain the atomic environments present in the structure using multi-dimensional ^{1}H and ^{29}Si solid-state NMR experiments. Next, we systematically generate well-defined structural motifs in agreement with the experimental constraints. These motifs are then assessed by comparing their calculated ^{1}H and ^{29}Si chemical shifts to experiment. We then use the accepted structural motifs as building blocks to generate an atomic-level structural model of amorphous C-S-H. Finally, the stability of the proposed structural model is verified using MD simulations.

3.2 The atomic-level structure of cementitious calcium silicate hydrate

This chapter has been adapted with permission from: Kumar, A.; Walder, B. J.; Kunhi Mohamed, A.; Hofstetter, A.; Srinivasan, B.; Rossini, A. J.; Scrivener, K.; Emsley, L.; Bowen, P., "The atomic-level structure of cementitious calcium silicate hydrate". *The Journal of Physical Chemistry C* **2017,** *121* (32), 17188-17196. ***(post-print)***

3.2.1 Introduction

Calcium silicate hydrate (C-S-H) is the primary binding component of concrete, forming about 50-60% by volume of hardened cement paste and making it one of the most common substances of the modern world. Because of its ubiquity, it is surprising that a complete description of its atomic-level structure remains the subject of debate,[286-287] and consequently its structure-property relationships are not well known. This makes it difficult to engineer C-S-H not only for its primary uses in construction, in which high reactivity and strength at low carbon footprints are desirable, but also for emerging applications such as dental filling and bone repair,[288-289] which require biocompatibility; waste water treatment,[290-291] which requires high specific surface areas; and encasement of nuclear waste,[292] which requires high structural integrity in the presence of significant radionuclide concentrations.

For Portland cements the precipitation of C-S-H occurs in conjunction with the precipitation of other material phases such as crystalline $Ca(OH)_2$, ettringite, and $CaCO_3$.[264-265] The C-S-H phases are known to be rich in calcium, with Ca:Si ratios exceeding 1.75 at early stages of hardening.[293] In contrast, synthetic C-S-H with Ca:Si ratios above ~1.5 are often observed in coexistence with a $Ca(OH)_2$ phase. Because of an inability to synthesize pure C-S-H with Ca:Si ratios above 1.5, many researchers believe that Ca-rich C-S-H systems are intrinsically a binary mixture of a chemically disordered single phase C-S-H material. In such a case, one phase consists of a "proper" C-S-H phase, with a layered silicate chain structure related to that of the naturally occurring calcium silicate hydrate mineral tobermorite and limited to Ca:Si ratios around 1.6. The other phase consists of nanocrystalline $Ca(OH)_2$, which is thought to occur in bulk form occupying pores in the proper C-S-H phase or as chemically distinct ribbons or sheets interwoven within the C-S-H structure itself.[286, 294-296] This interpretation has the support of thermodynamic and solubility data analyzing a multitude of C-S-H systems.[297] Furthermore, in spite of a vast amount of experimental data yielding partial characterization, the positions of the calcium atoms in the interlayer, which are the essential aspects of high Ca:Si ratios in C-S-H, remain undefined. Thermodynamic modeling and crystal chemical reasoning have been applied to propose complete C-S-H structural models at Ca:Si ratios greater than 1.5,[286] but for these compositions the focus has been on the binary C-S-H/$Ca(OH)_2$ representation, for which experimental validation is ongoing.[298]

Here, we introduce a method which achieves the synthesis of C-S-H possessing Ca:Si ratios between 1.0 and 2.0, maintaining a single phase composition even for C-S-H whose Ca:Si ratio exceeds 1.6. Aqueous calcium nitrate and sodium silicate solutions are reacted under conditions of high supersaturation and constant pH, the latter of which is set by the addition of a predetermined amount of alkali hydroxide. The production of a single phase composition at such Ca:Si ratios has not been achieved using conventional methods for C-S-H synthesis[299-303] [18–22] relying on combinations of dissolution and direct precipitation[287, 304-305] reactions that operate at either lower supersaturation or uncontrolled pH conditions. We also use ^{1}H-^{29}Si cross-polarization (CP) MAS NMR to measure populations of Q species, the connectivity between those species, and correlations between ^{29}Si and ^{1}H chemical shifts of the single-phase C-S-H produced using our rapid precipitation method. The greatest drawback of ^{29}Si solid-state NMR is its low sensitivity, which we circumvent by using modern dynamic nuclear polarization (DNP) strategies[306-308] that have been recently used to study the hydration of cementitious systems with tremendous success.[309] The Q species information allows us to quantify the extent of silicate polymerization in the structure. Finally, we use atomistic modeling to establish a connection between the measured ^{1}H chemical shifts and the atomic-level position of calcium atoms in the interlayer, allowing us to solve the three-dimensional atomic-level structure of synthetic cementitious C-S-H.

3.2.2 Methods

Synthesis

The pH governs the type of silicates species available for precipitation of C-S-H. The Ca:Si ratio attained in the solid phase was found to depend on the pH of the solution. Thermodynamic modeling[310-311] also predicts that Ca:Si ratios above 1.5 can only be produced under high pH conditions, as occurs in the hydration of real Portland cement systems, in order to ensure that the electrostatically stable monomeric $SiO_2(OH)_2^{2-}$ species remains in abundance at high supersaturation and rapid precipitation conditions.

To maintain the desired supersaturation, pH, and mixing conditions, and to avoid carbonation, we developed a synthetic apparatus for controlling the reaction conditions to the degree of precision required, aided by real-time acquisition of kinetic data such as Ca^{2+} ion concentration, pH and conductivity. Details regarding its construction are given in the **Appendix IV**.

All reaction solutions were prepared in decarbonized, demineralized ultrapure water. The reaction chamber was kept under an inert nitrogen atmosphere in order to prevent carbonation. C-S-H precipitates were collected after a duration of 3 hours and again after 24 hours. The products were separated from mother liquor using vacuum filtration over a 20 nm organic filter and later washed with ethanol and water to remove salts and unwanted ions from the surfaces of C-S-H. We produced five different C-S-H powders with nominal Ca:Si ratios of 1.0, 1.25, 1.5, 1.75 and 2.0. The precise experimental conditions for the precipitation of the different stoichiometry of the C-S-H were determined using thermodynamic modelling,[310-312] with the exclusion of calcium hydroxide, as there was no experimental evidence for its formation. Additional details are given in the **Appendix IV**.

Dynamic nuclear polarization

DNP solid-state NMR experiments were carried out on the aqueous suspensions of freshly prepared C-S-H nanoparticles with added impregnation agent and were not dried. The impregnation agent used was 22 mM AMUPol in 65:35 v:v d_8-glycerol:D_2O, which was purged of dissolved oxygen by bubbling with N_2 gas for roughly five minutes. The addition of the radical polarizing agent further dilutes the samples by about 20%, but simple drying steps to increase the concentration of C-S-H led to sample deterioration (see **Appendix IV**). About 25 mg of the impregnated gels were worked into a 3.2 mm OD sapphire rotor and plugged with a PTFE insert. The drive caps were zirconia. The DNP enhanced NMR experiments were carried out at a nominal field strength of 9.4 T using a commercial Bruker AV I 400 MHz/263 GHz DNP NMR spectrometer.[313] The samples were rapidly transferred into the stator of the NMR probe which was pre-cooled to 100 K to promote glass formation. Proton DNP enhancements were found to exceed 35 for all samples.

High resolution electron microscopy (HRSEM)

HRSEM micrographs were obtained by coating the samples with 6 nm of osmium (gas phase coating). The metallization reduces charging and provides enhanced image contrast. High resolution SEM analysis was performed on a Zeiss Merlin, equipped with the GEMINI II column which combines ultra-fast analytics with high resolution imaging using advanced detection modes. Osmium coated samples were analyzed with acceleration voltage of 1 kV with probing current of 300 nA. On-axis in-lens secondary electron detection mode was employed for imaging. The instrument provides up to 0.6 nm resolution in STEM mode. In TEM mode, the samples were imaged at room temperature using a Tecnai F20 (FEI, The Netherlands) operating at an acceleration voltage of 100kV LaB_6 gun with a line resolution of 0.34 nm, with images being recorded on a high sensitivity 4k x 4k pixel CCD camera. For SEM and TEM analysis, 50 mg of sample was dispersed in 40 mL of isopropanol. A drop of the suspended liquid was allowed to dry on a copper grid (200 mesh grids). The copper grids were glow discharged prior to sample disposition.

Fourier-transform infrared spectroscopy (FTIR)

Freshly prepared samples were analyzed with a PerkinElmer FTIR spectrometer, with a resolution of 0.5 cm^{-1} to 64 cm^{-1}. Wavelength accuracy was about 0.1 cm^{-1} at 1600 cm^{-1}. FTIR measurements were performed with an attenuated total reflectance (ATR) unit and data was recoded and processed using Spectrum One software. The ATR unit included a diamond crystal and a clamp for pressing solid materials onto the crystal with constant pressure. The transmittance results of 256 scans were recorded between 4000 and 450 cm^{-1}, with individual measurements taken every 2 cm^{-1}. For the solid gels, air was used as the background.

Raman spectroscopy

Non-invasive Raman microscopy was carried out using a Renishaw inVia Reflex spectrometer equipped with a 785 nm diode laser. The power delivered to the sample was 164 mW at a full power specification. The grating size was 1200 lines/mm with an edge filter for Rayleigh rejection. $Ca(OH)_2$ and $CaCO_3$ standards were measured at 5% power with a single 10 s accumulation period. Freshly prepared C-S-H was measured with multiple accumulation periods, each of 13 s exposure.

Molecular Dynamics (MD) simulations

Classical MD simulation with force field potentials were used to test the structural stability of the proposed structures. The force field parameters used are known to describe well cementitious material systems.[314] Simulations were done in a constant pressure ensemble at 300 K and a time step of 0.7 fs using Velocity Verlet integration algorithms implemented in DLPOLY.[315] Ewald summation was used to take into account the long range forces above a cutoff distance of 8.5 Å.

NMR chemical shift calculations

Atomic positions and unit cell parameters were optimized as described in **Appendix IV**. The chemical shielding σ_{calc} was calculated using the generalized gradient approximation (GGA) functional PBE[205] within the Quantum Espresso code[188] and the GIPAW method.[63] In every calculation a plane-wave maximum cutoff energy of 80 Ry, and a Monkhorst-Pack grid of k-points[220] corresponding to 0.03 $Å^{-1}$ - 0.04 $Å^{-1}$ in reciprocal space was employed. The chemical shielding was converted into calculated chemical shifts δ_{calc} by the relation $\delta_{\mathrm{calc}} = \sigma_{\mathrm{ref}} - \sigma_{\mathrm{calc}}$, with the value of σ_{ref} determined by a linear regression between the calculated and experimental values for the calcium hydroxide structure (^1H chemical shifts) and the unperturbed tobermorite structure[109] (^{29}Si chemical shifts).

3.2.3 Results and Discussion

Morphology

Two typical morphologies were seen by electron microscopy: “nanoglobules”, for the Ca:Si ratio of 1.00; and “nanofoils”, for Ca:Si ratios ≥ 1.25, which is the morphology shown in **Figure 3-2a**. The foil morphology is very similar to morphologies for C-S-H seen in Portland cement systems with high alkaline contents.[310, 316] Thicknesses of the foil-like structures are generally between 6 nm and 10 nm. The pure phase C-S-H systems were all shown by high-resolution analytical transmission electron microscopy (TEM) to be uniform for Ca:Si ratios between 1.0 and 2.0 at less than a 9 nm^2 pixel size. This is also supported by X-ray diffraction (XRD) and scanning TEM with energy dispersive X-ray analysis (STEM-EDX), as described in **Appendix IV**. No secondary phases such as $Ca(OH)_2$ were detected by IR or thermogravimetric analysis (TGA), as shown in **Figure 3-2b**; however, long exposure of C-S-H sample to open air (for example in TGA or XRD analysis) does eventually lead to the formation of $CaCO_3$. This phenomenon manifests well in the Raman spectra of **Figure 3-2c**, showing that $CaCO_3$ forms during prolonged measurements in air, whereas the signature of $Ca(OH)_2$ is never observed regardless of measurement duration. ζ-potential measurements on the show negative potential surfaces indicating that calcium does not reside at the surface but is incorporated into the particles.

Characterization by DNP NMR

C-S-H is a poorly ordered material, making atomic level structural determination using conventional X-ray and neutron diffraction methods challenging, especially for non-dried samples. Solid-state magic-angle spinning (MAS) NMR is a powerful method for studying disordered systems, and has been extensively used to study the molecular structure of C-S-H and related mineral phases.[317] Previous ^{29}Si MAS NMR[19, 52, 109, 318-320] and diffraction studies, often on dried materials, have established that the silicate chains in C-S-H are arranged according to the “dreierketten” model,[52, 319, 321-322] which specifies a repeating unit for the chains comprised of a bridging-type $Q^{(2b)}$ silicate tetrahedron flanked by pairing-type $Q^{(2p)}$ silicate tetrahedrons, highlighted in the tobermorite structure shown in **Figure 3-2d**. The silicate chains are flanked by a calcium oxide layer and a hydrous interlayer. Each silicate tetrahedron shares two O atoms with other silicate tetrahedrons and on this basis are both classified as $Q^{(2)}$ species. The pairing-type $Q^{(2p)}$ species direct the other two O atoms toward the main calcium layer whereas the bridging-type $Q^{(2b)}$ species direct them toward the hydrous interlayer. Defects occur through the removal of a $Q^{(2b)}$ SiO_2 unit, breaking up the idealized infinite silicate chains of tobermorite into finite segments consisting of (3n+2) silicate tetrahedrons, as illustrated in **Figure 3-2e**. The segments are terminated by $Q^{(1)}$ silicate species.

The interlayer calcium and water present in the original 14 Å tobermorite are Ca_I and H_2O respectively whereas the Ca_B, Ca_A and OH_I are only present in the defective structures. Ca_B sites replace bridging silicate tetrahedrons, Ca_A sites are additional calcium atoms in the interlayer, and OH_I are additional hydroxyl groups in the interlayer to charge compensate the additional Ca ions needed to reach high Ca:Si ratios. Silicate dimers (n = 0) have been observed by ^{29}Si-^{29}Si correlation NMR experiments to be the dominant species for systems with Ca:Si ~ 1.5, both for synthetic C-S-H systems and during the initial formation of C-S-H in hydrating tricalcium silicate.[52, 323]

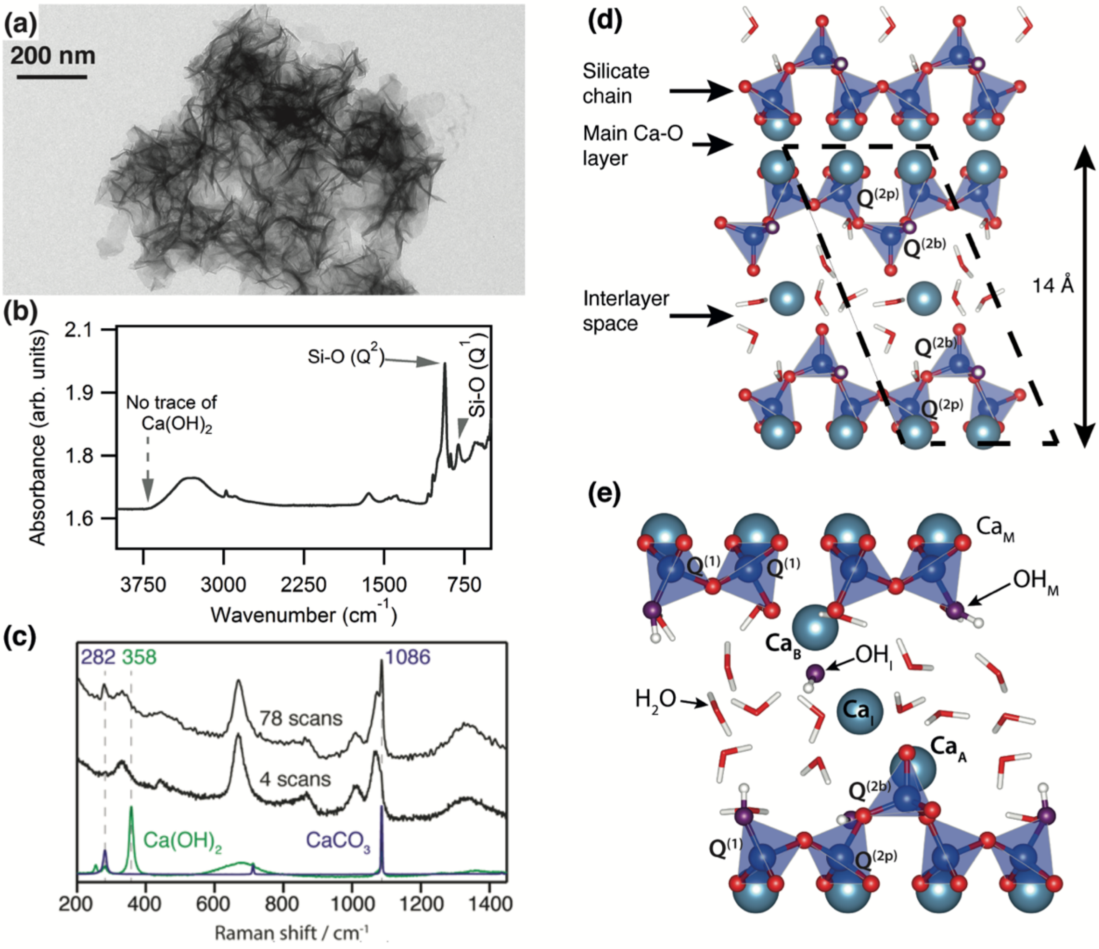


Figure 3-2. Structural elements of C-S-H. **(a)** High-resolution TEM image of pure C-S-H with Ca:Si ratio of 2.00, showing its "nanofoil" morphology. **(b)** Fourier transform IR spectroscopy showed no evidence of phases other than the C-S-H, including $Ca(OH)_2$. **(c)** Comparison of Raman spectra of $Ca(OH)_2$ (green), $CaCO_3$ (blue), a sample of C-S-H with Ca:Si = 2.0 after 4 scans (lower black), and a sample of C-S-H with Ca:Si = 2.0 after 78 scans (upper black). **(d)** Chain topology in the layered 14 Å tobermorite (Ca:Si = 0.83). **(d)** Defective and short dreierketten chains in C-S-H, showing two dimers (*n* = 0) and one pentamer (*n* = 1).

To overcome the low sensitivity of ^{29}Si MAS NMR at natural isotopic abundance we use modern DNP strategies.[306-308] DNP is based on the transfer of large unpaired electron spin polarization to nearby protons by saturation of the electron spin transitions with microwaves, followed by CP transfer of the enhanced polarization to the ^{29}Si nuclei. The electron polarization is provided here by the organic biradical AMUPol[324] that is added to the wet C-S-H as a minimal amount of d_8-glycerol/D_2O solution before the NMR sample is rapidly cooled to 100 K for the experiments.[308, 325-327] The cryogenic temperatures are required to maximize the sensitivity enhancements by DNP, but are also important here to quench proton exchange and prevent the C-S-H from degrading during the experiments. Efficient DNP occurs only for those parts of the sample that have successfully passed through the glass transition. We also note that pore water is susceptible to glass formation when rapidly inserted into the pre-cooled NMR probe even without the addition of a glassing agent such as glycerol.[328] We therefore do not expect the C-S-H structure to be disrupted by our experimental conditions; furthermore, even if pore water does crystallize in parts of the sample, inefficient DNP will suppress the NMR signal from these regions.

The polarizing agent contains labile deuterons, which can lead to the formation of calcium silicate deuterate through isotope exchange. At most, 40 mol% of labile hydrogen in the impregnated C-S-H gels (C-S-H hydrogen, D_2O, and the -OD groups of the d_8-glycerol) are deuterons given our DNP sample formulation and estimated C-S-H composition. If a reasonable allowance for excess pore and adsorbed water is made, this falls to about 25%. In fact, this upper limit is almost certainly never reached. Small-angle neutron scattering studies have shown that deuteron exchange into the gel is a diffusion driven process providing full isotope exchange on the time scale of tens of hours.[329] Since the impregnated sample never spent more than 1.25 h, and usually just 0.25 h, at room temperature prior to experiments, we expect the highest degree of partial deuteration to be surface based and the NMR signal should be representative of fully protonated bulk C-S-H. Moreover, there is little in the way of evidence in the small-angle neutron scattering literature to suggest that isotope exchange modifies C-S-H in any structurally significant way.

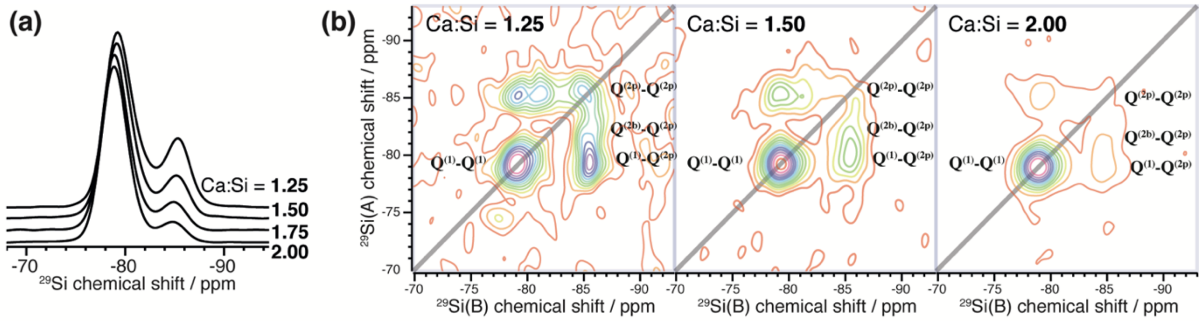


Figure 3-3. One- and two- dimensional DNP enhanced ^{29}Si CP MAS spectra of C-S-H samples for quantification of silicate chain distributions. **(a)** 1D spectra across the compositional series. **(b)** Experimental 2D refocused INADEQUATE spectra for three of the C-S-H compositions studied (the spectra have been sheared to produce a COSY-like representation). Contours are drawn in 10% intervals beginning at 5% of the maximum signal intensity.

One-dimensional ^{1}H-^{29}Si DNP CP echo spectra for the five compositions are shown in **Figure 3-3a**. With the exception of the Ca:Si = 1.00 composition, good fits to the line shapes are obtained by modeling each of the constituent Q sites as a Gaussian function, whose amplitudes are used to determine the relative populations of the Q species. Relative signal intensities in DNP enhanced CP MAS experiments are not usually in proportion to the relative populations of the nuclei generating the signal as they often are in experiments using direct excitation without hyperpolarization unless we assume that 1) the length scale of hyperpolarization non-uniformity is larger than the unit cell of the particle, and 2) cross-polarization kinetics can be measured and used to adjust the signal intensities appropriately.

The C-S-H particles are sufficiently small and have a proton density sufficient for nearly uniform polarization of the particles over the recycle period. To the second point, we performed cross-polarization measurements for different values of the cross-polarization contact time. This data was fit to a simple IS model of CP kinetics for each site[330]. A detailed description of the fitting procedure and the Q populations determined by this method are given in **Appendix IV**. We note here that the failure of the Ca:Si = 1.00 composition to fit well to the three-Gaussian model suggests a different molecular structure.

The ^{29}Si-^{29}Si connectivity is measured using 2D refocused INADEQUATE experiments,[331] whose application to cementitious systems has hitherto not been feasible without isotopic enrichment.[52, 323] In the ^{29}Si-^{29}Si INADEQUATE spectrum only signals from covalently bonded ^{29}Si – O – ^{29}Si pairs are retained. For linear silicate chains at natural isotopic abundance, these constitute at most 0.5% of all Si – O – Si pairs. The improvement in NMR sensitivity provided by DNP makes it possible to obtain such spectra,[27] as shown in **Figure 3-3b**. Autocorrelation peaks corresponding to $Q^{(1)}$-$Q^{(1)}$ dimer and $Q^{(2p)}$-$Q^{(2p)}$ extender units are observed, but peaks corresponding to $Q^{(2b)}$-$Q^{(2b)}$ are always absent, consistent with the dreierketten model. Remarkably, the usually dominant $Q^{(1)}$-$Q^{(1)}$ autocorrelation peak is entirely absent for the Ca:Si = 1.00 composition (see **Appendix IV**) suggesting that this composition does not contain silicate dimers. Cross peaks from all three Q sites to $Q^{(2p)}$ are also observed. Using the chemical shift constraints from the deconvolution of the 1D CP echo spectra, the INADEQUATE spectra are decomposed using 2D Gaussian line shapes to model each of the six possible correlation peaks. This line shape generates reasonably good fits (see **Appendix IV**), suggesting that the chemical disorder is very local. The 2D peak intensities are fit simultaneously across the four compositions for a conditional probability P(A|B) that Q site A is connected to Q site B.

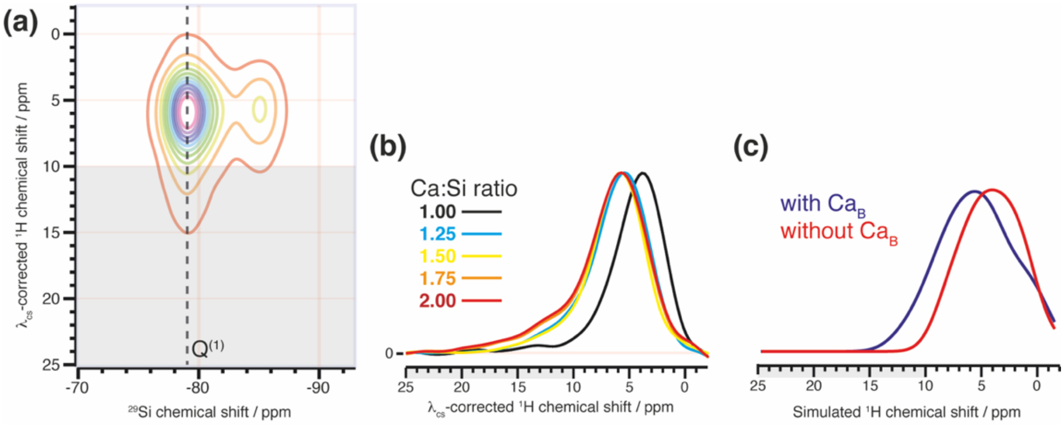


Figure 3-4. DNP enhanced 2D ^1H-^{29}Si HETCOR correlating ^1H spectra to specific Si sites. **(a)** The 2D correlation spectrum for the Ca:Si = 1.50 composition acquired with a 7 ms CP contact time. **(b)** 1D cross sections parallel to the ^1H dimension extracted at the position of the dashed line in the 2D spectrum, representing ^1H spectra correlated to $Q^{(1)}$. **(c)** Simulated ^1H chemical shift spectra aggregated over C-S-H substructures that either possess (blue) or lack (red) the bridging calcium site Ca_B. The intensity of these spectra is normalized with respect to the maximum of the $Q^{(1)}$ peak. The region downfield of 10 ppm is shaded to indicate the domain of strongly hydrogen bonded species.

2D ^1H-^{29}Si HETCOR experiments were used to correlate ^1H chemical shifts with the ^{29}Si chemical shifts. Measurements were made using CP contact times of 0.7 ms and 7 ms for each sample. The use of a short contact time biases the contribution to the NMR signal from those protons that are close to the correlating ^{29}Si nuclei, as compared to longer range correlations observed in the long contact time experiment, which samples proton environments out to ~1 nm.

The line shape in the 2D ^1H-^{29}Si HETCOR spectrum shown in **Figure 3-4a** is dominated by inhomogeneous broadening resulting from chemical disorder, which prevents an accurate line shape deconvolution on the basis of proton site. Cross sections of these spectra yield ^1H chemical shift spectra correlated to specific Q sites, as shown in **Figure 3-4b** for the $Q^{(1)}$ correlation and in the **Appendix IV** for the others. We find that the intensity of the of the $Q^{(1)}$ site relative to the $Q^{(2)}$ sites is greater at shorter contact time, implying that $Q^{(1)}$ species are located in a relatively hydrogen rich environment. We also see that the ^1H chemical shift profiles for the Ca:Si ≥ 1.25 ratios possess a significant contribution above 10 ppm, indicative of strong hydrogen bonding.[332] A comparison to HETCOR spectra taken at short contact time (see **Appendix IV**) reveals that the prominence of the downfield region for the $Q^{(1)}$ correlated cross sections increases significantly at short contact time, a feature which is not shared by the $Q^{(2b)}$ and $Q^{(2p)}$ cross sections. This suggests that the strong hydrogen bonding occurs primarily in association with $Q^{(1)}$ sites. We note that the signature of strong hydrogen bonding is almost entirely absent from the HETCOR spectrum of the Ca:Si = 1.00 composition, once again producing a spectrum deviating substantially from its relatively calcium rich counterparts.

The line shapes lack any significant features near 2 ppm, where basic hydroxide protons would be prominent, suggesting any secondary amorphous or crystalline $Ca(OH)_2$ phase, if present, is not intimately mixed with the C-S-H structure. Such a signal was previously reported for C-S-H compositions with Ca:Si ratios up to 1.5.[52, 323] It may be that the C-S-H/$Ca(OH)_2$ nanocomposite results from excessive drying and aging of the sample. Indeed, a recent high energy X-ray study lending support for a secondary phase of $Ca(OH)_2$ nanosheets interwoven into the C-S-H interlayer suggests that the $Ca(OH)_2$ phase grows as C-S-H ages.[298]

Structural determination

It is known that C-S-H resembles a defective tobermorite.[319, 333] In contrast to previous structural modeling studies for C-S-H, which consider random defects in tobermorite systems containing hundreds of atoms,[319, 334] we adopt a methodology that focuses on the systematic creation of structurally well-defined defects. The defective substructures are then used as building blocks to represent C-S-H at higher Ca:Si ratios.

A suitable base structure is required to begin. Tobermorite structures are generally named after their characteristic interlayer distances; namely, 9 Å, 11 Å, or 14 Å tobermorite.[334-336]The choice of base structure for modeling depends the Ca:Si ratio[337] and drying conditions.[286, 296] A dataset compiled by Richardson[286] shows that the interlayer distance in C-S-H decreases from ~13-14 Å at Ca:Si = 0.8 to ~10 Å at Ca:Si = 1.5. Recently, Roosz *et al.*[338] have shown that sample preparation and relative humidity significantly affect the interlayer distance measurement. The interlayer distance measured for a C-S-H of Ca:Si = 1.2 using XRD in dry and fully hydrated states were 9.5 and 12.3 Å, respectively. Since our samples are hydrated, we choose 14 Å tobermorite (**Figure 3-2d**) as the base motif for constructing our atomic-level model of C-S-H.

Table 3-1. Dimer mole fraction x_0 and mean repeat index for the four compositions analyzed.

Sample	x_0	$\sum_{n=0} x_n n$
Ca:Si = 1.25	0.751	0.450
Ca:Si = 1.50	0.816	0.285
Ca:Si = 1.75	0.873	0.185
Ca:Si = 2.00	0.900	0.136

A defect is introduced by the removal of an SiO_2 unit from a $Q^{(2b)}$ unit. The extent to which we need to create defects is determined by the distribution of silicate chain lengths. With the Q species populations and connectivities we can determine the distribution of chain lengths for each composition, as described in the **Appendix IV** and given in **Table 3-1**, to find

$$\sum_{n=0} x_n n = \frac{P(\mathrm{Q}^{(2p)})}{P(\mathrm{Q}^{(1)})},$$

(3-1)

where x_n is the mole fraction of dreierketten chain species with repeat index n, and

$$x_0 = P\left(\mathrm{Q}^{(1)} \middle| \mathrm{Q}^{(1)}\right),$$

(3-2)

as the mole fraction of dimers. The quantitative NMR results thereby provide three independent constraints for calculating the distribution of silicate chains for each C-S-H composition. Using these constraints, we adopt a Monte Carlo method to predict the mole fraction distribution for chains up to $n = 10$, which we report in the **Appendix IV** for each composition.

Defect creation transforms the silicate tetrahedrons adjacent to the removed $Q^{(2b)}$ site into $Q^{(1)}$ sites, requiring the addition of H^+ and $CaOH^+$ to satisfy requirements of local charge balance. Additional molecular units of H_2O and $Ca(OH)_2$ can also be incorporated into the structure. The defective motif is deemed acceptable if correct atomic bond distances, coordination numbers, and local charge balance remain satisfied after structural relaxation using density functional theory (DFT), leading to a series of substructures which are classified based on defect geometry. Reduced unit cells are constructed by connecting the defect units through an aqueous interlayer or an aqueous interlayer with a Ca_I and additional OH^- for charge balance. To study medium range effects, we also consider different ways to combine the reduced unit cells, resulting in chain, dimer, and pentamer motifs.

We study the effect of these different defect structures on the ^1H chemical shifts. A set of reduced unit cells are chosen to ensure a wide variety of different local defect environments as represented by the defect classification scheme described in **Appendix IV**. In **Figure 3-4c**, we show two calculated ^1H chemical shift spectra composed by summing over substructures that either possess or lack Ca_B. In comparison with the experimental ^1H spectra in **Figure 3-4b** the calculated spectra suggest that Ca_B is responsible for generating ^1H NMR signals downfield of 10 ppm. Furthermore, the association between downfield shifted protons and hydrogen bonding leads us to infer that bridging calcium holds terminating chains together by coordinating to the defect site and promoting the formation of strong hydrogen bonds. On this basis we might also conjecture that bridging calcium is preferentially associated with silicate dimers, as suggested by the fact that both strong hydrogen bonds and dimers are lost when crossing under to the Ca:Si = 1.00 composition, though without further evidence this remains speculative.

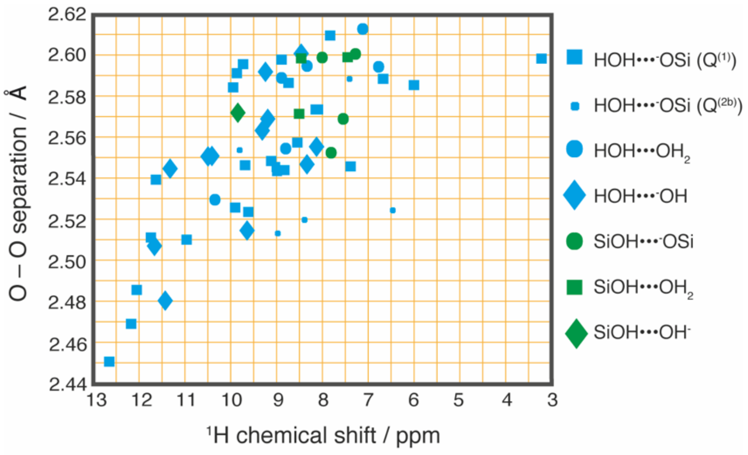


Figure 3-5. *Scatter plot showing the correlation between the O–O distances and the chemical shifts of protons participating in the different types of hydroxyl-oxygen interactions occurring in the C-S-H substructures.*

The proton chemical shift calculations provide additional structural insight regarding the nature of the hydrogen bonding interactions. As **Figure 3-5** shows, there is a linear correlation between the calculated ^{1}H chemical shift and the O – O separation of the species engaged in electrostatic hydrogen-oxygen interactions, a well-established trend for inorganic oxide systems.[60] In particular, we observe that interlayer water protons that interact with interlayer hydroxide ions and the oxygen atoms of Q$^{(1)}$ sites dominate in their contribution to the ^{1}H chemical shift signal above 10 ppm. The key observation here is that each of these types of protons are located within 3 to 4 Å of Ca$_B$. Furthermore, we may consider that the protons involved in hydrogen bonding between interlayer water and a Q$^{(1)}$ oxygen atom are less than a 3 Å from the Q$^{(1)}$ silicon atom and are therefore favored in the HETCOR experiments at short contact time. For only two of the substructures analyzed, one of which lacks Ca$_B$ entirely, the proton from the strongest OH$_2$–OH$^-$ group is located greater than 5 Å away from a Q$^{(1)}$. We infer that it is these types of protons which explains the prominence of the region downfield of 10 ppm in the Q$^{(1)}$ correlated proton spectrum, and that their association with bridging calcium in the structures that we have analyzed strengthens the confidence of our association.

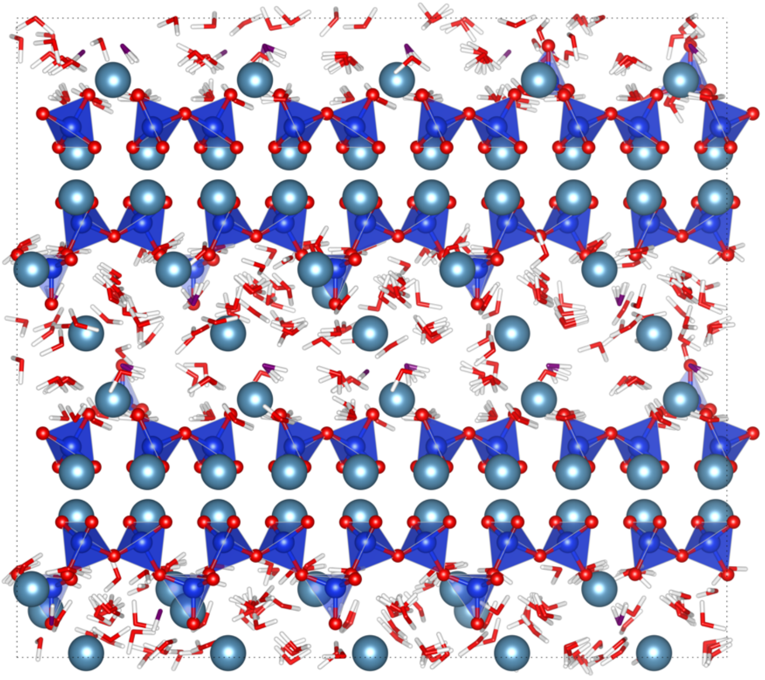


Figure 3-6. *The structure determined here of C-S-H for a Ca:Si ratio of 1.5, viewed along the [A] axis. The relative proportions of dimers, pentamers, octamers, undecamers, and tetradecamers are 81%, 14%, 3% 1%, and 1%, respectively. The chemical composition of this structure is* $Ca_{1.5}SiO_{3.35}(OH)_{0.3}$•$2H_2O$. *The relative positions of hydroxyls and water molecules have been relaxed keeping all other atoms frozen for ease of visualization.*

Construction of structures that are representative of C-S-H proceeds by drawing from these defective substructures and the defect-free motif and tessellating them in a way that satisfies both the constraints of stoichiometry and the chain distribution determined by the ^{29}Si NMR results. High Ca:Si ratios are obtained by deprotonation of a $Q^{(2b)}$ silanol and adding $CaOH^+$ and $Ca(OH)_2$ in the form of Ca_A to the interlayer (**Figure 3-2e**). Our representative C-S-H unit cell is a tessellation of sixty such substructures coming to roughly 3 nm on each side, consistent with the degree of uniformity found by high-resolution analytical TEM. One such bulk C-S-H structure permitted by the ensemble of experimental NMR constraints determined for the Ca:Si ratio of 1.50 is shown in **Figure 3-6**. A 2 ns MD simulation at constant pressure and temperature (300 K) shows that the resulting structures are stable, with realistic bond lengths and coordination geometries predicted. The C-S-H structures we propose for each of the four compositions are given in **Appendix IV**. Unlike previously proposed structures based upon defective tobermorite,[286, 319, 339-340]our computational methodology specifies unambiguously the positions and coordination of calcium in the interlayer, rather than leaving them undefined or relegating its existence to a second phase, as in the tobermorite/$Ca(OH)_2$ model. We do not claim that these structures represent the most energetically stable configurations; rather, we locate a viable, locally minimized configuration satisfying the NMR constraints. The proposed bulk structures are representative of a series of similar structures with similar defect concentrations and slightly different atomic arrangements. This should not change the average properties, but does explain why there is very little structural order seen in X-ray powder diffraction of non-dried C-S-H.

3.2.4 Conclusion

We introduce a new synthetic method for C-S-H which controls pH throughout the process, and we produced uniform C-S-H with controlled Ca:Si ratios up to 2.0 for the first time. High sensitivity DNP solid-state NMR techniques have been used to characterize unique highly uniform synthetic C-S-H particles with high Ca:Si ratios. In conjunction with atomistic scale modeling, atomic-level structures of defective tobermorite coherent over Ca:Si ratios from 1.25 to 2.00 have been determined without invoking secondary phases or glassy structures as confirmed by the clear absence of a signal from basic Ca-OH units in the 2D ^{1}H-^{29}Si HETCOR experiments. To interpret this data, we developed a computational approach which explores defective tobermorite sub-structural candidates, combining them in a manner satisfying our experimental constraints to build a full 3D structure which provides an accurate representation of structural and chemical environments in C-S-H for Ca:Si ratios up to 2.0. essential aspect of these structures is the inclusion of a calcium site in the interlayer which bridges chain terminating silicate $Q^{(1)}$ sites. This site is associated with an environment of strong hydrogen bonding which stabilizes the structure and, consequently, promotes high Ca:Si ratios in C-S-H. This thus establishes a clear relation between the atomic-level defect structure and the high Ca:Si ratio in C-S-H. This knowledge of the defect structure is a prerequisite for overcoming the self-limiting growth of C-S-H and to better understand growth mechanisms and kinetics. Such knowledge can further help formulate new classes of sustainable cements capable of exhibiting strong chain-bridging hydrogen bonding features while ensuring the early age strength development of the material.

3.2.5 Appendix IV

Supporting Analysis

XRD. X-ray diffraction data was collected with a Bruker D8 Discover X-Ray diffractometer using double bounced monochromatic CuK alpha radiation (λ=1.54 Å) with a fixed divergence slit size 0.5° and rotating sample stage. Freshly prepared C-S-H collected after washing with a water-ethanol solution followed by vacuum filtration was placed onto the sample stage and XRD patterns were recorded.

STEM EDX. Uniformity of the C-S-H was proved by chemical mapping or EDX measurements in STEM mode, using a FEI Tecnai Osiris analytical TEM instrument optimized for speed and sensitivity. The four windowless Super-X SDD EDX detectors integrated into the pole piece allow detection of 200,000 X-ray counts/s over a 0.9 rad solid angle. A high brightness XFEG gun allows EDX maps to be acquired in seconds to minutes. With a 11 Mpx Gatan Orius CCD camera, the microscope is also suitable for conventional BF/DF and high resolution TEM imaging. A BF, two ADF, and an HAADF STEM detector provide a wide range of diffraction and Z-contrast conditions. It operates with 200 kV high brightness XFEG with a point resolution of 0.24 nm and a probe current of 2 nA for EDX studies. The sample was prepared by dispersing 50 mg of C-S-H in 40 mL of isopropanol. A drop of the suspended liquid was allowed to dry on a 300 mesh copper grid.

XRF. In order to cross check the ICP results the samples were analyzed using X-Ray fluorescence spectroscopy (Optim'X 9900 Ceram XRF model). 20 g of hydrated sample was dried at 105 °C for 24 hours and ignited at 950 °C for 1 hour. 7.7 g of lithium tetraborate ($Li_2B_4O_7$) was added to the 0.7 g of calcinated sample to make a fused bead.

TGA. Samples were heated at 10 °C/min from 30 °C to 1000 °C to record the weight losses in setup from Mettler Toledo AG (TGA/SDTA851e). The total water bound in C-S-H was quantified from the total water loss between 30 and 250 °C. The amount of portlandite is quantified from the water loss around of the peak in the range from 400 – 480 °C and calcium carbonate was around 630 – 710 °C. No prior sample preparation involved.

ICP. ICP was performed on an ICPE-9000 series (Shimadzu) instrument, a multi-type ICP emission spectrometer with a near ppb detection limit. The sample compositions were analyzed using Optical Emission Spectroscopy mode (ICP-OES). 7 mL of 65% HNO_3 was added to a 0.25 g sample of C-S-H, then another 5 mL of fuming 100% ultra-pure HNO_3 was added to ensure complete dissolution. Each analysis consists of verification at further levels of 1-, 10-, and 100-fold dilution in pure water, with the 10-fold dilution affording concentrations best situated in the calibrated range of the instrument. Each analysis was repeated three times to check consistency.

Synthetic apparatus

The reaction system was fabricated in-house for the synthesis of C-S-H. The construction material is poly(methyl methacrylate), which is chemically stable under acidic or basic conditions. **Figure 3-7** shows a schematic of the reactor. It has four main parts – base, cylindrical wall, lid and micromixer unit. Calcium ion selective, conductivity, and pH measurement electrodes are inserted into the lid for real-time monitoring of the reaction conditions. There are also channels that allow for a purging flow of nitrogen gas across the main reaction chamber and an opening used for withdrawing small amounts of sample for kinetic analysis. A micromixer system is mounted on top of the vessel, consisting of three channels emerging form a central vertical column. The length of the column is fitted with a spiral static mixer to combine the reactant solutions prior to admission into the reaction chamber.

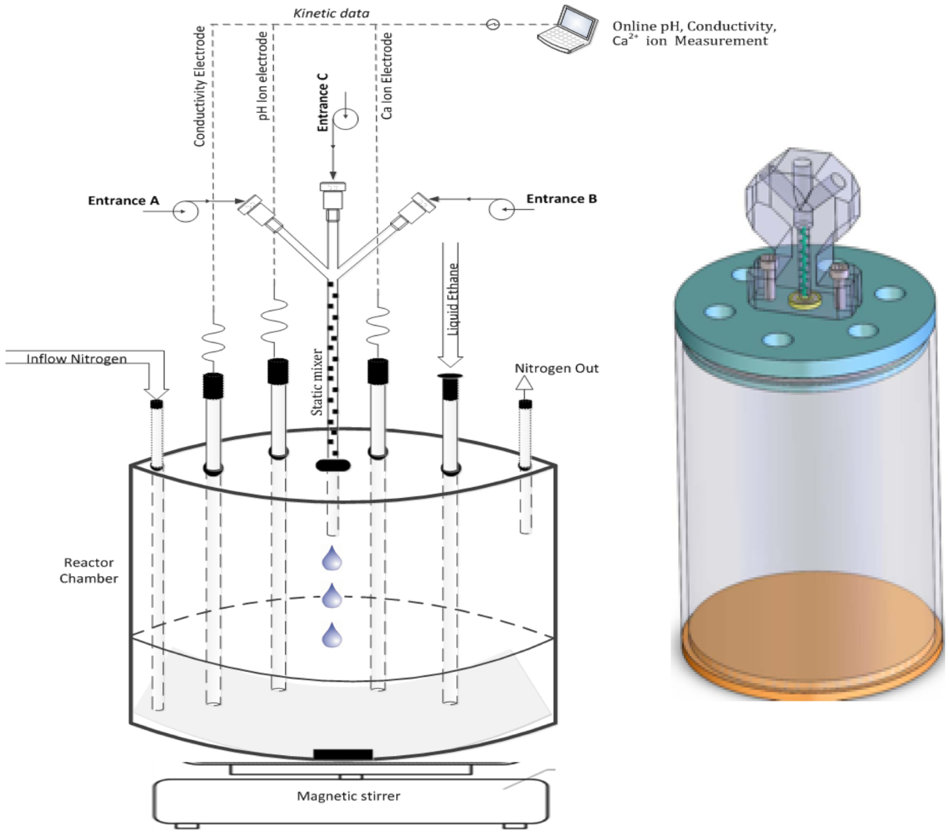


Figure 3-7. Schematic of the reaction vessel. A low pulsation piston pump was used to feed the reactants into the channels A, B, and C at rates between 0.01 mL/min to 5 mL/min. The off-axis reactant channels join the mixing column at an angle of 60°. The stirring rate was 700-800 rpm. Calcium ion selective, conductivity, and pH measurement electrodes are inserted into the lid for real-time monitoring of the reaction conditions. There are also channels that allow for a purging flow of nitrogen gas across the main reaction chamber and an opening used for withdrawing small amounts of sample for kinetic analysis. Nitrogen gas flowing at a rate of 20 mL/min was used to purge the chamber over the course of the reaction. Data was Recorded on a PC using LabX software (Mettler-Toledo).

Preparation and recovery

Solutions of calcium nitrate and sodium silicate were prepared in decarbonized water by boiling demineralized ultra-pure water (milliQ) for one hour and cooling in an ice bath. Solutions were immediately prepared after cooling. The quantity of solute used was measured with high accuracy. Measuring electrodes were calibrated twice before each synthesis. To avoid premature nucleation, all chemical glassware was washed and dried under laminar flow hood (Skanair®, Scan AG). After crystallization, the precipitated solids were recovered by washing and vacuum filtration. For each 200 mL aliquot, an equal amount of ultra-pure water mixed with ethanol (50:50 v:v), followed by pure ethanol, was used for the wash. Vacuum filtration was done on 20 nm filter paper (Whatman™, GE health care, ø 50 mm) to recover the washed C-S-H. The precipitated gel was carefully taken off the filter paper and stored in an airtight container. For characterization by TGA and XRD, drying of the filtered solid was necessary. This was performed under nitrogen flow at 70°C for 3 hours or 6 hours. All other characterizations were carried out in the native gel form.

Synthesis and characterization

For the current synthetic system, pH is a determining parameter for precipitation and ultimately controls the Ca:Si ratio and morphology. This is a consequence of how pH determines the predominant type of silicate species available in solution for reaction. Orthosilicic acid ($Si(OH)_4$) resists hydrolyzation even near neutral pH conditions owing to its small ionic radius (0.42 Å) and is therefore the predominant solution species below pH 7. In addition to pH, the silicate species which appear in an aqueous system is a sensitive function of cation type and concentration, such that the presence of small quantities of impurities can yield different synthetic results. In general, hydrolysis proceeds according to the following reaction to produce anionic species:

$$Si(OH)_x\,(aq) \rightarrow SiO_x(OH)_{4-x}^{x-} + xH^+. \tag{3-3}$$

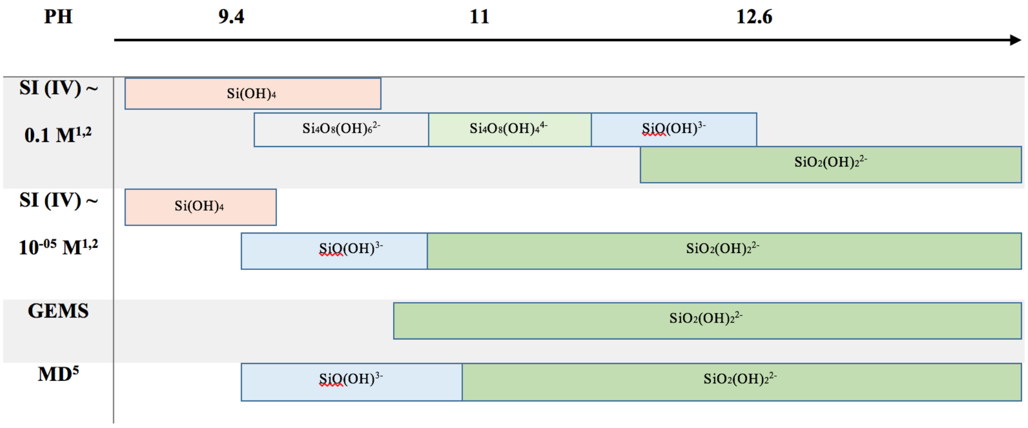


Figure 3-8. Predominant silicate species in aqueous solution as a function of pH according to different conditions and methods.[341-342]

Sodium silicate solutions at high pH are likely to contain silicates such as $SiO_2(OH)_2^{2-}$. Gibbs energy minimization software[312] (GEMS) predicts the same species in solution under these conditions. In conjunction with molecular dynamics (MD)[343], we summarize the presumable possible silicate species in solution as a function of pH in **Figure 3-8**. As long as an appropriate target pH range (pH > 11) is maintained, a chemical equilibrium favoring the silicate species $SiO_2(OH)_2^{2-}$ can be achieved under a wide variety of chemical conditions even at high silicate concentrations. In other words, regardless of whether or not an initially high concentration (high supersaturation) or low concentration (low supersaturation) of aqueous silicates is used, the pH can be used to favor high concentrations of the important silicate species $SiO_2(OH)_2^{2-}$, leading to the production of pure uniform product C-S-H, so long as the mixing is adequate.

We begin by setting a 2:1 ratio of calcium to silicon in the starting solution using equal volumes of 0.2 M and 0.1 M calcium nitrate to sodium silicate. GEMS[311-312] was used to calculate the pH required to achieve different Ca:Si ratios, and this pH was achieved during synthesis by adding an appropriate amount of concentrated NaOH, which is given in Precipitation was allowed to occur for 24 hours before the product was collected and analyzed. The pH calculated according to GEMS agrees with the experimentally measured pH. We see the amount of OH^- added in the system leads to a consistently increasing Ca:Si ratio in the solid precipitating phase.

Table 3-2. Amount of NaOH added to the reaction to achieve the pH necessary to produce the targeted Ca:Si ratio according to GEMS. The actual pH during the reaction is given in the final column.

Target Ca:Si (GEMS)	NaOH (GEMS)	pH (GEMS)	pH (Experiment)
1.0	0.05 mL	10.87	11.1
1.25	5.16 mL	11.47	12.5
1.5	10.58 mL	12.05	12.6
1.75	16.62 mL	12.55	12.7
2	20.00 mL	12.81	12.8

TEM analysis (**Figure 3-9**) shows that the morphology of the precipitated particles changes at pH 11 and a Ca:Si ratio of 1.25. The morphology resembles foils (*nanofoils*) for pH ≥ 11, Ca:Si ≥ 1.25; and globules (*nanoglobules*) for pH < 11, Ca:Si < 1.25. Repeat analysis confirms that these results can be easily replicated by our synthetic apparatus.

The composition of C-S-H produced by the rapid precipitation method is summarized in Elemental analysis for calcium, silicon, and sodium by ICP-OES indicates that the rapid precipitation method described here succeeds in synthesizing C-S-H with the targeted Ca:Si ratios. Despite the low measured sodium concentration, we cannot completely exclude the possibility that inclusion of some sodium may affect the structure. Nevertheless, when KOH is used as the pH regulator, we observe the formation of C-S-H globules for the Ca:Si = 1.00 composition and C-S-H foils for the Ca:Si = 2.00 composition. These products have the same morphological properties as the product obtained when NaOH is used as the pH regulator, as shown in **Figure 3-9**. On the other hand, the presence of cations such as Mg^{2+} or Ba^{2+} leads to the formation of a heterogeneous mixture of products. This strongly suggests that alkali cations are not critical structure determining factors, and that they serve primarily as charge balancing spectators.

Furthermore, our key structural insight is the necessity of the bridging calcium, Ca_B, which we propose is attendant to almost every defect site at high Ca:Si ratios. Considering sodium substitution of Ca_B, we calculate the Na:defect ratio, $2(n_{\mathrm{Na}}/n_{\mathrm{Si}})/\mathrm{P}(\mathrm{Q}^{(1)})$, where $n_{\mathrm{Na}}/n_{\mathrm{Si}}$ is the Na:Si mole ratio by ICP-OES and $\mathrm{P}(\mathrm{Q}^{(1)})$ is the population of $Q^{(1)}$ sites determined by NMR. We find this ratio is between 10 mol% and 30 mol% for each of the compositions with Ca:Si mole ratios at or above 1.25. This means that even in the worst-case scenario, in which every sodium atom substitutes a bridging calcium in a one-to-one fashion (for which we see no driving force), there is not enough sodium to accommodate every defect. In consideration of these matters, we remain confident that the key structural properties of our C-S-H systems can be analyzed in neglect of the small residual alkali content.

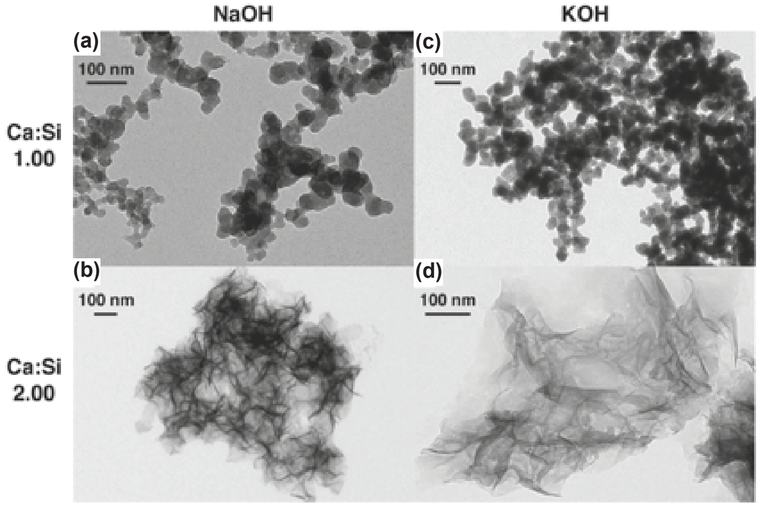


Figure 3-9. TEM imagery showing the morphology of the C-S-H produced for the Ca:Si ratio extremes when different alkali cations are present in the reaction. **(a)** Globule morphology produced for the Ca:Si = 1.00 composition using NaOH as pH regulator. **(b)** Foil morphology produced in the NaOH regulated reaction for the Ca:Si = 2.00 composition. **(c,d)** Same as **(a)** and **(b)**, respectively, but for the KOH regulated reactions.

Table 3-3. Mole ratios of important C-S-H components determined by various characterization methods.

Nominal Ca:Si	Ca:Si (XRF)	Ca:Si (ICP-OES)	Na:Ca (ICP-OES)	Na:defect (ICP-OES/NMR)
1.00	1.04	1.01 ± 0.03	0.13 ± 0.01	0.88 ± 0.13
1.25	1.21	1.24 ± 0.01	0.05 ± 0.02	0.20 ± 0.07
1.50	1.51	1.51 ± 0.03	0.02 ± 0.01	0.09 ± 0.06
1.75	1.77	1.78 ± 0.04	0.07 ± 0.02	0.30 ± 0.08
2.00	1.94	2.00 ± 0.07	0.05 ± 0.01	0.25 ± 0.04

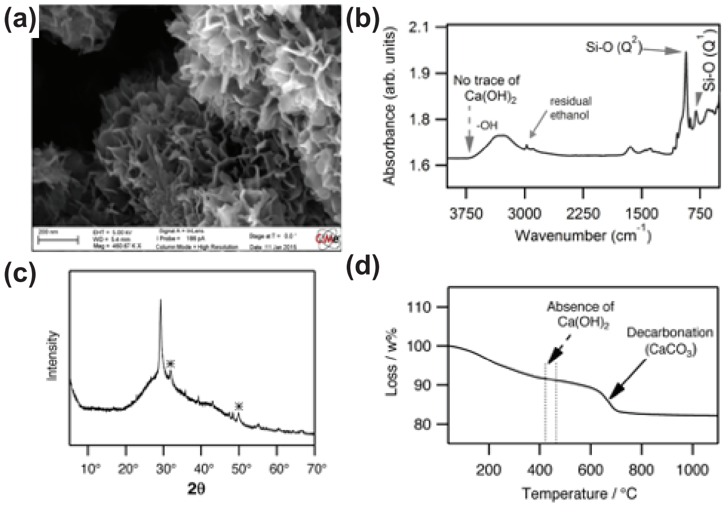
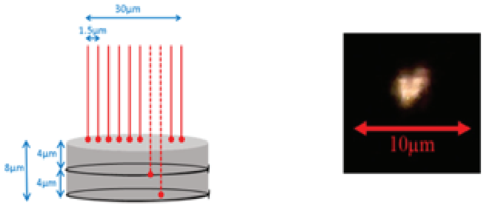


Figure 3-10. Characterization of freshly prepared C-S-H for Ca:Si ratio of 2. **(a)** SEM image showing foil morphology. **(b)** FTIR analysis.**(c)** XRD analysis. Resolved peaks corresponding to C-S-H are indicated with stars. Minor peaks correspond to calcium carbonate, which also contributes to the major peak at 29° where it overlaps a C-S-H peak. **(d)** TGA analysis. A calcium hydroxide phase is never observed, but XRD and TGA reveal that C-S-H is susceptible to the formation of calcium carbonate after prolonged air exposure.

These results confirm that our synthetic procedure yields particles of C-S-H with the targeted Ca:Si ratios. Importantly, the formation of $Ca(OH)_2$ is never observed, as illustrated by the FTIR, TGA and XRD analyses of the Ca:Si = 2.00 sample shown in **Figure 3-10**. It is worth noting, however, that long exposure of fresh C-S-H samples in open air (for example in TGA or XRD analysis) does eventually lead to the formation of $CaCO_3$.

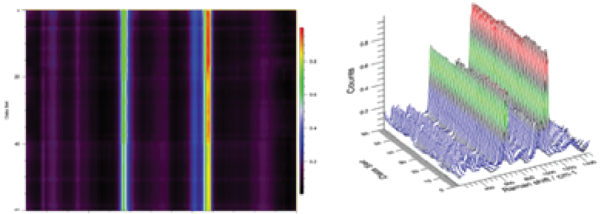
Sample uniformity

Determination by non-invasive Raman microscopy. We demonstrated compositional uniformity on pellet of C-S-H with a smoothed surface. The spot analysis (1 µm^2) analyzed more than 30 points on the particle surface. At depths of 4 µm and 8 µm the characteristic peaks positions in the C-S-H do not change, indicating the chemical environment uniformity of the sample at the micron level. A visual overview of the sampling and the results are given in **Figure 3-11**.



(a) Sample area & scanning points

(b) Spot size 4µm2



(c) Top view of the raman spectra

(d) 2-D view of the raman spectra

Figure 3-11. Raman microscopic analysis: **(a)** Sample pellet preparation, **(b)** Spot size used for analysis, **(c)** intensity plot comparing Raman spectra of all 30 spots, **(d)** stacked plot comparing Raman spectra of all 30 spots.

Determination by STEM-EDX. To prove the uniformity and consistency of the C-S-H samples to an even greater degree of spatial resolution, EDX in STEM mode was performed on the predefined grids. Once STEM micrographs are obtained, post-processing is performed using Bruker Esprit 1.8 software to obtain the corresponding chemical maps for the samples. The exported STEM image is processed for several parameters like detector effect corrections, Bremsstrahlung background, and Cliff-Lorimer quantification. The major constituents of our C-S-H system are defined for elemental identification. The maps are binned after defining the evaluation methods. As shown in **Figure 3-11**, about 50 – 60 points are analyzed individually from the chemical maps. Each spot corresponds to one pixel whose size is 2.34 nm x 2.34 nm. The signal obtained from each spot is processed to arrive at the final Ca:Si ratio at these points. A large background contribution to the signal is removed throughout the signal range spectra deconvolution is to be performed to address overlapped lines in the spectrum. The final quantification results of Ca:Si for each of these spots are recorded. For each sample, the standard deviation of the Ca:Si ratios measurements is less than 1%. EDX analysis provides us with useful information on the consistency of the Ca:Si ratio within the structure but due to the difficulty of accurately calibrating the instrument the actual Ca:Si determined ratio systematically less than that obtained from the XRF and ICP methods.

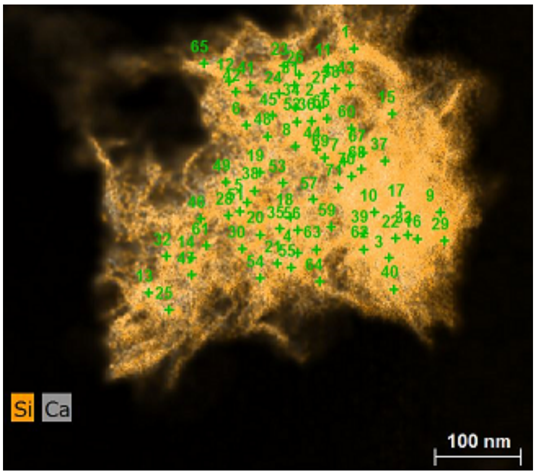


Figure 3-12. C-S-H chemical map revealing the spots used for the EDX analysis.

DNP enhanced NMR experiments

Sample preparation. **Table 3-4** describes the formulation of the samples, which were prepared as described in the Methods section. The C-S-H gels do not have an indefinite shelf life and are observed to harden over several weeks to months even in airtight containers. Driving off supernatant water from the gels accelerates this process. By drying the gels on a watch glass for about half an hour, very high DNP enhancements approaching 100 could be obtained, but the line shape would exhibit comparatively large $Q^{(2)}$ signals. Occasionally, signals from $Q^{(3)}$ and $Q^{(4)}$ species were observed, confirming that silicate polymerization accompanied the drying process.

Table 3-4. Formulation of samples used for DNP experiments. m_{gel} gives the mass of gel mixed with m_{agent} amount of DNP polarization agent. m_{in} is the amount of DNP ready C-S-H slurry that was put into in the rotor. t_{prep} is the estimated out of time between release of the C-S-H from storage in a saturated atmosphere to insertion of the sample into the DNP probe at 100 K.

Sample	m_{gel} / mg	m_{agent} / mg	m_{in} / mg	t_{prep} / min
Ca:Si = 1.00	124.1	25	-	60
Ca:Si = 1.25	133.3	33.3	26.6	75
Ca:Si = 1.50	119.3	31.8	23.8	15
Ca:Si = 1.75	114.1	27.0	23.4	15
Ca:Si = 2.00	121.0	30.6	25.4	15

NMR parameters. Table 3-5 gives the list of experimental parameters common to all NMR experiments, unless otherwise noted. All processing for the spectra presented here was performed using *RMN*.[344] Line shape analysis was performed using *gnuplot*. 1D CP MAS shifted echo experiments were performed using the sequence shown in **Figure 3-12**. In the presence of significant inhomogeneous broadening, advantages of the shifted echo experiment over conventional CP-detect are an improvement in sensitivity and improved accuracy of phase correction procedures.

Table 3-5. Parameters common to all NMR experiments.

MAS rate	**12.5 kHz**
^{1}H contact rf	60 kHz
^{1}H pulse/dec rf	100 kHz
^{1}H ramp profile	0.9 → 1.0
X contact RF	46 kHz
X pulse rf	66 kHz
Recycle delay[a]	3.0 s

[a]Recycle delay of 1.5 s used for 2D experiments on the Ca:Si = 1.00 sample.

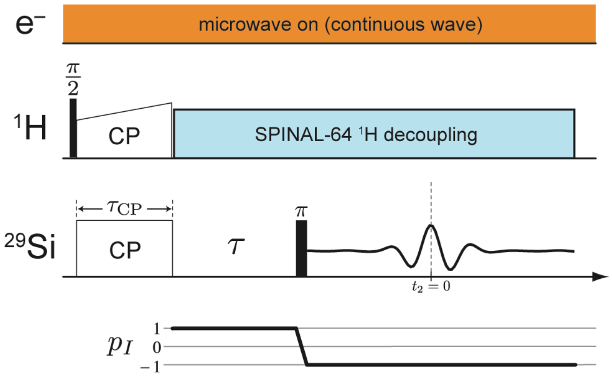


Figure 3-13. DNP enhanced CP MAS shifted echo pulse sequence used in this work.

For each sample, τ = 9.6 ms, τ_{CP} = 7 ms, and 32 transients were collected for a total experiment time of 1.6 min each. Gaussian apodization with a σ of 4.243 ms was applied to the t_2 signal envelope.

This experiment formed the basis of the variable contact time experiments, which nonuniformly sampled 49 different values of τ_{CP}: 200 µs to 2 ms (200 µs increment), 2.5 ms to 12 ms (500 µs increment), 14 ms to 30 ms (2 ms increment), and 35 ms to 80 ms (5 ms increment). 2D CP MAS refocused whole echo INADEQUATE experiments[331] were performed using the sequence shown in **Figure 3-14.** In addition to the use of hyper complex acquisition[345] to collect echo and anti-echo pathways, acquisition was initiated after the final π pulse in order to collect the entire signal envelope. This improves the sensitivity of the experiment and minimizes phasing artifacts during processing.

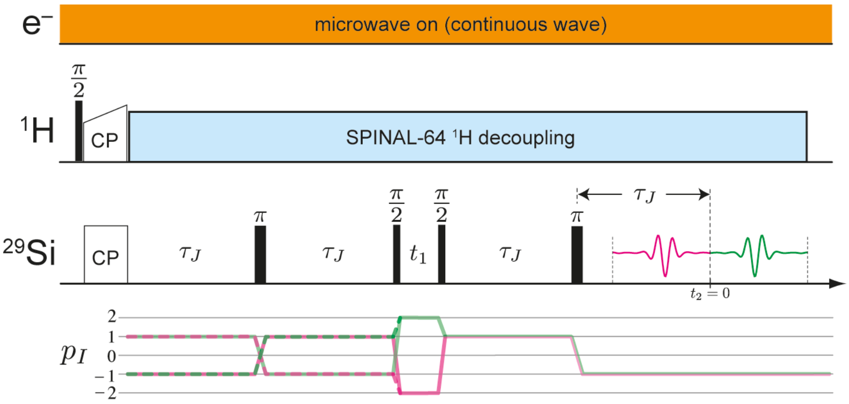


Figure 3-14. DNP enhanced CP MAS refocused whole echo INADEQUATE pulse sequence used in this work. The two p_I symmetry pathways (dashed and solid green) and anti-pathways (dashed and solid magenta) were collected and processed using hypercomplex acquisition. Whole signal envelopes were acquired during t_2 for path and anti-pathways.

For each sample, τ_J = 36 ms, τ_{CP} = 7 ms. The t_1 increment used was 240 µs. 16 complex t_1 points were collected. Other acquisition parameters are given in **Table 3-6** below.

Table 3-6. Acquisition parameters for 2D refocused INADEQUATE experiments.

Sample	Transients	Experiment Time
Ca:Si = 1.00[a]	640	9.6 h
Ca:Si = 1.25	320	9.0 h
Ca:Si = 1.50	320	9.0 h
Ca:Si = 1.75	512	14.5 h
Ca:Si = 2.00	512	14.5 h

[a] A gyrotron outage, lasting about an hour, occurred near the end of the experiment. The spectrum is qualitatively unaffected.

A shearing transformation was used to create a representation of the 2D INADEQUATE data that correlates two independent single-quantum dimensions.[346] Gaussian apodization with σ of 6 ms and 3 ms were applied to the t_2 and t_1 signal envelopes, respectively. The HETCOR echo sequence was performed using the sequence shown in **Figure 3-15**, utilizing the eDUMBO-22 homonuclear decoupling scheme[347-348] to suppress the line broadening from ^{1}H – ^{1}H dipolar interactions. This also scales the chemical shift and introduces an additional offset into the spectrum which were determined by comparison to a reference HETCOR spectrum of L-alanine. These values were used to present a corrected ^{1}H chemical shift dimension for the spectra shown in **Figure 3-20** and **Figure 3-21**, as well as **Figure 3-4**.

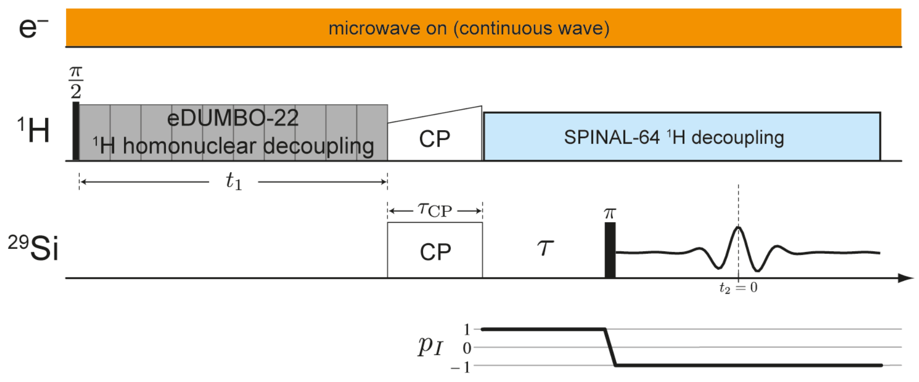


Figure 3-15. DNP enhanced HETCOR echo sequences used in this work. Hypercomplex acquisition was used to collect path and anti-pathways for t_1 evolution. Homonuclear decoupling was applied during t_1. Whole signal envelopes were acquired during t_2 for path and anti-pathways.

For each sample, τ = 9.6 ms. The eDUMBO pulse length was 32 µs. Other acquisition parameters are given in **Table 3-7**

Table 3-7. Acquisition parameters for HETCOR experiments.

Sample	τ_{CP}	Complex t_1 points	Δt_1	Transients	Experiment Time
Ca:Si = 1.00	0.7 ms	48	32 µs	24	61 min
	7 ms	48	32 µs	8	20 min
Ca:Si = 1.25	0.7 ms	48	32 µs	12	59 min
	7 ms	48	32 µs	4	20 min
Ca:Si = 1.50	0.7 ms	44	32 µs	12	55 min
	7 ms	48	32 µs	4	20 min
Ca:Si = 1.75	0.7 ms	20	64 µs	32	66 min
	7 ms	20	64 µs	16	33 min
Ca:Si = 2.00	0.7 ms	20	64 µs	32	66 min
	7 ms	20	64 µs	16	33 min

Gaussian apodization with decay constant of 4.243 ms and 1.2 ms were applied to the t_2 and t_1 signal envelopes, respectively. The apodization was applied to the t_1 dimension prior to multiplying the sampling interval by the chemical shift correction factor λ_{cs} = 0.57.

Sensitivity of DNP

For each sample, ^{1}H spectra were acquired both in the presence and absence of microwaves to measure the DNP enhancement of the protons. The enhancement level could not be determined accurately on the basis of the ^{1}H spectra alone due to a nonuniform enhancement of the broad line shape. The estimated proton enhancements $\varepsilon_{DNP}(^1H)$ are shown in **Table 24** below. Whereas a non-exponential recovery was observed for a ^{1}H saturation recovery experiment with approximate $T_{DNP}(^1H)$ = 1.3 s, a ^{29}Si CP saturation recovery experiment revealed a nearly exponential buildup with $T_{DNP}(^1H\text{-}^{29}Si)$ = 2.4 s. This suggests polarization relay into C-S-H particles with a steady state polarization reached after about ten seconds.

Table 3-8. Proton signal enhancements.

Sample	$\varepsilon_{DNP}(^1H)$
Ca:Si = 1.00	40
Ca:Si = 1.25	70
Ca:Si = 1.50	40
Ca:Si = 1.75	45
Ca:Si = 2.00	35

The sensitivity enhancement for DNP is called $\Sigma^{\dagger}$ and can be written as the product of several factors,[349]

$$\Sigma^{\dagger} = \varepsilon_{DNP}\theta d_{formulation}\left(\frac{S_{100K}}{S_{298K}}\right)\sqrt{\frac{T_1}{T_{DNP}}},$$

(3-4)

where θ is the fraction of observable nuclei in the sample, which is less than unity due to depolarization and quenching by the radical. $d_{\mathrm{formulation}}$ is a dilution factor related to the fact that additional of the polarization agent may reduce the amount of sample that can be placed into the rotor. The ratio S_{100K} / S_{298K} is generally accounts for the improvement in sensitivity gained by going to 100 K due to the ~2.8 improvement in the Boltzmann polarization as well as, e.g., an improvement in the probe quality factors. T_{DNP} is the approximate polarization build up time of the protons under DNP, and is to be compared room temperature proton T_1 values for C-S-H measured to be around 0.2 s.[52]

Equation 3-4 applies strictly only to signal from the polarizing agent and surface signals. Because the proton polarization is relayed into the C-S-H nanoparticles by proton spin diffusion, it is only of approximate validity. Nonetheless, taking $\theta \approx 1$ (signal is dominated by bulk C-S-H), $d_{\mathrm{formulation}} \approx 0.8$ (on the basis of **Table 3-4**), $S_{100K} / S_{298K} \approx 5$, and $(T_1 / T_{\mathrm{DNP}})^{½} \approx 0.25$, and the proton enhancements measured in **Table 3-8**, the sensitivity enhancement by DNP is generally the same as $\varepsilon_{\mathrm{DNP}}(^1\mathrm{H})$, indicating reduction of corresponding cross-polarization experiment times by $(\varepsilon_{\mathrm{DNP}}(^1\mathrm{H}))^2$, or about three orders of magnitude.

Quantification of Q species populations

Relative signal intensities in DNP enhanced CP MAS experiments are not usually in proportion to the relative populations of the nuclei generating the signal as they often are in experiments using direct excitation without hyperpolarization. Nonetheless, we can still use these signals for site quantification provided we assume that:

1. The length scale of hyperpolarization nonuniformity is larger than the unit cell of the particle, and
2. Cross-polarization kinetics can be measured and used to adjust the signal intensities appropriately.

The size of the C-S-H particles are sufficiently small (characteristic length ~100 nm) and have a proton density sufficient for nearly uniform polarization of the particles over the recycle period. To the second point, we performed cross-polarization measurements for different values of the cross-polarization contact time τ_{CP}, as shown in the first column of **Figure 3-16**. This data was fit to a simple *IS* model of CP kinetics for each site[330]. For our kinetic model, the signal intensities due to cross-polarization are given as a function of the cross-polarization contact time τ_{CP} by

$$I(\tau_{CP}) = I_0 \frac{e^{-\frac{\tau_{CP}}{T_{1\rho}}} - e^{-\frac{\tau_{CP}}{T_{IS}}}}{1 - \frac{T_{IS}}{T_{1\rho}}},$$

(3-5)

where $T_{1\rho}$ is the spin-lattice relaxation constant during rf irradiation and T_{IS} is the cross-relaxation time. I_0 is the base intensity, proportional to the equilibrium magnetization and hence number of nuclei generating the NMR signal for the given site. The 1D CP echo line shape was used in an initial unconstrained fit to three independent Gaussian functions, each representing the $Q^{(1)}$, $Q^{(2b)}$, and $Q^{(2p)}$ contributions. From this a set of mean Gaussian shift (δ) and widths (σ) for the frequency spectrum was determined and used to constrain the fit to the variable contact time data for the cross-polarization kinetic parameters. Stack plots representing the best fit and residual plots to this data are shown as the second and third columns of **Figure 3-16**. The cross-polarization kinetic parameters we determine from this analysis is given in **Table 3-9.**

Table 3-9. Cross-polarization kinetic parameters determined by the variable contact time experiments.

Sample	$Q^{(1)}$		$Q^{(2b)}$		$Q^{(2p)}$	
	$T_{1\rho}$ / ms	T_{IS} / ms	$T_{1\rho}$ / ms	T_{IS} / ms	$T_{1\rho}$ / ms	T_{IS} / ms
Ca:Si = 1.00	32.5 ± 0.6	1.81 ± 0.04	25.6 ± 0.8	1.09 ± 0.04	44.4 ± 0.6	4.07 ± 0.06
Ca:Si = 1.25	27.1 ± 0.3	2.31 ± 0.02	26.1 ± 0.9	1.41 ± 0.05	38.2 ± 0.9	5.14 ± 0.13
Ca:Si = 1.50	33.8 ± 0.3	2.19 ± 0.02	34.6 ± 2.1	1.24 ± 0.09	45.6 ± 1.7	4.78 ± 0.18
Ca:Si = 1.75	25.9 ± 0.2	2.22 ± 0.02	28.6 ± 1.9	1.49 ± 0.12	38.1 ± 1.9	4.62 ± 0.24
Ca:Si = 2.00	28.0 ± 0.2	2.40 ± 0.02	30.9 ± 2.7	1.29 ± 0.14	40.3 ± 2.5	4.91 ± 0.32

To complete the quantification, the 1D CP echo data was refit using **Equation 3-5** for the base intensities as well as new Gaussian shift parameters. The previously determined $T_{1\rho}$, T_{IS}, and Gaussian width parameters, averaged across the compositions with Ca:Si ≥ 1.25 for each site, were used for determination of the base intensities. The exception was the Ca:Si = 1.00 composition, where its own $T_{1\rho}$ and T_{IS} parameters were used. In accordance with the *dreierketten* model, the additional constraint $I_0(Q^{(2p)}) = 2\ I_0(Q^{(2b)})$ was enforced. The 1D CP echo spectra, best fit to this constrained 1D model, and best fit residuals are shown in **Figure 3-16**. Associated Gaussian shift and width parameters are given in **Table 3-10**.

Table 3-10. Shift (δ) and width (σ) parameters determined by the three Gaussian fit to the 1D CP MAS shifted echo data. The δ parameters were found in a fit subject to the constraint $I_0(Q^{(2p)}) = 2\ I_0(Q^{(2b)})$; σ parameters were carried over from a prior unconstrained fit.

Sample	$Q^{(1)}$		$Q^{(2b)}$		$Q^{(2p)}$	
	δ / ppm	σ / ppm	δ / ppm	σ / ppm	δ / ppm	σ / ppm
Ca:Si = 1.00	-79.71 ± 0.08	1.34 ± 0.05	-82.72 ± 0.08	1.08 ± 0.11	-85.77 ± 0.04	1.29 ± 0.03
Ca:Si = 1.25	-79.17 ± 0.02	1.25 ± 0.03	-81.85 ± 0.08	1.42 ± 0.31	-85.33 ± 0.03	1.25 ± 0.04
Ca:Si = 1.50	-79.10 ± 0.01	1.31 ± 0.02	-81.64 ± 0.07	1.03 ± 0.21	-85.17 ± 0.03	1.27 ± 0.06
Ca:Si = 1.75	-78.90 ± 0.01	1.27 ± 0.01	-81.54 ± 0.07	1.20 ± 0.19	-84.90 ± 0.03	1.24 ± 0.05
Ca:Si = 2.00	-78.87 ± 0.01	1.27 ± 0.01	-81.53 ± 0.08	1.11 ± 0.15	-84.81 ± 0.03	1.30 ± 0.05

By normalizing the sum of the base intensities to unity, we determine the Q species populations, reported in **Table 3-11**. As the residuals in **Figure 3-17** indicate, the analysis is not valid for the Ca:Si = 1.00 composition.

Table 3-11. Q species populations, subject to the constraint $P(Q^{(2p)}) = 2\ P(Q^{(2b)})$.

Sample	$P(Q^{(1)})$	$P(Q^{(2b)})$	$P(Q^{(2p)})$
Ca:Si = 1.00	0.290 ± 0.027	0.237 ± 0.009	0.473 ± 0.018
Ca:Si = 1.25	0.597 ± 0.107	0.134 ± 0.036	0.269 ± 0.071
Ca:Si = 1.50	0.700 ± 0.051	0.100 ± 0.017	0.200 ± 0.034
Ca:Si = 1.75	0.783 ± 0.053	0.072 ± 0.018	0.145 ± 0.035
Ca:Si = 2.00	0.830 ± 0.036	0.057 ± 0.012	0.113 ± 0.024

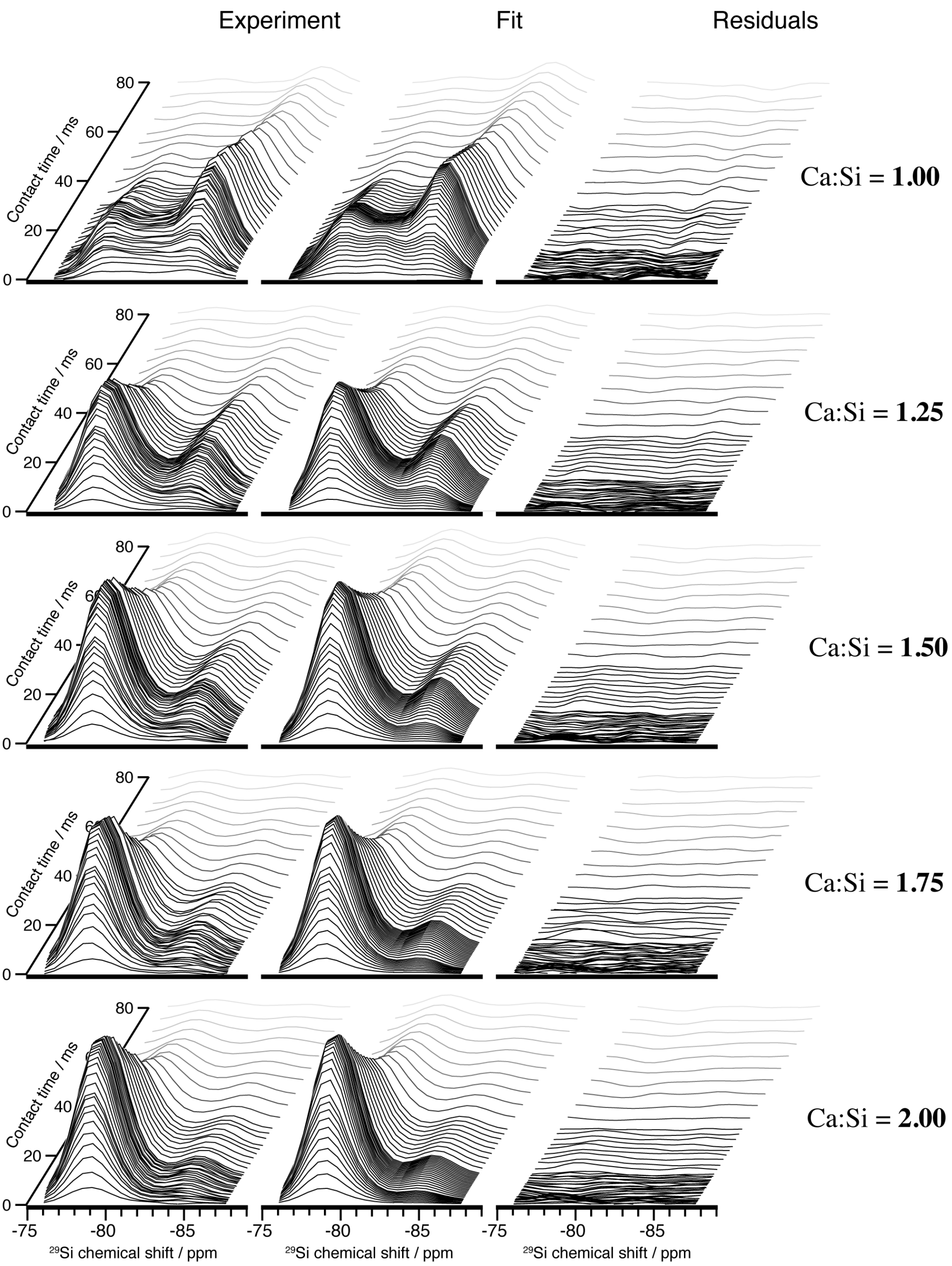


Figure 3-16. Stacked plots for the variable contact time spectra, best fit using the kinetic model, and the best fit residuals.

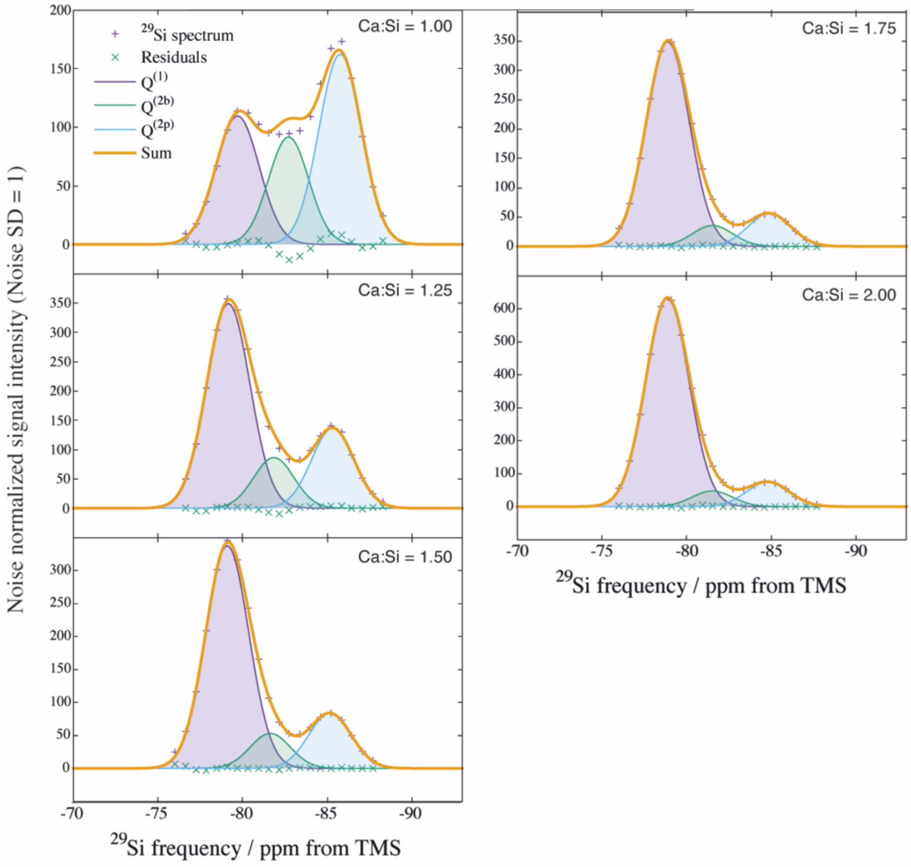


Figure 3-17. Deconvolution of the line shapes obtained in the DNP enhanced 1D CP MAS shifted echo experiments using the three Gaussian model described in the main text. The intensities are subject to the constraint $I_0(Q^{(2p)}) = 2\, I_0(Q^{(2b)})$.

Quantification of chain distributions

Each peak in the A-B chemical shift correlation line shapes presented by the INADEQUATE spectra in the first column of **Figure 3-18** were modeled by a 2D Gaussian function with zero correlation between independent A and B chemical shift dimensions. The shifts of the Gaussian functions along each dimension was constrained to the values shown in . The Gaussian width parameters were fixed to the same values for each fit, which were obtained by fitting the 1D projection onto the A chemical shift axis to three independent 1D Gaussian functions for the Ca:Si ≥ 1.25 compositions and taking the mean for each corresponding Q site. The 2D line shape model permits up to nine independent 2D Gaussian functions to be used; however, the functions corresponding to the $Q^{(1)}$-$Q^{(2b)}$, $Q^{(2b)}$-$Q^{(1)}$, and $Q^{(2b)}$-$Q^{(2b)}$ correlation peaks were omitted on the basis of the *dreierketten* model and validated by the absence of significant signal in the corresponding regions of the INADEQUATE spectra. The 2D experimental line was then fit for the intensities of the six constituent 2D Gaussian functions. The second and third columns of **Figure 3-18** shows the best fit results and residuals. **Table 3-12** gives the unnormalized peak intensities.

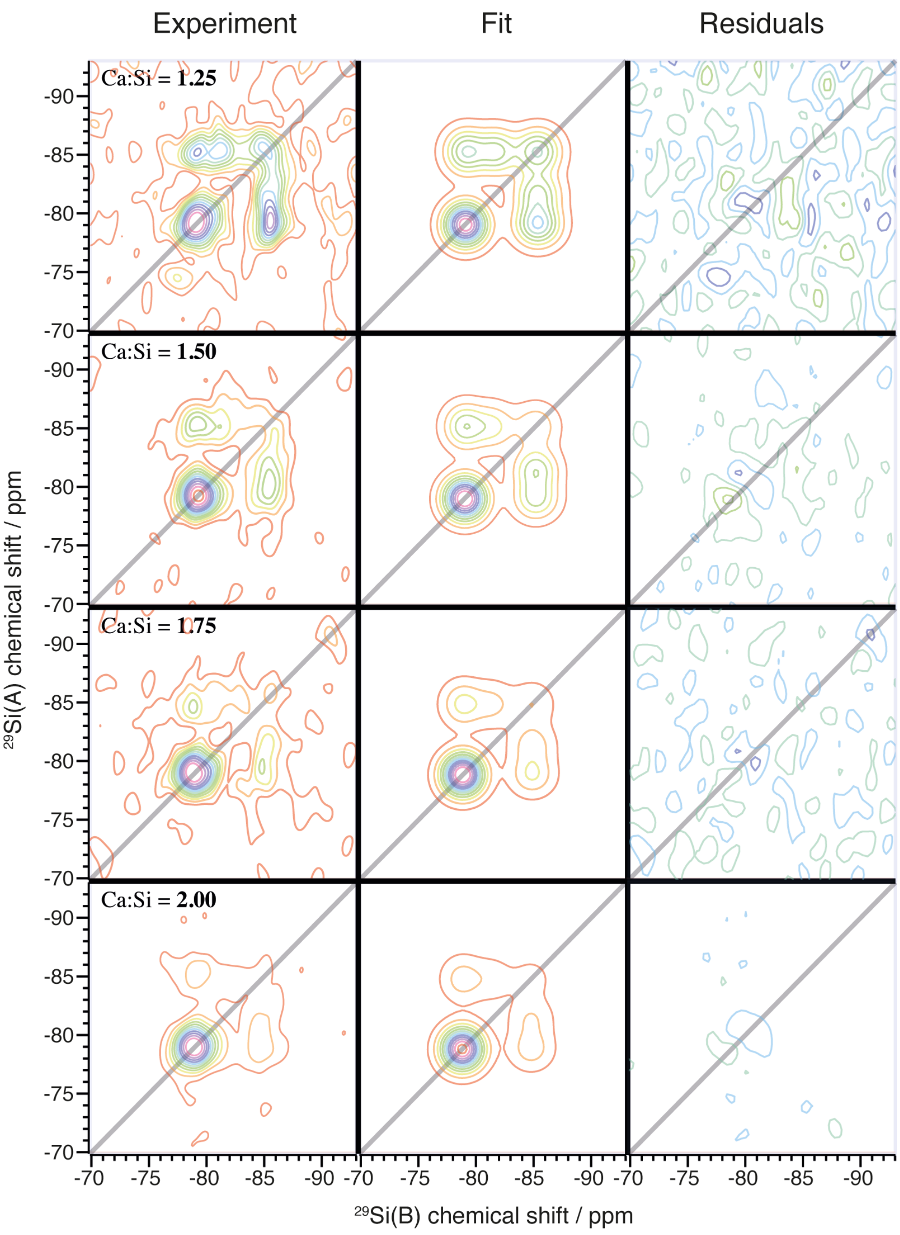


Figure 3-18. Experimental A-B correlated 2D refocused INADEQUATE spectra, best fit to the 2D Gaussian model, and best fit residuals for the C-S-H compositions with Ca:Si ≤ 1.25. Contours are drawn in 10% intervals beginning at 5% of the maximum signal intensity; the residual plots are relative to the experimental maximum and both positive (blue) and negative (green) contours are shown.

Table 3-12. Unnormalized best fit intensities of the A-B correlation peaks of the 2D refocused INADEQUATE spectra to the 2D Gaussian line shape model.

Sample	$I(Q^{(1)}\|Q^{(2p)})$	$I(Q^{(2b)}\|Q^{(2p)})$	$I(Q^{(2p)}\|Q^{(2p)})$	$I(Q^{(2p)}\|Q^{(2b)})$	$I(Q^{(1)}\|Q^{(1)})$	$I(Q^{(2p)}\|Q^{(1)})$
Ca:Si = 1.25	78.95	59.84	86.91	58.78	153.31	94.12
Ca:Si = 1.50	135.28	72.49	83.05	113.60	347.60	122.16
Ca:Si = 1.75	64.82	34.32	36.35	40.77	247.07	59.04
Ca:Si = 2.00	94.99	44.51	44.13	60.77	498.70	95.04

The intensity of an A-B correlation peak, denoted I(B|A), is given by,

$$I(B|A) = f(B|A)P(B|A)P_w(A).$$

(3-6)

We solve for the conditional probability P(B|A): the probability that a ^{29}Si nucleus of species B was detected given that it evolved with partner ^{29}Si nucleus of species A. They are normalized,

$$\sum_B P(B|A) = 1,$$

(3-7)

and Baye's theorem relates P(B|A) to P(A|B):

$$P(B|A) = \frac{P(A|B)P_w(B)}{P_w(A)}.$$

(3-8)

P_w(A) is the population of species A weighted for pair participation. At the sparse 4.7% natural abundance of ^{29}Si, the $Q^{(2)}$ sites are nearly twice as likely to have a ^{29}Si partner; therefore, $P_w(Q^{(2b)})$ and $P_w(Q^{(2p)})$ are obtained from the populations measured in the 1D experiments by doubling the population measured from the 1D experiments and renormalizing. Note that the sparse labeling simplifies the weighting analysis since the entire NMR signal is assumed to be derived only from isolated pairs and not triplets, etc. Finally, f(B|A) is an amplitude transfer factor that accounts for Q site differences in e.g. CP efficiency, T_2' relaxation, and J-coupling distributions, and were assumed not to change as a function of Ca:Si ratio.

The experimental intensities were normalized for each composition by dividing out $I(Q^{(1)}|Q^{(1)})$. Through the laws given above and the constraints imposed by the *dreierketten* model, any other conditional probability can be determined once $P(Q^{(1)}|Q^{(1)})$ is known. Upon substitution of **Equation 3-6** for each composition and using P_w(A) values determined from the 1D quantitative analysis, the five transfer factor ratios (**Table 3-13**) and $P(Q^{(1)}|Q^{(1)})$ for each composition were determined through a simultaneous fit of the twenty intensity ratios (five for each composition).

Table 3-13. Transfer factors determined for each type of correlation peak.

$f(Q^{(1)}\|Q^{(2p)})$	$f(Q^{(2b)}\|Q^{(2p)})$	$f(Q^{(2p)}\|Q^{(2p)})$	$f(Q^{(2p)}\|Q^{(2b)})$	$f(Q^{(1)}\|Q^{(1)})$	$f(Q^{(2p)}\|Q^{(1)})$
1.64	0.64	2.09	0.72	1 (defined)	1.74

The conditional probabilities are related to the distribution of chain species by

$$P\left(\mathrm{Q}^{(1)}\middle|\mathrm{Q}^{(1)}\right) = x_0, \tag{3-9}$$

$$P\left(\mathrm{Q}^{(1)}\middle|\mathrm{Q}^{(2\mathrm{p})}\right) = \frac{\sum_{n=1} x_n}{\sum_{n=1} x_n (2n)}, \tag{3-10}$$

$$P\left(\mathrm{Q}^{(2\mathrm{b})}\middle|\mathrm{Q}^{(2\mathrm{p})}\right) = \frac{1}{2}, \tag{3-11}$$

$$P\left(\mathrm{Q}^{(2\mathrm{p})}\middle|\mathrm{Q}^{(2\mathrm{p})}\right) = \frac{\sum_{n=1} x_n (n-1)}{\sum_{n=1} x_n (2n)}, \tag{3-12}$$

where the mole fractions of chains with repeat index *n* is denoted x_n. Application of the laws of conditional probability lead to the constraints reported in above. The parameters determined by our analysis are given in **Table 3-1**.

Recalling that previous studies have generally focused on Ca:Si < 1.50, which are not relevant to industrial formulations, we highlight that the Ca:Si = 1.00 composition is remarkable in that silicate dimers appear to be completely absent ($x_0 = 0$), as noted by the lack of a prominent $\mathrm{Q}^{(1)}$-$\mathrm{Q}^{(1)}$ correlation peak observed for all of the other C-S-H compositions. This is shown in **Figure 3-19**.

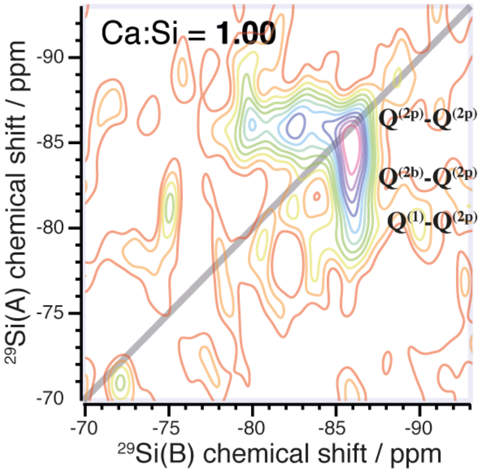


Figure 3-19. Experimental A-B correlated 2D refocused INADEQUATE spectrum for Ca:Si = 1.00. A gyrotron outage, lasting about an hour, occurred near the end of the experiment. The spectrum is qualitatively unaffected.

Heteronuclear ^{1}H-^{29}Si correlation

For each composition, a 2D HETCOR experiment using the pulse sequence described in **Figure 3-15** was performed for both a short (0.7 ms) and long (7 ms) values of τ_{CP}. The use of a short contact time biases the contribution to the NMR signal from those protons that are close to the correlating ^{29}Si nuclei, though without significant proton density fewer than three bonds away from the Si nuclei, the notion of a well-defined cutoff distance for the signals which appear in the correlation spectrum loses significance.[330]

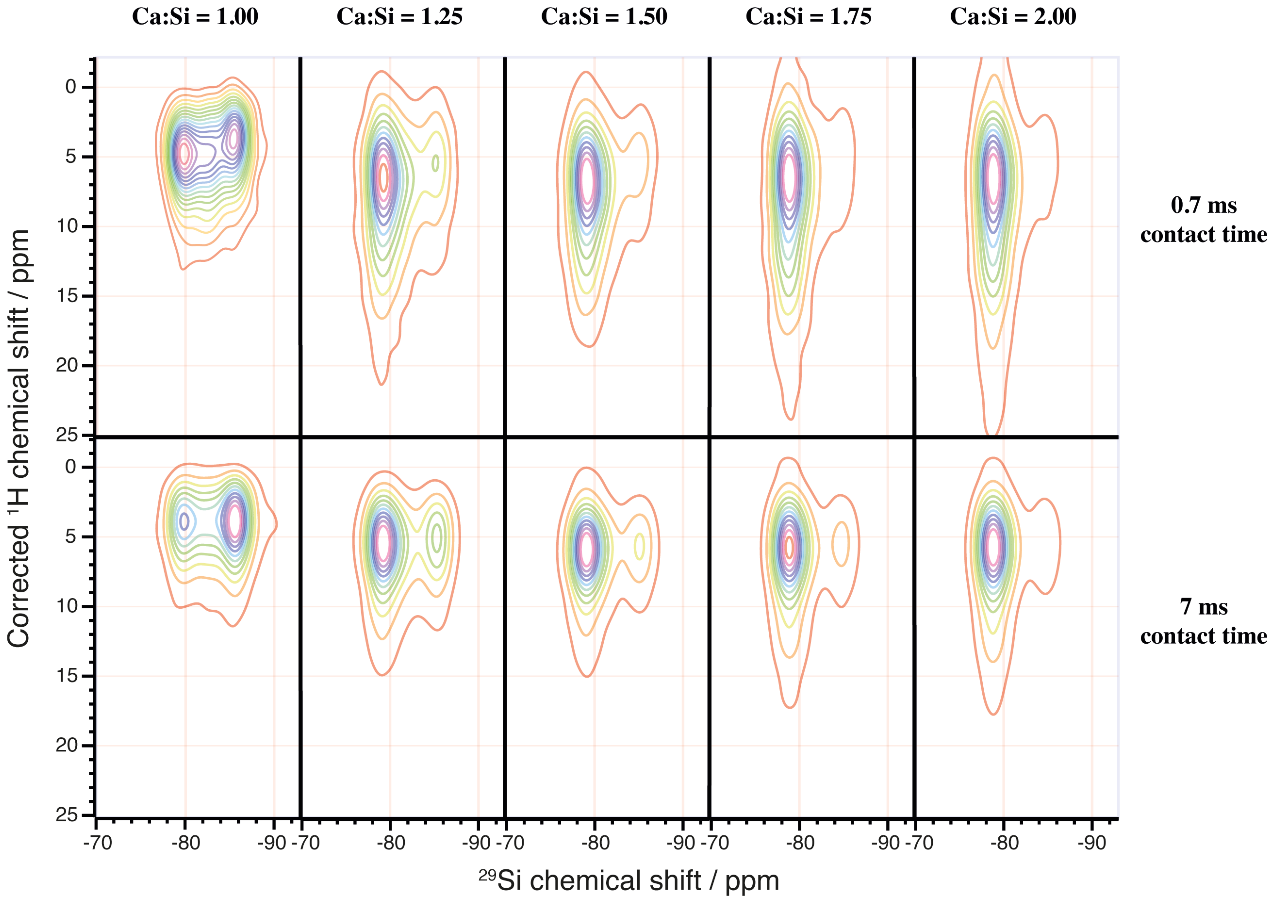


Figure 3-20. Complete series of DNP enhanced HETCOR spectra at both short and long contact times for all compositions studied.

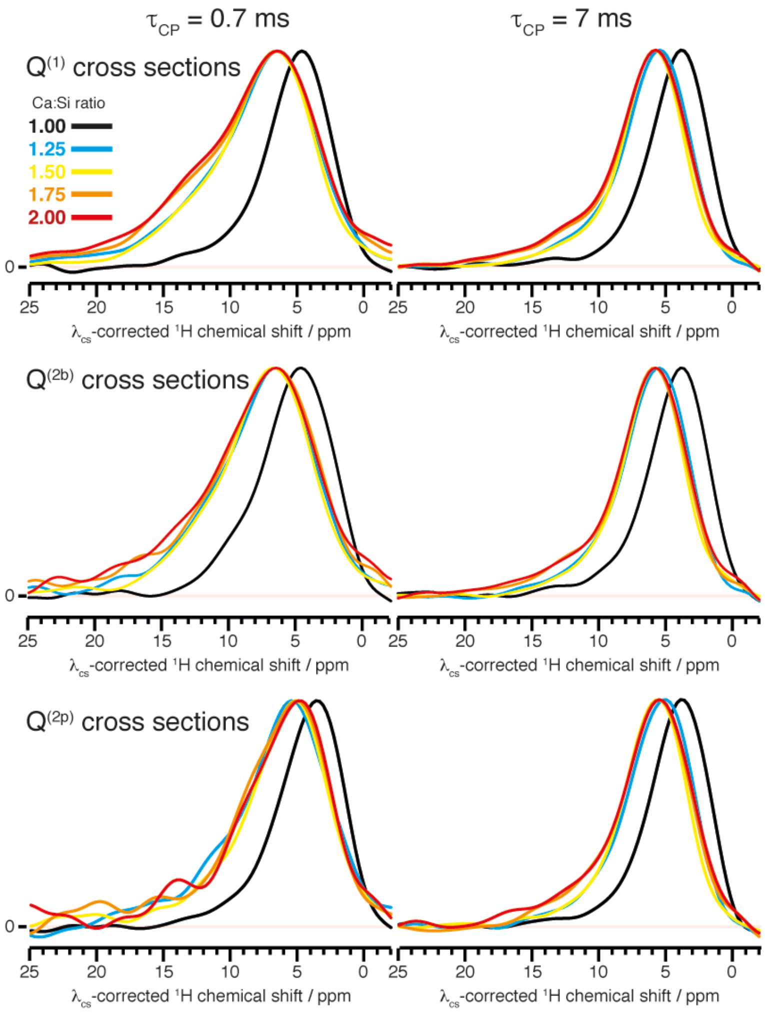


Figure 3-21. ^{29}Si site correlated ^{1}H spectra taken as cross sections from the full 2D HETCOR spectra at the appropriate ^{29}Si chemical shifts.

Structural model

It is known that C-S-H resembles a defective tobermorite.[319, 333] To create a structure based on defective tobermorite that possesses high Ca:Si ratios, we build substructures of C-S-H according to the following procedure:

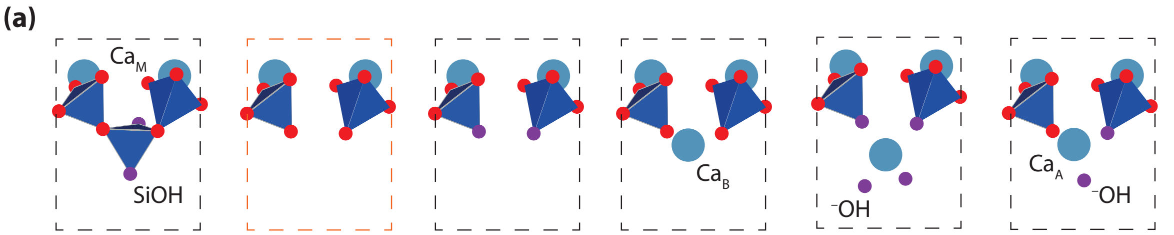
- Deprotonate silanol in the bridging tetrahedrons and replace it with a $CaOH^+$ ion in the interlayer.
- Remove a bridging silicate tetrahedron, performing charge compensation by adding two protons or a proton and a $CaOH^+$ ion or addition of a Ca^{2+} to coordinate the bridging site (Ca_B site in **Figure 3-22**).
- Add $Ca(OH)_2$ units in the interlayer space (Ca_I and Ca_A) to obtain higher Ca:Si ratios.

We study the effect of these different defect units (**Figure 3-22a**) on the ^{1}H chemical shifts. Reduced unit cells are constructed by connecting the defect units through an aqueous interlayer or an aqueous interlayer with a Ca_I and additional OH^- for charge balance (**Figure 3-22b**). In order to study medium range effects, we also consider different ways to combine the reduced unit cells, resulting in chain, dimer, and pentamer motifs (**Figure 3-22c**).

All the structures are first partially relaxed with energy minimization using METADISE[350] with a force field potential previously used for cementitious materials.[314] If the atomic bond distances, calcium coordination and local charge neutrality are satisfactory then they are relaxed using density function theory (DFT). For the former two criteria, we require specifically that Ca-O bonds are between 2.2 Å and 2.9 Å and that calcium coordination numbers are near six. The condition of local charge neutrality is implemented as systems with large distances between charged species consistently exhibit higher energies than systems for which this is not the case. Additional water molecules can be added to the interlayer to help satisfy these criteria. Depending on the initial atomic coordinates, especially those that specify the positioning of the interlayer water, the reduced unit cells may relax into different structures with the same defect classification.

These structures are again checked for the calcium coordination, lack of disruption of the main layer calcium-silicate backbone chain, and local charge neutrality. Once all the criteria are met, ^{1}H and ^{29}Si chemical shift calculations are performed on the candidates. The chemical shielding σ_{calc} was calculated using the generalized gradient approximation (GGA) functional PBE[205] within the Quantum Espresso code[188] and the GIPAW method.[63] For each calculation a plane-wave maximum cutoff energy of 80 Ry, and a Monkhorst-Pack grid of k-points[220] corresponding to 0.033 Å^{-1} in reciprocal space was employed. These values were tested for convergence of calculated energy and chemical shielding.

The convergence criteria for force, energy and pressure for structural relaxation were set to 10^{-3} E_h/a_0, 10^{-4} E_h, and 500 bar respectively. The final pressure of each relaxed structure was less than 150 bar. For structures which contain Ca_I, the final pressure was usually below 50 bar. To ensure this 500 bar threshold was sufficient, we performed an additional DFT relaxation of the structure based upon the ACcaV2 motif, setting a cell pressure threshold of 0.01 bar. Because of this stricter convergence criteria, O – O distances throughout the structure change by 0.05 - 0.1 Å, resulting in a ^{1}H chemical shift RMSD of 0.59 ppm and a ^{29}Si chemical shift RMSD of 0.34 ppm relative to the structure calculated with the higher convergence threshold for pressure. The higher ^{1}H chemical shift RMSD corresponds to the fact that proton chemical shifts are more sensitive to changes in the hydrogen bonding network than ^{29}Si. In NMR crystallography, two systems are considered identical if the ^{1}H chemical shift RMSD is below 0.5 ppm.[18] We justify a slightly higher limit for the C-S-H considering that most of the protons of weakly bonded interlayer species have lower barriers to conformational rearrangement relative to crystals of small organic molecules. Indeed, there is a correlation between the largest ^{1}H chemical shift changes occur for species near 0 ppm, as shown in **Figure 3-23**. If the proton chemical shifts corresponding to these non-hydrogen bonded H_2O are excluded from the comparison, we calculate a ^{1}H chemical shift RMSD of 0.38 ppm, which is well below the cutoff of 0.5 ppm. Therefore, a stricter convergence criterion for the DFT relaxation does not affect our interpretation of the ^{1}H chemical shifts nor the conclusions drawn from them.



A: "intact" Q^{2b} site

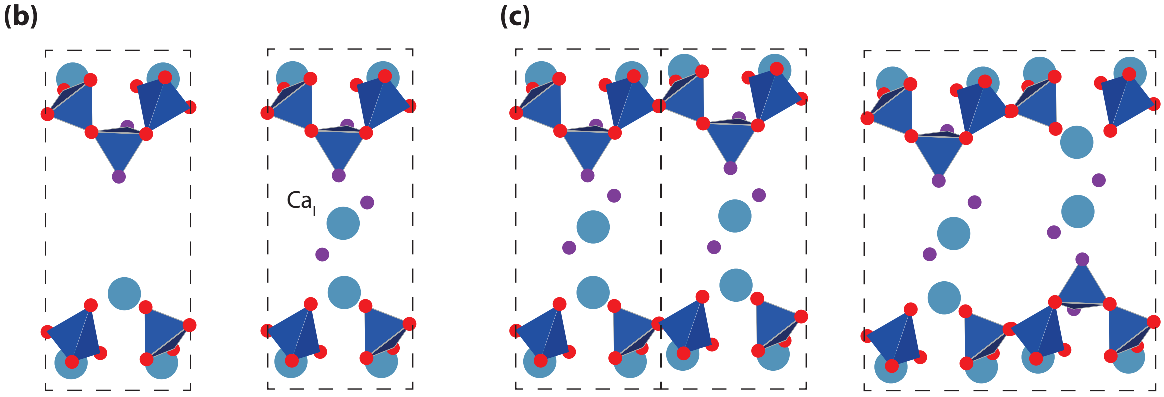
D: defect site, negatively charged

B: two Silanols at defect site

C: Ca$_B$ bridging defect site

H: two Silanols at defect site with Ca$_A$

G: Ca$_A$, Silanol and HO$^-$ at defect site



AC: A and B motive connected trough an aqueous interlayer.

ACca: A and B motif connected trough an aqueous interlayer with a Ca$_I$ and charge balanced by additional $^-$OH.

ACca: used as a reduced unit cell, resulting in a chain and dimer motif.

ACcaCAca: two reduced unit cells combined by rotating the second unit cell about 180° around the horizontal axis, resulting in a double pentamer motif.

Figure 3-22. Defect classification. (A) Simple defect units. (B) Simple defect units are combined with added interlayer water to form reduced unit cells. ^{1}H chemical shifts are calculated for structurally viable reduced unit cells. (C) Two possible ways of combining two reduced unit cells, showing how infinite chain, dimer, and pentamer motifs can be generated. The water in the aqueous interlayer and the hydrogen atoms are not shown.

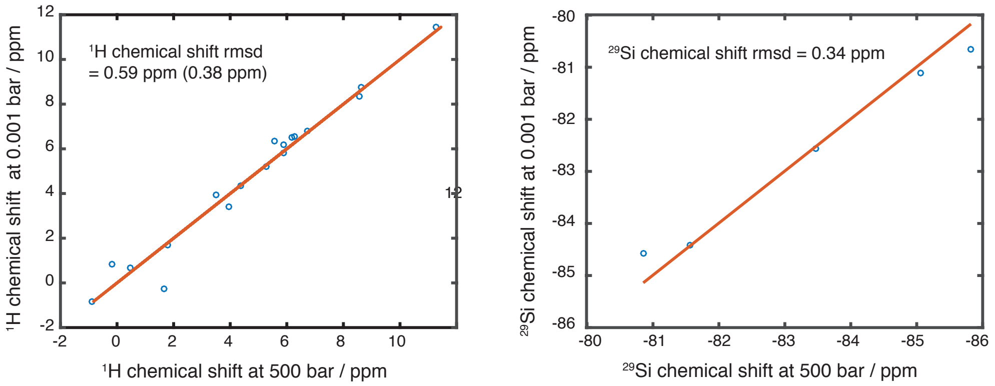


Figure 3-23. Calculated chemical shift correlations between DFT structures of C-S-H based upon the ACcaV2 motif at 500 bar and 0.001 bar.

Using the constraints from 1D ^{29}Si NMR and INADEQUATE experiments, we have calculated the number of dimers and the mean repeat index of the distribution. These two values are then used to fit a chain distribution, which was determined using the following Monte Carlo procedure:

1. We define a cutoff of n = 10 for the repeat index ($x_n = 0$ for $n \geq 11$).
2. For $n \geq 2$, the mole fractions are generated by a random number that is uniformly distributed between 0 to its theoretical maximum value given by the contribution to the $Q^{(2p)}$-$Q^{(2p)}$ correlation for that Ca:Si ratio:

$$x_{n\geq 2} = r \text{ where } 0 \leq r \leq \frac{1-x_0}{n-1}$$

3. Pentamers constitute the remaining fraction.
4. A chain distribution is accepted only if the difference between mean repeat index ($\sum_n x_n n$) obtained from the distribution and that calculated from the NMR constraints is less than 0.0005.
5. This procedure is iterated and the average fractions are stored.
6. The iteration is continued until the average values of the distribution converge to a unique distribution.

The random chain distributions calculated for each Ca:Si ratio are shown in **Figure 3-25**. For constructing our representative C-S-H structures, the longest chain used is a tetradecamer (n = 4), as indicated in **Figure 3-25c**.

The reduced unit cells deemed likely structural elements (see above) are permuted and stacked in the directions of the crystal axes in order to build a three-dimensional crystal structure satisfying all of our experimental NMR constraints. The proposed structures are shown in **Figure 3-24** and their silicate species distributions are compared with the experimental values in **Figure 3-25a-b**.

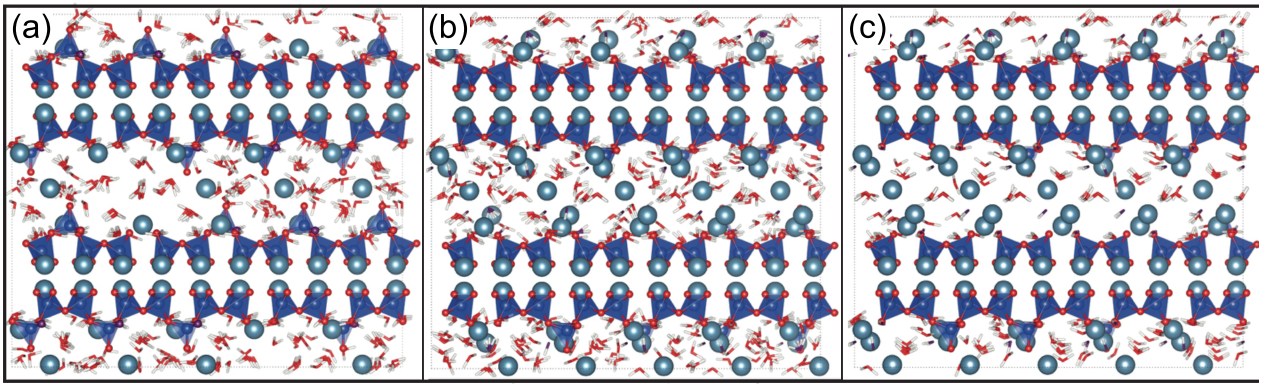


Figure 3-24. Proposed structures satisfying the NMR constraints for Ca:Si = 1.25 **(a)**, Ca:Si = 1.75 **(b)** and Ca:Si = 2.00 **(c)** viewed along the [100] direction. The relative positions of hydroxyls and water molecules have been relaxed with energy minimization at 0 K. Corresponding relaxed structures using MD are shown in **Figure 3-26**.

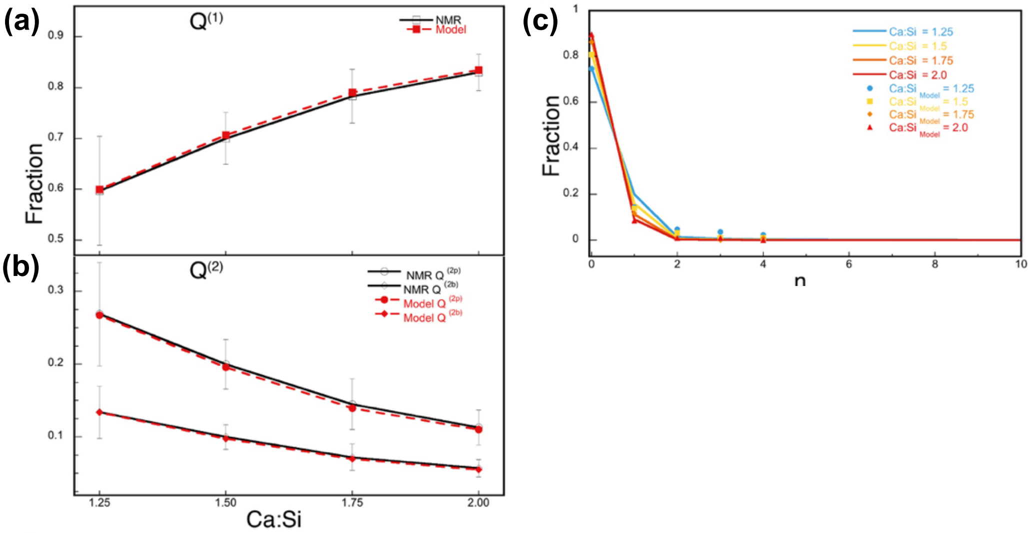


Figure 3-25. Distribution of silicate species determined by NMR compared to those predicted by the random distribution model. **(a)** Comparison between $Q^{(1)}$ populations and **(b)** $Q^{(2)}$ populations. The experimental values are shown in unfilled markers connected by solid lines whereas the corresponding values in our proposed structures are shown in filled markers connected by dashed lines. **(c)** Distribution of silicate chains according to the random distribution model. The mole fractions (up to $n = 4$) used in our representative C-S-H structures are shown as markers.

Structural Relaxation

Initial structural relaxation was performed with classical molecular dynamics using force field potentials. The force field parameters used are known to describe well cementitious material systems.[314] Simulations were done using a constant pressure ensemble at 300 K and a time step of 0.7 fs using Velocity Verlet integration algorithms implemented in DLPOLY.[315] Ewald summation was used to take into account the long range forces above a cutoff distance of 8.5 Å. Snapshots after 2 ns of molecular dynamics simulation of each structure are shown in **Figure 3-26** and are found to be structurally stable. Stoichiometry of the structures, bond distances and average calcium coordination numbers of bulk structures minimized after 2 ns are presented in **Table 3-14**. The bond distances from MD simulations are realistic. Histograms showing the distribution of coordination numbers for main phase calcium, interlayer calcium, and grand total of all calcium in these bulk C-S-H representations are shown in **Figure 3-27**. A systematic shift of the coordination number toward lower values is inevitable due to anharmonic vibrational motion of the atoms with respect to their proper equilibrium positions, an effect which is a function of the choice of force field used for the simulations. To estimate the magnitude of this shift for these systems, we carried out MD simulations on the known structure of 14 Å tobermorite for which 20% of the calcium are six coordinate and 80% are seven coordinate. The 2 ns MD snapshot of 14 Å tobermorite indicates roughly 30% fivefold coordination and 70% six fold coordination. Therefore, we expect the results in **Figure 3-27** to systematically underestimate a proper coordination number by nearly one.

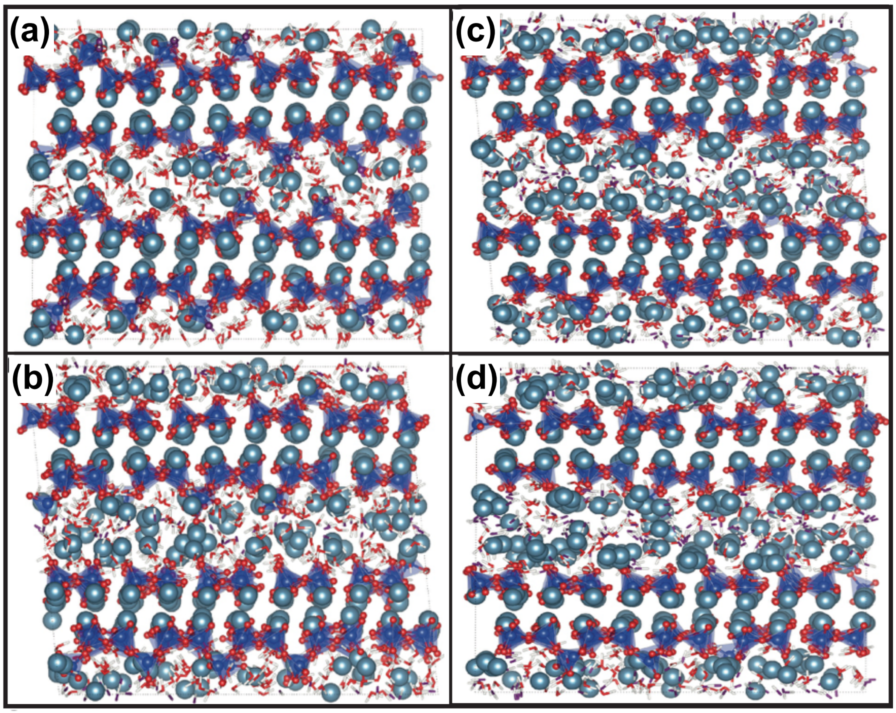


Figure 3-26. Snapshots of bulk structures relaxed for 2 ns using classical MD simulations. The structures shown are **(a)** Ca:Si = 1.25, **(b)** Ca:Si = 1.5, **(c)** Ca:Si = 1.75 and **(d)** Ca:Si = 2.0 respectively viewed along the [100] axes. All simulations produced structurally stable defective tobermorite features.

Table 3-14. Structural characteristics of the representative C-S-H structures. These values are given for MD structures relaxed for 2 ns. These values show that the chemical and physical environment in the structures are realistic. Ca-OH/Ca indicates the percentage of Ca atoms charge compensated by hydroxyl ions. The errors on the force field were estimated to be around 5% on distances.[314]

Ca:Si	**Chemical formula**	$Ca-OH/Ca$ **[%]**	$Ca-O$ **[Å]**	$Si-O$ **[Å]**	CN **(Ca-O)**
1.25	$Ca_{1.25}\,Si\,O_{3.2}\,(OH)_{0.1}(H_2O)\,_{1.82}$	0	2.3 ± 0.12	1.55 ± 0.08	5.9
1.50	$Ca_{1.5}\,Si\,O_{3.35}\,(OH)_{0.30}(H_2O)\,_{1.91}$	10	2.3 ± 0.12	1.55 ± 0.08	5.9
1.75	$Ca_{1.75}\,Si\,O_{3.39}\,(OH)_{0.71}(H_2O)\,_{1.72}$	20.1	2.3 ± 0.12	1.55 ± 0.08	5.8
2.00	$Ca_2\,Si\,O_{3.41}\,(OH)_{1.18}(H_2O)\,_{1.31}$	29.4	2.3 ± 0.12	1.55 ± 0.08	5.8

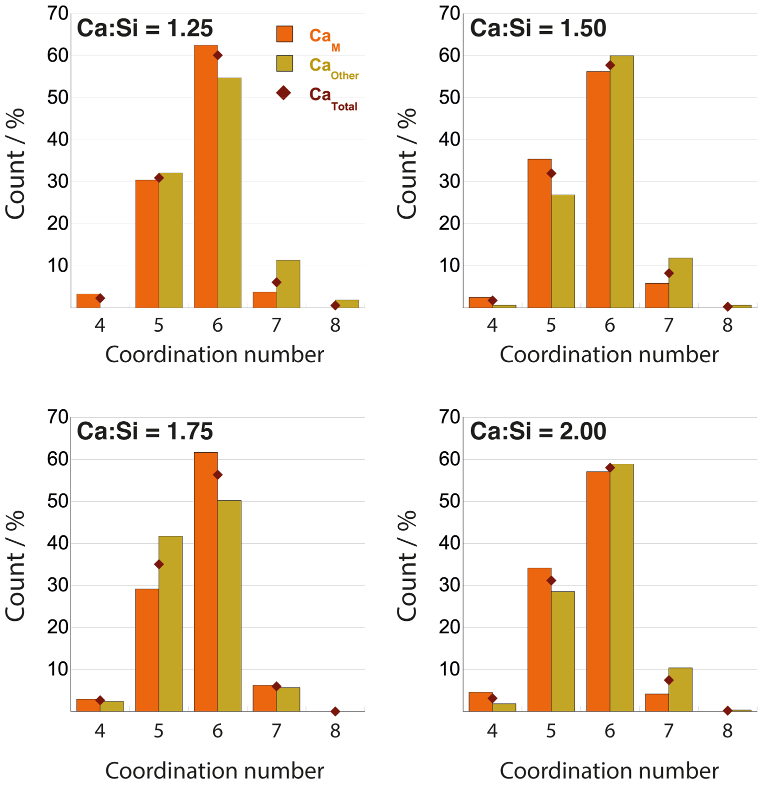


Figure 3-27. Histograms showing populations of coordination numbers for each of the representative C-S-H structures. These values are given for MD structures relaxed for 2 ns. Orange and green bars indicate coordination of main phase and all other calcium, defined as Ca_M and Ca_{Other}. The black markers indicate the coordination over all calcium in the structure (Ca_{Total}). Owing to positional bias in the MD simulated structures, the populations are systematically shifted toward lower coordination number by nearly one.

We also find that in the Ca:Si = 1.75 structure 20% of Ca atoms are charge compensated by hydroxyl ions. Thomas *et al.*[137] calculated this value to be 23% in C-S-H with Ca:Si = 1.7 in hydrated cement samples and argued that such a bonding is possible only if a structural motif resembling jennite is present. Our results show that the jennite structural motif is not required to give this hydroxyl charge compensation – a highly defective tobermorite is sufficient. We have not considered any structures with a defective jennite motif, in which a missing dimer is replaced by two OH^- groups. Pentamers, octamers, undecamers and tetradecamers are the only non-dimers in our proposed structures limited by the box size considered. Generally, the interlayer separation distance shrinks up to 2 Å (down from 14 Å) upon structural relaxation for Ca:Si ≤ 1.5, affirming our choice of 14 Å tobermorite as a reasonable base structure. Clino-tobermorite or other orthotobermorites can also be treated as the base structure satisfying the ^{29}Si and ^{1}H NMR constraints but without additional information describing the calcium environment in C-S-H it is difficult to evaluate which form of tobermorite would serve as the best base structure.

Proton chemical shift calculations

The ^{1}H chemical shift calculations are performed on the set of reduced unit cells displayed in **Figure 3-28**. These reduced unit cells are selected to ensure a wide variety of different local defect environments, classified according to **Figure 3-22a**, are captured. We also probe the influence of Ca_I in the aqueous interlayer and perform a test of the influence of medium range interactions by studying the containing pentamers rather than infinitely long silicate chains and dimers, which are the only types of chains possible without juxtaposition of different reduced unit cells. Calculated proton chemical shift spectra for each of these structural candidates are shown in **Figure 3-29**. Structures that are not distinguishable on the basis of defect classification may have different arrangements of water molecules in the interlayer, representing viable structures with different local energy minima and indicated as different "versions" in **Figure 3-29**.

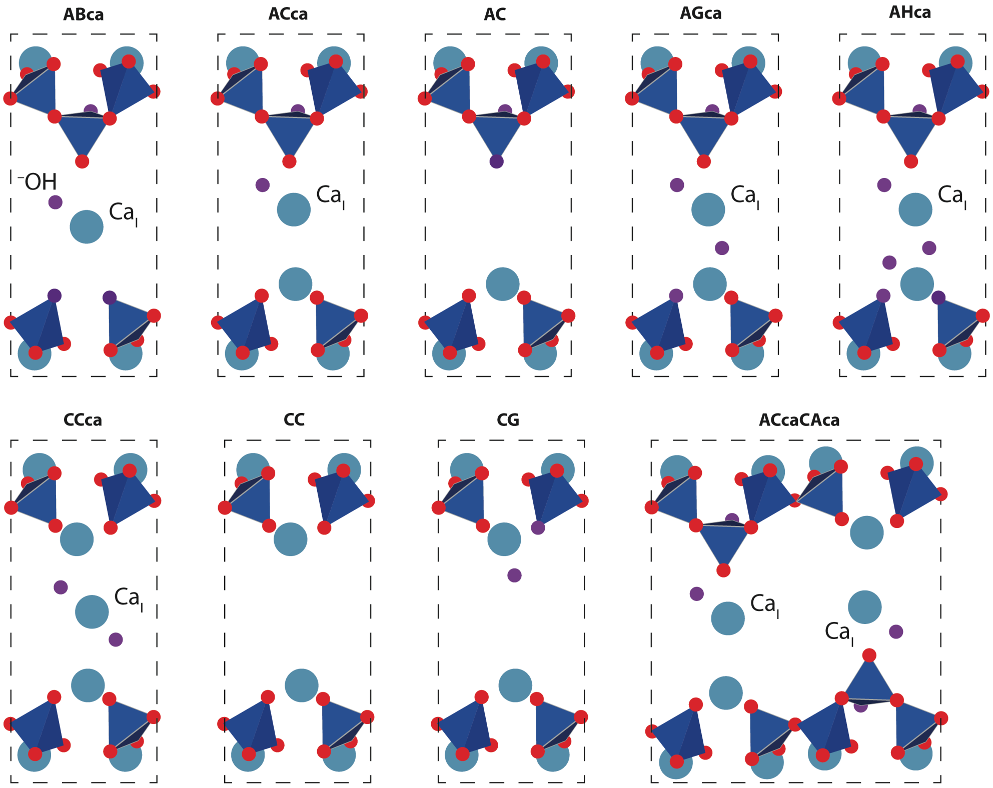


Figure 3-28. Reduced unit cells used in ^{1}H and ^{29}Si chemical shift calculations. Interlayer water molecules are not shown.

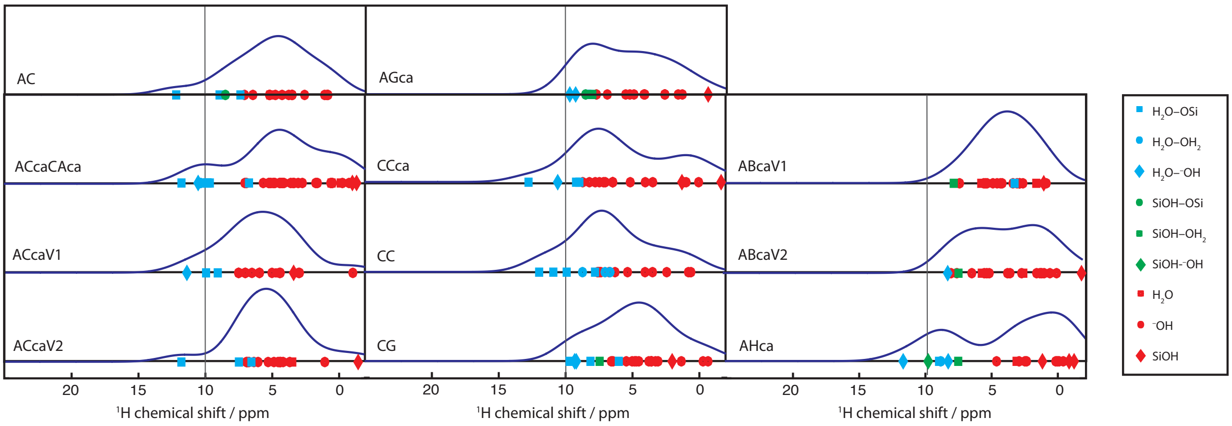


Figure 3-29. Calculated spectra of ^{1}H GIPAW isotropic magnetic shifts for the investigated reduced unit cells of C-S-H. The line-shapes $S(\delta)$ are extrapolated from the calculated chemical shifts δ_{calc} as $S(\delta) = \frac{1}{\sqrt{2\,\pi R^2}} \exp\left[-\frac{1}{2}\left(\frac{\delta-\delta_{calc}}{R}\right)^2\right]$ with R = 1.5 ppm. In general, structures with Ca_B at the bridging site (types AC, AH, CC, CG) better reproduce the characteristic tail in the ^{1}H line shape above 10 ppm. Structures that are identical according to our defect classification scheme but possess different arrangements of water molecules in the interlayer are distinguished by V1 or V2.

^{29}Si chemical shift calculations

In addition to the ^{1}H chemical shift calculations, we also calculate ^{29}Si chemical shift parameters (**Figure 3-30**) for all structures used in **Figure 3-29**. The calculated ^{29}Si chemical shifts are compared to previous calculations[109] and to our experimental results. To the level of intrinsic accuracy of ^{29}Si chemical shift calculations,[62-63, 70, 107] there is good agreement between the three datasets, allowing us to conclude that the C-S-H models proposed here are a good approximation of the studied systems.

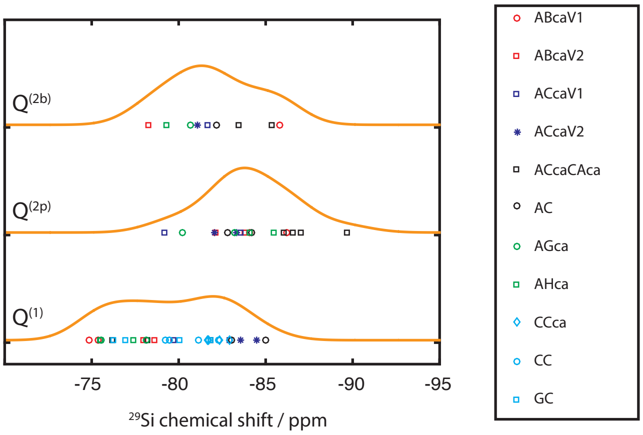


Figure 3-30. Overlap of calculated ^{29}Si GIPAW isotropic magnetic shift spectra for each different Si site in the calculated structures shown in **Figure 3-29.** The line-shapes $S(\delta)$ are extrapolated from the calculated chemical shifts δ_{calc} as $S(\delta) = \frac{1}{\sqrt{2\pi R^2}} \exp\left[-\frac{1}{2}\left(\frac{\delta-\delta_{calc}}{R}\right)^2\right]$ with R = 1.5 ppm.

3.3 Conclusion and Outlook

In conclusion, we determine the atomic-level structure of amorphous C-S-H using an approach that combines the abundant electronic structure information contained in the ^{1}H and ^{29}Si chemical shifts with constraints extracted from multidimensional ^{1}H and ^{29}Si NMR experiments and various other spectroscopic methods. The developed computational approach first uses experimental constraints to restrict the structural search space. From within this constrained space, local structural motifs are then explored and combined in a manner satisfying ^{1}H and ^{29}Si chemical shift constraints in order to build a full 3D structure, which provides an accurate representation of structural and chemical environments in C-S-H.

Note, that the inherently disordered nature of amorphous materials makes their structural characterization much less straightforward than for microcrystalline powders, as demonstrated in **Chapter 2**. For microcrystalline powders it is mainly the size of the crystallites that hinders their characterization by X-ray or other diffraction methods. However, the structures still exhibit a high degree of long-range order, which can be characterized through CSP-NMRX structure determination methods. Amorphous materials, on the other hand, are characterized through a lack of such high degree of long-range order. Here, we characterize C-S-H based on an ensemble of structural defect motifs assembled in a large structural model. However, the approach we present here possesses a few limitations which must be addressed to allow for widespread adoption as has been seen for NMRX of microcrystalline powders.

In the presented approach, we use structural motifs in agreement with the experimental constraints as building blocks to generate an atomic-level structural model of amorphous C-S-H. Note, that while the energetic stability of the combined model was evaluated using MD simulations on a structure containing several hundreds of atoms, the NMR chemical shifts were only evaluated for structures consisting of up to two of the local structural motifs. Further, for the DFT chemical shift calculations the local structural motifs were embedded in a fully periodic structure, which might further influence their calculated chemical shifts. Here, we propose two methods to overcome this limitation.

Hartmann *et al.*[83-84, 111, 351-352] have recently presented and benchmarked a fragment based DFT method to accurately calculate chemical shifts in molecular solids. The method is based on the calculation of pairwise interactions using a locally dense basis and embedded charges to model the extended chemical environments within a solid. We propose to use this method to calculate chemical shifts of structural motifs extracted directly from snapshots of large MD simulations.

Another method to overcome the limits of periodic DFT calculations, would be to extend the machine learning method presented in **Chapter 2.3**. Based on the work presented in **Chapter 2.2**, Cuny *et al.*[101] and Chaker *et al.*[102] have very recently demonstrated that it is possible to accurately and efficiently predict ^{17}O and ^{29}Si chemical shifts of glassy solids. Here, both methodologies can easily be adapted to various classes of amorphous materials by the choice of an appropriate training set. After the ML model has been trained, it could be used to directly calculate the chemical shifts of snapshots extracted from large MD simulations. Note, that the ML method could be combined with the fragment-based approach described above. The speed and efficiency of the ML method would allow for a large-scale screening of various MD simulations, while the fragment-based approach could then be used to more accurately calculate the chemical shifts of interesting structural motifs identified by the ML chemical shift calculations.

Note that, for C-S-H we constrained the search space of the local structural motifs through the use of multidimensional ^{1}H and ^{29}Si NMR experiments in combination with various other spectroscopic methods. However, for C-S-H, the initial structural search space is tremendous and we were only able to interpret the constraints due to the extensive prior chemical knowledge (e.g tobermorite being the base structure of amorphous C-S-H) available for these systems. Thus, the presented method needs to be extended and generalized to be applicable for other systems, where such prior knowledge might not exist or not be readily available and / or interpretable. A possible approach to extend and generalize the presented method would be to use ab-initio random structure searching (AIRSS)[353] to systematically and automatically screen the possible structural space. Candidate structures could then be selected based on their energetic properties as well as the evaluation of their calculated chemical shifts, either by DFT or ML. Also note that, prior information on the investigated structure could be incorporated into the AIRSS approach at many different levels. For example, for C-S-H, the AIRSS approach could be constrained to the known calcium-silicate backbone chain in combination with random Ca^+, Si^+, HO^- and H_2O defects.

Chapter 4 Defective and doped solids

4.1 Introduction

Doping is a key technology for tuning electrical and structural properties in industrial materials such as organic and silicon-based semiconductors,[354-356] oxide materials,[357-358] diamonds,[359-361] graphene[362] and perovskites.[135, 363-379] Doping has been reported to improve crystallinity,[375-376] enhance stability,[374] affect the optical and electric properties[354, 356, 358-361, 375-378] and improve the photocatalytic[357] and photovoltaic[135, 363-373, 375-376] performance. However, while several hypotheses have been put forward to explain these results, there often exists no full atomic-level characterization of the defective and doped materials. This is because diffraction-based methods, such as powder XRD which is currently the method of choice to investigate extended solids, lack information about the non-crystalline and disordered regions of the sample. In contrast, solid-state NMR can directly probe the local atomic environment around a dopant or defect site and is capable of detecting all species of a given spin-active nucleus that are present in the sample, regardless of the degree of crystallinity or the extent of phase segregation.

Here, we investigate the doping mechanism for a set of photovoltaic lead halide perovskite materials. For these materials different doping mechanisms resulting from interstitial defects,[375, 379] replacement of A-site cations,[374] or phase separation and passivation of grain boundaries[376] were suggested. In analogy to the procedure for microcrystalline solids (see **Chapter 2**) and amorphous materials (see **Chapter 3**) we evaluate different structural hypotheses by creating structural models for which we calculate the chemical shifts of different probe nuclei and compare the results to experiment. However, contrary to the procedure for microcrystalline solids (see **Chapter 2**) and amorphous materials (see **Chapter 3**), where NMRX provides a full and detailed atomic-level structure, here we only need to compare well defined reference structures, since the overall possible inorganic structures are well known.

The main computational challenge for the investigated systems is the presence of heavy atoms (e.g. ^{127}I, ^{133}Cs and ^{207}Pb) within the lead-halide perovskite structure. The heavy atomic cores significantly affect the electronic structure in the very vicinity of the nuclei. Therefore, molecular and structural properties depending on the electronic structure around the core often require full relativistic treatment.[380] For heavy nuclei of the 6th row of the periodic table it has been demonstrated that a full relativistic treatment, including scalar relativistic effects and spin-orbit coupling, significantly improves the calculated chemical shift accuracy.[123-125] For heavy nuclei of the 5th row of the periodic table the spin-orbit contribution is less pronounced and depends on the oxidation state and the stereochemistry of the atomic site.[122, 125-127] Additionally, the presence of the heavy atoms (HA) does not only influence their own shielding through the HAHA effect[381] but can also influence the shielding of neighboring light atoms (LA) through the HALA effect.[127, 382] Note, that for systems containing heavy atoms the use of hybrid Hartree-Fock-DFT functionals has also been shown to improve the calculated chemical shift accuracy.[122-127] However, in the investigated systems this effect is less pronounced compared to the full relativistic correction and mainly improves the slope of the correlation between calculated magnetic shielding and experimental chemical shift. To further investigate the required level of theory for the chemical shift calculation of these lead halide perovskite materials we calculate the ^{1}H, ^{13}C, ^{15}N, ^{39}K, ^{87}Rb and ^{133}Cs chemical shifts of organic molecules within the lead halide cage at different levels of theory (see **Chapter 4.2**).

Note that, the requirement of full relativistic calculations leads to a drastic increase in computational resources and thus strongly limits the system size and the number of possible local defects which can be evaluated computationally. Also note that, neither the full relativistic calculation of chemical shifts nor the use of hybrid functionals is currently available for periodic systems. As a consequence, the chemical shift calculations are generally performed using a cluster-based approach. However, for small dopant concentrations the number of atoms required to quantitatively model the defect site typically extends over multiple unit-cells and is thus often prohibitive for full relativistic DFT approaches. Here, we propose a divide and conquer approach, where we selectively investigate possible doping mechanisms (e.g. interstitial defects, replacement of A-site cations or phase separation and passivation of grain boundaries) by approximating the NMR parameters through computationally accessible limit cases. As an example, for the structural hypothesis of defect incorporation we investigate the limit case of an isolated defect not affecting the perovskite packing and the limit case of a dense defect area, where a dopant is incorporated into every 2nd to 3rd unit-cell.

In **Chapters 4.3** and **4.4** we use NMRX together with a set of other characterization methods to investigate the doping mechanism of a set of hybrid organic-inorganic multi-cation lead halide perovskites using different cation dopants (^{39}K, ^{133}Cs and ^{87}Rb). For this, we propose a set of structural hypotheses (interstitial defects, replacement of A-site cations and phase separation) for which we then calculate chemical shifts shift of different probe nuclei and compare the results to experiment.

4.2 Methods

As starting structures, we used the crystal structures of black and yellow formamidinium lead iodide (FAPbI$_3$), black methylammonium lead iodide (MAPbI$_3$) and cubic and tetragonal guanidinium lead iodide (GUAPbI$_3$). For all structures we optimized the hydrogen positions using Quantum Espresso (QE)[188] with the same parametrization as in **Chapters 4.3.5** and **4.4.5**. For the chemical shift calculations, we extracted small clusters from the crystal structures containing one XPbI$_3$ motif (X=MA, FA, GUA, K, Rb or Cs). Example clusters are shown in **Figure 4-1**. We then calculated the ^{1}H, ^{13}C, ^{15}N, ^{39}K, ^{87}Rb and ^{133}Cs chemical shifts using the Amsterdam Density Functional (ADF) **[383-384]** at the scalar-relativistic level and the same parametrization as in **Chapters 4.3.2** and **4.4.2**, unless otherwise specified.

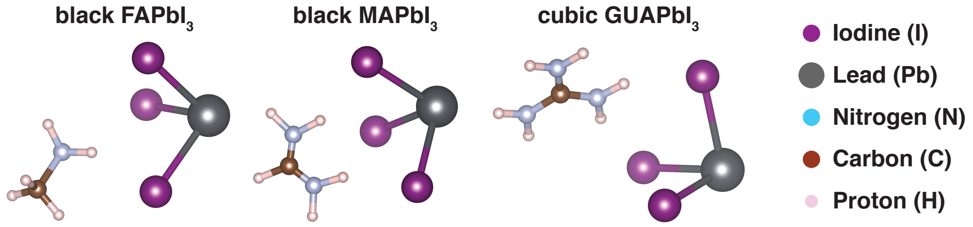


Figure 4-1. Example clusters used in DFT chemical shift evaluation.

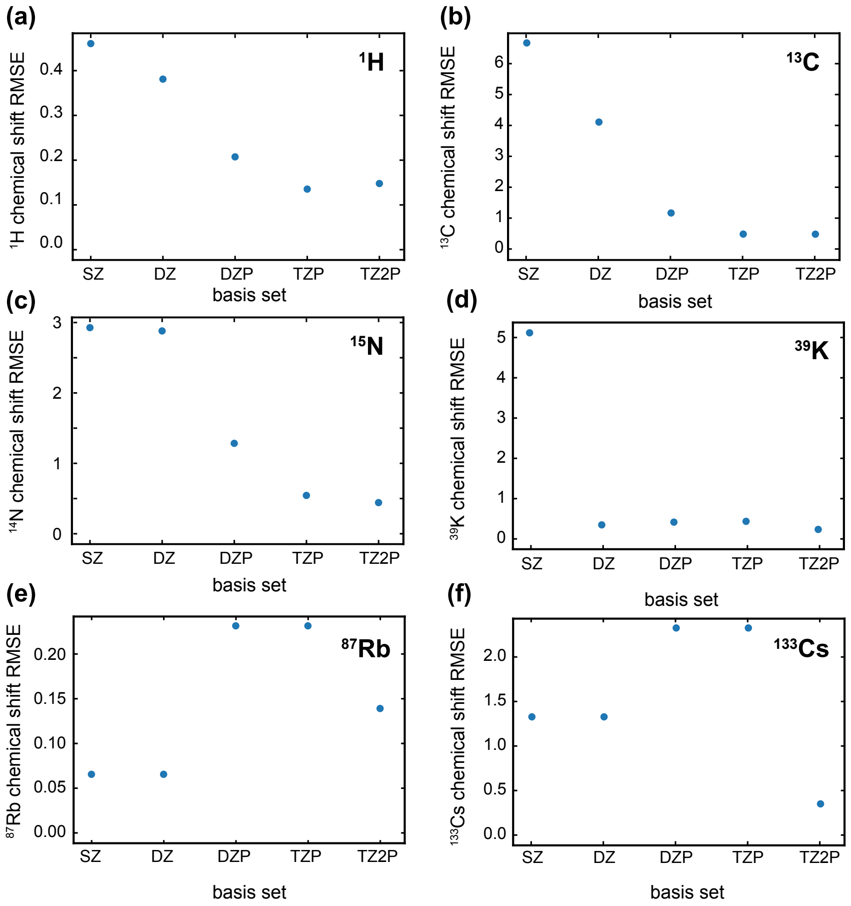


Figure 4-2. Chemical shift convergence with respect to the used basis-set for ^{1}H **(a)**, ^{13}C **(b)**,^{15}N **(c)**, ^{39}K **(d)**, ^{87}Rb **(e)** and ^{133}Cs **(f)**. The chemical shift RMSE is calculated with respect to the chemical shifts calculated with the largest basis -set used here (QZ4P).

For the ^{1}H, ^{13}C, ^{15}N, ^{39}K, ^{87}Rb and ^{133}Cs chemical shifts we investigated the size of the used basis-set (single zeta (SZ), double zeta (DZ), double zeta polarized (DZP), triple zeta polarized (TZP), triple zeta with two polarization functions (TZ2P) and quadruple zeta with four set of polarization functions (QZ4P)) and the used DFT functional, where we looked at a set of GGA functionals (PW91, PBE, BP86 and BLYP) and two Hybrid (B3LYP and PBE0) functionals with different amount of Hartree-Fock exchange (20% - 50%). Finally, we also investigated the use of the relativistic corrections (non-relativistic, scalar-relativistic and full relativistic, including spin-orbit coupling).

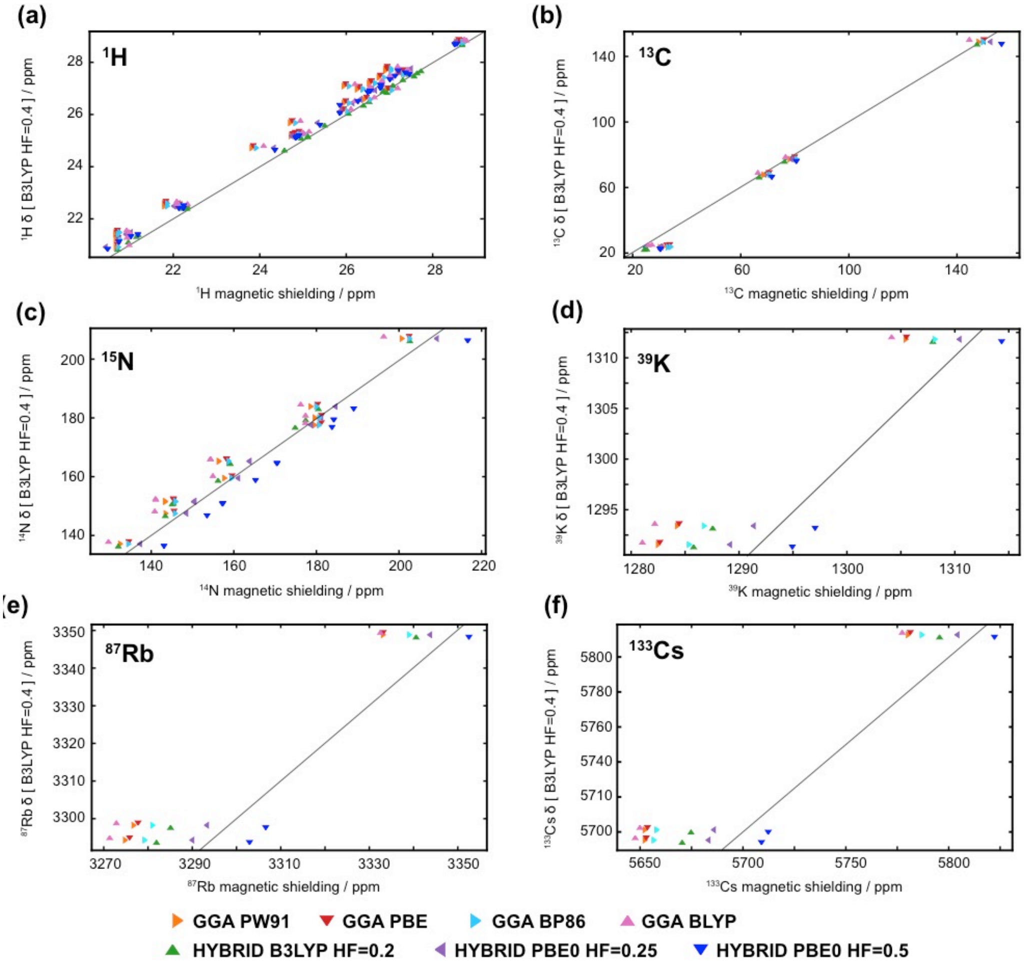


Figure 4-3. Magnetic shielding correlation with respect to the used DFT functional for ^{1}H **(a)**, ^{13}C **(b)**,^{15}N **(c)**, ^{39}K **(d)**, ^{87}Rb **(e)** and ^{133}Cs **(f)**. The magnetic shielding is plotted against the magnetic shieldings calculated with the hybrid B3LYP functional including 40% Hartree-Fock. The grey diagonal line shows a perfect linear correlation.

For the investigated basis sets we find that the ^{1}H, ^{13}C and ^{15}N chemical shifts start to converge at the TZP level. The ^{39}K chemical shifts seem reasonably converged at the DZ level. However, at the TZP level we obtain a slope closer to unity for the linear regression. For the ^{87}Rb we find a good convergence already at the SZ level, while for the ^{133}Cs chemical shifts the TZ2P level is needed for a good convergence (see **Figure 4-2** and **Table 4-1**).

Table 4-1. Chemical shift convergence with respect to the used basis-set. The correlation is calculated with respect to the chemical shifts calculated with the largest basis -set used here (QZ4P).

		RMSE / ppm	**offset (a) / ppm**	**slope (b) / ppm**
^{1}H				
	SZ	0.460	15.2	0.469
	DZ	0.381	13.0	0.579
	DZP	0.207	0.20	1.01
	TZP	0.135	0.43	0.998
	TZ2P	0.148	-1.05	1.04
^{13}C				
	SZ	6.67	105.2	0.55
	DZ	4.11	31.3	1.00
	DZP	1.17	26.6	0.95
	TZP	0.48	7.2	0.995
	TZ2P	0.48	7.5	0.988
^{15}N				
	SZ	2.93	161.6	0.48
	DZ	2.88	44.7	0.93
	DZP	1.28	37.5	0.92
	TZP	0.54	8.97	0.97
	TZ2P	0.44	4.9	1.00
^{39}K				
	SZ	5.12	1144.8	0.13
	DZ	0.39	116.1	0.91
	DZP	0.42	94.7	0.93
	TZP	0.44	44.7	0.97
	TZ2P	0.24	58.3	0.96
^{87}Rb				
	SZ	0.07	-69.4	1.02
	DZ	0.07	-69.4	1.02
	DZP	0.23	37.1	0.99
	TZP	0.23	37.1	0.99
	TZ2P	0.14	97.7	0.97
^{133}Cs				
	SZ	1.32	145.23	0.98
	DZ	1.32	145.23	0.98
	DZP	2.31	-7.86	1.00
	TZP	2.31	-7.86	1.00
	TZ2P	0.34	-0.11	1.00

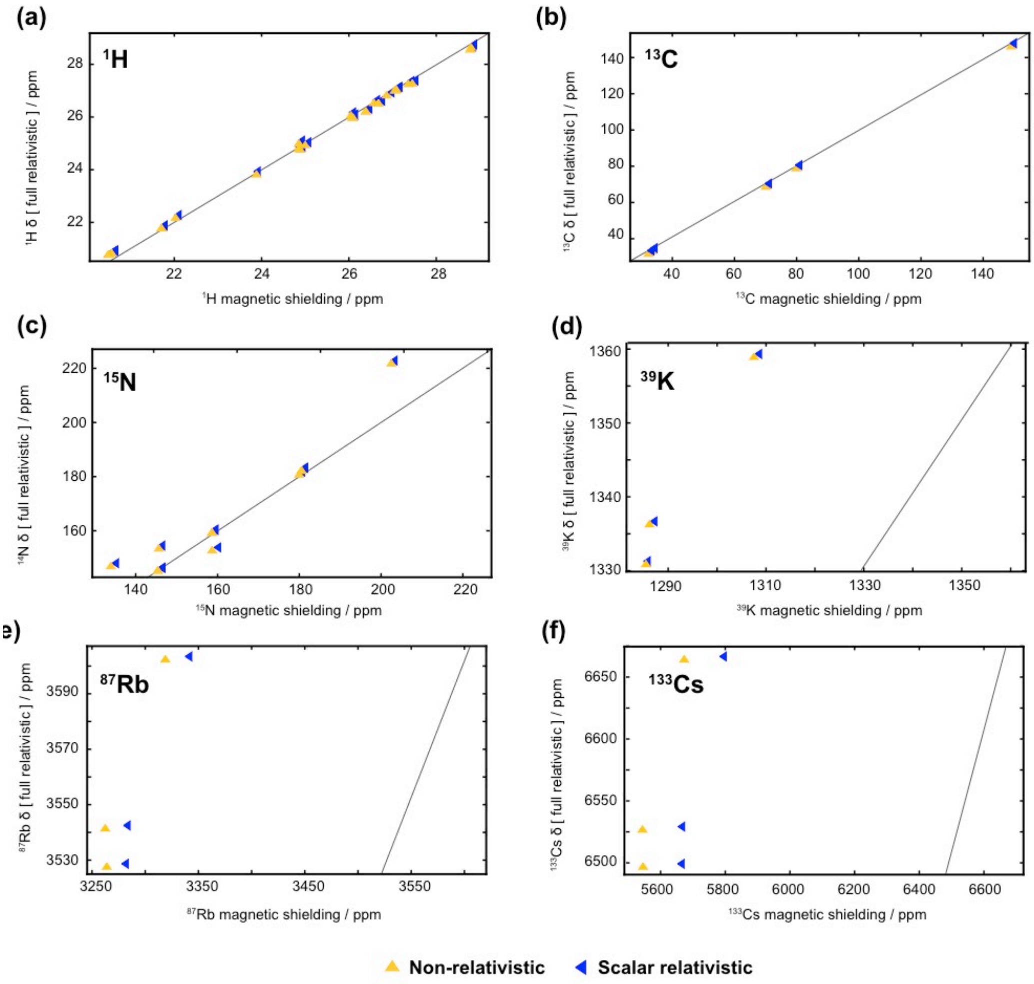


Figure 4-4. Magnetic shielding correlation with respect to the used relativistic correction for ^{1}H **(a)**, ^{13}C **(b)**,^{15}N **(c),** ^{39}K **(d)**, ^{87}Rb **(e)** and ^{133}Cs **(f)**. The magnetic shielding is plotted against the magnetic shieldings calculated at the full relativistic level, including scalar relativistic and spin-orbit coupling effects. The grey diagonal line shows a perfect linear correlation.

For the investigated DFT functionals we find that the ^{1}H and ^{15}N chemical shifts are nearly identical within a given family of functionals (GGA and Hybrid functionals) and that the amount of Hartree-Fock included does not lead to a systematic change in the chemical shifts (**Figure 4-3** and **Table 4-2**). Note that also the difference between the two families of functionals is below the expected DFT chemical shift accuracy for ^{1}H (0.33-0.43 ppm) and ^{15}N (5.4 ppm). [18, 83] For ^{13}C the GGA functionals perform very similar to the Hybrid functionals and the chemical shifts appear to depend mostly on the DFT contribution (BLYP), see **Figure 4-3** and **Table 4-2**. Note that also for ^{13}C the differences are below the expected DFT chemical shift accuracy (1.9-2.2 ppm). [18, 83] For the ^{39}K chemical shifts we find that the GGA and Hybrid functionals give a similar RMSE, with the exception of the BP86 and BLYP functional which give a slightly higher RMSE. We also note, that the reference value (here the offset (a)) steadily increases with the amount of Hartee-Fock (**Table 4-2**). For the ^{87}Rb and ^{133}Cs chemical shifts we observe that the RMSE is about half as low for the Hybrid functionals. However, also for the GGA functionals the RMSE is still very low (around 1-2%) compared to the full chemical shift range investigated here (around 75 ppm for ^{87}Rb and around 167 ppm for ^{133}Cs). Similar to the ^{39}K chemical shifts we note that the reference value (here the offset (a)) steadily increases with the amount of Hartee-Fock (**Figure 4-3** and **Table 4-2**).

Table 4-2. Chemical shift convergence with respect to the used DFT functional. The correlation is calculated with respect to the chemical shifts calculated with the hybrid B3LYP functional including 40% Hartree-Fock.

		RMSE / ppm	**offset (a) / ppm**	**slope (b) / ppm**
^{1}H				
	PW91	0.27	-0.88	1.01
	PBE	0.27	-0.92	1.02
	BP86	0.28	-1.12	1.02
	BLYP	0.26	-0.53	1.01
	B3LYP HF=0.2	0.05	-0.83	1.03
	PBE0 HF=0.25	0.05	-1.04	1.03
	PBE0 HF=0.5	0.09	-0.45	1.03
^{13}C				
	PW91	1.60	8.35	0.93
	PBE	1.75	9.88	0.93
	BP86	1.57	9.75	0.93
	BLYP	0.73	3.71	0.94
	B3LYP HF=0.2	0.58	1.18	0.98
	PBE0 HF=0.25	1.61	6.83	0.97
	PBE0 HF=0.5	1.70	5.16	1.01
^{15}N				
	PW91	3.42	-9.38	1.03
	PBE	3.36	-7.03	1.03
	BP86	3.04	-6.26	1.02
	BLYP	3.51	-10.00	1.02
	B3LYP HF=0.2	1.28	-9.44	1.03
	PBE0 HF=0.25	1.09	-4.25	1.03
	PBE0 HF=0.5	1.18	1.38	1.03
^{39}K				
	PW91	0.14	-189.6	1.14
	PBE	0.13	-182.5	1.13
	BP86	0.30	-185.8	1.14
	BLYP	0.41	-218.5	1.16
	B3LYP HF=0.2	0.11	-134.2	1.10
	PBE0 HF=0.25	0.11	-58.8	1.04
	PBE0 HF=0.5	0.14	-68.6	0.95

^{87}Rb				
	PW91	0.97	-310.0	1.09
	PBE	0.95	-247.7	1.07
	BP86	1.10	-401.5	1.12
	BLYP	1.23	-506.1	1.15
	B3LYP HF=0.2	0.45	-290.0	1.09
	PBE0 HF=0.25	0.25	28.7	0.99
	PBE0 HF=0.5	0.03	310.8	0.91
^{133}Cs				
	PW91	2.68	-706.3	1.12
	PBE	2.45	-712.6	1.12
	BP86	2.15	-788.6	1.13
	BLYP	1.75	-749.8	1.12
	B3LYP HF=0.2	0.82	-447.9	1.07
	PBE0 HF=0.25	1.45	-269.9	1.05
	PBE0 HF=0.5	1.00	156.6	0.97

We find that for the ^{1}H and ^{13}C chemical shifts the inclusion of relativistic corrections (both at the scalar relativistic and spin-orbit coupling level) does not have a significant effect and is well below the expected DFT accuracy. However, for ^{15}N chemical shifts we observe that full relativistic corrections at the spin-orbit coupling level must be considered to reach the expected DFT accuracy of around 5.4 ppm.[18, 83] Note that, inclusion of only scalar relativistic corrections does not significantly affect the ^{15}N chemical shifts (**Figure 4-4** and **Table 4-3**). For the ^{39}K, ^{87}Rb and ^{133}Cs chemical shifts we find that the exclusion of the full relativistic correction leads to the largest observed RMSE for the individual species. For these species we also do not observe a significant improvement upon inclusion of only the scalar relativistic correction. We also note that, for the ^{39}K, ^{87}Rb and ^{133}Cs chemical shifts, without the full relativistic correction, we find a relatively large offset (a) and a slope (b) relatively far from unity (**Figure 4-4** and **Table 4-3**). This agrees very well with the previous studies on similar systems[122-127, 380-382] and we expect the observed trend to increase for even "heavier" nuclei.

Table 4-3. Chemical shift convergence with respect to the used relativistic correction. The correlation is calculated with respect to the chemical shifts calculated at the full relativistic level, including scalar relativistic and spin-orbit coupling effects.

		RMSE / ppm	offset (a) / ppm	slope (b) / ppm
^{1}H				
	Non-relativistic	0.07	-0.42	1.02
	Scalar relativistic	0.07	-0.39	1.02
^{13}C				
	Non-relativistic	0.23	-1.89	1.02
	Scalar relativistic	0.23	-1.91	1.02
^{15}N				
	Non-relativistic	5.79	21.57	0.84
	Scalar relativistic	5.79	22.76	0.84
^{39}K				
	Non-relativistic	1.56	174.9	0.83
	Scalar relativistic	1.33	151.6	0.85
^{87}Rb				
	Non-relativistic	5.13	459.6	0.79
	Scalar relativistic	4.12	293.8	0.84
^{133}Cs				
	Non-relativistic	10.83	242.9	0.81
	Scalar relativistic	9.71	266.9	0.83

4.3 Phase Segregation in Cs-, Rb- and K-Doped Mixed-Cation $(MA)_x(FA)_{1-x}PbI_3$ Hybrid Perovskites

This chapter has been adapted with permission from: Kubicki, D. J.; Prochowicz, D.; Hofstetter, A.; Zakeeruddin, S. M.; Grätzel, M.; Emsley, L., "Phase Segregation in Cs-, Rb-and K-Doped Mixed-Cation $(MA)_x$ $(FA)_{1-x}$ PbI_3 Hybrid Perovskites from Solid-State NMR". *Journal of the American Chemical Society* **2017,** ***139*** (40), 14173-14180. ***(post-print)***

4.3.1 Introduction

Hybrid organic-inorganic multi-cation lead halide perovskites (HOPs) have taken the field of photovoltaics by storm since their first successful application as sensitizers for solar cells.[385] They generate intense interest as a conceivable alternative to traditional silicon solar cells, as they can be processed using various vapor[386]- and solution-based[387-389], techniques. Since the first report, power conversion efficiencies (PCE) have increased from 3.8% to about 22%.[371] Key to this remarkable progress was the notion of alloying structurally similar perovskites into multi-cation and multi-anion lead HOPs.[371]

A generic HOP can be represented by an ABX_3 formula, in which *A* stands for a monovalent cation such as methylammonium, ($CH_3NH_3^+$, MA), formamidinium ($CH_3(NH_2)_2^+$, FA), cesium or rubidium. *A* cations are confined within a cubo-octahedral cage formed by $[BX_3]^-$ octahedra. *B* is typically a divalent metal such as Pb^{2+}, Sn^{2+} or Ge^{2+} and *X* is a halide: I^-, Br^- or Cl^-. Current champion HOP materials, in terms of their photovoltaic performance and light/moisture stability, are double- (MA/FA[363-364, 366], Cs/FA[135, 365, 367-368], Rb/FA[372], K/MA[375]), triple- (Cs/MA/FA[369], Rb/MA/FA[370, 373]) and quadruple-cation (Rb/Cs/MA/FA)[371] lead halide solid alloys with one (I) or two (I, Br) halides. They are all based on FA as the majority cation owing to the fact that the black α-$FAPbI_3$ phase has a bandgap of 1.40 eV, which is close to the Shockley–Queisser limit (1.34 eV), a factor crucial in the design of efficient PV materials.[390] However, the α phase of $FAPbI_3$ is thermodynamically unstable under ambient conditions and it spontaneously transforms into photo-inactive yellow δ-$FAPbI_3$. Incorporation of MA, Cs and Rb was found to alleviate the problem of phase stability, but the consequences reach well beyond that, since devices based on mixed-cation phases consistently exhibit higher open-circuit voltage (V_{OC}), short-circuit current (J_{SC}), fill factor (FF), PCE and long-term stability towards light and moisture.

While several hypotheses have been put forward to explain these results, there is still no satisfactory understanding of the microscopic structure in these mixed-cation systems. For example, powder X-Ray diffraction is currently the method of choice to assess whether the incorporation of an ancillary cation was successful. This is typically inferred from a shift (on the order of 0.05°) of the main reflection of the α-$FAPbI_3$ phase (14.00°) to higher angles, indicative of a decrease in lattice constant, and accompanied by a shift in photoluminescence (PL) spectra.[365, 369] However, diffraction-based methods lack information about the non-crystalline and disordered regions of the sample, and they are not quantitative. When we started this work, solid-state NMR, on the other hand, seemed to be perfectly suited for the task. It had been used in several recent examples to probe perovskites. [391-396] Not only does it provide quantitative information but it is also capable of detecting all species of a given nucleus that are present in the sample, regardless of the degree of crystallinity. For instance, recently Rossini *et al.*[393] had shown that ^{207}Pb NMR chemical shifts and line shapes are a sensitive probe of the halogen coordination in pure and mixed-halogen HOPs. Our group had very recently used solid-state NMR to elucidate microscopic phase composition and segregation in MA/FA HOPs.[397]

Here we show that in Cs/FA solid alloys, cesium is incorporated into the perovskite lattice as Cs^+, and can take up to 15 mol% of the A site. Above this ratio, it separates into a mixture of disordered $CsPbI_3$, and free $CsPbI_3$. Similarly, we confirm incorporation of Cs^+ into the state-of-the-art triple- (Cs/MA/FA) and quadruple-cation (Rb/Cs/MA/FA) PV perovskites. In contrast, we find that Rb^+ is not incorporated into the 3D perovskite lattice at any composition studied here. Rather, it separates into $RbPbI_3$ (in Rb-doped systems with only iodine), mixed cesium-rubidium lead iodide (in Cs- and Rb-doped systems with only iodine) or a mixture of rubidium halides, mixed cesium-rubidium lead iodide and various rubidium lead bromides (in Rb/Cs/MA/FA systems with bromine and iodine). The improved performance of the Rb containing materials is thus not due to incorporation into the main perovskite lattice. We suggest that the performance is improved since the Rb compounds present can potentially act as a passivation layer. In the case of K/MA, pure $MAPbI_3$ is formed, accompanied by unreacted KI.

All above results were obtained for samples prepared by mechanochemistry which has emerged as an appealing method for synthesizing large quantities of high-quality perovskites for PV applications.[398-401] We thus address the question of whether bulk mechanochemically synthesized perovskites are a good representation of the thin films used in PV devices. Comparison of NMR spectra between a bulk mechanochemical triple-cation Cs/MA/FA perovskite and a thin film prepared by spin-coating[388] shows no significant differences between the two materials, validating that bulk mechanochemical perovskites can be used to obtain structural information about newly developed HOP systems.

4.3.2 Methods

Perovskite synthesis and sample preparation.

We focus on the following perovskite materials of practical importance: $Cs_xFA_{1-x}PbI_3$ (x=0.10, 0.15, 0.20, 0.30, abbreviated as "Cs_xFA_{1-x}"); $Cs_{0.10}(MA_{0.17}FA_{0.83})_{0.9}Pb(I_{0.83}Br_{0.17})_3$ ("CsMAFA", prepared according to Saliba *et al.*[369]); $Rb_xFA_{1-x}PbI_3$ (x=0.1, 0.2, abbreviated as "Rb_xFA"); a Rb/Cs/MA/FA/Pb/Br/I material prepared according to Saliba *et al.* ("RbCsMAFA(Br,I)")[371], and $K_{0.10}MA_{0.90}PbI_3$.[375] We also prepared the following materials with only iodine as counterion: $Rb_{0.05}Cs_{0.10}FA_{0.85}PbI_3$, $Rb_{0.05}MA_{0.25}FA_{0.70}PbI_3$, $Rb_{0.05}Cs_{0.10}MA_{0.25}FA_{0.60}PbI_3$, abbreviated respectively as RbCsFA(I), RbMAFA(I) and RbCsMAFA(I). Further, we made the following compounds to use as references: δ-$CsPbI_3$ (yellow), δ-$RbPbI_3$ (yellow); $Cs_{0.5}Rb_{0.5}PbI_3$ (pale yellow) and $RbPb_2Br_5$ (white). We attempted to prepare Rb_4PbBr_6[402] but instead we obtained a mixture of RbBr and an unknown rubidium lead bromide whose pXRD pattern did not correspond to any known Rb/Pb/Br phase in the ICDD database. We designate this composition as "phase X" and report its pXRD pattern and NMR parameters (single Rb site with C_Q=3.4 MHz) in the SI. pXRD patterns of all the materials are given in **Appendix V**.

All materials were prepared by mechanochemistry, as described previously by Prochowicz *et al.*, and annealed at 140 °C for 10 minutes to reproduce the thin-film synthetic procedure.[398, 403] The thin film of CsMAFA was prepared according to the procedure described previously, except an uncoated glass substrate was used instead of FTO-coated glass.[369] Samples were packed into 3.2 mm rotors under inert dry nitrogen atmosphere.

Thin film preparation.

The CsMAFA(Br,I) perovskite precursor solution was prepared according to the previously published recipe.[369] The solution was deposited onto a glass substrate (3.5 cm^2) by spin coating in a two-step program at 1000 and 6000 rpm for 10 and 20 s, respectively. During the second step, 100 µL of chlorobenzene was dripped onto the spinning substrate 10 s prior to the end of the program. The substrates were then annealed at 100 °C for 30 min in a dry box. The films were then scratched off the glass substrates using a razor. 12 glass substrates were used in total (42 cm^2) yielding about 1.5 mg of a solid perovskite which was then immediately transferred into an NMR rotor.

NMR measurements.

Variable-temperature ^{133}Cs (65.6 MHz), ^{87}Rb (163.6 MHz), ^{14}N (32.1 MHz), ^{39}K (23.4 MHz), ^{13}C (125.7 MHz) and ^{1}H (500.0 MHz) NMR spectra were recorded on a Bruker Avance III 11.7 T spectrometer equipped with a 3.2 mm low-temperature CPMAS probe. ^{133}Cs, ^{87}Rb and ^{39}K shifts were referenced to 1 M aqueous solutions of the respective alkali metal chlorides, using solid CsI (δ=271.05 ppm), RbI (δ=177.08 ppm) and KI (δ=59.3 ppm) as secondary references.[404]

^{133}Cs and ^{87}Rb chemical shift calculations.

The perovskite (Cs/Rb/FA)PbI_3 clusters and the reference (Rb/Cs)I clusters were generated as described in **Appendix V**. Chemical shift calculations were performed at DFT level using the GGA BP86[405-406] functional with all-electron TZ2P basis functions (triple-ζ in the valence with two polarization functions) including relativistic effects (up to spin-orbit coupling) with the ZORA[407-409] approximation and the Grimme[206] dispersion correction implemented within the Amsterdam Density Functional (ADF)[383-384] suite.

The calculated chemical shieldings were converted to chemical shifts by a linear correlation.

$$\delta_{exp} = \sigma_{ref} + b\,\sigma_{calc}. \tag{4-1}$$

For the linear correlation only the experimental and calculated chemical shifts of the reference (Cs/Rb)I and the hexagonal (yellow) (Cs/Rb)PbI_3 structures were used, leading to a reference shielding and a slope of σ_{ref} = 2653 , b = -0.79 for Rb and σ_{ref} = 3490, b = -0.54 for Cs. In both cases, we ignored second-order quadrupolar contributions to the shift since they are zero in the cubic compounds (CsI, RbI) and negligible in $CsPbI_3$ (calculated C_Q of 0.4 MHz leading to a shift of <1 ppm) and $RbPbI_3$ (at most 4 ppm given the fitted C_Q of around 2 MHz).

4.3.3 Results and Discussion

Figure 4-5 shows a schematic representation of the crystal structures of the studied materials. The starting point for all solid-alloys investigated in this study is the perfect cubic perovskite structure of α-$FAPbI_3$ (**Figure 4-5a**).[390] Solid alloys can be formed by replacing some FA cations inside the cubo-octahedral cages by MA and conceivably Cs and Rb (**Figure 4-5b**), accompanied by gradual departure from cubic symmetry. Excess Cs^+ and Rb^+ ions can separate into a thermodynamically stable, yellow, non-perovskite (orthorhombic, Pnma space group) phase: δ-$CsPbI_3$ or δ-$RbPbI_3$, respectively (**Figure 4-5c**). We note that to date there is only two single-crystal studies reported on mixed-cation (MA/FA)[410] and (Cs/FA)[411] systems.

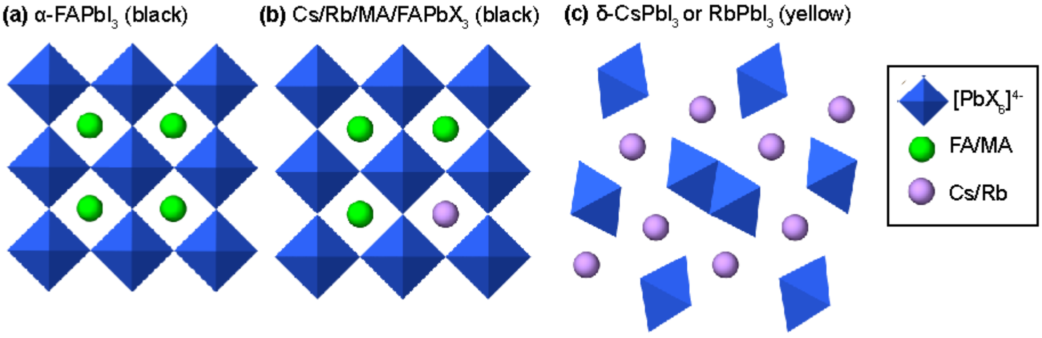


Figure 4-5. Schematic representation of structural motifs investigated in this study: **(a)** black single-cation α-$FAPbI_3$, **(b)** black double- (CsFA, RbFA), triple- (CsMAFA) or quadruple-cation (RbCsMAFA) compositions (X=I, Br), **(c)** yellow non-perovskite δ-$FAPbI_3$,

Cesium phases from ^{133}Cs MAS NMR.

In order to determine cesium incorporation into PV perovskites, we performed ^{133}Cs MAS NMR on the most prominent cesium-containing materials recently reported in the literature (**Figure 4-6**). The spectrum of δ-$CsPbI_3$, (**Figure 4-6a**) contains one relatively narrow (FWHM : ~350 Hz) peak centered at 240 ppm, accompanied by a manifold of spinning sidebands (SSB), spaced by the MAS frequency. The longitudinal relaxation time (T_1) of this species is about 100 s.

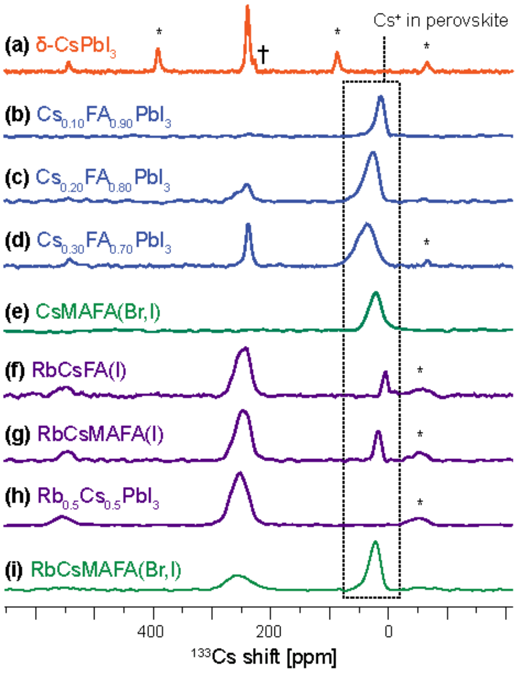


Figure 4-6. Quantitative ^{133}Cs echo-detected MAS spectra of various $(Cs/Rb/MA/FA)Pb(Br/I)_3$ systems at 298 K and a) 10 kHz MAS, b-j) 20 kHz MAS acquired within 1 hour after annealing. Asterisks indicate spinning sidebands and † is a transmitter artefact.

Moving on to the Cs_xFA_{1-x} solid alloys (**Figure 4-6b-e**) one sees a new, much broader peak whose position and linewidth depend on cesium content (shifts: 13, 18, 26 and 37 ppm, FWHM: 1169±21, 858±15, 1477±51 and 2261±51 Hz for Cs mole ratio x=0.10, 0.15, 0.20 and 0.30, respectively). This new species is peculiar in that its ^{133}Cs signal position and relaxation time are a strong function of temperature.

Figure 4-7 shows the temperature dependence of the ^{133}Cs shift and line shapes in $Cs_{0.20}FA_{0.80}$ between 100 and 330 K. The corresponding smooth change in the ^{133}Cs shift in this temperature range covers about 100 ppm, and is accompanied by a change in relaxation time from 26 s (at 298 K) to 3 s (at 103 K). This behavior is consistent with the Cs^+ cation being incorporated into the cubo-octahedral space and interacting strongly with the $[PbI_3]^-$ lattice. The change in relaxation time is caused by the change dynamics of the nearby nuclei, and/or a change in the ^{133}Cs quadrupolar coupling as the lattice changes with temperature. Indeed, upon cooling the lattice undergoes successive first- and second-order displacive phase transitions attributed to gradual freezing of phonon modes associated with the rotational movement of the $[PbI_3]^-$ octahedra.[412-413] The reason for the progressive broadening of the resonances is most likely caused by a distribution of sites with slightly different chemical environments that is created upon the freezing of $[PbI_4]^-$ liberations. Conversely, no such behavior is present in pure δ-$CsPbI_3$ (or CsI) which preserve sharp lines across the whole temperature range, indicating no phase transitions (**Figure 4-19**).

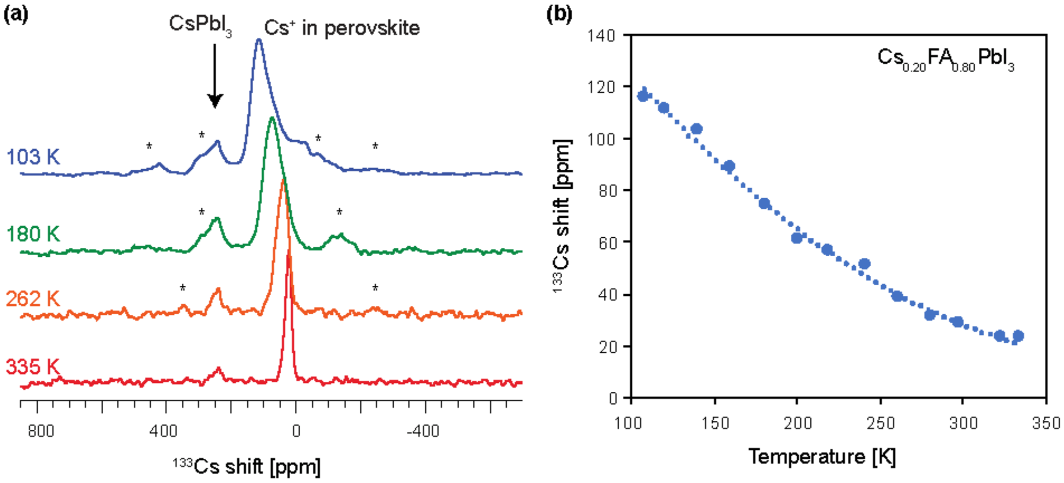


Figure 4-7. **(a)** Variable-temperature solid-state ^{133}Cs MAS NMR spectra of $Cs_{0.20}FA_{0.80}PbI_3$. **(b)** Temperature dependence of the ^{133}Cs shift (measured at the maximum of the most intense peak). Spinning sidebands are marked with asterisks (*).

In attempt to elucidate the change in shifts, we carried out fully-relativistic DFT ^{133}Cs chemical shift calculations for two $FAPbI_3$ lattices in which one FA was replaced by Cs in (a) a perfectly cubic and (b) a tetragonal perovskite lattice arrangement. We have found that distorting the lattice from cubic to tetragonal leads to an increase in ^{133}Cs shift of around the same magnitude as that observed experimentally (**Table 4-9**). That said, this result is only qualitative since the $Cs_{0.20}FA_{0.80}$ lattice, unlike that of $FAPbI_3$, is not perfectly cubic. This comes about because incorporation of cesium leads to lattice distortions, and in turn to reduction in symmetry of the environment in which the FA cation is reorienting. We have previously shown that ^{14}N MAS NMR is very sensitive to such distortions owing to the interaction of its quadrupole moment with the electric field gradient created by the distorted lattice, with higher asymmetry leading to broader ^{14}N spectral envelopes.[397] Cs-induced lattice distortion is indeed clearly evidenced by ^{14}N MAS spectra of the two materials, with cesium incorporation leading to a spectral envelope nearly 4 times broader than that of the pure $FAPbI_3$ phase (**Figure 4-20**).

To corroborate that the signal close to 30 ppm at 298 K originates from Cs^+ incorporated inside the perovskite lattice, we carried out a through-space heteronuclear correlation experiment (HETCOR), which maps all cesium chemical environments that are in the immediate spatial vicinity of any protons (**Figure 4-8**). The experiment was carried out at 100 K to take advantage of the faster proton relaxation at low temperature.[397] The cross-peak can be easily assigned, since there is only one source of protons in the sample, to Cs^+ dipolar coupled to FA. It is thus Cs^+ inside the 3D perovskite lattice, and which correlates with the nearby FA protons.[397]

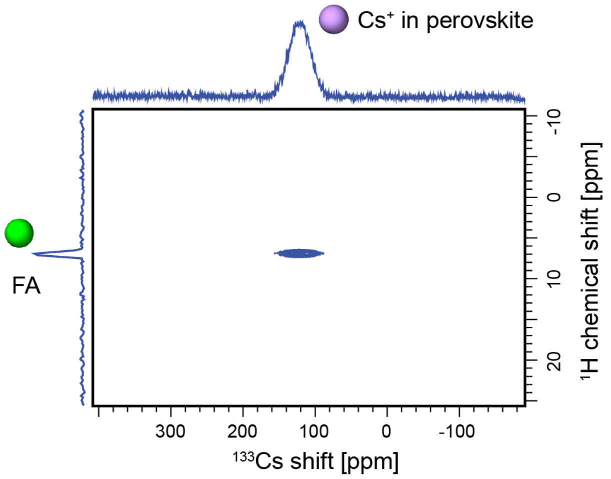


Figure 4-8. A ^{1}H-^{133}Cs heteronuclear through-space correlation experiment (HETCOR) of $Cs_{0.20}FA_{0.80}$ at 100 K and 12 kHz MAS.

Another characteristic feature of the spectra in **Figure 4-6b-e** is the resonance around 240 ppm corresponding to the δ-$CsPbI_3$ phase. In $Cs_{0.10}FA_{0.90}$ and $Cs_{0.15}FA_{0.85}$ it is absent, whereas in $Cs_{0.20}FA_{0.80}$ and $Cs_{0.30}FA_{0.70}$ it is clearly present, confirming phase separation taking place in these systems above 10% doping. Note that this resonance has a slightly broader component shifted to higher values, visible in **Figure 4-6d**. This broadened signal can tentatively be assigned to a disordered interface region between the CsFA alloy and pure δ-$CsPbI_3$.

A comment is in order regarding the stability of Cs_xFA_{1-x} compositions. Photovoltaic parameters measured on devices fabricated using $Cs_{0.15}FA_{0.85}PbI_3$ have been monitored over the course of 14 days and found stable during that period.[135] Nazarenko *et al.* have reported that single crystals of $Cs_xFA_{1-x}PbI_3$ compositions are stable up to 20 days after which time the presence of hexagonal δ-$FAPbI_3$ can be detected.[411] Here we find that mechanochemically prepared Cs_xFA_{1-x} compositions are thermodynamically unstable and give off δ-$CsPbI_3$ over time. For example, the composition denoted $Cs_{0.10}FA_{0.90}$ is, based on the quantitative ^{133}Cs spectrum acquired immediately after annealing, a phase pure perovskite but separates into a mixture of $Cs_{0.07}FA_{0.93}PbI_3$ and $CsPbI_3$ after 24 hours. Similarly, $Cs_{0.20}FA_{0.80}$ after annealing is a mixture of $Cs_{0.16}FA_{0.84}PbI_3$ and $CsPbI_3$ but the same preparation after 5 days contains $Cs_{0.14}FA_{0.86}PbI_3$ and a correspondingly larger amount of $CsPbI_3$. The $Cs_{0.30}FA_{0.70}$ composition is particularly unstable reproducibly yielding a transitory $Cs_{0.23}FA_{0.77}PbI_3$ perovskite (within 30 minutes from annealing) which quickly loses the incorporated cesium in favor of $CsPbI_3$ and becomes $Cs_{0.15}FA_{0.85}PbI_3$ (after 1h), $Cs_{0.08}FA_{0.92}PbI_3$ (after 2h) finally stabilizing as $Cs_{0.03}FA_{0.97}PbI_3$ after 5h. Note that the shortest quantitative spectrum takes 30 minutes to acquire so it possible that in this sample more cesium is transiently incorporated during annealing, leading to lattice instability and, as a consequence, rapid cesium release. We did not further investigate the reasons behind this instability. The fact that its timescale is much faster than that observed in single crystals suggests it may be related to grain boundaries, with smaller crystallites promoting cesium loss from the 3D perovskite lattice. Notably, this process stops at 100 K which indicates its reliance on the lattice phonon modes.

The performance of Cs containing materials continues to increase as loadings increase to 15%, consistent with full cesium incorporation in the $Cs_{0.15}FA_{0.85}$ composition (**Figure 4-6c**).[135] Lee *et al.* have reported enhanced photo- and moisture stability of Cs_xFA_{1-x} solid alloys, which they attributed to stronger interaction between FA and I^- in the perovskite.[365] Other studies have confirmed increased stability both experimentally and theoretically, by rationalizing through entropic stabilization of the cubic α-$FAPbI_3$ structure.[135, 368] Poor stability of the pristine α-$FAPbI_3$ phase at ambient conditions towards humidity, as well as against elevated temperature, has been explained by its propensity to decompose into ammonia and *sym*-triazine.[413] Further, the presence of excess $CsPbI_3$ explains the consistently poorer photovoltaic parameters measured on Cs_xFA_{1-x} devices with x>0.15.[135] It is noteworthy that an opposite effect has been reported for excess PbI_2 which typically led to improved photovoltaic parameters but has been shown to be detrimental to device stability.[414]

Cesium has been shown to improve PV parameters and stability in triple and quadruple-cation compositions in a similar way. **Figures 4-6f** and **4-6j** show ^{133}Cs spectra of two of the currently best performing solid alloys, CsMAFA(Br,I) and RbCsMAFA(Br,I), respectively. In both cases a broad peak of Cs^+ incorporated into the perovskite lattice is present. RbCsMAFA(Br,I) exhibits an additional broad peak (δ=255.4±0.3 ppm, FWHM=2662±55 Hz) making up 47% of the whole amount of cesium in this sample, markedly different in appearance from that of δ-$CsPbI_3$ (δ=239.32±0.03 ppm, FWHM=367±7 Hz). Given the similarity between the hexagonal lattices of δ-$CsPbI_3$ and δ-$RbPbI_3$ we suggest it might belong to a mixed cesium-rubidium lead iodide phase.

This was confirmed by preparing pure $Cs_{0.5}Rb_{0.5}PbI_3$ (**Figure 4-6i**) which yielded a very similar signal (δ=253.2±0.2 ppm, FWHM=2034±27 Hz). We note that the exact shift and linewidth are expected to vary depending on the exact Rb/Cs ratio in such 1D mixed-cation hexagonal phase. To exclude the possibility of this peak being due to a bromine-containing species, we prepared two more Cs/Rb compositions (**Figures 4-6g** and **4-6h**) featuring only iodine as counterion, both of which gave the same resonance (RbCsFA(I): δ=247.3±0.3 ppm, FWHM=1468±66 Hz, RbCsMAFA(I): δ=248.4±0.3 ppm, FWHM=1592±63 Hz), confirming the assignment to $Cs_{0.5}Rb_{0.5}PbI_3$. This finding implies that rubidium competes with cesium incorporation into the perovskite lattice by forming a stable hexagonal mixed Cs/Rb phase. In fact, in the case of pure iodides (RbCsFA(I) and RbCsMAFA(I)) there is more cesium bound in the mixed cesium-rubidium hexagonal lead iodide (92 and 84%, respectively) than there is cesium incorporated into the perovskite (8 and 16%, respectively) (**Figure 4-6g**, **h**). The addition of bromine (in RbCsMAFA(Br,I)) alleviates this effect to certain extent (**Figure 4-6j**).

Rubidium phases from ^{87}Rb MAS NMR

We now investigate the fate of rubidium in rubidium-doped multi-cation perovskites. **Figure 4-9** shows solid-state ^{87}Rb MAS NMR spectra of ten compositions studied here. The spectra of Rb_xFA_{1-x} and RbMAFA(I) perfectly match that of $RbPbI_3$, indicating that the only form in which Rb^+ exists in these systems is a separate $RbPbI_3$ phase. Rb is not incorporated into the MAFA perovskite lattice. This finding challenges previous reports on rubidium incorporation into the perovskite lattice which were based on shifts observed in pXRD diffractograms and photoluminescence spectra.[370, 372-373] A very recent work by Hu *et al.* explains these shifts using EDX in terms of rubidium-induced bromide extraction, which is in excellent agreement with our findings described in the next paragraph.[415] Similarly to the Cs-doped HOPs, Rb-doped materials also exhibit improved long-term stability under high humidity conditions and light irradiation.[370, 372-373] We suggest that this can be explained by passivation of the perovskite phase by a fully inorganic $RbPbI_3$ layer, less prone to decomposition.

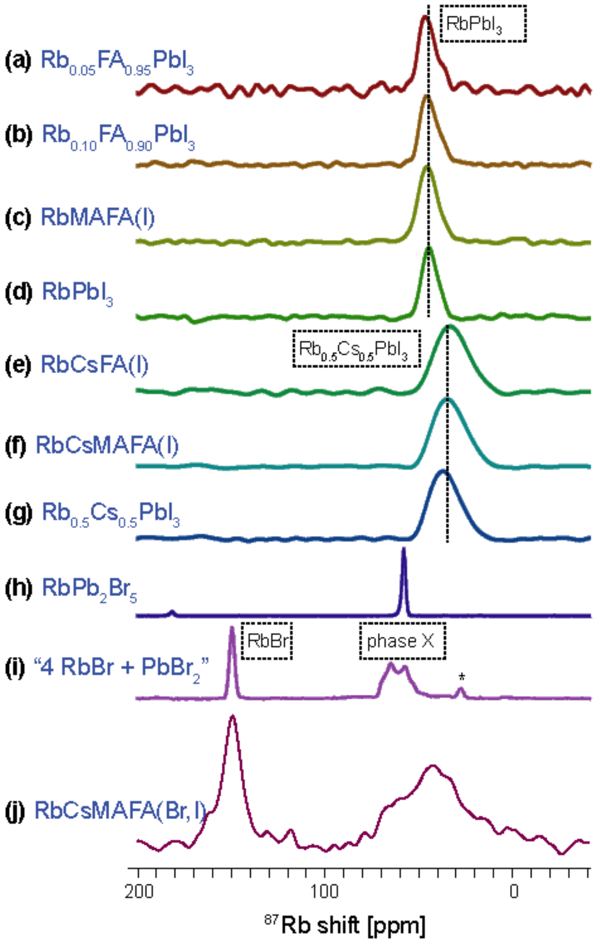


Figure 4-9. 11.7 T Solid-state ^{87}Rb echo-detected MAS (20 kHz, 298 K) spectra of various (Cs/Rb/MA/FA)Pb(Br/I)$_3$ systems. The corresponding 100 K ^{13}C CP MAS spectra of **a-c**, **e-f** and **j** show only one FA signal corresponding to its being in a 3D perovskite environment (**Figure 4-21**).

As mentioned above, we find that cesium tends to form a stable $Cs_{0.5}Rb_{0.5}PbI_3$ phase in the presence of rubidium. This is confirmed here, as the ^{87}Rb spectra of RbCsFA(I) and RbCsMAFA(I) both match that of $Cs_{0.5}Rb_{0.5}PbI_3$ (**Figure 4-9e-g**). To ensure this is not simply a sheer coincidence, we measured the same spectrum at 100 K. If this rubidium species were to be incorporated inside the perovskite lattice one should expect their shift to be strongly temperature dependent, as was the case for cesium (**Figure 4-7**). On the contrary, we observed only a small shift of ~6 ppm, consistent with ordinary lattice shrinkage at low temperatures (**Figure 4-19b,d**).[416] In addition, we carried out a fully-relativistic DFT calculation of the ^{87}Rb shift expected for a rubidium cation incorporated into the α-$FAPbI_3$ lattice, using the known RbI and δ-$RbPbI_3$ shifts as a reference (see the **Appendix V** for details). We obtained a value of -110 ppm (**Table 4-10**), which is very different from the shift observed experimentally (**Figure 4-9e,f,j**).

Figure 4-9j shows a ^{87}Rb MAS spectrum of the state-of-the-art quadruple-cation composition developed by Saliba *et al.*[371] Again, there is no evidence for incorporation of the Rb into the CsMAFA perovskite lattice. In this case, since this composition also contains bromide anions, rubidium can be expected to form both iodide- and bromide-containing species. The spectrum in **Figure 4-9** exhibits a relatively sharp peak at 150 ppm which corresponds to a pure RbBr phase.[404] Pure RbI is expected at 177 ppm[404] and in this sample is not present. That said, rubidium is known to form a continuum of mixed $RbI_{1-x}Br_x$ phases,[417] which explains the distribution of shifts in the region, delimited by the values of pure RbI and RbBr (150-177 ppm). The mixed $RbI_{1-x}Br_x$ phases make up 38% of rubidium content in this sample and are responsible for bromide depletion from the perovskite, the reason behind the previously observed XRD and PL shifts, at the time ascribed to rubidium incorporation into the perovskite lattice.[415] The other, much broader peak centered around 50 ppm can be attributed to a mixture of rubidium lead halides. Its breadth is consistent with the presence of $RbPbI_3$, $Cs_{0.5}Rb_{0.5}PbI_3$ (**Figure 4-9d** and **g**) and "phase X" (**Figure 4-9i**). The presence of $RbPb_2Br_5$ cannot be excluded as its sharp signal is overlapping with the broad peak of RbCsMAFA(Br,I). The only other known rubidium lead bromide is Rb_4PbBr_6[402], and since we did not succeed in synthesizing it by mechanochemistry, its presence in this composition is unlikely. As before, also in this case, the two ^{87}Rb signals in RbCsMAFA(Br,I) do not broaden or shift significantly between 298 and 100 K (**Figure 4-19**), which provides further evidence that these rubidium species are not involved in the displacive phase transition of the perovskite lattice, as was the case for incorporated Cs^+ ions.

The hypothesis that rubidium-rich phases may act as a passivation layer is supported by a recent XPS study which has found unexpectedly high (with respect to a theoretical homogeneous distribution) concentration of Cs and Rb in the 18 nm surface layer of a RbCsMAFA(Br,I) thin film.[418] Taken with the NMR result suggesting the formation of δ-$Cs_{0.5}Rb_{0.5}PbI_3$ it indicates that the mixed rubidium/cesium hexagonal phase has a propensity to form at the top of the perovskite film during solution processing, thereby isolating it from ambient humidity.

In summary, **Table 4-4** rounds up the capacity for incorporation of Cs^+ and Rb^+ into perovskite lattices found here.

Table 4-4. Incorporation capacity of Cs^+ and Rb^+ into $FAPbI_3$-based perovskite lattices.

perovskite	**incorporation into lattice**		**separate phases**
	Cs	**Rb**	
Cs_xFA_{1-x}(I)	✓		**δ-$CsPbI_3$ (for >10% Cs)**
CsMAFA(Br,I)	✓		-
RbFA(I)		✗	**δ-$RbPbI_3$**
RbMAFA(I)		✗	**δ-$CsPbI_3$**
RbCsFA(I)	✓	✗	**δ-$Cs_{0.5}Rb_{0.5}PbI_3$**
RbCsMAFA(I)	✓	✗	**δ-$Cs_{0.5}Rb_{0.5}PbI_3$**
RbCsMAFA(Br,I)	✓	✗	**RbI_xBr_{1-x}** **δ-$Cs_{0.5}Rb_{0.5}PbI_3$** **$Rb_xPb_yBr_z$**

Potassium has an atomic radius similar to that of rubidium, and its incorporation has recently attracted attention as a means of improving PV performance of perovskite materials.[374-375] Here we investigate the simplest case of $K_{0.10}MA_{0.90}PbI_3$. **Figure 4-10a-b** show a comparison between ^{13}C and ^{14}N spectra of $MAPbI_3$ and $K_{0.10}MA_{0.90}PbI_3$. The spectra are, to within error, identical, and indicate that no potassium incorporation into the $MAPbI_3$ lattice takes place. Further, the ^{39}K spectrum of $K_{0.10}MA_{0.90}PbI_3$ acquired over 12 hours shows only the presence of unreacted potassium iodide used as a precursor. Given the similarity of the atomic radii of Rb and K and in light of the above discussion, it is not surprising that no potassium incorporation takes place.

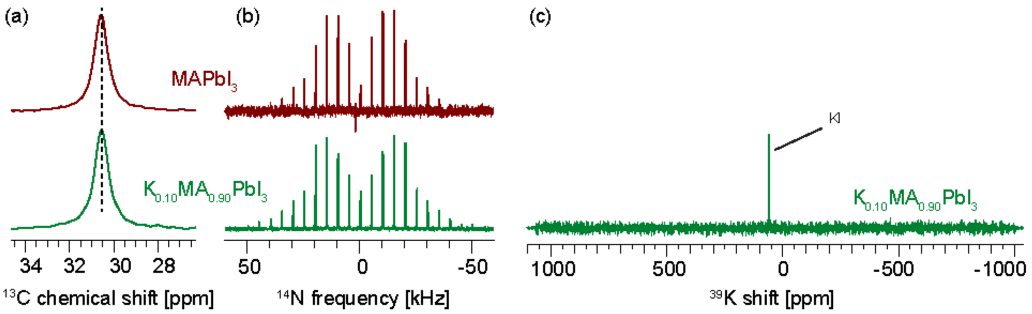


Figure 4-10. (a) Low-temperature (100 K) ^{13}C CP MAS spectra, **(b)** echo-detected ^{14}N MAS spectra at 300 K and 5 kHz MAS of $MAPbI_3$ (top) and $K_{0.10}MA_{0.90}PbI_3$ (bottom), **(c)** echo-detected ^{39}K spectrum of $K_{0.10}MA_{0.90}PbI_3$ at 300 K and 20 kHz MAS (20 s recycle delay, 12 h total acquisition time).

Bulk microstructure matches that of thin films. The bulk perovskites synthesized by means of mechanochemistry studied here are also potentially a convenient source of material for scaling up the production of PV perovskites.[398, 403] However, so far it has been unclear whether their microscopic structure corresponds to that of thin films prepared by solution processing. In order to address this, we prepared a mechanochemical bulk sample of CsMAFA(Br,I) and compared it with a spin-coated CsMAFA(Br,I) thin film.[369]

Figure 4-11 shows solid-state ^{133}Cs, ^{13}C CP and ^{14}N MAS NMR spectra of the two samples. The low-temperature ^{133}Cs spectra are essentially identical and contain one broad peak corresponding to Cs^+ incorporated into the perovskite lattice, analogous to the one observed for $Cs_{0.20}FA_{0.80}$ (**Figure 4-7a**, 103 K). The experiment was carried out at 100 K to take advantage of the shorter recycle delay and improve the overall sensitivity. The low-temperature ^{13}C CP spectra (**Figure 4-11b**) indicate that only the black phase of FA is present in both cases.[397] The two spectra have no significant differences and their appearance corresponds to that of the MAFA system, given for reference at the top of fig. **Figure 4-11b**.

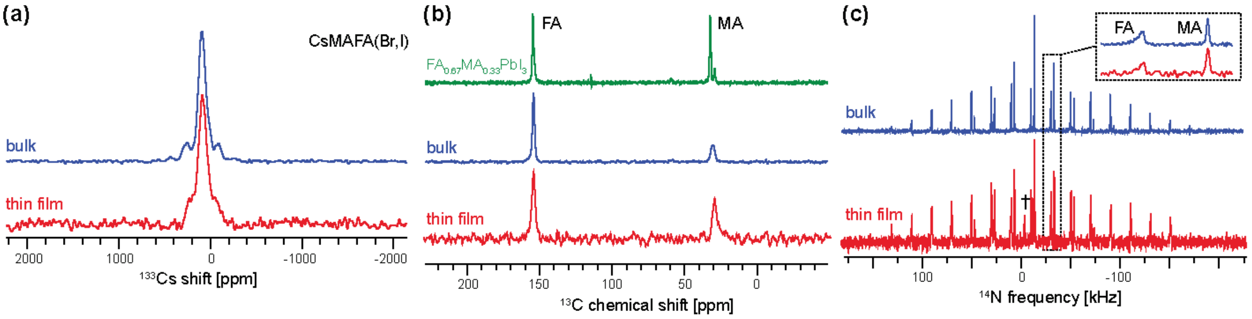


Figure 4-11. Solid-state MAS NMR spectra of CsMAFA(Br,I) in bulk (blue) and prepared as thin film on glass (red). **(a)** Echo-detected ^{133}Cs spectra at 100 K and 12 kHz MAS (**Figure 4-6f** is the corresponding 298 K spectrum of the bulk material), **(b)** ^{13}C CP at 100 K and 12 kHz MAS and **(c)** ^{14}N echo-detected spectra at 298 K and 20 kHz MAS (acquisition times: bulk 20 h, thin film 60 h). The isotropic signal marked "†" most likely comes from traces of DMF used during spin-coating.

We have previously shown that ^{14}N MAS spectra of mixed-cation phases are a sensitive probe of the cation reorientation dynamics which is encoded in the spectral envelope and linewidths.[397] Here, the two ^{14}N spectra (**Figure 4-11c**) again have very similar envelopes and linewidths. However, the observed linewidths are in this case determined by inhomogeneous effects (disorder), as evidenced by the fact that they are not Lorentzian in shape and do not change with increasing the temperature, thus preventing us from extracting quantitative information on cation reorientation. On the other hand, the similarity of the two spectral envelopes indicates that the two cations in both cases reorient in a potential of similar symmetry, pointing to a similar extent of lattice distortion in the two materials.

4.3.4 Conclusion

In summary, we have shown that ^{133}Cs and ^{87}Rb solid-state NMR offers a robust way of identifying cesium and rubidium species in multi-cation perovskite materials relevant to photovoltaics.

In particular, we have found that cesium is readily incorporated into the perovskite lattice of FA-based materials up to around 15 mol%. Above 15 mol% a second $CsPbI_3$ phase is observed. Rubidium, on the other hand, does not form a solid alloy with FA in any of the studied compositions. Rather, it separates into a mixture of rubidium-rich phases ($RbPbI_3$ mixed cesium-rubidium lead iodides, mixture of rubidium halides, various rubidium lead bromides, depending on the exact composition). All these rubidium-rich phases potentially act as a passivation layer for the perovskite material. We have also found that potassium, which has a size similar to rubidium, is not incorporated into the $MAPbI_3$ lattice.

Further, we have shown that the microscopic composition, as probed by 1D ^{133}Cs, ^{13}C and ^{14}N MAS NMR, of a bulk mechanochemical perovskite preparation, here CsMAFA(Br,I), is indistinguishable from that of a thin film prepared using the two-step solution process.

4.3.5 Appendix V

Perovskite synthesis

Perovskite powders were synthesized by grinding the substrates in an electric ball mill (Retsch Ball Mill MM-200, a grinding jar (10 ml) and a ball with ⌀10 mm) for 30 min at 30 Hz. Substrates were packed into the jar inside a glove box under argon. The resulting perovskite powders were annealed at 140 °C for 10 minutes to reproduce the thin-film synthetic procedure.[403]

Table 4-5. Synthesis of mixed-cation and mixed-halide lead perovskites.

Mixed-cation and mixed-halide lead perovskites
MAFA(Br,I) perovskite The double cation mixed-halide perovskite was fabricated according to the previously published procedure.[403] 0.172 g of FAI (1 mmol), 0.507 g of PbI_2 (1.1 mmol), 0.022 g of MABr (0.2 mmol) and 0.073 g of $PbBr_2$ (0.2 mmol) were milled to prepare the MAFA_(Br,I) black powder.
CsMAFA(Br,I) perovskite The triple cation perovskite was fabricated according to the previously published recipe.[369] 0.172 g of FAI (1 mmol), 0.507 g of PbI_2 (1.1 mmol), 0.022 g of MABr (0.2 mmol), 0.080 g of $PbBr_2$ (0.22 mmol) and 0.014 g of CsI (0.055 mmol) were milled to prepare the CsMAFA black powder.
RbCsMAFA(Br,I) perovskite The quadruple cation perovskite was fabricated according to the previously published recipe.[371] 0.172 g of FAI (1 mmol), 0.507 g of PbI_2 (1.1 mmol), 0.022 g of MABr (0.2 mmol), 0.080 g of $PbBr_2$ (0.22 mmol), 0.014 g of CsI (0.055 mmol) and 0.011 g of RbI (0.055 mmol) were milled to prepare the RbCsMAFA black powder.

Table 4-6. Synthesis of mixed-cation lead iodide perovskites

Mixed-cation lead iodide perovskites

FA/Cs perovskite

0.154 g of FAI (0.90 mmol), 0.026 g of CsI (0.10 mmol) and 0.461 g of PbI_2 (1.00 mmol) were mixed to prepare the $(FA)_{0.90}(Cs)_{0.10}PbI_3$ black powder.

0.137 g of FAI (0.80 mmol), 0.052 g of CsI (0.20 mmol) and 0.461 g of PbI_2 (1.00 mmol) were mixed to prepare the $(FA)_{0.80}(Cs)_{0.20}PbI_3$ black powder.

0.120 g of FAI (0.70 mmol), 0.078 g of CsI (0.30 mmol) and 0.461 g of PbI_2 (1.00 mmol) were mixed to prepare the $(FA)_{0.70}(Cs)_{0.30}PbI_3$ black powder.

FA/Rb perovskite

0.154 g of FAI (0.90 mmol), 0.021 g of RbI (0.30 mmol) and 0.461 g (1.00 mmol) of PbI_2 were mixed to prepare the $(FA)_{0.90}(Rb)_{0.10}PbI_3$ black powder.

RbMAFA(I) perovskite

0.039 g of MAI (0.25 mmol), 0.120 g of FAI (0.70 mmol), 0.010 g of RbI (0.05 mmol) and 0.461 g of PbI_2 (1 mmol) were mixed to prepare the $(Rb)_{0.05}(MA)_{0.25}(FA)_{0.70}PbI_3$ black powder.

RbCsFA(I) perovskite

0.146 g of FAI (0.85 mmol), 0.010 g of RbI (0.05 mmol), 0.026 g of CsI (0.10 mmol) and 0.461 g of PbI_2 (1 mmol) were mixed to prepare the $(Rb)_{0.05}(Cs)_{0.10}(FA)_{0.85}PbI_3$ black powder.

RbCsMAFA(I) perovskite

0.039 g of MAI (0.25 mmol), 0.103 g of FAI (0.60 mmol), 0.010 g of RbI (0.05 mmol), 0.026 g of CsI (0.10 mmol) and 0.461 g of PbI_2 (1 mmol) were mixed to prepare the $(Rb)_{0.05}(Cs)_{0.10}(MA)_{0.25}(FA)_{0.60}PbI_3$ black powder.

$K_{0.10}MA_{0.90}PbI_3$ perovskite

0.016 g of KI (0.10 mmol), 0.143 g of MA (0.90 mmol) and 0.461 g (1.00 mmol) of PbI_2 were mixed to prepare the $(K)_{0.10}(MA)_{0.90}PbI_3$ black powder.

Table 4-7. Synthesis of rubidium lead bromides.

Rubidium lead bromides

$RbPb_2Br_5$

0.082 g of RbBr (0.5 mmol) and 0.367 g of $PbBr_2$ (1 mmol) were mixed and annealed at 150°C for 15 min to prepare the $RbPb_2Br_5$ white powder.

Rb_4PbBr_6

0.165 g of RbBr (1 mmol) and 0.091 g of $PbBr_2$ (0.25 mmol) were mixed and annealed at 150°C for 15 min.

$Cs_{0.50}Rb_{0.50}PbI_3$

0.128 g of CsI (0.50 mmol), 0.106 g of RbI (0.50 mmol) and 0.461 g of PbI2 (1 mmol) were mixed to prepare the $Cs_{0.50}Rb_{0.50}$PbI3 alloy.

Powder X-ray Diffraction

Diffractograms were recorded on an X'Pert MPD PRO (Panalytical) diffractometer equipped with a ceramic tube (Cu anode, λ = 1.54060 Å), a secondary graphite (002) monochromator and an RTMS X'Celerator (Panalytical) in an angle range of 2θ = 5° to 40°, by step scanning with a step of 0.02 degree.

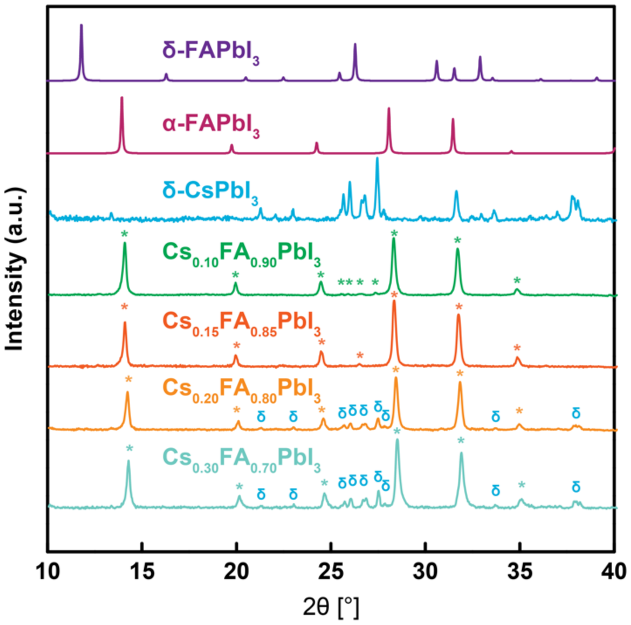


Figure 4-12. pXRD pattern for the Cs$_x$FA$_{1-x}$ compositions. Asterisks (*) indicate the primary phases. Deltas (δ) indicate the phase separated δ-CsPbI3.

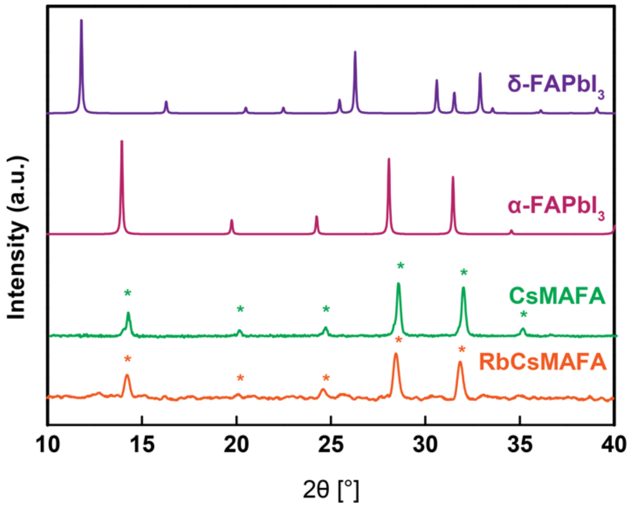


Figure 4-13. pXRD pattern for the CsFAMA(Br,I) and RbCsMAFA(Br,I) compositions.

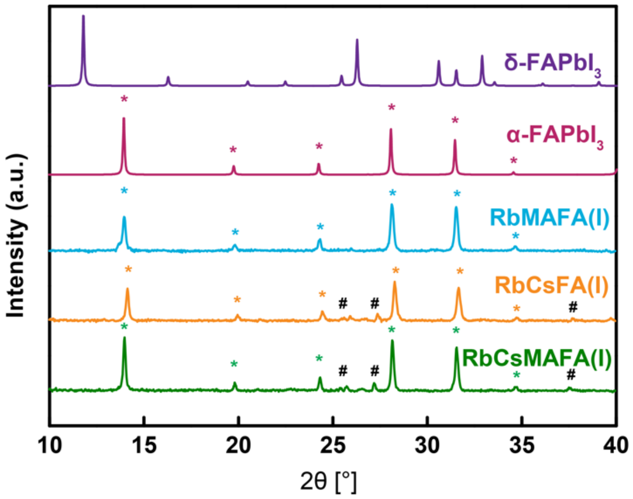


Figure 4-14. pXRD pattern for the RbMAFA(I), RbCsFA(I) and RbCsMAFA(I) compositions. Asterisks (*) indicate the primary phases. Hashes (#) indicate the mixed $Cs_{0.5}Rb_{0.5}PbI_3$ phase.

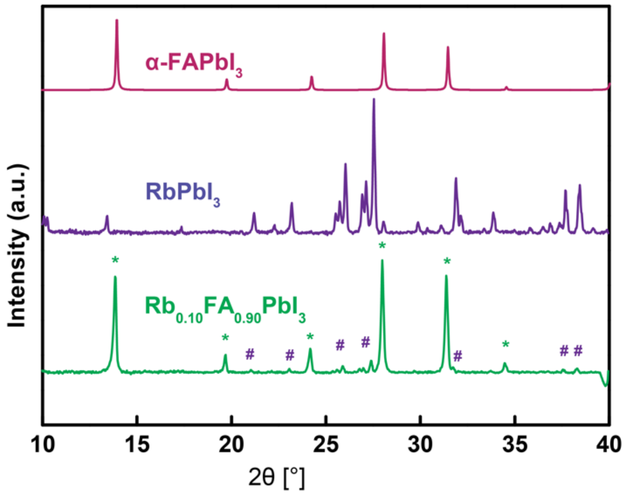


Figure 4-15. pXRD pattern for the $RbPbI_3$ and $Rb_{0.10}FA_{0.90}$ compositions. Asterisks (*) indicate the primary phase, hashes (#) indicate the phase separated $RbPbI_3$.

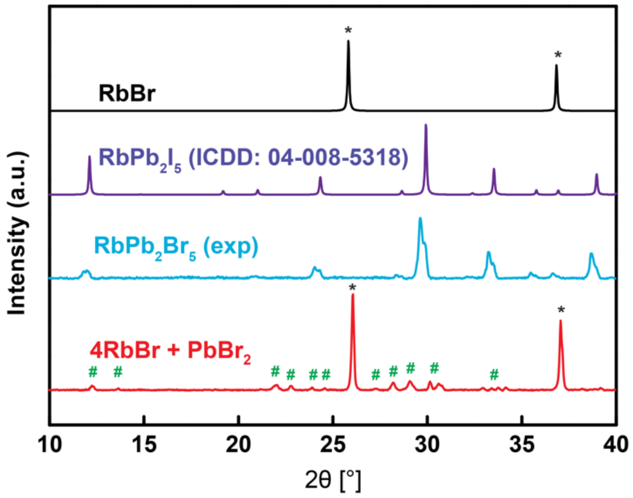


Figure 4-16. pXRD pattern for the rubidium lead bromides. Asterisks (*) indicate RbBr, hashes (#) indicate “phase X”.

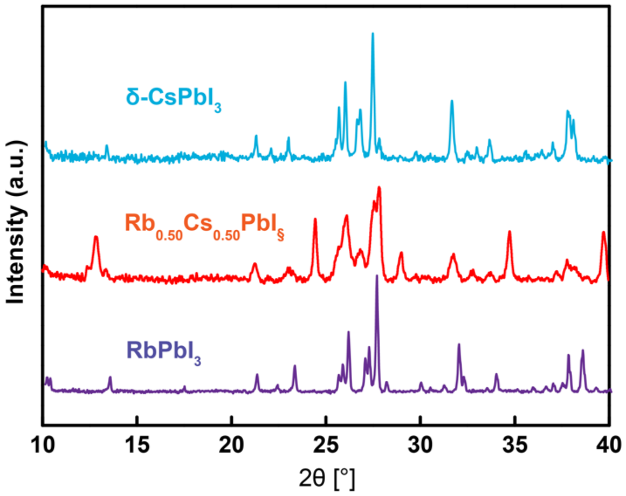


Figure 4-17. pXRD data for $Cs_{0.50}Rb_{0.50}PbI_3$.

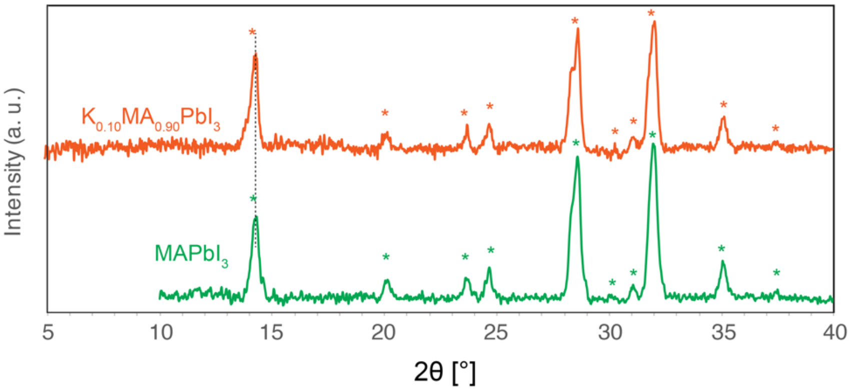


Figure 4-18. pXRD data for $K_{0.10}MA_{0.90}PbI_3$. Asterisks (*) indicate the primary $MAPbI_3$ phase.

NMR measurements

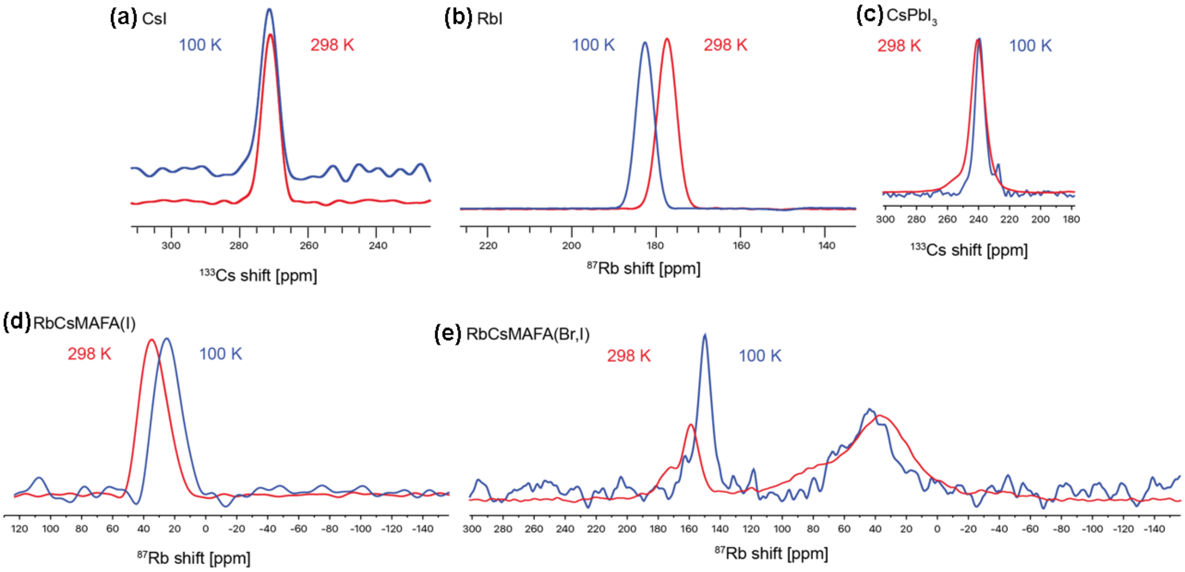


Figure 4-19. A comparison between 298 K and 100 K MAS spectra of **(a)** CsI, **(b)** RbI, **(c)** δ-$CsPbI_3$, **(d)** RbCsMAFA(I), **(e)** RbCsMAFA(Br,I).

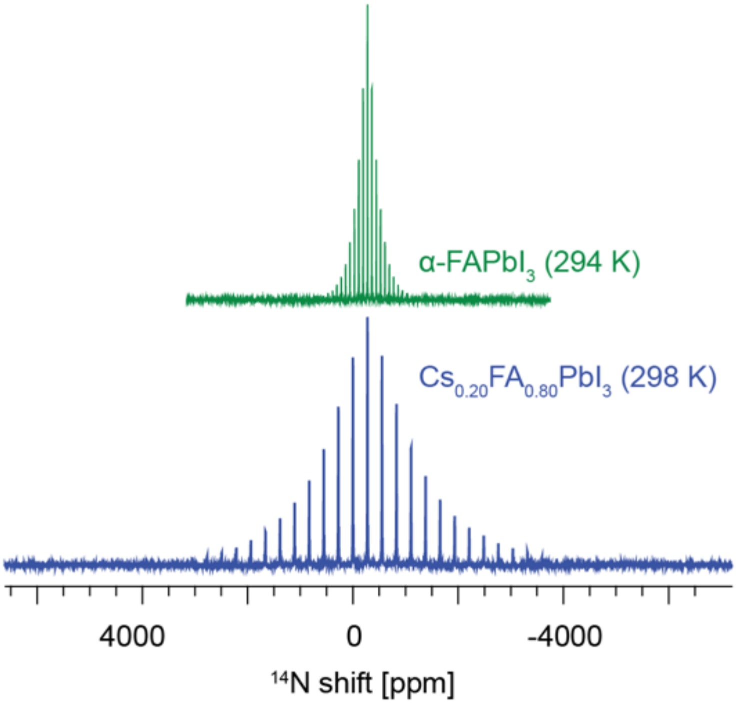


Figure 4-20. ^{14}N MAS spectra of α-FAPbI$_3$ and Cs$_{0.20}$FA.

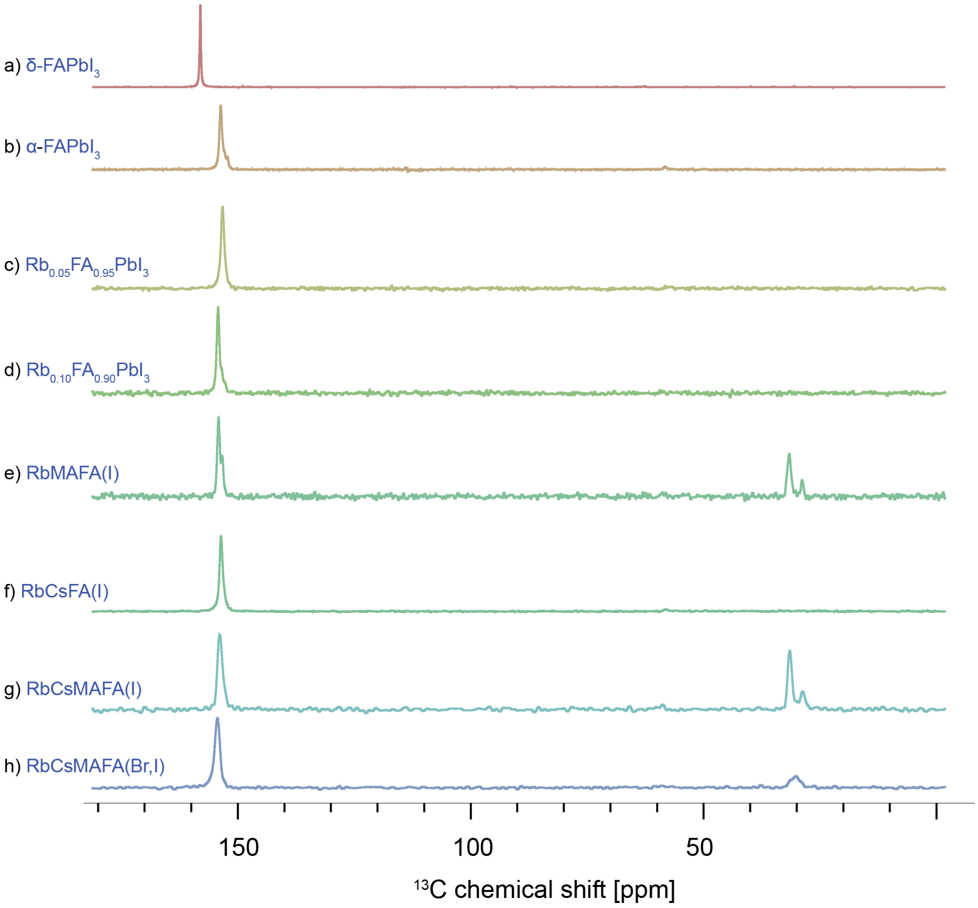


Figure 4-21.. Low-temperature (100 K) ^{13}C CP MAS spectra of the materials studied in this work.

Details of DFT calculations of ^{133}Cs and ^{87}Rb shifts

The crystal structures of CsI,[419] RbI,[420] cubic (black) and hexagonal (yellow) $FAPbI_3$[390], tetragonal (black) $MAPbI_3$[421], hexagonal (yellow) $CsPbI_3$[422] and hexagonal (yellow) $RbPbI_3$[422] were used as a starting point for the clusters. The remaining crystal structures (cubic $RbPbI_3$, cubic $CsPbI_3$ and tetragonal $CsPbI_3$) were generated by replacing the FA/MA cations of the corresponding cubic/tetragonal crystal structure by Cs/Rb cations.

Next, the proton positions in the periodic black $FAPbI_3$ structure as well as the Cs/Rb positions in the substituted periodic structures were optimized using density functional theory (DFT) at the generalized gradient approximation (GGA) level with the PBE[205] functional including relativistic effects (with spin-orbit coupling) and the Grimme[206] dispersion correction within the Quantum Espresso suite.[188] In every calculation a plane-wave maximum cutoff energy of 90 Ry and a 3x3x3 Monkhorst-Pack[423] grid of k-points was employed. Note, that we assume the doping doesn't lead to a change in the Perovskite lattice.

The final clusters were generated as a central cation surrounded by a PbI_3 cage representing the asymmetric unit of the periodic crystal structure. To ensure charge compensation and to represent the solid state, additional cations surrounding the PbI_3 cage were included, resulting in symmetry-adapted clusters[110] containing the non-translational-symmetry elements from the perspective of the central molecule. Generic models of the generated clusters are depicted in **Figure 4-22**. The procedure described above leads to the cluster $Cs_{32}I_{32}$, $Rb_{14}I_{14}$, cubic and tetragonal $Cs_{20}Pb_8I_{36}$, cubic $XFA_{19}Pb_8I_{36}$ and hexagonal $X_{18}Pb_6I_{30}$ (with X = Rb/Cs), see **Table 36**. For the hexagonal $RbPbI_3$ structure the chemical shifts were also calculated with a larger cluster ($Rb_{20}Pb_8I_{36}$). They were within 1 ppm agreement of the shifts calculated for the smaller cluster ($Rb_{18}Pb_6I_{30}$). In general, the setup of the clusters, with respect to level of theory, charge compensation and cluster symmetry, was done according to recent studies on calculations of electronic and magnetic properties of heavy atoms.[123-124, 424-427] All the calculated NMR and EFG parameters are given in **Tables 4-9** and **4-10**.

Table 4-8. Source and modifications of the cluster structures used in the DFT calculations.

Structure name	Original structure	Modifications in periodic system (with Quantum Espresso)[188]	Modifications of cluster (with ADF)[383-384]
$Rb_{32}I_{32}$	RbI[420]	-	-
$Cs_{32}I_{32}$	CsI[419]	-	-
$Cs_{18}Pb_6I_{30}$ (hexagonal)	hexagonal (yellow) $CsPbI_3$[422]	-	-
$Rb_{18}Pb_6I_{30}$ (hexagonal)	hexagonal (yellow) $RbPbI_3$[422]	-	-
$Cs_{20}Pb_8I_{36}$ (cubic)	cubic (black) $FAPbI_3$[390]	Optimization of ^{133}Cs positions	Symmetric Replacement of all FA^+ to Cs^+
$Cs_{20}Pb_8I_{36}$ (tetragonal)	tetragonal (black) $MAPbI_3$[421]	Optimization of ^{87}Cs positions	Symmetric Replacement of all MA^+ to Cs^+
$CsFA_{19}Pb_8I_{36}$ (cubic)	cubic (black) $FAPbI_3$[390]	Optimization of ^{1}H and ^{133}Cs positions	Symmetric Replacement of central FA^+ to Cs^+
$Rb_{20}Pb_8I_{36}$ (cubic)	cubic (black) $FAPbI_3$[390]	Optimization of ^{1}H and ^{87}Rb positions	Symmetric Replacement of central FA^+ to Rb^+

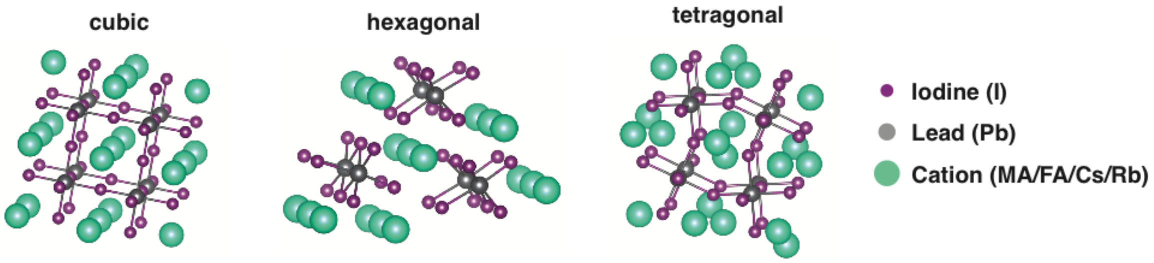


Figure 4-22. Example clusters used in DFT chemical shift calculations.

Table 4-9. ^{133}Cs DFT calculated and experimental magnetic shieldings, chemical shifts and EFG tensor parameters.

Structure	**DFT chemical shielding** [ppm]	**Experimental shifts** [ppm]	**DFT shifts** (σ_{ref}=3490, b=0.54) [ppm]	**DFT shifts** (σ_{ref}=6225, b=1.0) [ppm] (RMSE = 24.13 ppm)	**C_Q** [MHz]	**η**	**V_{zz}** 10^{21} [Vm^{-2}]
$Cs_{32}I_{32}$	5940.4	271.05	282.48	284.6	1.4E-3	2.0E-1	-5.1E-3
$Cs_{18}Pb_6I_{30}$ *(hexagonal)*	5997.7	240.0	251.24	227.3	-4.0E-1	5.0E-1	3.0E-1
$Cs_{20}Pb_8I_{36}$ *(cubic)*	6468.5		-2.99	-243.5	-2.1E-2	1.3E-3	2.5E-2
$Cs_{20}Pb_8I_{36}$ *(tetragonal)*	6093.9		199.3	131.1	-4.4E-1	2.4E-1	5.5E-1
$CsFA_{19}Pb_8I_{36}$ *(cubic)*	6456.7		3.38	231.7	-5.2E-2	7.4E-1	6.5E-2

Table 4-10. ^{87}Rb DFT calculated and experimental magnetic shieldings, chemical shifts and EFG tensor parameters.

Structure	**DFT chemical shielding** [ppm]	**Experimental shifts** [ppm]	**DFT shifts** (σ_{ref}=2653, b=0.79) [ppm]	**DFT shifts** (σ_{ref}=3335, b=1.0) [ppm] (RMSE = 18.56)	**C_Q** [MHz]	**η**	**V_{zz}** 10^{21} [Vm^{-2}]
$Rb_{32}I_{32}$	3140.8	177.08	171.8	194.2	-1.9E-2	1.5E-2	-3.0E-4
$Rb_{18}Pb_6I_{30}$ *(hexagonal)*	3302.0	50.0	44.42	33.0	13.0	7.7E-1	2.0E-1
$Rb\ FA_{19}Pb_8I_{36}$ *(cubic)*	3497.6		-110.1	-162.6	2.2	7.4E-1	3.4E-2

4.4 Phase Segregation in Potassium-Doped Lead Halide Perovskites

This chapter has been adapted with permission from: Kubicki, D. J.; Prochowicz, D.; Hofstetter, A.; Zakeeruddin, S. M.; Grätzel, M.; Emsley, L., "Phase Segregation in Potassium-Doped Lead Halide Perovskites from ^{39}K solid-state NMR at 21.1 T". *Journal of the American Chemical Society* **2018**, 140 (23), 7232-7238. ***(post-print)***

4.4.1 Introduction

The field of photovoltaics based on organic-inorganic lead halide perovskites is thriving (see **Chapter 4.3.1**), owing to their long charge carrier lifetimes and mobilities and the ease with which they can be processed and with which their properties can be tuned.[428-429] The key photovoltaic metrics include open-circuit voltage (V_{OC}), short-circuit current (J_{SC}), fill factor (FF) and power conversion efficiency (PCE). PCE is determined experimentally by measuring photocurrent as a function of the applied bias, and typically plotted in the form of a J-V curve. However, the shape of the J-V curve is often significantly different depending on the scanning direction of the voltage. This effect, known as J-V hysteresis, makes it difficult to determine the correct value for the PCE and to compare intrinsic efficiencies of different perovskite light absorbers, hence it has recently been a subject of intense investigation.[430] The microscopic origins of hysteresis are still a subject of a debate, and include polarization of the perovskite layer caused by ion migration under illumination[431-434] and capacitive charging effects due to carrier trapping in surface states.[435]

Several strategies have been proposed to eliminate hysteresis based on modification of the electron transport layer (ETL) or the perovskite itself. The use of chlorine-capped TiO_2 or SnO_2 nanocrystals,[436-437] and lithium-doped mesoporous TiO_2[438] as an ETL has been shown to significantly reduce hysteresis. Very recently, potassium doping of the perovskite layer has been reported as a straightforward and universal way of alleviating hysteresis, although the reports by Tang *et al.* and Son *et al.* differed in the optimal dopant concentration (5 and 1 mol% relative to Pb, respectively, in a double-cation (FA/MA) mixed-halide (I/Br) material).[379, 439] Shortly after, these findings were contradicted in a study by Jacobsson *et al.*, who showed that potassium doping leads to anomalously large hysteresis (for 6 mol% K^+ in double-cation (FA/MA) and 3 mol% K^+ in triple cation(FA/MA/Cs) mixed-halide (I/Br) materials), compared to the undoped perovskites.[377] Considering these discrepancies, and the current effort put into understanding of the root causes of hysteresis, there is a need to subject new protocols to reduce hysteresis to atomic-level scrutiny.

Further, potassium doping has been reported to improve crystallinity,[375-376] enhance stability,[374] lead to longer charge carrier lifetimes,[375-377] and modify the band gap.[375, 377-378] In these works potassium was suggested to either form interstitial defects,[375, 379] replace A-site cations,[374] or passivate grain boundaries[376]. These conclusions were drawn based on XRD and XPS measurements of perovskite thin films. However, we note that XPS is not a phase-specific method, while XRD can suffer from specimen displacement errors (see **Chapter 4.3.1**). The latter problem occurs when the thin film is not aligned precisely on the focusing circle of the diffractometer, and leads to 2θ errors on the order of 0.04° for displacements as small as 70 µm, which is *larger* than the XRD shifts typically given as evidence for potassium incorporation.[440]

Solid-state NMR, on the other hand, has become the primary tool to study the atomic-level microstructure and phase composition of lead halide perovskites.[391-393, 395-396, 441-442] In particular, we have shown (see **Chapter 4.3**) that cesium and rubidium incorporation, phase separation phenomena and cation dynamics can be easily studied by solid-state NMR using local ^{13}C, ^{14}N, ^{2}H ^{133}Cs and ^{87}Rb nuclear probes.[397, 443-444] We note that ^{207}Pb is a sensitive probe of the halide environment but is far less sensitive to the A-site cation: e.g. with shifts of 1423 ppm in $MAPbI_3$ and 1495 ppm in α-$FAPbI_3$ which given the fwhm of ~250 ppm makes them essentially indistinguishable.[393, 442, 445]

^{39}K (I=3/2, 93.3% abundant) solid-state NMR has been used to characterize a wide range of inorganic materials including potassium salts,[446] oxides, fulleride superconductors,[447] potassium-containing clays,[448] glasses[449] and microporous solids,[450] and a variety of biological,[451-452] organic[453-455] and organometallic[11] systems. Even though the receptivity of ^{39}K is 2.8 times higher than that of ^{13}C, sensitivity is the key challenge in ^{39}K solid-state NMR as the quadrupolar coupling constants can reach up to 5 MHz, leading to central transitions (CT, $-½\leftrightarrow+½$) spanning hundreds of ppm.[446] One of the most efficient strategies to reduce the effect of second-order quadrupolar broadening on the CT is to use high magnetic field strength, B_0, to which the broadening is inversely proportional. Magic angle spinning (MAS) provides another factor of 3 reduction in the CT linewidth.

In terms of structure, higher asymmetry of the potassium site translates to larger quadrupolar coupling constant (C_Q) and thus to broader CT. There exist two experimental regimes that allow one to optimize sensitivity when dealing with quadrupolar nuclei, and they are distinguished based on the relative strengths of C_Q and the radiofrequency (RF) excitation. If the RF strength is much larger (e.g. for symmetric K sites in KI or KBr with C_Q≈0 kHz and a typical RF strength of 30 kHz) the excitation is called nonselective. If applied to a K site with a large C_Q, it would lead to complex interactions between the different transitions and in turn to an intractable pattern of overlapping sidebands. In the case when C_Q dominates significantly (e.g. for asymmetric K sites in $KMnO_4$ with C_Q=1190 kHz), CT-selective excitation can be used to overcome this problem and selectively manipulate the ½⟷+½ transition leading to optimal sensitivity and clean spectra. The results we present here are based on regimes that experimentally proved best on a case-by-case basis.

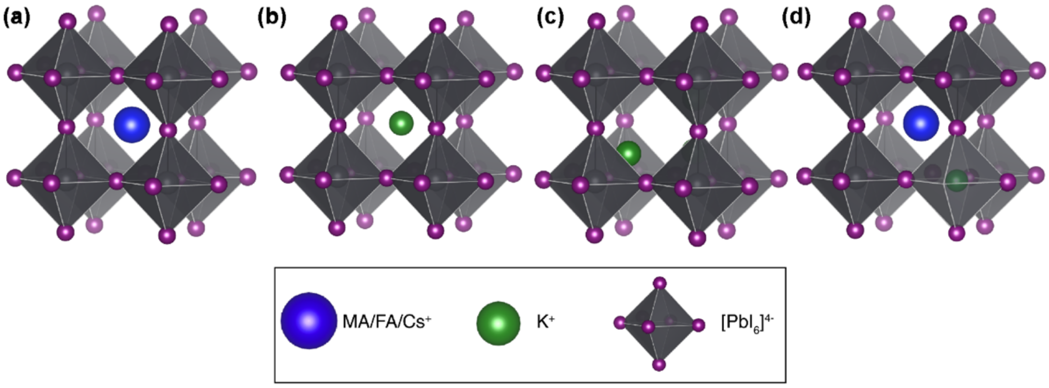


Figure 4-23. Schematic representation of hypothetical scenarios for potassium incorporation into the perovskite lattice: **(a)** parent $APbI_3$ lattice (A=MA, FA, Cs^+), **(b)** A-site replacement, **(c)** interstitial K + A-site vacancy, **(d)** B-site replacement + X-site vacancy.

Figure 4-23. shows three hypothetical ways in which potassium could dope lead halide perovskite lattices. An A-site cation (MA, FA or Cs^+) of the parent lattice (**Figure 4-23a**) could be replaced by K^+ either with preserving its original crystallographic position (**Figure 4-23b**) or by assuming a normally unoccupied site in the perovskite structure and forming an interstitial defect along with an A-site vacancy (**Figure 4-23c**). Potassium could also conceivably replace a B-site cation leading to an X-site vacancy (**Figure 4-23d**). The latter scenario is very unlikely owing to the large difference in electronegativity between potassium (0.8) and lead (1.9) which would lead to ionic rather than coordinate covalent bonds with the iodides and in turn to a collapse of the octahedron.

Here we apply ^{39}K solid-state NMR at 21.1. T to characterize the atomic-level microstructure of phases that are formed when bulk mechanochemical lead halide perovskites [400, 403] are doped with KI. We show that under typical annealing conditions KI partly reacts with the perovskite components to form non-perovskite $KPbI_3$ (for iodide-based materials), a mixture of KI and KBr (in mixed iodide-bromide perovskites) or a non-perovskite mixed-K/Cs lead iodide phase (in compositions containing Cs). We find no evidence of potassium incorporation into the perovskite lattice in any of these compositions, nor in any of the modes shown in **Figure 4-23**, which suggests that the root causes of potassium-induced reduction of J-V hysteresis should be sought elsewhere. These results also explain the XRD and PL peak shifts observed upon doping with KI, which were previously interpreted as evidence for potassium incorporation into lead-halide perovskite phases.

4.4.2 Methods

Materials.

The following materials were used: methylammonium iodide (DyeSol), formamidinium iodide (DyeSol), PbI_2 (TCI, 99.99%), $PbBr_2$ (TCI), KI (abcr, 99.998%), KBr (Sigma, 99.999%), CsI (Sigma, 99.999%).

Perovskite mechanosynthesis.

Starting materials were stored inside a glove box under argon. Perovskite powders were synthesized by grinding the reactants in an electric ball mill (Retsch Ball Mill MM-200 using a grinding jar (10 ml) and a ball (⌀10 mm) for 30 min at 25 Hz. The resulting perovskite powders were annealed at 140 °C (280 °C in the case of $K_{0.075}Cs_{0.925}PbI_2Br$ and $CsPbI_2Br$) for 10 minutes to reproduce the thin-film synthetic procedure. The amounts of reagents taken into the synthesis are given in **Appendix VI**.

NMR measurements.

Solid-state MAS NMR spectra of ^{39}K (23.4 MHz at 11.7 T and 42 MHz at 21.1 T) and ^{133}Cs (52.5 MHz at 9.4 T), were recorded on Bruker Avance III 9.4 T and 11.7 T and Avance IV 21.1 T spectrometers equipped with 3.2 mm CPMAS probes. ^{133}Cs and ^{39}K shifts were referenced to 1 M aqueous solutions of the respective alkali metal chlorides, using solid CsI (δ=271.05 ppm) and KI (δ=59.3 ppm) as secondary references.[404] Typically recycle delays between 3 and 12 seconds were used, based on the measured T_1 values of KI (~9 s), KBr (~8 s) and $KPbI_3$ (~1.4 s). Further experimental details are given in **Appendix VI**.

EFG tensor and NMR chemical shift calculations

The Amsterdam Density Functional (ADF) suite[383-384] was used to perform the EFG tensor and NMR chemical shift calculations within the DFT framework. For the calculations the GGA BP86[405-406] functional including the Grimme dispersion correction[206] and relativistic effects up to spin-orbit couplings within the ZORA[408-409, 456] approximation were used. All-electron triple-ζ basis sets with two polarization functions (TZ2P) were used in the calculations. Both the cluster generation and the EFG tensor calculation are set up analogue to the previous paper by Kubicki *et al.*[443] (see **Appendix VI**) and in accordance with recent computational studies of systems including heavy atoms.[123, 424-426, 457-458]

4.4.3 Results and discussion

In order to estimate whether non-selective or central-transition selective excitation[270] should be optimal to detect potassium inside the perovskite lattice, we carried out fully-relativistic DFT calculations of NMR and EFG parameters for potassium incorporated on the A-site or in the interstitial site (details in **Appendix VI**). The calculations suggest that K^+ sites inside a perovskite lattice should have C_Q values between 68 and 243 kHz, which given the experimental RF strength of 29 kHz (2-8 smaller than the calculated C_Q), points to an intermediate nutation regime where either nonselective and CT-selective excitation might prove more efficient. We thus carried out the measurements using both regimes.

Figure 4-24 shows experimental **(a-l)** and calculated **(m-n)** ^{39}K NMR spectra at 21.1 T and 20 kHz MAS of reference (blue) and KI-doped perovskite phases (black). KI (**Figure 4-24a**) exhibits one narrow (fwhm 45 Hz) peak at 59.3 ppm, consistent with a single symmetric potassium site in a cubic lattice and a C_Q close to 0 kHz (the non-zero C_Q value is due to the presence of defects and finite crystallite sizes which lead to breaking of the perfect cubic point symmetry). Fitting of the KI signal leads to a C_Q of at most 230 kHz for η=0.6 (or less for η<0.6 or η>0.6). An equimolar mixture of KI and PbI_2 yields a single narrow peak at 5.6 ppm (**Figure 4-24b**) and a C_Q of at most 230 kHz for η=0.6, or less for η<0.6 or η>0.6 (the line width might be dominated by inhomogeneous broadening), corresponding to a single potassium site in a highly symmetrical environment. The XRD pattern of this phase (**Figure 4-29**) does not correspond to any of the $KPbI_3$ or K_2PbI_4 patterns deposited in the ICDD database, which suggests it might be a different polymorph than those previously reported. Notably, the available ICDD reference patterns are annotated with low-precision quality marks. We have so far not been able to solve the structure of this phase from powder-XRD, and, and we therefore report the fitted peak positions and the corresponding d-spacings in **Table 4-15**. Since the quantitative ^{39}K spectrum (**Figure 4-39**) as well as XRD data (**Figure 4-30**) indicate there is no unreacted KI or PbI_2 in this phase, it seems reasonable to assume that its stoichiometry corresponds to $KPbI_3$. In what follows we therefore refer to it as $KPbI_3$.

Figure 4-24c shows a single-cation (MA) lead iodide doped with KI, a material reported by P. Zhao *et al.* to exhibit full potassium incorporation into the perovskite lattice based on XRD and PL shifts.[375] W. Zhao *et al.*, on the other hand, suggested that K^+ passivates grain boundaries of the perovskite in KI-doped $MAPbI_3$, unfortunately these authors did not specify the doping level used.[376] The ^{39}K spectrum clearly shows that potassium exists in this material as a mixture of unreacted KI and $KPbI_3$, consistent with our previous preliminary report.[443] This is also the case for a the FA-based (**Figure 4-24d**) and double-cation (MA/FA) lead iodide doped with KI (**Figure 4-24e**). The corresponding XRD patterns are given in **Figures 4-28** and **4-30**.

Figure 4-24g shows a double-cation (MA/FA) mixed-halide (I/Br) perovskite doped with KI, similar to those reported by Tang *et al.*[378, 439] and Son *et al.*[379] as exhibiting potassium incorporation based on XRD, PL and UPS. In this case, KI does partially react with the perovskite components, yielding KBr (**Figure 4-24f**), thus changing the iodide-to-bromide ratio in the final perovskite composition, and in turn the band gap and lattice parameters, relative to the parent material. The change in the I/Br ratio will lead to XRD shifts significantly larger than those observed for purported cation incorporation. For instance, the main perovskite peak shifts from about 14° (2θ) in $MAPbI_3$ to about 15° in $MAPbBr_3$ and takes on intermediate values in mixed-halide $MAPb(I,Br)_3$ compositions, roughly 0.1° per every +10% change in the halide ratio.[459] This effect is even more pronounced for higher order reflections (e.g. 40.4° and 43.2° in $MAPbI_3$ and $MAPbBr_3$, respectively (0.28° per every +10% change in the halide ratio). Interestingly, no potassium-rich lead bromide or mixed bromide-iodide phases are formed (**Figure 4-24h** and **i**) in this case, suggesting higher thermodynamical stability of simple potassium halides under these experimental conditions. There is no incorporation of potassium into the perovskite phase. The corresponding XRD pattern is given in **Figure 4-28f** and only shows the main perovskite phase.

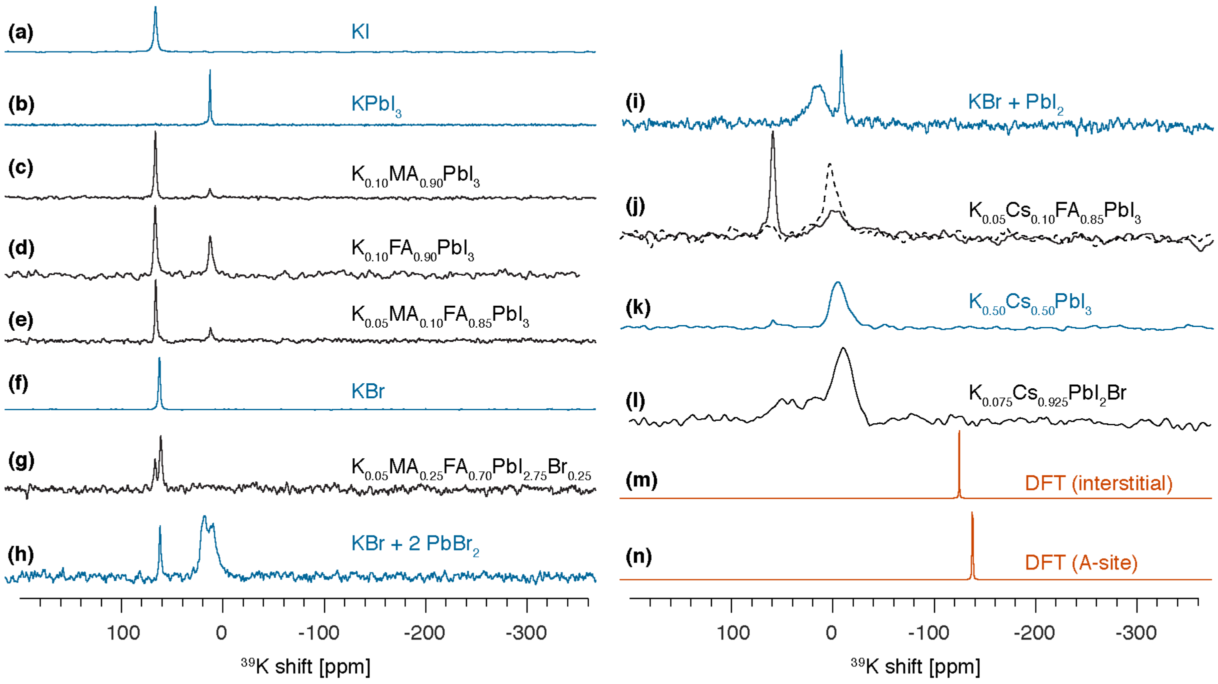


Figure 4-24. ^{39}K solid-state NMR spectra at 21.1 T and 20 kHz MAS and 298 K of reference (blue) and perovskite (black) compositions. The excitation regime used is given in parentheses. **(a)** KI (nonselective), **(b)** $KPbI_3$ (CT selective), **(c)** $K_{0.10}MA_{0.90}PbI_3$ (nonselective), **(d)** $K_{0.10}FA_{0.90}PbI_3$ (nonselective), **(e)** $K_{0.05}MA_{0.10}FA_{0.85}PbI_3$ (nonselective), **(f)** KBr (nonselective), **(g)** "KMAFAPb(I,Br)" (nonselective, see the **Appendix VI** for the exact stoichiometry), **(h)** KBr + $2PbBr_2$ (CT selective), (i) KBr + PbI_2 (CT selective), **(j)** $K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ (solid line: nonselective, dashed line: CT-selective) **(k)** $K_{0.50}Cs_{0.50}PbI_3$ (CT selective), **(l)** $K_{0.075}Cs_{0.925}PbI_2Br$ (CT selective). Spectra simulated using parameters from DFT: **(m)** for K^+ in an interstitial position (structure "h" in **Table 4-20**), **(n)** for K^+ at A-site (structure "g" in **Table 4-20**). The CT-selective spectra for compositions in panels **c**, **e** and **g** are given in **Figure 4-36**. Apodization parameters (leading to slightly different apparent fwhm for KI signals in different spectra) are given in **Table 4-17**.

Phase segregation in potassium and cesium doped perovskites.

Mixed-cation perovskites containing Cs and FA doped with KI have also been reported.[379, 460] We illustrate this case using a double-cation (Cs/FA) lead iodide doped with 5 mol% of KI (**Figure 4-24j**). The ^{39}K spectrum of this material shows two peaks, one corresponding to unreacted KI (its larger apparent width in this case is only due to apodization applied during processing) and another one, significantly broader (δ=-2 ppm, fwhm about 700 Hz) and shifted to high-field. In a CT-selective spectrum the position and shape of this peak change slightly (δ=-4 ppm, fwhm about 500 Hz) suggesting it has several components with different C_Q values. We hypothesize it might correspond to a mixed Cs/K non-perovskite lead iodide, since similar mixed K/Rb and Cs/Rb phases are known.[443, 461] We confirm this hypothesis by preparing a series of $K_xCs_{1-x}PbI_3$ (x=0, 0.1, 0.5, 0.9, 1.0) phases (**Figure 4-25**). The ^{39}K spectra of these phases (**Figure 4-25a-d**) strongly depend on the K/Cs ratio with the signal broadening and shifting to the right for decreasing K/Cs ratios. An analogous trend is observed in the ^{133}Cs spectra (**Figure 4-25g-j**) as well as XRD patterns (**Figure 123**) of this series of mixed K/Cs phases. The broad component in the ^{39}K spectrum of $K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ (**Figure 4-24j** and **Figure 4-25e**) matches well the shift exhibited by these mixed Cs/K-rich phases. For instance, the spectrum of $K_{0.50}Cs_{0.50}PbI_3$ (**Figure 4-24k** and **Figure 4-25c**) exhibits a very similar signal (δ=-2 ppm, fwhm about 700 Hz).

The hypothesis of phase segregation into secondary Cs/K-rich phases was further confirmed by acquiring a ^{133}Cs spectrum of this perovskite material, which shows a peak from Cs^+ inside the perovskite lattice (**Figure 4-25k**, dashed box) and a second, broader peak from a non-perovskite Cs-rich δ phase.[443] We have recently shown that at 10 mol% Cs^+ is fully incorporated into the α-$FAPbI_3$ lattice (**Figure 4-25**).[443] The presence of these secondary phases itself for 10 mol% Cs^+ doping, as well as the resemblance of their resonance to those of the $K_xCs_{1-x}PbI_3$ phases (**Figure 4-25g-j**) corroborates the formation of non-perovskite Cs/K-rich lead iodide δ phases. This in turn decreases the amount of Cs incorporated into the perovskite lattice relative to the perovskite undoped with KI. This, again, is expected to change XRD and PL shifts. Quantification of the ^{133}Cs spectrum of $K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ indicates that 32% of the Cs^+ is incorporated into the perovskite while 68% forms separate Cs/K-rich lead iodide phases. The XRD pattern of $K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ is given in **Figure 4-28d.**

Figure 4-24l shows an all-inorganic $K_{0.075}Cs_{0.925}PbI_2Br$ composition reported by Nam *et al.* to feature potassium incorporation into the perovskite lattice based on XRD and XPS.[374] The ^{39}K spectrum shows that this is not the case. As previously, mostly a mixed-Cs/K lead halide phase (δ=-9 ppm, fwhm about 900 Hz) is formed along with other potassium-rich lead iodide-bromide phases. The ^{133}Cs spectrum of this material (**Figure 4-25m**, solid line) contains a component corresponding to the parent $CsPbI_2Br$ perovskite (**Figure 4-25m**, dashed line) as well as a second broad component similar to $K_{0.50}Cs_{0.50}PbI_3$. In this case the secondary phase can also conceivably contain bromine. The XRD patterns of these two materials are given in **Figure 4-32d-e**. We note that phase segregation into potassium- and cesium-rich lead iodide phases could not be studied by ^{207}Pb NMR owing to insufficient resolution (**Figure 4-37**).

Finally, the DFT results suggest a shift in the range between -119 and -143 ppm and fwhm of at most 110 Hz (if limited by the C_Q) for potassium incorporated into the perovskite lattice (**Figure 4-24m** and **n**, details in the **Appendix VI**). No such signals were found in the experimental spectra.

Estimation of the ^{39}K detection limit.

In order to ensure that the experiments are capable of detecting the small amounts of potassium present in the materials, we carried out a measurement on $KHCO_3$, a compound with a relatively large, well-defined C_Q value of 1490 kHz (**Figure 4-26a**) and a T_1 comparable to that of $KPbI_3$ (**Table 4-16**). Its C_Q is between 3.9 to 213 times larger than the C_Q values predicted by DFT for K^+ incorporated into the perovskite lattice (**Table 4-20**).The measurement was performed using 1.2 mg (12 µmol) of $KHCO_3$ which is comparable to the amount of potassium present in the $K_{0.10}MA_{0.90}PbI_3$ and $K_{0.10}FA_{0.90}PbI_3$ samples in fig. **Figure 4-24c** and **d** (~12 µmoles of K^+ inside a rotor, total sample mass ~75 mg) and using the same recycle delay of 3 s. The resulting spectrum had a signal-to-noise ratio of 6 (higher resolution, visible quadrupolar pattern, **Figure 4-26b**) or 11 (processed with a matched filter of 1 kHz, **Figure 4-26c**) after 20 hours, confirming that any potassium environment with a C_Q comparable or lower should also be readily detected.

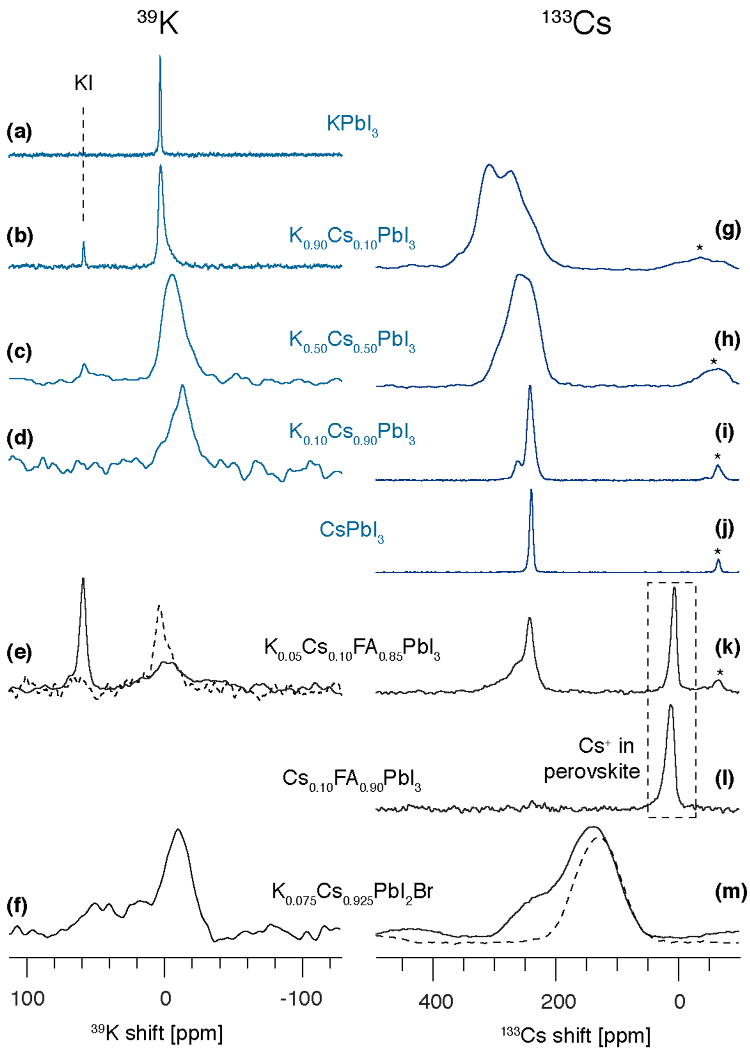


Figure 4-25. ^{39}K (at 21.1 T, a-f) and ^{133}Cs (at 11.7 T, g-m) solid-state NMR spectra at 20 kHz MAS and 298 K of $K_xCs_{1-x}PbI_3$ phases: **(a)** x=1 (nonselective), **(b,g)** x=0.9 (^{39}K nonselective), **(c,h)** x=0.5 (^{39}K CT selective), **(d,i)** x=0.1 (^{39}K CT selective), **(j)** x=0, and perovskite compositions: **(e,k)** $K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ (solid line: ^{39}K nonselective, dashed line: ^{39}K CT-selective), **(f,m)** $K_{0.075}Cs_{0.925}PbI_2Br$ (^{39}K CT selective), dashed line: $CsPbI_2Br$, **(l)** $Cs_{0.10}FA_{0.90}PbI_3$. Asterisks indicate spinning sidebands. The dashed box indicates signals from Cs$^+$ inside the perovskite lattice.

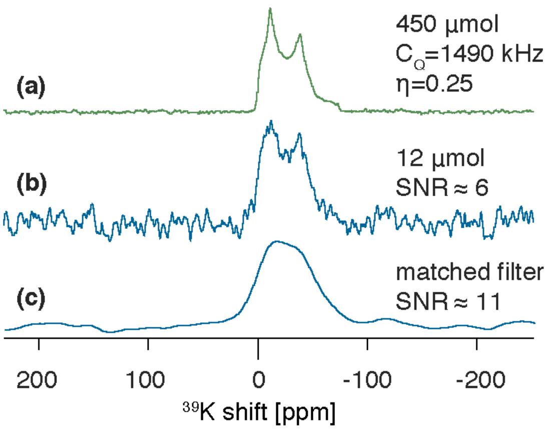


Figure 4-26. ^{39}K solid-state NMR spectra at 21.1 T, 20 kHz MAS and 298 K of **(a)** bulk sample of $KHCO_3$ (45 mg, 450 µmol), recycle delay: 1 s, acquisition time: 17 min., Lorentzian apodization of 50 Hz **(b)** 1.2 mg (12 µmol) of $KHCO_3$, topped with 30 mg of TiO_2 to ensure stable spinning, recycle delay: 3 s, acquisition time: 20 h., Lorentzian apodization of 50 Hz, **(c)** as **(b)** but with Lorentzian apodization of 1 kHz (matched filter) to maximize the SNR.

Comparison between the mechanochemical and solution synthetic route.

We have previously shown that ^{13}C, ^{14}N and ^{133}Cs NMR spectra of mechanoperovskites are essentially indistinguishable from those prepared as thin films by spin-coating from solution.[443-444] In the case of ^{39}K NMR the amount that could be recovered from thin films (~1 mg/15 films) would not be sufficient to achieve satisfactory sensitivity (typical perovskite sample masses in our study are about 75 mg). That said, we have prepared bulk $K_{0.10}MA_{0.90}PbI_3$ and $K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ perovskites using well-established solution processing (dissolution of precursors in a 4:1 v/v mixture of DMF:DMSO, followed by solvent evaporation and vacuum drying, details in the **Appendix VI**).[462] We have found no major differences in XRD patterns (**Figures 4-30d,e** and **4-31d,e**) and ^{133}Cs and ^{39}K spectra (**Figure 4-27**) between the samples prepared using the solution- and solid-state synthetic route. The main difference is visible in the ^{39}K spectra whereby the signals in the solution-prepared materials are significantly broadened but not shifted (**Figure 4-27b,d**) compared to those from the mechanoperovskites (**Figure 4-27a,d**). This is likely due to the remaining strongly coordinated solvent forming Lewis base adducts[463-464] with the $[PbI_6]^{4-}$ sublattice, which we found impossible to remove using vacuum drying at 120 °C. The quantitative ^{1}H MAS NMR spectrum of the solution-prepared $K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ material indicates that after drying the solvent amounts to about 4% of protons in the sample (**Figure 4-38**). The approximate shift of these broad ^{39}K signals spans a range similar to that observed in the mechanochemical analogues. The quantitative ^{133}Cs spectrum of the solution-processed $K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ (**Figure 4-27f**) exhibits signals from the analogous potassium-rich K/Cs lead iodide phases as in the mechanochemical sample, (**Figure 4-27e**) although the relative amounts of the secondary phases to the Cs^+ incorporated into the perovskite are different (mechanochemical: 68% as secondary phases, 32% incorporated, solution-processed: 25% as secondary phases, 75% incorporated).

This highlights that the solution-based and mechanochemical routes qualitatively lead to similar phases although the local atomic environment and the resulting spectral appearance can be complicated by the presence of residual solvents and their coordination to the secondary phases. For comparison, this effect was not appreciable in the case of pure-phase perovskites with cesium[443] and guanidinium[444] incorporation on thin films. As regards particle size, morphology and crystallinity, there are no significant differences between mechanoperovskites, perovskites made using solution processing (present study) and perovskite thin films (typical XRD fwhm <0.3° 2θ for the main perovskite reflections, apparent particle size between 200 and 500 nm, as measured by SEM).[398]

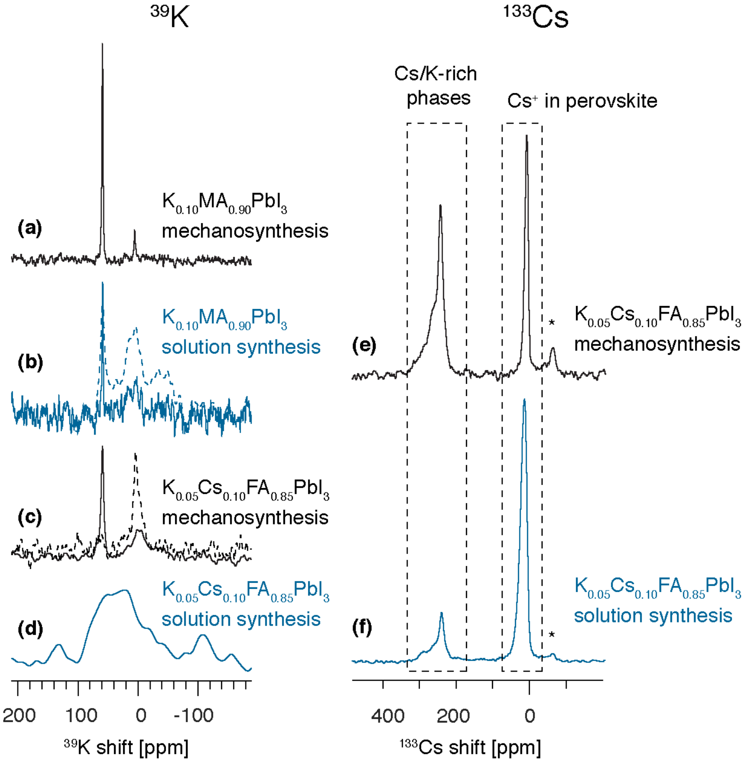


Figure 4-27. ^{39}K (at 21.1 T, **a-d**) and ^{133}Cs (at 11.7 T, **e-f**) solid-state NMR spectra at 20 kHz MAS and 298 K of **(a)** $K_{0.10}MA_{0.90}PbI_3$ (nonselective, mechanochemical), **(b)** $K_{0.10}MA_{0.90}PbI_3$ (nonselective, solution-processed), solid and dashed lines: Lorentzian apodization of 50 Hz and 300 Hz, respectively, **(c)** $K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ (mechanochemical, solid line: nonselective, dashed line: CT-selective), **(d)** $K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ (CT-selective, solution-processed). Quantitative (recycle delay of 450 s) ^{133}Cs spectra of $K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ prepared by **(e)** mechanosynthesis and **(f)** in solution. Asterisks indicate spinning sidebands.

Other possible reasons for the effect of potassium on hysteresis.

Suppression of J-V hysteresis could conceivably originate from K-doping of the mesoporous TiO_2 scaffold in a full PV device, similar to the effect observed upon Li-doping.[438] To verify this hypothesis we mechanochemically prepared samples of TiO_2 activated at 300 °C mixed with anhydrous KI in a 10:1 molar ratio. The CT-selective ^{39}K MAS NMR spectrum of this mixture acquired at 11.7 T and 20 kHz MAS over the course of 117 hours exhibits only a peak from the unreacted KI (**Figure 4-35**) indicating that no reaction took place. A study by Abdi-Jalebi *et al.* appeared when our manuscript was under review that corroborated our conclusion of no potassium incorporation.[465] The authors suggested that potassium preferentially occupies surface sites on perovskite grains. Since the concentration of surface atoms in typical microcrystalline solids is on the order of 10 µmol/g[327] and only a small fraction of these (surface defects) is suggested to be passivated with K^+, it can easily be estimated that the amount of such K^+ sites in a typical NMR sample would be on the order of nanomoles, three orders of magnitude below the detection limit in the present study. Dynamic Nuclear Polarization (DNP) studies are underway to increase sensitivity and lower the detection limit of ^{39}K to further explore the microscopic effect of potassium addition to different components of perovskite-based photovoltaic devices.

4.4.4 Conclusion

In conclusion, we have investigated a number of perovskite materials that have been recently reported to exhibit superior photovoltaic performance after doping with potassium iodide. ^{39}K solid-state NMR shows that K^+ is not incorporated into the perovskite lattice of these materials. Rather, it exists as a mixture of unreacted KI and $KPbI_3$ (in MA and MA/FA lead iodides), KBr (in MA/FA mixed iodide/bromide perovskites) or non-perovskite mixed-Cs/K lead iodide phases (in cesium-containing perovskites). We have found no qualitative differences between materials prepared by solid-state and solution-based synthetic routes. The formation of these secondary non-perovskite phases leads to changes in the composition of the parent perovskite material, and in turn to shifts in diffraction patterns and PL, XPS and UPS spectra. This highlights the essential need for atomic-level characterization of photovoltaic perovskite materials developed through new doping strategies.

4.4.5 Appendix VI

Perovskite synthesis

Table 4-11. Perovskites prepared by the solution route.

Perovskites prepared by the solution route
Polycrystalline powder of the $K_{10}MA_{90}PbI_3$ compositions was synthesized as follows. 0.016 g of KI (0.10 mmol), 0.143 g of MA·HI (0.90 mmol) and 0.461 g of PbI_2 (1.00 mmol) were dissolved in 1 ml of DMF/DMSO mixture (4:1, v:v), and then drop-cast on a petri dish and heated at 120°C in air. The resulting powder was scratched from the glass and dried under vacuum at 120°C for 12 h. For the synthesis of $K_{0.05}Cs_{0.1}FA_{0.85}PbI_3$ powder, the same procedure was followed using 0.008 g of KI (0.05 mmol), 0.026 g of CsI (0.10 mmol), 0.146 g of FAI (0.85 mmol) and 0.461 g (1.00 mmol) of PbI_2.

Table 4-12. Mixed-cation and mixed-halide lead perovskites.

Mixed-cation and mixed-halide lead perovskites
$CsPbI_2Br$ 0.106 g of CsBr (0.50 mmol) and 0.23 g of PbI_2 (0.50 mmol) were mixed to prepare the $CsPbI_2Br$ powder.
$K_{0.075}Cs_{0.925}PbI_2Br$ 0.012 g of KI (0.075 mmol), 0.239 g of CsI (0.925 mmol), 0.183 g of $PbBr_2$ (0.50 mmol) and 0.23 g of PbI_2 (0.50 mmol) were mixed to prepare the $K_{0.075}Cs_{0.925}PbI_2Br$ powder
KMAFA(Br,I) The triple cation perovskite was fabricated according to the previously published recipe using KI instead of CsI.[443] 0.172 g of FA·HI (1 mmol), 0.507 g of PbI_2 (1.1 mmol), 0.022 g of MA·HBr (0.2 mmol), 0.080 g of $PbBr_2$ (0.22 mmol) and 0.009 g of KI (0.055 mmol) were milled to prepare the KMAFA black powder.

Table 4-13. Synthesis of mixed-cation lead iodide perovskites.

Mixed-cation lead iodide perovskites
$K_{0.10}MA_{0.90}PbI_3$ 0.016 g of KI (0.10 mmol), 0.143 g of MA·HI (0.90 mmol) and 0.461 g of PbI_2 (1.00 mmol) were mixed to prepare the $K_{0.10}MA_{0.90}PbI_3$ powder.
$K_{0.10}FA_{0.90}PbI_3$ 0.016 g of KI (0.10 mmol), 0.154 g of FA·HI (0.90 mmol) and 0.461 g of PbI_2 (1.00 mmol) were mixed to prepare the $K_{0.10}FA_{0.90}PbI_3$ powder.
$K_{0.05}MA_{0.1}FA_{0.85}PbI_3$ 0.008 g of KI (0.05 mmol), 0.016 g of MA·HI (0.10 mmol), 0.146 g of FAI (0.85 mmol) and 0.461 g of PbI_2 (1.00 mmol) were mixed to prepare the $K_{0.05}MA_{0.1}FA_{0.85}PbI_3$ powder.
$K_{0.05}Cs_{0.1}FA_{0.85}PbI_3$ 0.008 g of KI (0.05 mmol), 0.026 g of CsI (0.10 mmol), 0.146 g of FA·HI (0.85 mmol) and 0.461 g (1.00 mmol) of PbI_2 were mixed to prepare the $K_{0.05}Cs_{0.1}FA_{0.85}PbI_3$ powder.
$K_{0.1}Cs_{0.9}PbI_3$ 0.017 g of KI (0.10 mmol), 0.232 g of CsI (0.90 mmol) and 0.461 g of PbI_2 (1.00 mmol) were mixed to prepare the $K_{0.1}Cs_{0.9}PbI_3$ alloy.
$K_{0.5}Cs_{0.5}PbI_3$ 0.083 g of KI (0.50 mmol), 0.129 g of CsI (0.50 mmol) and 0.461 g of PbI_2 (1.00 mmol) were mixed to prepare the $K_{0.5}Cs_{0.5}PbI_3$ alloy.
$K_{0.9}Cs_{0.1}PbI_3$ 0.148 g of KI (0.90 mmol), 0.026 g of CsI (0.10 mmol) and 0.461 g of PbI_2 (1.00 mmol) were mixed to prepare the $K_{0.9}Cs_{0.1}PbI_3$ alloy.

Table 4-14. Synthesis of potassium and cesium lead iodides and bromides.

Potassium and cesium lead iodides and bromides
$KPbI_3$ *perovskite* 0.083 g of KI (0.50 mmol) and 0.23 g (0.50 mmol) of PbI_2 were mixed to prepare the $KPbI_3$ powder.
KPb_2Br_5 0.059 g of KBr (0.50 mmol) and 0.367 g of $PbBr_2$ (1.00 mmol) were mixed to prepare the KPb_2Br_5 powder.
$KPbI_2Br$ 0.059 g of KI (1.00 mmol), 0.23 g of PbI_2 (0.50 mmol) and 0.183 g of $PbBr_2$ (0.50 mmol) were mixed to prepare the $KPbI_2Br$ powder.
$CsPbI_3$ 0.260 g of CsI (1.00 mmol) and 0.461 g of PbI_2 (1.00 mmol) were mixed to prepare $CsPbI_3$.

XRD patterns

Diffractograms were recorded on an X'Pert MPD PRO (Panalytical) diffractometer equipped with a ceramic tube (Cu anode, λ = 1.54060 Å), a secondary graphite (002) monochromator and an RTMS X'Celerator (Panalytical) in an angle range of 2θ = 5° to 40°, by step scanning with a step of 0.02 degree.

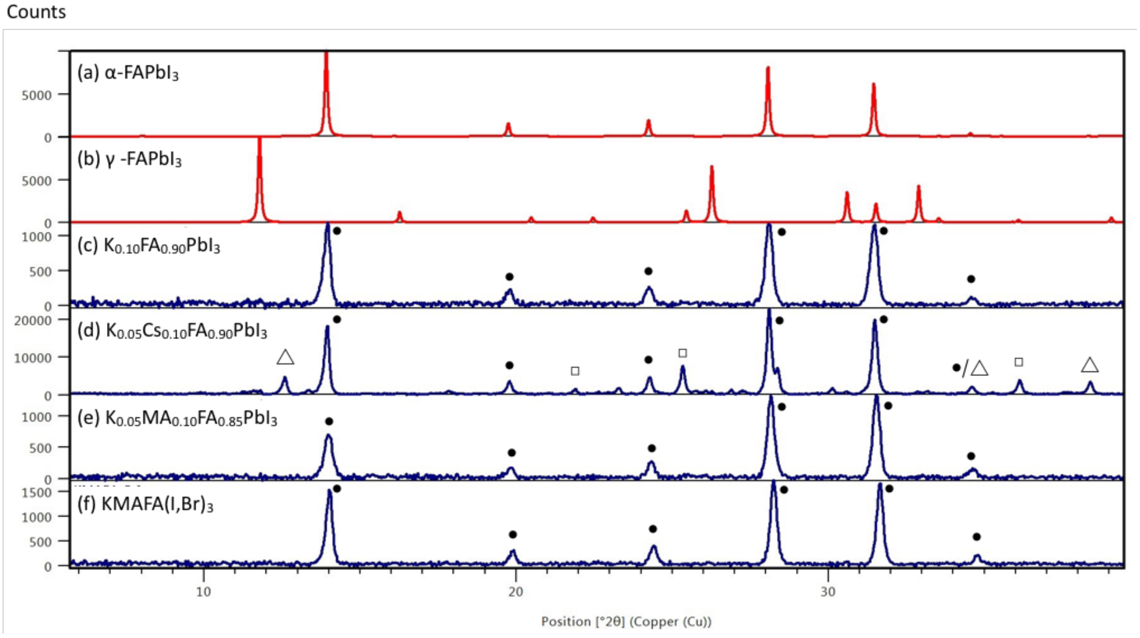


Figure 4-28. XRD patterns of the materials reported above. Simulated patterns: **(a)** α-FAPbI$_3$ (black 3D perovskite) **(b)** γ-FAPbI$_3$, (yellow, hexagonal non-perovskite phase). Experimental patterns of mechanochemical perovskite preparations: **(c)** K$_{0.10}$FA$_{0.90}$PbI$_3$, **(d)** K$_{0.05}$Cs$_{0.10}$FA$_{0.90}$PbI$_3$, **(e)** K$_{0.05}$MA$_{0.10}$FA$_{0.85}$PbI$_3$, **(f)** KMAFA(I,Br)$_3$. •, ∆, □ and indicate the main perovskite phase, PbI$_2$ and KI peaks, respectively. For K$_{0.10}$FA$_{0.90}$PbI$_3$ there is no measurable shift of the main perovskite peaks with respect to the undoped α-FAPbI$_3$.

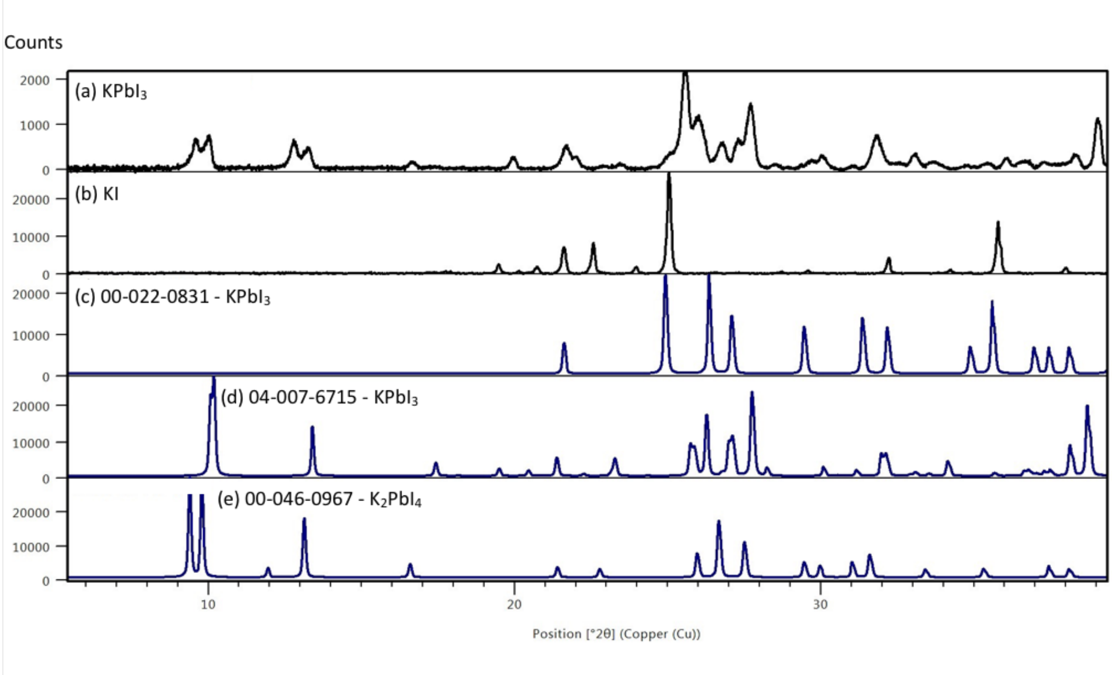


Figure 4-29. Experimental XRD pattern of the **(a)** mechanochemical annealed KI:PbI$_2$ (1:1 mol/mol) material ("KPbI$_3$"), **(b)** KI. ICDD database reference patterns (the numbers are ICDD database reference codes): **(c)** 00-022-0831, KPbI$_3$ (ICDD quality mark: low precision), **(d)** 04-007-6715, KPbI$_3$ (ICDD quality mark: prototype), 00-046-0967, K$_2$PbI$_4$ **(e)** (ICDD quality mark: low precision).

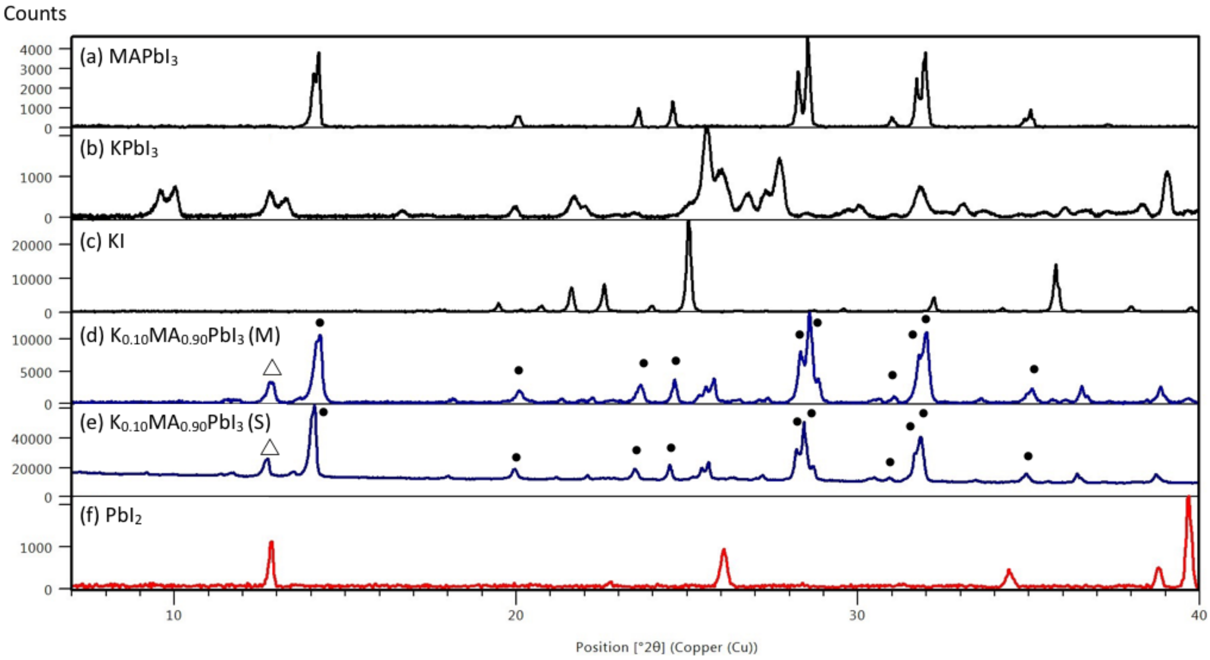


Figure 4-30. Experimental XRD patterns of **(a)** mechanochemical $MAPbI_3$,[398] **(b)** mechanochemical, annealed KI:PbI_2 (1:1 mol/mol) material ("$KPbI_3$"), **(c)** KI, **(d)** mechanochemical ("M") $K_{0.10}MA_{0.90}PbI_3$, **(e)** $K_{0.10}MA_{0.90}PbI_3$ prepared by the solution ("S") route, **(f)** PbI_2. • and △ indicate the main perovskite phase and PbI_2 peaks, respectively. For $K_{0.10}MA_{0.90}PbI_3$ there is no measurable shift of the main perovskite peaks with respect to the undoped $MAPbI_3$.

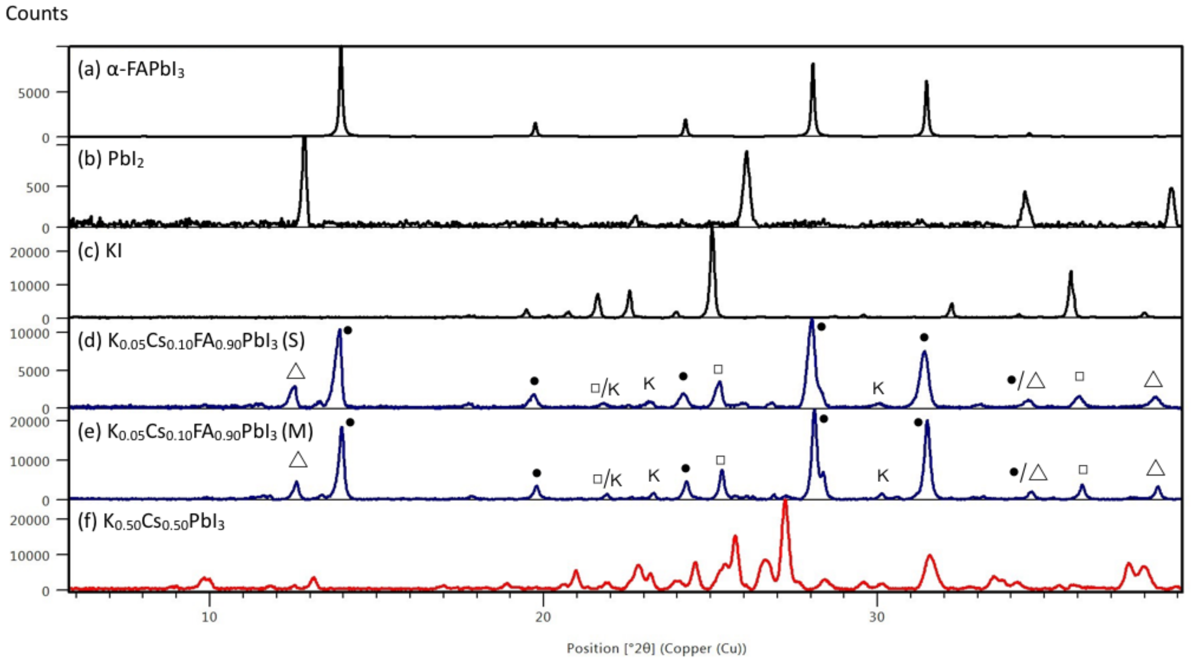


Figure 4-31. (a) Simulated XRD pattern of α-$FAPbI_3$ (black 3D perovskite). Experimental XRD patterns of **(b)** PbI_2, **(c)** KI, **(d)** mechanochemical ("M") $K_{0.05}Cs_{0.10}FA_{0.90}PbI_3$, **(e)** $K_{0.05}Cs_{0.10}FA_{0.90}PbI_3$ prepared by the solution ("S") route, (f) mechanochemical $K_{0.50}Cs_{0.50}PbI_3$. •, △, □ and κ indicate the main perovskite phase, PbI_2, KI and a K-rich non-perovskite phase peaks, respectively.

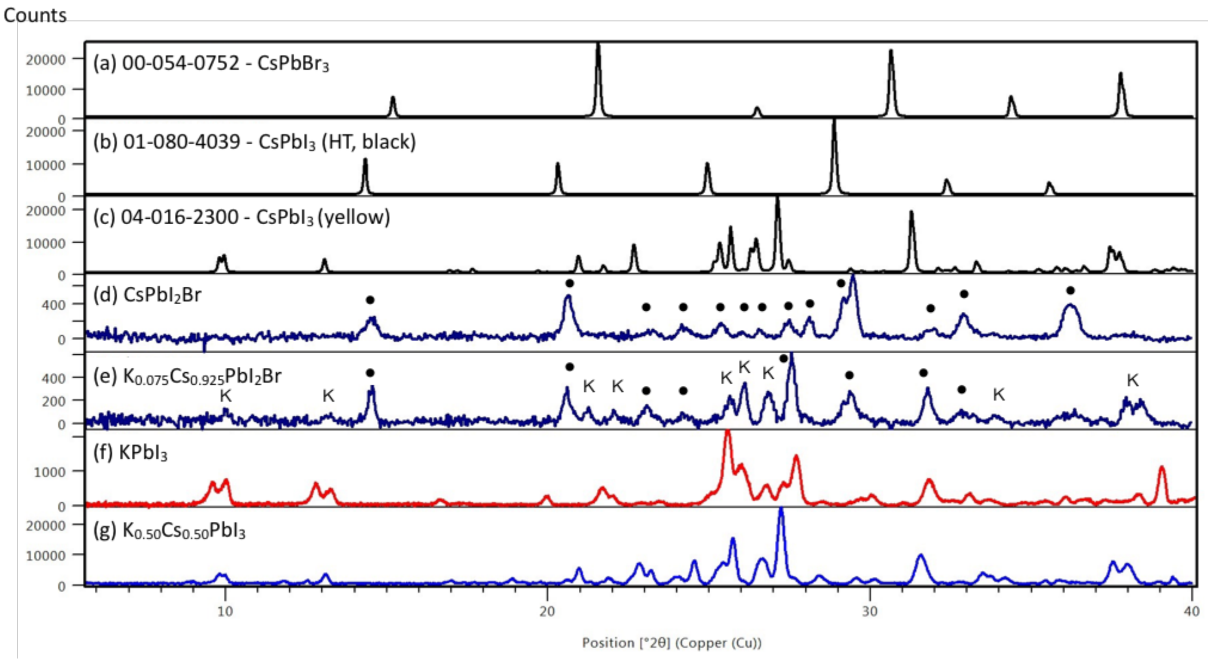


Figure 4-32. ICDD database reference patterns (the numbers are ICDD database reference codes): **(a)** 00-054-0752 - $CsPbBr_3$, **(b)** 01-080-4039 - $CsPbI_3$ (black, high-temperature phase), **(c)** 04-016-2300 - $CsPbI_3$ (yellow, room-temperature phase). Experimental XRD patterns of mechanochemical perovskite preparations **(d)** $CsPbBrI_2Br$, **(e)** $K_{0.075}Cs_{0.975}PbBrI_2Br$, **(f)** $KPbI_3$, (g) $K_{0.50}Cs_{0.50}PbI_3$. • and κ indicate the main perovskite phase and a K-rich non-perovskite phase peaks, respectively. In the case of $K_{0.075}Cs_{0.925}PbI_2Br$ there is no appreciable shift with respect to $CsPbI_2Br$ either (beside a shift of +0.05° 2θ for the 29.2° peak after K^+ doping which can be caused by **(a)** a change in I/Br ratio, **(b)** specimen displacement errors, as discussed in the manuscript)

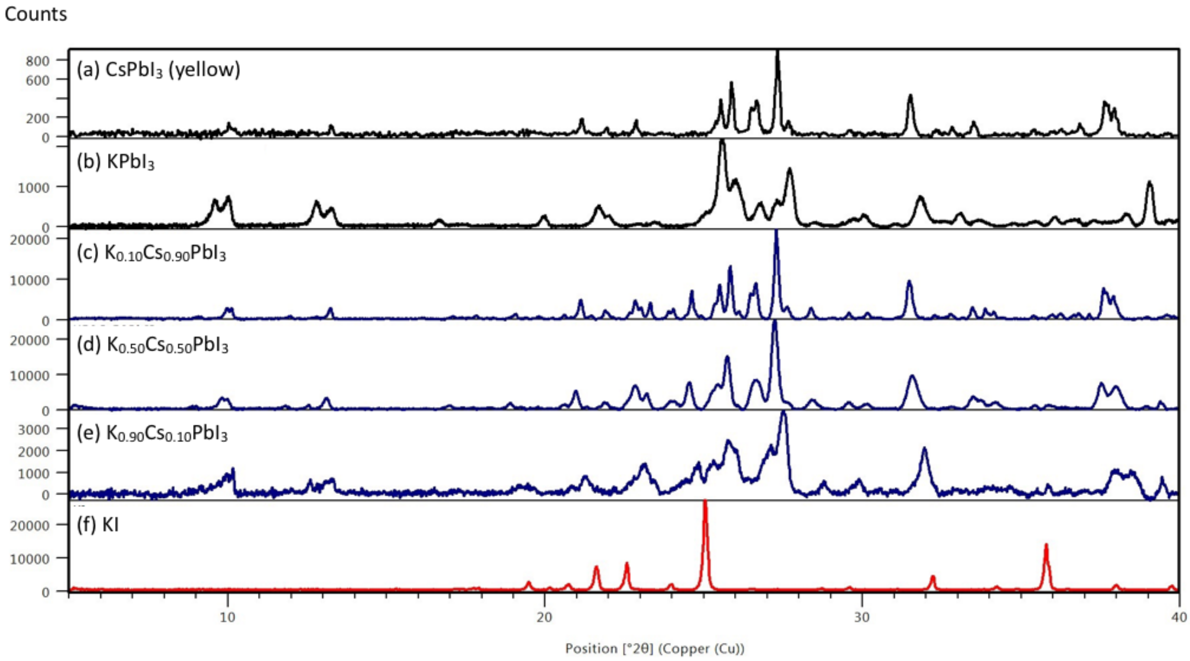


Figure 4-33. Experimental XRD patterns of mechanochemical **(a)** $CsPbI_3$ (yellow, room-temperature phase), **(b)** $KPbI_3$, **(c)** $K_{0.10}Cs_{0.90}PbI_3$, **(d)** $K_{0.50}Cs_{0.50}PbI_3$, **(e)** $K_{0.90}Cs_{0.10}PbI_3$, **(f)** KI.

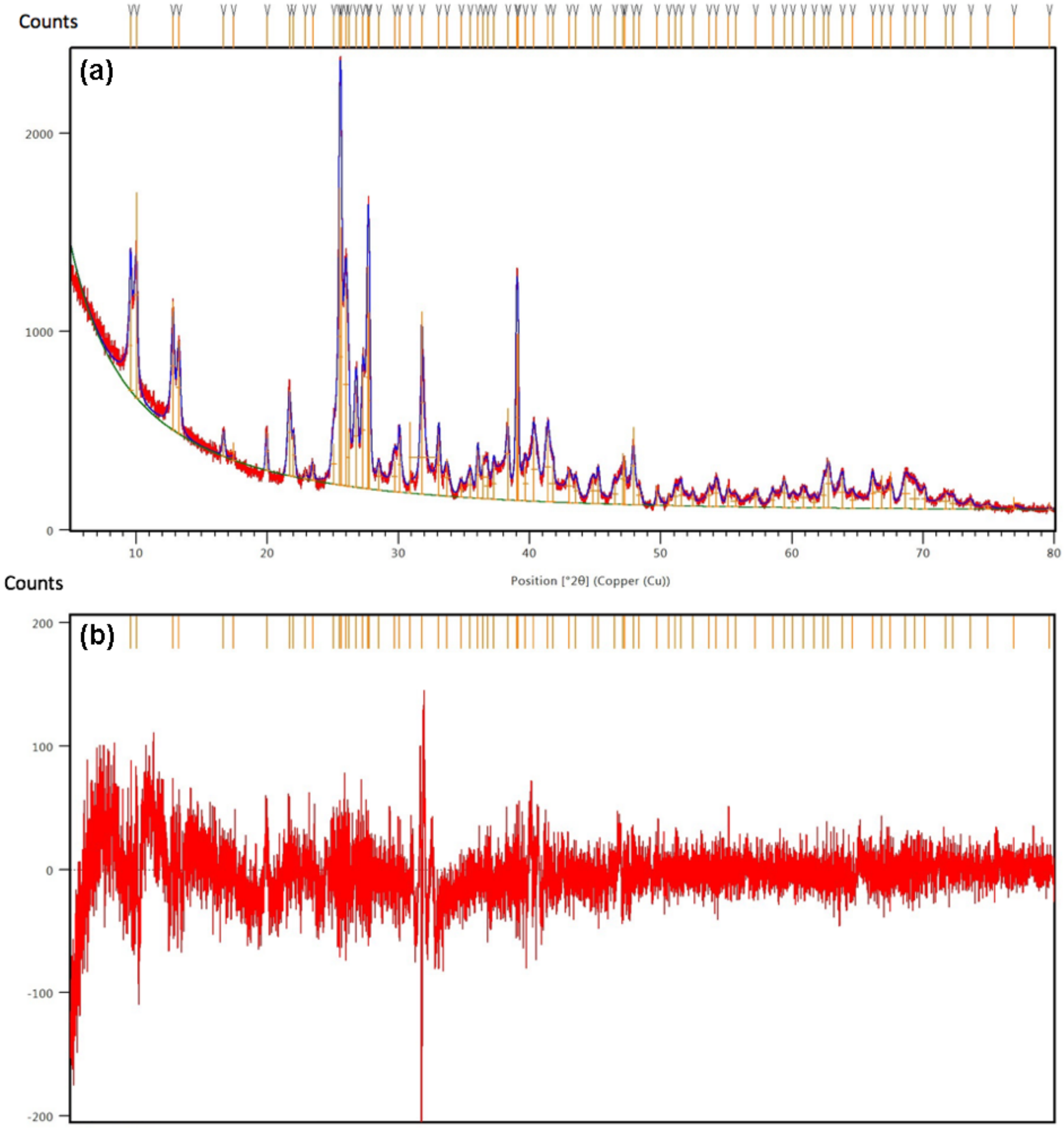


Figure 4-34. (a) Experimental XRD pattern of mechanochemical $KPbI_3$ (red) and its best fit (blue). The fit residual is given in **(b)**. The weighted-profile *R*-factor R_{wp} = 6.6566.

Table 4-15. Numerical results of the fit in **Figure 4-34a**.

No.	Pos. [°2θ]	d-spacing [Å]	Height [cts]
1	9.587(8)	9.21827	466.11
2	10.047(6)	8.79655	1056.87
3	12.812(9)	6.90392	661.66
4	13.255(9)	6.67429	479.45
5	16.67(2)	5.31339	127.19
6	17.43(4)	5.08493	89.46
7	19.984(9)	4.43944	194.89
8	21.71(1)	4.08959	442.13
9	22.01(2)	4.03519	223.62
10	22.90(3)	3.88	47.78
11	23.46(2)	3.7889	87.47
12	25.06(2)	3.55056	207.06
13	25.53(2)	3.48561	1529.39
14	25.65(3)	3.47009	1325.42
15	26.01(4)	3.42237	1043.88
16	26.24(3)	3.393	297.38
17	26.80(2)	3.32417	524.04
18	27.29(1)	3.26487	598.39
19	27.68(4)	3.22029	1145.79
20	27.80(3)	3.20708	910.57
21	28.49(2)	3.13027	143.93

22	29.73(3)	3.00285	159.47
23	30.08(1)	2.96818	283.45
24	30.92(8)	2.88998	367.09
25	31.791(3)	2.8125	940.07
26	33.09(1)	2.70537	313.5
27	33.68(2)	2.65867	192.68
28	34.77(4)	2.57771	75.97
29	35.47(5)	2.52905	136.41
30	36.07(4)	2.48833	245.68
31	36.48(5)	2.46099	224.9
32	36.82(3)	2.43896	139.1
33	37.29(3)	2.40961	162.83
34	38.35(1)	2.345	471.69
35	39.02(5)	2.30667	859.03
36	39.12(5)	2.30061	859.42
37	39.68(3)	2.26977	176.9
38	40.29(1)	2.23654	333.14
39	41.41(2)	2.1789	361.58
40	41.80(5)	2.15936	192.92
41	43.04(5)	2.10012	176.68
42	43.52(3)	2.07805	123.91
43	44.84(7)	2.01978	135.25
44	45.25(3)	2.00248	132.06

45	46.48(3)	1.95224	104.46
46	47.15(4)	1.92618	263.64
47	47.25(5)	1.92213	247.33
48	47.93(2)	1.89646	399.8
49	48.35(4)	1.88101	98.34
50	49.71(2)	1.83254	100.5
51	50.61(4)	1.80198	67.49
52	51.10(4)	1.78594	94.08
53	51.56(3)	1.77112	107.52
54	52.47(3)	1.74256	76.7
55	53.70(8)	1.70557	110.99
56	54.24(5)	1.6897	138.23
57	55.12(4)	1.66488	107.42
58	55.71(9)	1.64857	60.16
59	57.21(3)	1.60896	102.07
60	58.56(3)	1.57495	70.91
61	59.41(5)	1.55438	148.58
62	60.1(1)	1.53937	61.51
63	60.90(5)	1.52002	108.06
64	61.70(9)	1.50212	54.95
65	62.4(2)	1.48697	125.86
66	62.8(1)	1.47872	247.05
67	63.89(2)	1.45593	162.21

68	64.64(6)	1.44079	80.65
69	66.15(2)	1.41152	153.11
70	66.83(7)	1.39873	170.02
71	67.53(4)	1.38602	191.3
72	68.7(1)	1.36588	155.43
73	69.4(2)	1.35381	104.9
74	70.14(3)	1.34069	107.16
75	71.75(5)	1.31452	81
76	72.28(5)	1.30606	52.92
77	73.63(6)	1.28545	44.71
78	74.94(6)	1.26624	28.32
79	76.93(3)	1.23833	63.16
80	79.66(2)	1.20268	33.49

Details of NMR measurements

Table 4-16. Nuclear ^{39}K T_1 values measured using a saturation-recovery sequence and fitted using a monoexponential function (unless otherwise stated). The uncertainties of fits given are one standard deviation.

compound	^{39}K T_1 [s]
KI	9.03 ± 0.02
KBr	7.73 ± 0.06
$KPbI_3$	1.44 ± 0.03
$KHCO_3$	0.115 ± 0.007 0.607 ± 0.007 (biexponential)

Table 4-17. Acquisition and processing parameters used for the ^{39}K spectra in **Figures 115-118**.

^{39}K spectra

composition	recycle delay [s]	number of scans	acquisition time [h]	Lorentzian apodization [Hz]
KI	10	4	0.01	20
$KPbI_3$	3	256	0.2	0
$K_{0.10}MA_{0.90}PbI_3$	60	1053	17.6	50
$K_{0.10}FA_{0.90}PbI_3$	3	2607	2.2	100
$K_{0.05}MA_{0.10}FA_{0.85}PbI_3$	3	19922	16.6	50
KBr	10	4	0.01	50
KMAFAPb(I,Br)	12	7765	25.9	50
KBr + $2PbBr_2$	3	1125	0.9	50
KBr + PbI_2	3	256	0.2	50
$K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ (CT selective)	3	22875	19.1	100
$K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ (non-selective)	3	50840	42.4	200
$K_{0.10}Cs_{0.90}PbI_3$	1	19892	5.5	200
$K_{0.50}Cs_{0.50}PbI_3$	1	5120	1.4	200
$K_{0.90}Cs_{0.10}PbI_3$	1	1024	0.3	20
$K_{0.075}Cs_{0.925}PbI_2Br$	3	47541	39.6	200

Table 4-18. Acquisition and processing parameters used for the ^{133}Cs spectra in **Figures 116** and **118**.

^{133}Cs spectra

composition	recycle delay [s]	number of scans	acquisition time [h]	Lorentzian apodization [Hz]
$CsPbI_3$	450	4	0.5	50
$K_{0.10}Cs_{0.90}PbI_3$	450	4	0.5	50
$K_{0.50}Cs_{0.50}PbI_3$	450	4	0.5	400
$K_{0.90}Cs_{0.10}PbI_3$	450	120	15	500
$K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ (mechanosynthesis)	450	24	3	250
$K_{0.05}Cs_{0.10}FA_{0.85}PbI_3$ (solution synthesis)	450	128	16	250
$Cs_{0.10}FA_{0.90}PbI_3$	450	4	0.5	200
$K_{0.075}Cs_{0.925}PbI_2Br$	450	8	1	300
$CsPbI_2Br$	56	8	0.1	300

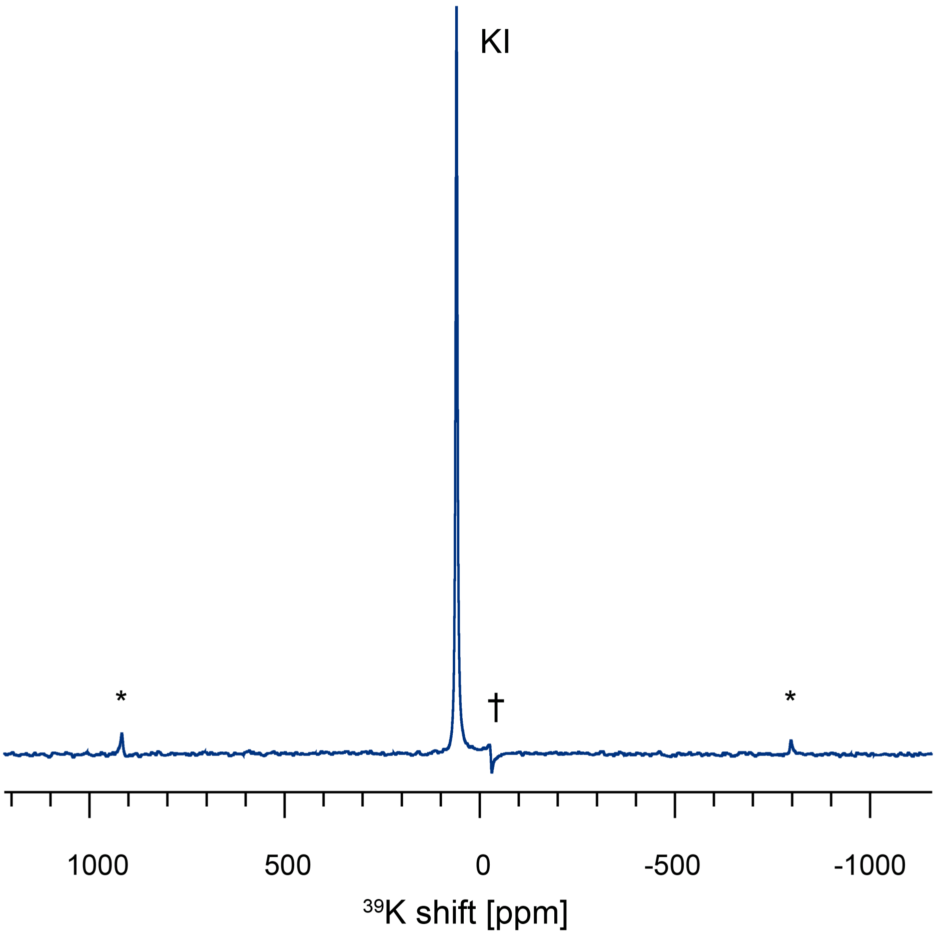


Figure 4-35. CT-selective ^{39}K MAS NMR spectrum of TiO_2-KI, at 11.7 T at 20 kHz MAS and 300 K (recycle delay: 10 s, acquisition time: 117 h). TiO_2 was activated at 300 °C and mechanochemically ground with anhydrous KI in a 10:1 molar ratio. Asterisks indicate spinning sidebands. The † symbol indicates a quadrature detection artefact.

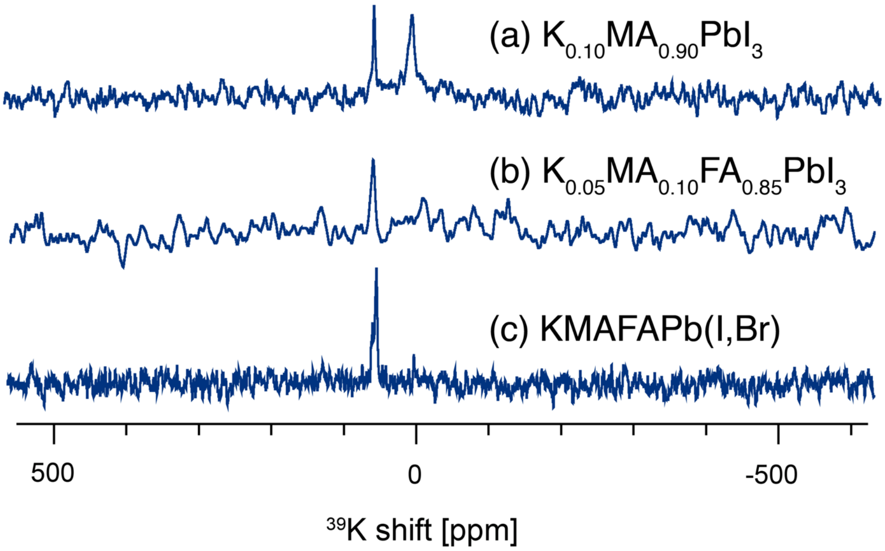


Figure 4-36. CT-selective spectra for compositions in **Figure 4-24c**, **e** and **g** in the main text. Acquisition parameters: **(a)** number of scans: 4096, recycle delay: 12 s, acquisition time: ~13.7 h **(b)** number of scans: 10240, recycle delay: 1 s, acquisition time: ~3 h, **(c)** number of scans: 23604, recycle delay: 3 s, acquisition time: ~20 h.

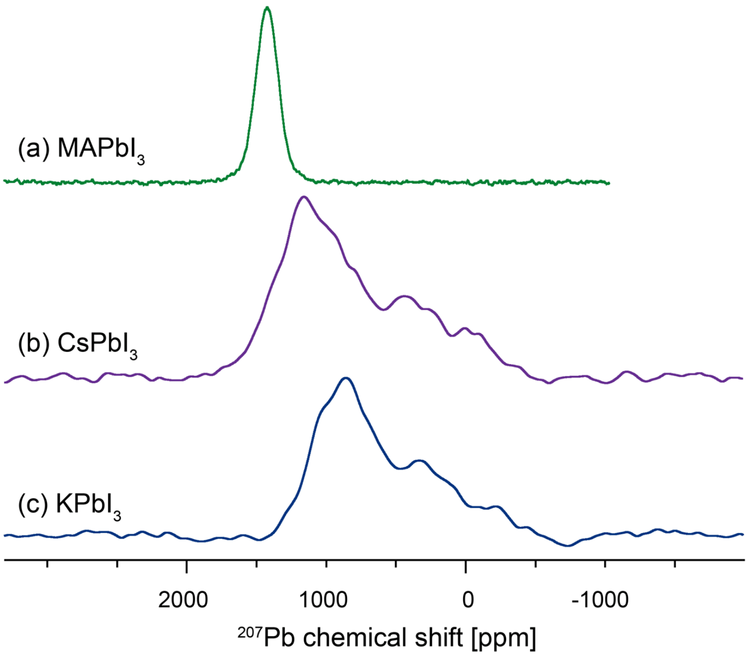


Figure 4-37. ^{207}Pb MAS NMR spectra at 11.7 T, 298 K and 20 kHz MAS of **(a)** MAPbI$_3$, number of scans: 27280, recycle delay: 0.1 s, acquisition time: 45 min., processed with 1 kHz Lorentzian apodization, **(b)** δ-CsPbI$_3$, number of scans: 30364, recycle delay: 0.1 s, acquisition time: 51 min., processed with 10 kHz Lorentzian apodization, **(c)** KPbI$_3$, number of scans: 102400, recycle delay: 0.02 s, acquisition time: 34 min, processed with 10 kHz Lorentzian apodization. ^{207}Pb chemical shifts were referenced to Pb(CH$_3$)$_4$ using solid Pb(NO$_3$)$_2$ as a secondary reference (-2961 ± 1) ppm.[457]

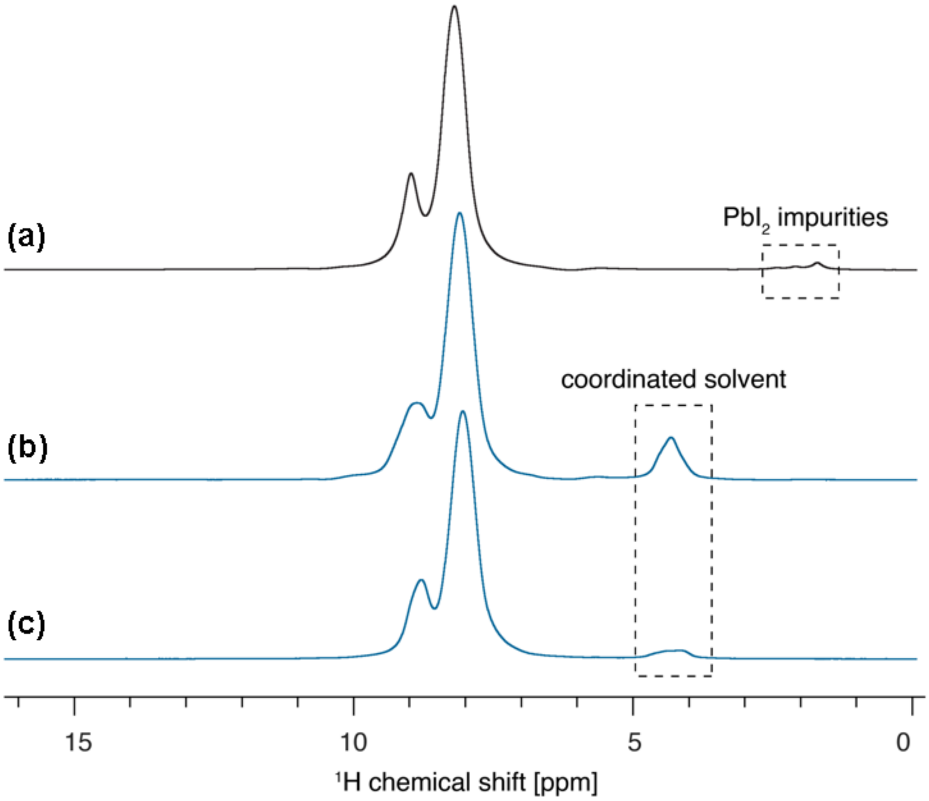


Figure 4-38. ^{1}H MAS NMR spectra at 21.1 T, 298 K and 20 kHz MAS of **(a)** mechanochemical K$_{0.05}$Cs$_{0.10}$FA$_{0.85}$PbI$_3$ (proton-containing impurities originating from the supplied PbI$_2$ and amounting to ~1% of protons in the sample are indicated). K$_{0.05}$Cs$_{0.10}$FA$_{0.85}$PbI$_3$ prepared by the solution route: **(b)** after initial annealing in air, **(c)** after 12 h of vacuum drying.

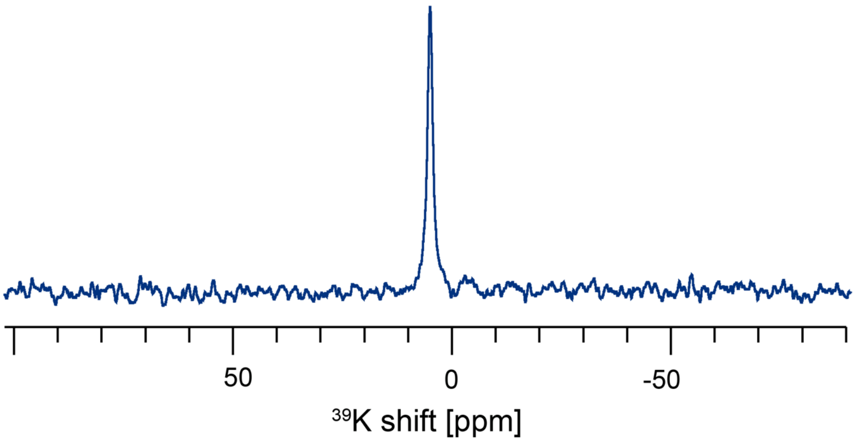


Figure 4-39. Quantitative ^{39}K MAS NMR spectrum at 21.1 T, 298 K and 20 kHz MAS of the mechanochemical $KPbI_3$ preparation. The spectrum was acquired using a recycle delay of 45 s, hence it is quantitative with respect to KI (T_1=9 s) and confirms there is no unreacted KI (δ=59.3 ppm) in the $KPbI_3$ phase. Number of scans: 32. Acquisition time: 24 minutes.

Details of DFT calculations

Cluster generation. We start from the assumption, that K is incorporated into the $FAPbI_3$ lattice, replacing the FA cation, without significantly changing the perovskite lattice formed by the $[PbI_6]^{4-}$ octahedra. Thus, in a first step we replace the FA ions of the cubic (black) $FAPbI_3$ crystal structures by K^+ cations. For the pure $FAPbI_3$ phase the ^1H positions inside the $[PbI_6]^{4-}$ cage were optimized using a periodic system within the density functional theory (DFT) framework and the generalized gradient approximation (GGA) functional PBE[205] within the Quantum Espresso suite.[188] The DFT optimization includes the Grimme dispersion correction[206] and relativistic effects up to spin-orbit couplings. For every calculation we use a plane-wave maximum cut-off energy of 100 Ry and a 2x2x2 Monkhorst-Pack grid of k-points.[423] The K^+ cations are either positioned at the geometrical center of the replaced FA ion, leading to a slightly interstitial K^+, or at the center of the surrounding $[PbI_6]^{4-}$ cage.

We assemble the final clusters from the relaxed structures as one central $[PbI_6]^{4-}$ cage with surrounding A^+ ions as $KA_{19}Pb_8I_{36}$ analogue to the ones used in the previous paper by Kubicki *et al.*[443], ensuring charge compensation and representing the solid-state by using symmetry-adapted clusters[110] containing the non-translational-symmetry elements from the perspective of the central molecule. The effect of the surrounding A^+ ions is investigated by either using FA^+ or Cs^+ as A^+ ions.

Additionally we investigate the ^{39}K chemical shift and EFG tensor in the perovskite structures given in a recent by work by Kubicki *et al.*[443] where we replace the central A^+ ion by K^+.

Table 4-19. Source and modifications of the cluster structures used in the DFT calculations.

Structure name	*Original structure*	*Modifications in periodic system (with Quantum Espresso)*[188]	*Modifications of cluster (with ADF)*[383-384]
$KFA_{19}Pb_8I_{36}$ *(A-site replacement)*	Cubic $FAPbI_3$[390]	Optimization of ^1H positions	Symmetric Replacement of central FA^+ to K^+
$KFA_{19}Pb_8I_{36}$ *(interstitial K^+)*	Cubic $FAPbI_3$[390]	Optimization of ^1H positions	Asymmetric Replacement of central FA^+ to K^+
$KCs_{19}Pb_8I_{36}$ *(A-site replacement)*	Cubic $FAPbI_3$[390]	Optimization of ^1H positions	Symmetric Replacement of central FA^+ to K^+ and replacement of surrounding FA ions with Cs ions
$KCs_{19}Pb_8I_{36}$ *(interstitial K^+)*	Cubic $FAPbI_3$[390]	Optimization of ^1H positions	Asymmetric Replacement of central FA^+ to K^+ and replacement of surrounding FA ions with Cs ions
$KCs_{19}Pb_8I_{36}$ *(from cubic $FAPbI_3$)*	Cubic $FAPbI_3$[390]	Optimization of all atomic positions after replacement of FA ions with Cs ions	Symmetric Replacement of central Cs^+ to K^+
$KCs_{19}Pb_8I_{36}$ *(from tetragonal $MAPbI_3$)*	Tetragonal $MAPbI_3$[413]	Optimization of Cs position after replacement of MA ions with Cs ions	Symmetric Replacement of central Cs^+ to K^+
$KCs_{19}Pb_8I_{36}$ *(from hexagonal $CsPbI_3$)*	Hexagonal $CsPbI_3$[413]	None	Symmetric Replacement of central Cs^+ to K^+
$KFA_{19}Pb_8I_{36}$ *(from hexagonal $FAPbI_3$)*	Hexagonal $FAPbI_3$[413]	Optimization of ^1H positions	Symmetric Replacement of central FA^+ to K^+

NMR shift calculation. The chemical shieldings (σ_{DFT}) are transformed to chemical shifts (δ_{DFT}) using the linear relation $\delta_{exp} = \sigma_{ref} - b\sigma_{DFT}$, where σ_{ref} and b are fit using calculated chemical shieldings from known reference compounds (KI, KBr, KCN and KF). The slope b was either fixed at 1 or used as a fit parameter, resulting in $b = 1.248$.

We use KCN, KI, KBr, and KF as reference compounds with the experimental chemical shifts obtained from the work by Moudrakovski and Ripmeester.[446] The crystal structures of KCN, KI, KF and KBr were obtained from the works by Price *et al.*[466], Van Den Bosch *et al.*[420], Broch *et al.*[467] and Ott.[468]

The DFT accuracy for ^{39}K chemical shift calculations has been investigated by Wu *et al.*[452] and Shimoda *et al.*[469], where they report a ^{39}K chemical shift root-mean-square deviation between experiment and calculation of around 4-8 ppm.

To estimate the goodness of the fit parameters, we calculate the root-mean-square-deviation (RMSD) between the experimental and DFT calculated chemical shifts. Note, that in the KCN cluster we have two distinct ^{39}K sites depending on the relative orientation of the CN ion. If we calculate the average of the two ^{39}K shifts, assuming motional averaging due to CN rotations, we obtain a RMSD of 1.46 ppm and 3.63 ppm, for either a variable or fixed slope, between the experimental and DFT calculated chemical shifts. The reference chemical shifts and the linear regression models are shown in **Figure 130**.

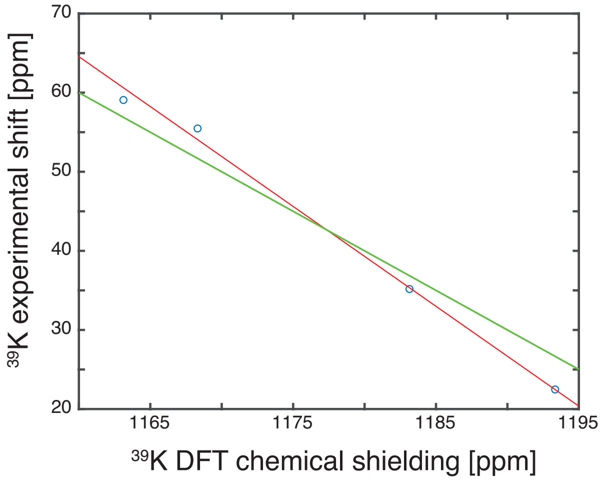


Figure 4-40. Scatterplot showing ^{39}K experimental shifts against DFT calculated chemical shielding. The solid lines show the different linear models used to convert chemical shielding to shifts (green: fixed slope (b=1), red: variable slope (b=1.248)).

Table 4-20. DFT calculated and experimental magnetic shieldings, chemical shifts and EFG tensor parameters.

Structure	*DFT chemical shielding* *[ppm]*	*Experimental shifts [ppm]*	*DFT shifts (σ_{ref}=1220, b=1.0) [ppm]*	*DFT shifts (σ_{ref}=1512, b=1.248) [ppm]*	*C_Q* *[MHz]*	*η*	*V_{zz}* *10^{21} [Vm^{-2}]*
(a) KF	1193.38	22.4	26.62	22.66	-0.518	0.0004	-0.037
(b) KBr	1168.35	55.4	51.65	53.90	-0.002	0.011	-0.00015
(c) KI	1163.16	59	56.84	60. 38	0.005	0.0065	0.0004
(d) KCN (site 1)	1120.18				0.33	0.397	0.024
(e) KCN (site 2)	1246.19				0.94	0.530	0.069
(f) KCN (avg.)	1183.19	35.1	36.8	35.39			
(g) $KFA_{19}Pb_8I_{36}$ *(A-site replacement)*	1321.45		-101.45	-137.17	-0.137	0.77	0.097
(h) $KFA_{19}Pb_8I_{36}$ *(interstitial K^+)*	1311.41		-91.41	-124.64	-0.242	0.81	-0.17
(i) $KCs_{19}Pb_8I_{36}$ *(A-site replacement)*	1316.6		-96.6	-131.12	-0.068	0.19	0.048
(k) $KCs_{19}Pb_8I_{36}$ *(interstitial K^+)*	1307.1		-87.1	-119.26	-0.243	0.55	-0.17
(l) $KCs_{19}Pb_8I_{36}$ *(from cubic $FAPbI_3$)*	1326.12		-106.1	-143.0	0.007	0.00012	0.0005
(m) $KCs_{19}Pb_8I_{36}$ *(from tetragonal $MAPbI_3$)*	1309.25		-89.25	-121.94	0.38	0.179	0.028
(n) $KCs_{19}Pb_8I_{36}$ *(from hexagonal $CsPbI_3$)*	1324.3		-91.23	-124.42	0.43	0.804	-0.53
(o) $KFA_{19}Pb_8I_{36}$ *(from hexagonal $FAPbI_3$)*	1324.3		-104.3	-140.73	-0.37	0.519	-0.027

4.5 Conclusion and Outlook

In conclusion, we have demonstrated how the calculation of solid-state NMR and EFG parameters can aid the atomic-level characterization of doped and disordered materials. For this set of materials diffraction-based methods lack information about the non-crystalline and disordered regions of the sample. However, solid-state NMR can directly probe the local atomic environment around a defect and thus allow for structural characterization. For a set of doped photovoltaic lead-halide perovskite materials we have demonstrated how computational methods can be used to aid the parametrization and interpretation of solid-state NMR experiments using different probe nuclei (here ^{39}K, ^{87}Rb, ^{133}Cs and ^{207}Pb). Thus, providing strong evidence for or against a given structural hypothesis.

Note, that this particular set of materials still poses a big challenge to computational methods for NMR crystallography and that the results have to be evaluated very critically. For the structural characterization of microcrystalline solids and amorphous powders described in **Chapters 2** and **3** the calculation of NMR parameters is reasonably accurate and has become fairly standard. However, for the doped perovskite materials described in **Chapter 4** the presence of heavy atoms (e.g. ^{127}I, ^{133}Cs and ^{207}Pb) has been shown to require fully relativistic DFT calculations. including spin-orbit coupling ,and hybrid-functionals if possible.[122-125, 425] These requirements lead to a drastic increase in computational resources and thus strongly limit the cluster size and the number of possible defect environments which can be evaluated computationally. Thus, due the expected accuracy of the DFT based NMR calculations and the non-extensiveness of the structural screening, the NMRX method used here is not sufficient to fully characterize the investigated materials. Instead NMRX is used in combination with other structural characterization methods, in order to provide evidence for or against a given structural hypothesis (e.g. incorporation vs. passivation or phase separation). Note that in many cases this information is already sufficient and complements the information that can be extracted from other characterization methods. Also note, that for disordered and doped materials containing no heavy atoms the approach outlined in **Chapter 3** can be easily adapted to fully characterize the atomic-level structure. Also note that, following the early work described here, NMR characterization has become a key part of the perovskite research, and has been used in many recent studies.[391-397, 441-442, 444-445, 470-481]

Moving forward, the chemical shift driven NMRX approach for disordered and doped materials containing heavy atoms could be improved in two main directions in order to be generally and routinely applicable.

First, the chemical shift calculations have to be extended such as to allow the calculation of larger and / or extended systems. The idea is that larger and / or extended systems will better model the electronic structure around the local defect environments and thus lead to a more reliable chemical shift prediction. In general, the same arguments concerning the system size as for amorphous systems (**Chapter 3**) are applicable. In order to extend the system size amendable to NMR calculations, three main ideas can be investigated. Periodic DFT in combination with a correction term calculated at higher level of theory for an isolated molecule or cluster could be used to extend the full relativistic and hybrid functional chemical shift calculations to periodic system.[112] Another possibility would be to use a fragment based DFT method with locally dense basis sets as demonstrated by Hartmann *et al.*[83-84, 111, 351-352] Such an approach would allow the calculation of larger clusters, since for atoms further away from the defect site only two body interactions are considered and smaller basis sets are used. Note, that for this method the use of the bond-valence method should be investigated in order to achieve convergence in the DFT calculations.[124] A third method to overcome the limits of DFT calculations would be to extend the ML method presented in **Chapter 2.3** to organo-metallic systems containing heavy atoms. However, the size of the required ML training-set might limit this approach. We propose to overcome this limitation by a combined DFT-ML delta-learning approach. It has been shown, that learning only the correction term between two different levels of theory requires significantly smaller training-sets than learning entire properties.[155, 482] It thus might be feasible to learn only the DFT spin-orbit coupling correction to the chemical shift term, which can then be used to improve the accuracy of scalar relativistic periodic DFT calculations for doped systems containing heavy atoms.

Second, the structural screening method used here should be extended in order to extensively sample the possible structural space. Here, we propose to generate an extensive ensemble of possible defect motifs by a step wise approach, conceptually similar to the CSP approach used for molecular crystals. In a first step classical or semi-empirical MD simulations can be used to screen a large ensemble of possible defect structures. Mixed techniques, such as ONIOM,[483-484] would allow for relatively large MD simulations while retaining high accuracy (possibly at quantum-mechanical level) around the defect site. Note, that it is also possible to parametrize the force-field using ML methods trained on comparable structures.[485-486] As a next step, the energetically most stable structures can be selected and evaluated using non-relativistic DFT. Next, the energetically most stable structures can again be selected and now evaluated using fully-relativistic DFT to generate an ensemble of the most probable defect structures. This ensemble can then finally be compared to experiment using fully-relativistic DFT chemical shift calculations.

Chapter 5 Conclusion

5.1 Achieved results

In summary, we have shown how chemical shift information extracted from solid-state NMR experiments in combination with advanced computational methods can be translated into information on the crystal structure. We have demonstrated the approach for microcrystalline powders, amorphous materials and disordered and doped solids, all of which are not amenable to resolution with diffraction methods.

Additionally, we have shown how the investigated material dictates the applicable computational method and the information content which can be extracted from this combined approach.

For molecular solids, an approach combining periodic DFT chemical shift calculations, or chemical shifts predicted using ML, combined with a constrained CSP search has been demonstrated. This approach allows for a full characterization and determination of a given crystal structure, even including a well-defined determination confidence as well as positional uncertainties on the individual atoms. Furthermore, we demonstrated a direct mapping, using a ML approach, between the structural information and the chemical shifts of a molecular solid, without the need to calculate the electronic structure.

For amorphous, disordered and doped materials, a cluster or fragment-based approach must be selected, which can be comparable in accuracy with the periodic calculations. However, the nature of these materials often does not allow for the full determination of the crystal structure. Instead the structure is characterized by a set of determined structural motifs. This information can then often be combined with other structure elucidation methods to fully characterize the given material.

For materials containing heavy elements the methods described above do not offer sufficient accuracy. Here, fully relativistic DFT chemical shift calculations, which are currently only accessible in a cluster-based approach, must be employed. This restriction currently limits the complexity of the structures which can be investigated. Additionally, even for the fully relativistic DFT approach, the calculated chemical shifts are still below the accuracy that is achievable for other systems. However, we have demonstrated that the information extracted from the chemical shift calculations can be used in combination with other experimental and computational methods to answer key-questions and deliver important new insight into such materials.

5.2 Future development.

Possible future developments for the structural characterization and determination based on solid-state NMR experiments in combination with advanced computational methods for the different materials and degrees of disorder has already been described in the previous chapters.

A point which has not yet been discussed in detail here is the temperature dependence of solid-state NMR parameters, especially of the chemical shift. Note, that the methods discussed above can be used to probe phase transitions and structural changes caused by a change in temperature. However, they only consider static snapshots of the investigated structures and local dynamics due to an effective temperature are not considered. In reality, the local dynamics due to an effective temperature within an experimental structure results in each atom sampling an ensemble of different local environments, and thus, experiencing an ensemble of corresponding chemical shifts. If the motion is sufficiently slow, compared to the chemical shift difference within the sampled environments, we experimentally observe a set of distinguishable chemical shifts, which can be mapped computationally to a discreet set of structures. However, for fast motions we experimentally observe an average chemical shift, which is usually compared to the calculated chemical shift of a static snapshot. Note, that this is only correct if the average dynamic structure indeed corresponds to the static snapshot and if the chemical shift is affected isotropically by the local dynamics. Additionally, the DFT optimized structures used in the CSP-NMRX approaches described in this thesis correspond to the 0° K structure, whereas the experimental chemical shifts are measured at a finite temperature. This is most commonly observed as a difference in lattice parameters and bond-lengths between the XRD and NMRX determined crystal structures (see **Chapter 2.6**).[487]

For more general cases, the NMR parameters including nuclear motion have been studied using a variety of quantum-mechanical methods.[64, 223, 280, 487-493] In these studies the dynamical effect on the NMR parameters was investigated either by molecular dynamics,[280, 487, 489-492] path-integral molecular dynamics,[64] perturbation theory[223, 493] or Monte-Carlo sampling.[488, 490] However, all of these approaches depend on the ab-initio calculation of chemical shifts for a set of perturbed structures and their computational expense prohibits their routine use in many applications. Here, we propose the use of ML predicted chemical shifts in combination with the above-mentioned sampling approaches to accurately describe the effects of nuclear motion on the chemical shifts. Note, that after the ML model has been trained, the computational expense to calculate the chemical shifts of a given structure is negligible compared to the other involved calculations and thus a large ensemble of motional snapshots can be sampled.

Additionally, we would like to briefly mention another research area, where the combination of solid-state NMR experiments with advanced computational methods seems very promising. Note, that the chemical shift of a structure is uniquely determined by its electronic structure, which is in turn uniquely determined by the crystal structure. In **Chapter 2.3** we have demonstrated how we can use ML methods to circumvent the calculation of the electronic structure and directly map the chemical shift information onto the crystal structure. Also note, that it is the strong and direct correlation between the chemical shifts and the electronic structure as well as between the electronic structure and the crystal structure, which allows for this direct mapping. Thus, as there exist other structural properties which are strongly correlated to and dependent on the electronic structure, we hypothesize that it should also be possible to directly map them on to the chemical shifts using machine learning methods. In other words, we propose that chemical shift information can be used to extract / infer information on structural properties and chemical activity, such as activity in ferroelectrics, which are not as easily accessible experimentally or computationally.

As an example, Corperet and co-workers have demonstrated a strong correlation between the chemical shielding tensor and the reactivity and polymerization ability of different materials.[494-495] We thus propose the usage of machine learning methods to exploit this correlation and to possibly use calculated and / or experimental chemical shift information to predict reactivity, polymerization effects and crystal structure formation.[398]

Bibliography

1. Rietveld, H. M., A Profile Refinement Method for Nuclear and Magnetic Structures. *J Appl Crystallogr* **1969,** *2*, 65-&.
2. Wollan, E. O.; Shull, C. G., The Diffraction of Neutrons by Crystalline Powders. *Physical Review* **1948,** *73* (8), 830-841.
3. Shull, C. G.; Wollan, E. O., X-Ray, Electron, and Neutron Diffraction. *Science* **1948,** *108* (2795), 69-75.
4. Gorelik, T. E.; Czech, C.; Hammer, S. M.; Schmidt, M. U., Crystal structure of disordered nanocrystalline alpha(vertical bar vertical bar)- quinacridone determined by electron diffraction. *Crystengcomm* **2016,** *18* (4), 529-535.
5. Das, P.; Mugnaioli, E.; Nicolopoulos, S.; Tossi, C.; Gemmi, M.; Galanis, A.; Borodi, G.; Pop, M., Crystal structures of two important pharmaceuticals solved by 3D precession electron diffraction tomography. *Organic Process Research & Development* **2018**.
6. Gruene, T.; Wennmacher, J. T. C.; Zaubitzer, C.; Holstein, J. J.; Heidler, J.; Fecteau-Lefebvre, A.; De Carlo, S.; Müller, E.; Goldie, K. N.; Regeni, I., Rapid structure determination of microcrystalline molecular compounds using electron diffraction. *Angewandte Chemie* **2018**.
7. Jones, C. G.; Martynowycz, M. W.; Hattne, J.; Fulton, T. J.; Stoltz, B. M.; Rodriguez, J. A.; Nelson, H.; Gonen, T., The CryoEM Method MicroED as a Powerful Tool for Small Molecule Structure Determination. **2018**.
8. Ochsenfeld, C.; Brown, S. P.; Schnell, I.; Gauss, J.; Spiess, H. W., Structure assignment in the solid state by the coupling of quantum chemical calculations with NMR experiments: a columnar hexabenzocoronene derivative. *J Am Chem Soc* **2001,** *123* (11), 2597-606.
9. Harper, J. K.; Mulgrew, A. E.; Li, J. Y.; Barich, D. H.; Strobel, G. A.; Grant, D. M., Characterization of stereochemistry and molecular conformation using solid-state NMR tensors. *Journal of the American Chemical Society* **2001,** *123* (40), 9837-9842.
10. Witter, R.; Priess, W.; Sternberg, U., Chemical shift driven geometry optimization. *J Comput Chem* **2002,** *23* (2), 298-305.
11. Widdifield, C. M.; Schurko, R. W., A Solid-State 39K and 13C NMR Study of Polymeric Potassium Metallocenes. *The Journal of Physical Chemistry A* **2005,** *109* (31), 6865-6876.
12. Harper, J. K.; Grant, D. M.; Zhang, Y. G.; Lee, P. L.; Von Dreele, R., Characterizing challenging microcrystalline solids with solid-state NMR shift tensor and synchrotron X-ray powder diffraction data: Structural analysis of ambuic acid. *Journal of the American Chemical Society* **2006,** *128* (5), 1547-1552.
13. Harper, J. K.; Grant, D. M., Enhancing crystal-structure prediction with NMR tensor data. *Cryst Growth Des* **2006,** *6* (10), 2315-2321.
14. Robbins, A. J.; Ng, W. T. K.; Jochym, D.; Keal, T. W.; Clark, S. J.; Tozer, D. J.; Hodgkinson, P., Combining insights from solid-state NMR and first principles calculation: applications to the 19F NMR of octafluoronaphthalene. *Physical Chemistry Chemical Physics* **2007,** *9* (19), 2389-2396.
15. Shen, Y.; Delaglio, F.; Cornilescu, G.; Bax, A., TALOS plus : a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *Journal of Biomolecular Nmr* **2009,** *44* (4), 213-223.
16. Kervern, G.; D'Aleo, A.; Toupet, L.; Maury, O.; Emsley, L.; Pintacuda, G., Crystal-Structure Determination of Powdered Paramagnetic Lanthanide Complexes by Proton NMR Spectroscopy. *Angew Chem Int Edit* **2009,** *48* (17), 3082-3086.
17. Webber, A. L.; Emsley, L.; Claramunt, R. M.; Brown, S. P., NMR crystallography of campho[2,3-c]pyrazole (Z' = 6): combining high-resolution 1H-13C solid-state MAS NMR spectroscopy and GIPAW chemical-shift calculations. *J Phys Chem A* **2010,** *114* (38), 10435-42.
18. Salager, E.; Day, G. M.; Stein, R. S.; Pickard, C. J.; Elena, B.; Emsley, L., Powder Crystallography by Combined Crystal Structure Prediction and High-Resolution H-1 Solid-State NMR Spectroscopy. *Journal of the American Chemical Society* **2010,** *132* (8), 2564-+.
19. Rawal, A.; Smith, B. J.; Athens, G. L.; Edwards, C. L.; Roberts, L.; Gupta, V.; Chmelka, B. F., Molecular Silicate and Aluminate Species in Anhydrous and Hydrated Cements. *Journal of the American Chemical Society* **2010,** *132* (21), 7321-7337.
20. Kim, J.; Middlemiss, D. S.; Chernova, N. A.; Zhu, B. Y. X.; Masquelier, C.; Grey, C. P., Linking Local Environments and Hyperfine Shifts: A Combined Experimental and Theoretical P-31 and Li-7 Solid-State NMR Study of Paramagnetic Fe(III) Phosphates. *Journal of the American Chemical Society* **2010,** *132* (47), 16825-16840.
21. Harris, R. K.; Hodgkinson, P.; Zorin, V.; Dumez, J. N.; Elena-Herrmann, B.; Emsley, L.; Salager, E.; Stein, R. S., Computation and NMR crystallography of terbutaline sulfate. *Magnetic Resonance in Chemistry* **2010,** *48*, S103-S112.
22. Lai, J. F.; Niks, D.; Wang, Y. C.; Domratcheva, T.; Barends, T. R. M.; Schwarz, F.; Olsen, R. A.; Elliott, D. W.; Fatmi, M. Q.; Chang, C. E. A.; Schlichting, I.; Dunn, M. F.; Mueller, L. J., X-ray and NMR Crystallography in an Enzyme Active Site: The Indoline Quinonoid Intermediate in Tryptophan Synthase. *Journal of the American Chemical Society* **2011,** *133* (1), 4-7.
23. Jiang, J. X.; Jorda, J. L.; Yu, J. H.; Baumes, L. A.; Mugnaioli, E.; Diaz-Cabanas, M. J.; Kolb, U.; Corma, A., Synthesis and Structure Determination of the Hierarchical Meso-Microporous Zeolite ITQ-43. *Science* **2011,** *333* (6046), 1131-1134.
24. Santos, S. M.; Rocha, J.; Mafra, L., NMR crystallography: Toward chemical shift-driven crystal structure determination of the β-lactam antibiotic amoxicillin trihydrate. *Crystal Growth and Design* **2013,** *13* (6).
25. Mueller, L. J.; Dunn, M. F., NMR Crystallography of Enzyme Active Sites: Probing Chemically Detailed, Three-Dimensional Structure in Tryptophan Synthase. *Acc. Chem. Res.* **2013,** *46* (9), 2008-2017.
26. 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* **2013,** *13* (12), 5391-5396.
27. Lee, D.; Monin, G.; Duong, N. T.; Lopez, I. Z.; Bardet, M.; Mareau, V.; Gonon, L.; De Paëpe, G., Untangling the Condensation Network of Organosiloxanes on Nanoparticles using 2D 29Si–29Si Solid-State NMR Enhanced by Dynamic Nuclear Polarization. *J. Am. Chem. Soc.* **2014,** *136* (39), 13781–13788.

28. Mollica, G.; Dekhil, M.; Ziarelli, F.; Thureau, P.; Viel, S. p. S., Quantitative structural constraints for organic powders at natural isotopic abundance using dynamic nuclear polarization solid-state NMR spectroscopy. *Angewandte Chemie - International Edition* **2015,** *54* (20), 6028-6031.
29. Ludeker, D.; Brunklaus, G., NMR crystallography of ezetimibe co-crystals. *Solid State Nucl Magn Reson* **2015,** *65*, 29-40.
30. Widdifield, C. M.; Robson, H.; Hodgkinson, P., Furosemide's one little hydrogen atom: NMR crystallography structure verification of powdered molecular organics. *Chem Commun* **2016,** *52* (40), 6685-8.
31. 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* **2016,** *16* (4), 1798-1804.
32. Hartman, J. D.; Day, G. M.; Beran, G. J. O., Enhanced NMR Discrimination of Pharmaceutically Relevant Molecular Crystal Forms through Fragment-Based Ab Initio Chemical Shift Predictions. *Crystal Growth and Design* **2016,** *16* (11).
33. Seymour, I. D.; Middlemiss, D. S.; Halat, D. M.; Trease, N. M.; Pell, A. J.; Grey, C. P., Characterizing Oxygen Local Environments in Paramagnetic Battery Materials via (17)O NMR and DFT Calculations. *J Am Chem Soc* **2016,** *138* (30), 9405-8.
34. Leclaire, J.; Poisson, G.; Ziarelli, F.; Pepe, G.; Fotiadu, F.; Paruzzo, F. M.; Rossini, A. J.; Dumez, J. N.; Elena-Herrmann, B.; Emsley, L., Structure elucidation of a complex CO2-based organic framework material by NMR crystallography. *Chem Sci* **2016,** *7* (7), 4379-4390.
35. Widdifield, C. M.; Nilsson Lill, S. O.; Broo, A.; Lindkvist, M.; Pettersen, A.; Svensk Ankarberg, A.; Aldred, P.; Schantz, S.; Emsley, L., 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* **2017,** *19* (25), 16650-16661.
36. Szell, P. M.; Gabriel, S. A.; Gill, R. D.; Wan, S. Y.; Gabidullin, B.; Bryce, D. L., (13)C and (19)F solid-state NMR and X-ray crystallographic study of halogen-bonded frameworks featuring nitrogen-containing heterocycles. *Acta Crystallogr C Struct Chem* **2017,** *73* (Pt 3), 157-167.
37. Goward, G. R.; Sebastiani, D.; Schnell, I.; Spiess, H. W.; Kim, H. D.; Ishida, H., Benzoxazine oligomers: evidence for a helical structure from solid-state NMR spectroscopy and DFT-based dynamics and chemical shift calculations. *J Am Chem Soc* **2003,** *125* (19), 5792-800.
38. Selent, M.; Nyman, J.; Roukala, J.; Ilczyszyn, M.; Oilunkaniemi, R.; Bygrave, P. J.; Laitinen, R.; Jokisaari, J.; Day, G. M.; Lantto, P., Clathrate Structure Determination by Combining Crystal Structure Prediction with Computational and Experimental (129) Xe NMR Spectroscopy. *Chemistry* **2017,** *23* (22), 5258-5269.
39. Mali, G., Ab initio crystal structure prediction of magnesium (poly)sulfides and calculation of their NMR parameters. *Acta Crystallogr C Struct Chem* **2017,** *73* (Pt 3), 229-233.
40. Gambuzzi, E.; Pedone, A.; Menziani, M. C.; Angeli, F.; Caurant, D.; Charpentier, T., Probing silicon and aluminium chemical environments in silicate and aluminosilicate glasses by solid state NMR spectroscopy and accurate first-principles calculations. *Geochimica Et Cosmochimica Acta* **2014,** *125*, 170-185.
41. Fyfe, C. A.; Brouwer, D. H.; Lewis, A. R.; Villaescusa, L. A.; Morris, R. E., Combined solid state NMR and X-ray diffraction investigation of the local structure of the five-coordinate silicon in fluoride-containing as-synthesized STF zeolite. *Journal of the American Chemical Society* **2002,** *124* (26), 7770-7778.
42. Forler, N.; Vasconcelos, F.; Cristol, S.; Paul, J.-F.; Montagne, L.; Charpentier, T.; Mauri, F.; Delevoye, L., New insights into oxygen environments generated during phosphate glass alteration: a combined O-17 MAS and MQMAS NMR and first principles calculations study. *Physical Chemistry Chemical Physics* **2010,** *12* (31), 9054-9063.
43. Florian, P.; Massiot, D., Beyond periodicity: probing disorder in crystalline materials by solid-state nuclear magnetic resonance spectroscopy. *Crystengcomm* **2013,** *15* (43), 8623-8626.
44. 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* **2015,** *15* (8), 3674-3683.
45. Farnan, I.; Grandinetti, P. J.; Baltisberger, J. H.; Stebbins, J. F.; Werner, U.; Eastman, M. A.; Pines, A., Quantification of the disorder in network-modified silicate-glasses. *Nature* **1992,** *358* (6381), 31-35.
46. Elena, B.; Pintacuda, G.; Mifsud, N.; Emsley, L., Molecular structure determination in powders by NMR crystallography from proton spin diffusion. *Journal of the American Chemical Society* **2006,** *128* (29), 9555-9560.
47. Dudenko, D.; Kiersnowski, A.; Shu, J.; Pisula, W.; Sebastiani, D.; Spiess, H. W.; Hansen, M. R., A strategy for revealing the packing in semicrystalline pi-conjugated polymers: crystal structure of bulk poly-3-hexyl-thiophene (P3HT). *Angew Chem Int Ed Engl* **2012,** *51* (44), 11068-72.
48. Dedios, A. C.; Pearson, J. G.; Oldfield, E., Secondary and Tertiary Structural Effects on Protein Nmr Chemical-Shifts - an Abinitio Approach. *Science* **1993,** *260* (5113), 1491-1496.
49. Clark, T. M.; Grandinetti, P. J.; Florian, P.; Stebbins, J. F., Correlated structural distributions in silica glass. *Phys Rev B* **2004,** *70* (6).
50. Castellani, F.; Rossum van, B.; Diehl, A.; Schubert, M.; Rehbein, K.; Oschkinat, H., Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. *Nature* **2002,** *420* (November), 98-102.
51. Cadars, S.; Allix, M.; Brouwer, D. H.; Shayib, R.; Suchomel, M.; Garaga, M. N.; Rakhmatullin, A.; Burton, A. W.; Zones, S. I.; Massiot, D.; Chmelka, B. F., Long- and Short-Range Constraints for the Structure Determination of Layered Silicates with Stacking Disorder. *Chemistry of Materials* **2014,** *26* (24), 6994-7008.
52. Brunet, F.; Bertani, P.; Charpentier, T.; Nonat, A.; Virlet, J., Application of 29Si Homonuclear and 1H−29Si Heteronuclear NMR Correlation to Structural Studies of Calcium Silicate Hydrates. *J. Phys. Chem. B* **2004,** *108* (40), 15494-15502.

53. Brouwer, D. H.; Darton, R. J.; Morris, R. E.; Levitt, M. H., A solid-state NMR method for solution of zeolite crystal structures. *Journal of the American Chemical Society* **2005,** *127* (29), 10365-10370.

54. Brouwer, D. H.; Cadars, S.; Eckert, J.; Liu, Z.; Terasaki, O.; Chmelka, B. F., A General Protocol for Determining the Structures of Molecularly Ordered but Noncrystalline Silicate Frameworks. *Journal of the American Chemical Society* **2013,** *135* (15), 5641-5655.

55. Brouwer, D. H., NMR crystallography of zeolites: Refinement of an NMR-solved crystal structure using ab initio calculations of Si-29 chemical shift tensors. *Journal of the American Chemical Society* **2008,** *130* (20), 6306-+.

56. Baias, M.; Widdifield, C. M.; Dumez, J.-N.; Thompson, H. P. G.; Cooper, T. G.; Salager, E.; Bassil, S.; Stein, R. S.; Lesage, A.; Day, G. M.; Emsley, L., Powder crystallography of pharmaceutical materials by combined crystal structure prediction and solid-state 1H NMR spectroscopy. *Physical Chemistry Chemical Physics* **2013,** *15* (21), 8069-8069.

57. Baias, M.; Lesage, A.; Aguado, S.; Canivet, J.; Moizan-Basle, V.; Audebrand, N.; Farrusseng, D.; Emsley, L., Superstructure of a substituted zeolitic imidazolate metal-organic framework determined by combining proton solid-state NMR spectroscopy and DFT calculations. *Angewandte Chemie - International Edition* **2015,** *54* (20), 5971-5976.

58. Baias, M.; Dumez, J. N.; Svensson, P. H.; Schantz, S.; Day, G. M.; Emsley, L., De novo determination of the crystal structure of a large drug molecule by crystal structure prediction-based powder NMR crystallography. *J Am Chem Soc* **2013,** *135* (46), 17501-7.

59. Ashbrook, S. E.; Cutajar, M.; Pickard, C. J.; Walton, R. I.; Wimperis, S., Structure and NMR assignment in calcined and as-synthesized forms of AlPO-14: a combined study by first-principles calculations and high- resolution Al-27-P-31 MAS NMR correlation. *Physical Chemistry Chemical Physics* **2008,** *10* (37), 5754-5764.

60. Ditchfield, R., Self-Consistent Perturbation-Theory of Diamagnetism .1. Gauge-Invariant Lcao Method for Nmr Chemical-Shifts. *Molecular Physics* **1974,** *27* (4), 789-807.

61. Wolinski, K.; Hinton, J. F.; Pulay, P., Efficient Implementation of the Gauge-Independent Atomic Orbital Method for Nmr Chemical-Shift Calculations. *Journal of the American Chemical Society* **1990,** *112* (23), 8251-8260.

62. Pickard, C. J.; Mauri, F., All-electron magnetic response with pseudopotentials: NMR chemical shifts. *Phys Rev B* **2001,** *63* (24).

63. Yates, J. R.; Pickard, C. J.; Mauri, F., Calculation of NMR chemical shifts for extended systems using ultrasoft pseudopotentials. *Phys Rev B* **2007,** *76* (2), 024401.

64. Dracinsky, M.; Hodgkinson, P., Effects of Quantum Nuclear Delocalisation on NMR Parameters from Path Integral Molecular Dynamics. *Chem-Eur J* **2014,** *20* (8), 2201-2207.

65. Brouwer, D. H., A structure refinement strategy for NMR crystallography: An improved crystal structure of silica-ZSM-12 zeolite from (29)Si chemical shift tensors. *Journal of Magnetic Resonance* **2008,** *194* (1), 136-146.

66. Tremayne, M.; Kariuki, B. M.; Harris, K. D. M., 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. *Angewandte Chemie-International Edition in English* **1997,** *36* (7), 770-772.

67. Meejoo, S.; Kariuki, B. M.; Kitchin, S. J.; Cheung, E. Y.; Albesa-Jove, D.; Harris, K. D. M., Structural aspects of the beta-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* **2003,** *86* (5), 1467-1477.

68. Brouwer, D. H.; Moudrakovski, I. L.; Udachin, K. A.; Enright, G. D.; Ripmeester, J. A., Guest loading and multiple phases in single crystals of the van der Waals host p-tert-butylealix[4]arene. *Cryst Growth Des* **2008,** *8* (6), 1878-1885.

69. Partridge, B. E.; Leowanawat, P.; Aqad, E.; Imam, M. R.; Sun, H. J.; Peterca, M.; Heiney, P. A.; Graf, R.; Spiess, H. W.; Zeng, X. B.; Ungar, G.; Percec, V., Increasing 3D Supramolecular Order by Decreasing Molecular Order. A Comparative Study of Helical Assemblies of Dendronized Nonchlorinated and Tetrachlorinated Perylene Bisimides. *Journal of the American Chemical Society* **2015,** *137* (15), 5210-5224.

70. Profeta, M.; Mauri, F.; Pickard, C. J., Accurate first principles prediction of O-17 NMR parameters in SiO2: Assignment of the zeolite ferrierite spectrum. *Journal of the American Chemical Society* **2003,** *125* (2), 541-548.

71. Facelli, J. C.; Grant, D. M., Determination of molecular symmetry in crystalline naphthalene using solid-state NMR. *Nature* **1993,** *365* (6444), 325-7.

72. Harris, R. K.; Joyce, S. A. 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.* **2006,** *8*, 137-143.

73. Salager, E.; Stein, R. S.; Pickard, C. J.; Elena, B.; Emsley, L., Powder NMR crystallography of thymol. *Phys Chem Chem Phys* **2009,** *11* (15), 2610-21.

74. Abraham, A.; Apperley, D. C.; Gelbrich, T.; Harris, R. K.; Griesser, U. J., NMR crystallography - Three polymorphs of phenobarbital. *Can J Chem* **2011,** *89* (7), 770-778.

75. London, F., The quantic theory of inter-atomic currents in aromatic combinations. *J Phys-Paris* **1937,** *8*, 397-409.

76. Ramsey, N. F., Magnetic shielding of nuclei in molecules. *Physical Review* **1950,** *78* (6), 699-703.

77. Mcweeny, R., Perturbation Theory for Fock-Dirac Density Matrix. *Physical Review* **1962,** *126* (3), 1028-+.

78. Ditchfield, R., Molecular-orbital theory of magnetic shielding and magnetic susceptibility. *J Chem Phys* **1972,** *56* (11), 5688-+.

79. Schindler, M.; Kutzelnigg, W., Theory of magnetic-susceptibilities and NMR chemical shifts in terms of localized quantities .3. Application to hydrocarbons and other organic-molecules. *Journal of the American Chemical Society* **1983,** *105* (5), 1360-1370.

80. Hansen, A. E.; Bouman, T. D., Localized orbital local origin method for calculation and analysis of NMR shieldings - Applications to C-13 shielding tensors. *J Chem Phys* **1985,** *82* (11), 5035-5047.

81. Blochl, P. E., Projector augmented-wave method. *Phys Rev B Condens Matter* **1994,** *50* (24), 17953-17979.

82. Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. J., A comparison of models for calculating nuclear magnetic resonance shielding tensors. *J Chem Phys* **1996,** *104* (14), 5497-5509.

83. Hartman, J. D.; Kudla, R. A.; Day, G. M.; Mueller, L. J.; Beran, G. J., Benchmark fragment-based (1)H, (13)C, (15)N and (17)O chemical shift predictions in molecular crystals. *Phys Chem Chem Phys* **2016,** *18* (31), 21686-709.

84. Hartman, J. D.; Beran, G. J. O., Accurate 13-C and 15-N molecular crystal chemical shielding tensors from fragment-based electronic structure theory. *Solid State Nucl Magn Reson* **2018,** *96*, 10-18.

85. Harper, J. K.; Barich, D. H.; Hu, J. Z.; Strobel, G. A.; Grant, D. M., Stereochemical analysis by solid-state NMR: Structural predictions in ambuic acid. *Journal of Organic Chemistry* **2003,** *68* (12), 4609-4614.

86. Harris, R. K.; Wasylishen, R. E.; Duer, M. J., *NMR Crystallography*. John Wiley & Sons Ltd: United Kingdom, 2009.

87. Wishart, D. S.; Watson, M. S.; Boyko, R. F.; Sykes, B. D., Automated H-1 and C-13 chemical shift prediction using the BioMagResBank. *Journal of Biomolecular Nmr* **1997,** *10* (4), 329-336.

88. 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* **1999,** *13* (3), 199-211.

89. Xu, X. P.; Case, D. A., Automated prediction of (15)N, (13)C(alpha), (13)C(beta) and (13)C ' chemical shifts in proteins using a density functional database. *Journal of Biomolecular Nmr* **2001,** *21* (4), 321-333.

90. 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. *Journal of Biomolecular Nmr* **2003,** *26* (3), 215-240.

91. Shen, Y.; Bax, A., Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J Biomol NMR* **2007,** *38* (4), 289-302.

92. Moon, S.; Case, D. A., A new model for chemical shifts of amide hydrogens in proteins. *J Biomol NMR* **2007,** *38* (2), 139-50.

93. 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* **2009,** *106* (40), 16972-7.

94. 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* **2009,** *131* (39), 13894-5.

95. Meiler, J., PROSHIFT: Protein chemical shift prediction using artificial neural networks. *Journal of Biomolecular Nmr* **2003,** *26* (1), 25-37.

96. Shen, Y.; Bax, A., SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J Biomol NMR* **2010,** *48* (1), 13-22.

97. Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S., SHIFTX2: significantly improved protein chemical shift prediction. *J Biomol NMR* **2011,** *50* (1), 43-57.

98. Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J., HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *Journal of the American Chemical Society* **2003,** *125* (7), 1731-1737.

99. Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M., Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci. U. S. A.* **2007,** *104* (23), 9615-9620.

100. Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J. M.; Liu, G. H.; Eletsky, A.; Wu, Y. B.; Singarapu, K. K.; Lemak, A.; Ignatchenko, A.; Arrowsmith, C. H.; Szyperski, T.; Montelione, G. T.; Baker, D.; Bax, A., Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci. U. S. A.* **2008,** *105* (12), 4685-4690.

101. 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* **2016,** *12* (2), 765-73.

102. Chaker, Z.; Salanne, M.; Delaye, J.-M.; Charpentier, T., NMR shifts in aluminosilicate glasses via machine learning. *Physical Chemistry Chemical Physics* **2019**.

103. Liu, S.; Li, J.; Bennett, K. C.; Ganoe, B.; Stauch, T.; Head-Gordon, M.; Hexemer, A.; Ushizima, D.; Head-Gordon, T., A Multi-Resolution 3D-DenseNet for Chemical Shift Prediction in NMR Crystallography. *arXiv preprint arXiv:1906.00102* **2019**.

104. Sebastiani, D.; Parrinello, M., A new ab-initio approach for NMR chemical shifts in periodic systems. *Journal of Physical Chemistry A* **2001,** *105* (10), 1951-1958.

105. Profeta, M.; Benoit, M.; Mauri, F.; Pickard, C. J., First-principles calculation of the 17O NMR parameters in Ca oxide and Ca aluminosilicates: the partially covalent nature of the Ca-O bond, a challenge for density functional theory. *J Am Chem Soc* **2004,** *126* (39), 12628-35.

106. Marques, M. A. L.; d'Avezac, M.; Mauri, F., Magnetic response and NMR spectra of carbon nanotubes from ab initio calculations. *Phys Rev B* **2006,** *73* (12).

107. Ashbrook, S. E.; Berry, A. J.; Frost, D. J.; Gregorovic, A.; Pickard, C. J.; Readman, J. E.; Wimperis, S., O-17 and Si-29 NMR parameters of MgSiO3 phases from high-resolution solid-state NMR spectroscopy and first-principles calculations. *Journal of the American Chemical Society* **2007,** *129* (43), 13213-13224.

108. Moudrakovski, I. L.; Alizadeh, R.; Beaudoin, J. J., Natural abundance high field (43)Ca solid state NMR in cement science. *Phys Chem Chem Phys* **2010,** *12* (26), 6961-9.

109. Rejmak, P.; Dolado, J. S.; Stott, M. J.; Ayuela, A., 29Si NMR in Cement: A Theoretical Study on Calcium Silicate Hydrates. *The Journal of Physical Chemistry C* **2012,** *116* (17), 9755-9761.

110. Holmes, S. T.; Iuliucci, R. J.; Mueller, K. T.; Dybowski, C., Critical Analysis of Cluster Models and Exchange-Correlation Functionals for Calculating Magnetic Shielding in Molecular Solids. *Journal of Chemical Theory and Computation* **2015,** *11* (11), 5229-5241.

111. Hartman, J. D.; Monaco, S.; Schatschneider, B.; Beran, G. J. O., Fragment-based C-13 nuclear magnetic resonance chemical shift predictions in molecular crystals: An alternative to planewave methods. *J Chem Phys* **2015,** *143* (10).

112. 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. *Physical Chemistry Chemical Physics* **2019**.

113. Reilly, A. M.; Cooper, R. I.; Adjiman, C. S.; Bhattacharya, S.; Boese, A. D.; Brandenburg, J. G.; Bygrave, P. J.; Bylsma, R.; Campbell, J. E.; Car, R.; Case, D. H.; Chadha, R.; Cole, J. C.; Cosburn, K.; Cuppen, H. M.; Curtis, F.; Day, G. M.; DiStasio, R. A.; Dzyabchenko, A.; Van Eijck, B. P.; Elking, D. M.; Van Den Ende, J. A.; Facelli, J. C.; Ferraro, M. B.; Fusti-Molnar, L.; Gatsiou, C. A.; Gee, T. S.; De Gelder, R.; Ghiringhelli, L. M.; Goto, H.; Grimme, S.; Guo, R.; Hofmann, D. W. M.; Hoja, J.; Hylton, R. K.; Iuzzolino, L.; Jankiewicz, W.; De Jong, D. l. T.; Kendrick, J.; De Klerk, N. J. J.; Ko, H. Y.; Kuleshova, L. N.; Li, X.; Lohani, S.; Leusen, F. J. J.; Lund, A. M.; Lv, J.; Ma, Y.; Marom, N.; Masunov, A. m. E.; McCabe, P.; McMahon, D. P.; Meekes, H.; Metz, M. P.; Misquitta, A. J.; Mohamed, S.; Monserrat, B.; Needs, R. J.; Neumann, M. A.; Nyman, J.; Obata, S.; Oberhofer, H.; Oganov, A. R.; Orendt, A. M.; Pagola, G. I.; Pantelides, C. C.; Pickard, C. J.; Podeszwa, R.; Price, L. S.; Price, S. L.; Pulido, A.; Read, M. G.; Reuter, K.; Schneider, E.; Schober, C.; Shields, G. P.; Singh, P.; Sugden, I. J.; Szalewicz, K.; Taylor, C. R.; Tkatchenko, A.; Tuckerman, M. E.; Vacarro, F.; Vasileiadis, M.; Vazquez-Mayagoitia, A.; Vogt, L.; Wang, Y.; Watson, R. E.; De Wijs, G. A.; Yang, J.; Zhu, Q.; Groom, C. R., Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2016,** *72* (4), 439-459.

114. Nilsson Lill, S. O.; Widdifield, C. M.; Pettersen, A.; Svensk Ankarberg, A.; Lindkvist, M.; Aldred, P.; Gracin, S.; Shankland, N.; Shankland, K.; Schantz, S.; Emsley, L., Elucidating an Amorphous Form Stabilization Mechanism for Tenapanor Hydrochloride: Crystal Structure Analysis Using X-ray Diffraction, NMR Crystallography, and Molecular Modeling. *Mol Pharm* **2018,** *15* (4), 1476-1487.

115. Charpentier, T.; Ispas, S.; Profeta, M.; Mauri, F.; Pickard, C. J., First-principles calculation of O-17, Si-29, and Na-23 NMR spectra of sodium silicate crystals and glasses. *Journal of Physical Chemistry B* **2004,** *108* (13), 4147-4161.

116. Charpentier, T.; Kroll, P.; Mauri, F., First-Principles Nuclear Magnetic Resonance Structural Analysis of Vitreous Silica. *Journal of Physical Chemistry C* **2009,** *113* (18), 7917-7929.

117. Ispas, S.; Charpentier, T.; Mauri, F.; Neuville, D. R., Structural properties of lithium and sodium tetrasilicate glasses: Molecular dynamics simulations versus NMR experimental and first-principles data. *Solid State Sciences* **2010,** *12* (2), 183-192.

118. Pedone, A.; Charpentier, T.; Menziani, M. C., Multinuclear NMR of CaSiO3 glass: simulation from first-principles. *Physical Chemistry Chemical Physics* **2010,** *12* (23), 6054-6066.

119. Soleilhavoup, A.; Delaye, J.-M.; Angeli, F.; Caurant, D.; Charpentier, T., Contribution of first-principles calculations to multinuclear NMR analysis of borosilicate glasses. *Magnetic Resonance in Chemistry* **2010,** *48*, S159-S170.

120. Sykina, K.; Bureau, B.; Le Polles, L.; Roiland, C.; Deschamps, M.; Pickard, C. J.; Furet, E., A combined Se-77 NMR and molecular dynamics contribution to the structural understanding of the chalcogenide glasses. *Physical Chemistry Chemical Physics* **2014,** *16* (33), 17975-17982.

121. Gambuzzi, E.; Pedone, A.; Menziani, M. C.; Angeli, F.; Florian, P.; Charpentier, T., Calcium environment in silicate and aluminosilicate glasses probed by Ca-43 MQMAS NMR experiments and MD-GIPAW calculations. *Solid State Nuclear Magnetic Resonance* **2015,** *68-69*, 31-36.

122. Bagno, A.; Casella, G.; Saielli, G., Relativistic DFT calculation of 119Sn chemical shifts and coupling constants in tin compounds. *Journal of Chemical Theory and Computation* **2006,** *2* (1), 37-46.

123. Alkan, F.; Dybowski, C., Calculation of chemical-shift tensors of heavy nuclei: a DFT/ZORA investigation of\n 199\n Hg chemical-shift tensors in solids, and the effects of cluster size and electronic-state approximations. *Phys. Chem. Chem. Phys.* **2014,** *16* (27), 14298-14308.

124. Alkan, F.; Dybowski, C., Chemical-shift tensors of heavy nuclei in network solids : a DFT / ZORA investigation of 207 Pb. *Physical Chemistry Chemical Physics* **2015,** *17*, 25014-25026.

125. Alkan, F.; Holmes, S. T.; Iuliucci, R. J.; Mueller, K. T.; Dybowski, C., Spin-orbit effects on the (119)Sn magnetic-shielding tensor in solids: a ZORA/DFT investigation. *Phys Chem Chem Phys* **2016,** *18* (28), 18914-22.

126. Alkan, F.; Dybowski, C., Spin-orbit effects on the (125)Te magnetic-shielding tensor: A cluster-based ZORA/DFT investigation. *Solid State Nucl Magn Reson* **2018,** *95*, 6-11.

127. Holmes, S. T.; Schurko, R. W., A DFT/ZORA Study of Cadmium Magnetic Shielding Tensors: Analysis of Relativistic Effects and Electronic-State Approximations. *J Chem Theory Comput* **2019,** *15* (3), 1785-1797.

128. Moon, S.; Patchkovskii, S., *First-principles calculations of paramagnetic NMR shifts, in Calculation of NMR and EPR parameters*. Wiley: Weinheim, 2004.

129. Hrobarik, P.; Reviakine, R.; Arbuznikov, A. V.; Malkina, O. L.; Malkin, V. G.; Kohler, F. H.; Kaupp, M., Density functional calculations of NMR shielding tensors for paramagnetic systems with arbitrary spin multiplicity: Validation on 3d metallocenes. *J Chem Phys* **2007,** *126* (2).

130. Autschbach, J.; Patchkovskii, S.; Pritchard, B., Calculation of Hyperfine Tensors and Paramagnetic NMR Shifts Using the Relativistic Zeroth-Order Regular Approximation and Density Functional Theory. *Journal of Chemical Theory and Computation* **2011,** *7* (7), 2175-2188.

131. Viger-Gravel, J.; Avalos, C. E.; Kubicki, D. J.; Gajan, D.; Lelli, M.; Ouari, O.; Lesage, A.; Emsley, L., 19F Magic Angle Spinning Dynamic Nuclear Polarization Enhanced NMR Spectroscopy. *Angewandte Chemie International Edition* **2019,** *58* (22), 7249-7253.

132. Brown, I. D.; Bergerhoff, G., Databases of Inorganic Crystal-Structures. *Abstr Pap Am Chem S* **1979,** (Sep), 36-36.

133. Grazulis, S.; Chateigner, D.; Downs, R. T.; Yokochi, A. F. T.; Quiros, M.; Lutterotti, L.; Manakova, E.; Butkus, J.; Moeck, P.; Le Bail, A., Crystallography Open Database - an open-access collection of crystal structures. *J Appl Crystallogr* **2009,** *42*, 726-729.

134. Grazulis, S.; Daskevic, A.; Merkys, A.; Chateigner, D.; Lutterotti, L.; Quiros, M.; Serebryanaya, N. R.; Moeck, P.; Downs, R. T.; Le Bail, A., Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Res* **2012,** *40* (D1), D420-D427.

135. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2016,** *72* (2).

136. Batisai, E.; Ayamine, A.; Kilinkissa, O. E. Y.; Bathori, N. B., Melting point-solubility-structure correlations in multicomponent crystals containing fumaric or adipic acid. *Crystengcomm* **2014,** *16* (43), 9992-9998.

137. Sternberg, U.; Koch, F.-T.; Prieß, W.; Witter, R., Crystal structure refinements of cellulose polymorphs using solid state 13C chemical shifts. *Cellulose* **2003,** *10* (3), 189-199.

138. Heider, E. M.; Harper, J. K.; Grant, D. M., Structural characterization of an anhydrous polymorph of paclitaxel by solid-state NMR. *Phys Chem Chem Phys* **2007,** *9* (46), 6083-97.

139. Zhu, L. Y.; Agarwal, A.; Lai, J. F.; Al-Kaysi, R. O.; Tham, F. S.; Ghaddar, T.; Mueller, L.; Bardeen, C. J., Solid-state photochemical and photomechanical properties of molecular crystal nanorods composed of anthracene ester derivatives. *J. Mater. Chem.* **2011,** *21* (17), 6258-6268.

140. Harris, R. K., NMR crystallography: the use of chemical shifts. *Solid State Sciences* **2004,** *6* (10), 1025-1037.

141. Harris, R. K., NMR studies of organic polymorphs and solvates. *Analyst* **2006,** *131* (3), 351-73.

142. Mifsud, N.; Elena, B.; Pickard, C. J.; Lesage, A.; Emsley, L., Assigning powders to crystal structures by high-resolution (1)H-(1)H double quantum and (1)H-(13)C J-INEPT solid-state NMR spectroscopy and first principles computation. A case study of penicillin G. *Phys Chem Chem Phys* **2006,** *8* (29), 3418-22.

143. Harris, R. K., Applications of solid-state NMR to pharmaceutical polymorphism and related matters. *J Pharm Pharmacol* **2007,** *59* (2), 225-39.

144. Othman, A.; Evans, J. S.; Evans, I. R.; Harris, R. K.; Hodgkinson, P., Structural study of polymorphs and solvates of finasteride. *J Pharm Sci-Us* **2007,** *96* (5), 1380-97.

145. Pawlak, T.; Jaworska, M.; Potrzebowski, M. J., NMR crystallography of alpha-poly(L-lactide). *Phys Chem Chem Phys* **2013,** *15* (9), 3137-45.

146. Koike, R.; Higashi, K.; Liu, N.; Limwikrant, W.; Yamamoto, K.; Moribe, K., Structural Determination of a Novel Polymorph of Sulfathiazole-Oxalic Acid Complex in Powder Form by Solid-State NMR Spectroscopy on the Basis of Crystallographic Structure of Another Polymorph. *Cryst Growth Des* **2014,** *14* (9), 4510-4518.

147. Paluch, P.; Pawlak, T.; Oszajca, 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 Reson* **2015,** *65*, 2-11.

148. 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* **2015,** *12* (11), 4146-53.

149. Kalakewich, K.; Iuliucci, R.; Mueller, K. T.; Eloranta, H.; Harper, J. K., Monitoring the refinement of crystal structures with N-15 solid-state NMR shift tensor data. *J Chem Phys* **2015,** *143* (19).

150. Neumann, M. A.; de Streek, J. V.; Fabbiani, F. P. A.; Hidber, P.; Grassmann, O., Combined crystal structure prediction and high-pressure crystallization in rational pharmaceutical polymorph screening. *Nat Commun* **2015,** *6*.

151. 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* **2012,** *108* (5), 058301.

152. Curtarolo, S.; Hart, G. L.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O., The high-throughput highway to computational materials design. *Nat Mater* **2013,** *12* (3), 191-201.

153. Xue, D.; Balachandran, P. V.; Hogden, J.; Theiler, J.; Xue, D.; Lookman, T., Accelerated search for materials with targeted properties by adaptive design. *Nat Commun* **2016,** *7*, 11241.

154. Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C., A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Comput Mater* **2016,** *2* (1).

155. Bartok, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csanyi, G.; Ceriotti, M., Machine learning unifies the modeling of materials and molecules. *Sci Adv* **2017,** *3* (12), e1701816.

156. Aires-de-Sousa, J.; Hemmer, M. C.; Gasteiger, J., Prediction of 1H NMR chemical shifts using neural networks. *Analytical Chemistry* **2002,** *74* (1), 80-90.

157. Blinov, K.; Smurnyy, Y.; Elyashberg, M.; Churanova, T.; Kvasha, M.; Steinbeck, C.; Lefebvre, B.; Williams, A., Performance validation of neural network based 13C NMR prediction using a publicly available data source. *Journal of chemical information and modeling* **2008,** *48* (3), 550-555.

158. Kuhn, S.; Egert, B.; Neumann, S.; Steinbeck, C., Building blocks for automated elucidation of metabolites: machine learning methods for NMR prediction. *BMC Bioinformatics* **2008,** *9* (1), 400.

159. 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* **2008,** *48* (1), 128-34.

160. Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O. A., Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J Phys Chem Lett* **2015,** *6* (16), 3309-3313.

161. Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L., Chemical shifts in molecular solids by machine learning. *Nat Commun* **2018,** *9* (1), 4501.

162. Thompson, H. P. G.; Day, G. M., Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape. *Chem Sci* **2014,** *5* (8), 3173-3182.

163. Chierotti, M. R.; Gobetto, R., NMR crystallography: the use of dipolar interactions in polymorph and co-crystal investigation. *Crystengcomm* **2013,** *15* (43).

164. Jacob, P.; Kalakewich, K.; Uribe-Romo, F. J.; Harper, J. K., Solid-state NMR and DFT predictions of differences in COOH hydrogen bonding in odd and even numbered n-alkyl fatty acids. *Physical Chemistry Chemical Physics* **2016,** *18*, 12541-12549.

165. 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. *Journal of the American Chemical Society* **1987,** *109* (14), 4163-4169.

166. Colombo, M.; Meier, B.; Ernst, R., Rotor-driven spin diffusion in natural-abundance 13C spin systems. *Chemical physics letters* **1988,** *146* (3-4), 189-196.

167. Raleigh, D.; Levitt, M.; Griffin, R., Rotational resonance in solid state NMR. *Chemical Physics Letters* **1988,** *146* (1-2), 71-76.

168. Van Rossum, B.-J.; De Groot, C.; Ladizhansky, V.; Vega, S.; De Groot, H., A method for measuring heteronuclear (1H− 13C) distances in high speed MAS NMR. *Journal of the American Chemical Society* **2000,** *122* (14), 3465-3472.

169. Seidel, K.; Etzkorn, M.; Sonnenberg, L.; Griesinger, C.; Sebald, A.; Baldus, M., Studying molecular 3D structure and dynamics by high-resolution solid-state NMR: Application to l-tyrosine-ethylester. *The Journal of Physical Chemistry A* **2005,** *109* (11), 2436-2442.

170. Dekhil, M.; Mollica, G.; Bonniot, T. T.; Ziarelli, F.; Thureau, P.; Viel, S., Determining carbon–carbon connectivities in natural abundance organic powders using dipolar couplings. *Chemical Communications* **2016,** *52* (55), 8565-8568.

171. Märker, K.; Paul, S.; Fernández-de-Alba, C.; Lee, D.; Mouesca, J.-M.; Hediger, S.; De Paëpe, G., Welcoming natural isotopic abundance in solid-state NMR: probing π-stacking and supramolecular structure of organic nanoassemblies using DNP. *Chem Sci* **2017,** *8* (2), 974-987.

172. Thureau, P.; Sturniolo, S.; Zilka, M.; Ziarelli, F.; Viel, S.; Yates, J.; Mollica, G., Reducing the computational cost of NMR crystallography of organic powders at natural isotopic abundance with the help of 13C-13C dipolar couplings. *Magnetic Resonance in Chemistry* **2019**.

173. 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−13C Heteronuclear Correlation Spectra. *Journal of the American Chemical Society* **1996,** *118* (4), 822-829.

174. Cerrioti, M.; De, S.; Meissner, R. H.; A., T. G. Sketch map package. https://github.com/cosmo-epfl/sketchmap/.

175. Ceriotti, M.; Tribello, G. A.; Parrinello, M., From the Cover: Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences* **2011,** *108* (32), 13023-13028.

176. De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M., Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016,** *18* (20), 13754-13769.

177. De, S.; Musil, F.; Ingram, T.; Baldauf, C.; Ceriotti, M., Mapping and classifying molecules from a high-throughput structural database. *JOURNAL OF CHEMINFORMATICS* **2017,** *9*.

178. Antzutkin, O. N.; Lee, Y. K.; Levitt, M. H., 13C and15N—Chemical Shift Anisotropy of Ampicillin and Penicillin-V Studied by 2D-PASS and CP/MAS NMR. *Journal of Magnetic Resonance* **1998,** *135* (1), 144-155.

179. Murthy, H. M. K.; Bhat, T. N.; Vijayan, M., Structural Studies of Analgesics and Their Interactions .9. Structure of a New Crystal Form of 2-((3-(Trifluoromethyl)Phenyl)Amino)Benzoic Acid (Flufenamic Acid). *Acta Crystallogr B* **1982,** *38* (Jan), 315-317.

180. Hrynchuk, R. J.; Barton, R. J.; Robertson, B. E., The crystal structure of free base cocaine, C 17 H 21 NO 4. *Canadian Journal of Chemistry* **1983,** *61*, 481-487.

181. Cense, J. M.; Agafonov, V.; Ceolin, R.; Ladure, P.; Rodier, N., Crystal and Molecular-Structure Analysis of Flutamide - Bifurcated Helicoidal C-H ... O Hydrogen-Bonds. *Struct Chem* **1994,** *5* (2), 79-84.

182. Hayashi, S.; Hayamizu, K., Chemical Shift Standards in High-Resolution Solid-State NMR (1) 13C, 29Si, and 1H Nuclei. *Bulletin of the Chemical Society of Japan* **1991,** *64* (2), 685-687.

183. Rasmussen, C. E.; Williams, C. K., *Gaussian processes for machine learning*. MIT press Cambridge: 2006; Vol. 1.

184. Bartók, A. P.; Kondor, R.; Csányi, G., On representing chemical environments. *Phys Rev B* **2013,** *87* (18), 1-16.

185. Grisafi, A.; Wilkins, D. M.; Csanyi, G.; Ceriotti, M., Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys Rev Lett* **2018,** *120* (3), 036002.

186. Ceriotti, M.; Tribello, G. A.; Parrinello, M., Demonstrating the Transferability and the Descriptive Power of Sketch-Map. *J Chem Theory Comput* **2013,** *9* (3), 1521-32.

187. Campello, R. J.; Moulavi, D.; Zimek, A.; Sander, J., Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **2015,** *10* (1), 5.

188. Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Davide, C.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; Corso, A. D.; Gironcoli, d. S.; Fabris, S.; Fratesi, G.; Gebauer, R.; Gerstmann, U.; Gougoussis, C.; Anton, K.; Lazzeri, M.; Martin-Samos, L.; Marzari, N.; Mauri, F.; Mazzarello, R.; Stefano, P.; Pasquarello, A.; Paulatto, L.; Sbraccia, C.; Scandolo, S.; Sclauzero, G.; Seitsonen, A. P.; Smogunov, A.; Umari, P.; Wentzcovitch, R. M., QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter* **2009,** *21* (39), 395502.

189. 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. *Computer Physics Communications* **2013,** *184* (8), 1827-1833.

190. Giannozzi, P.; Andreussi, O.; Brumme, T.; Bunau, O.; Buongiorno Nardelli, M.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Cococcioni, M.; Colonna, N.; Carnimeo, I.; Dal Corso, A.; de Gironcoli, S.; Delugas, P.; DiStasio, R. A.; Ferretti, A.; Floris, A.; Fratesi, G.; Fugallo, G.; Gebauer, R.; Gerstmann, U.; Giustino, F.; Gorni, T.; Jia, J.; Kawamura, M.; Ko, H. Y.; Kokalj, A.; Kucukbenli, E.; Lazzeri, M.; Marsili, M.; Marzari, N.; Mauri, F.; Nguyen, N. L.; Nguyen, H. V.; Otero-de-la-Roza, A.; Paulatto, L.; Ponce, S.; Rocca, D.; Sabatini, R.; Santra, B.; Schlipf, M.; Seitsonen, A. P.; Smogunov, A.; Timrov, I.; Thonhauser, T.; Umari, P.; Vast, N.; Wu, X.; Baroni, S., Advanced capabilities for materials modelling with Quantum ESPRESSO. *J Phys Condens Matter* **2017,** *29* (46), 465901.

191. Clark, S. J.; Segall, M. D.; Pickard, C. J.; Hasnip, P. J.; Probert, M. J.; Refson, K.; Payne, M. C., First principles methods using CASTEP. *Z Kristallogr* **2005,** *220* (5-6), 567-570.

192. Arico-Muendel, C. C.; Blanchette, H.; Benjamin, D. R.; Caiazzo, T. M.; Centrella, P. A.; DeLorey, J.; Doyle, E. G.; Johnson, S. R.; Labenski, M. T.; Morgan, B. A.; O'Donovan, G.; Sarjeant, A. A.; Skinner, S.; Thompson, C. D.; Griffin, S. T.; Westlin, W.; White, K. F., Orally active fumagillin analogues: transformations of a reactive warhead in the gastric environment. *Acs Med Chem Lett* **2013,** *4* (4), 381-6.

193. Dao, H. T.; Li, C.; Michaudel, Q.; Maxwell, B. D.; Baran, P. S., Hydromethylation of Unactivated Olefins. *J Am Chem Soc* **2015,** *137* (25), 8046-9.

194. Garozzo, D.; Gattuso, G.; Kohnke, F. H.; Notti, A.; Pappalardo, S.; Parisi, M. F.; Pisagatti, I.; White, A. J.; Williams, D. J., Inclusion networks of a calix[5]arene-based exoditopic receptor and long-chain alkyldiammonium ions. *Org Lett* **2003,** *5* (22), 4025-8.

195. Bats, J. W., *CSD Communication* **2010**.

196. Huang, G. B.; Liu, W. E.; Valkonen, A.; Yao, H.; Rissanen, K.; Jiang, W., 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* **2018,** *29* (1), 91-94.

197. 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* **2017,** *41* (4), 235-238.

198. Willatt, M. J.; Musil, F.; Ceriotti, M., Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Physical Chemistry Chemical Physics* **2018,** *20* (47), 29661-29668.

199. Csato, L.; Opper, M., Sparse on-line Gaussian processes. *Neural Comput* **2002,** *14* (3), 641-668.

200. Seeger, M.; Williams, C.; Lawrence, N., Fast Forward Selection to Speed Up Sparse Gaussian Process Regression. **2003**.

201. Musil, F.; Willatt, M. J.; Langovoy, M. A.; Ceriotti, M., Fast and Accurate Uncertainty Estimation in Chemical Machine Learning. *Journal of Chemical Theory and Computation* **2019,** *15* (2), 906-915.

202. Imbalzano, G.; Anelli, A.; Giofre, D.; Klees, S.; Behler, J.; Ceriotti, M., Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J Chem Phys* **2018,** *148* (24).

203. Day, G. M.; Motherwell, W. S.; Jones, W., A strategy for predicting the crystal structures of flexible molecules: the polymorphism of phenobarbital. *Physical Chemistry Chemical Physics* **2007,** *9* (14), 1693-1704.

204. Lejaeghere, K.; Bihlmayer, G.; Björkman, T.; Blaha, P.; Blügel, S.; Blum, V.; Caliste, D.; Castelli, I. E.; Clark, S. J.; Dal Corso, A., Reproducibility in density functional theory calculations of solids. *Science* **2016,** *351* (6280), aad3000.

205. Perdew, J. P.; Burke, K.; Ernzerhof, M., Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996,** *77* (18), 3865.

206. Grimme, S., Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of computational chemistry* **2006,** *27* (15), 1787-1799.

207. Monkhorst, H. J.; Pack, J. D., Special points for Brillouin-zone integrations. *Phys Rev B* **1976,** *13* (12), 5188.

208. Musil, F.; De, D.; Ceriotti, M. Glosim2 package. https://github.com/cosmo-epfl/glosim2.

209. Hofstetter, A.; Emsley, L., Positional Variance in NMR Crystallography. *J. Am. Chem. Soc* **2017,** *139* (7), 2573-2576.

210. 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. *The Journal of Physical Chemistry C* **2013,** *117* (34), 17731-17740.

211. Uldry, A.-C.; Griffin, J. M.; Yates, J. R.; Pérez-Torralba, M.; Santa María, M. D.; Webber, A. L.; Beaumont, M. L.; Samoson, A.; Claramunt, R. M.; Pickard, C. J., 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. *Journal of the American Chemical Society* **2008,** *130* (3), 945-954.

212. Sardo, M.; Santos, S. M.; Babaryk, A. A.; López, C.; Alkorta, I.; Elguero, J.; Claramunt, R. M.; Mafra, L., Diazole-based powdered cocrystal featuring a helical hydrogen-bonded network: Structure determination from PXRD, solid-state NMR and computer modeling. *Solid state nuclear magnetic resonance* **2015,** *65*, 49-63.

213. Trueblood, K. N.; Bürgi, H. B.; Burzlaff, H.; Dunitz, J. D.; Gramaccioli, C. M.; Schulz, H. H.; Shmueli, U.; Abrahams, S. C.; B\urgi, H.-B., Atomic Dispacement Parameter Nomenclature. Report of a Subcommittee on Atomic Displacement Parameter Nomenclature. *Acta Crystallographica Section A* **1996,** *52*, 770-781.

214. Hamilton, W. C., On the isotropic temperature factor equivalent to a given anisotropic temperature factor. *Acta Crystallographica* **1959,** *12*, 609-610.

215. Willis, B. T. M.; Howard, J. A. K., Do the ellipsoids of thermal vibration mean anything Analysis of neutron diffraction measurements on hexamethylenetetramine. *Acta Crystallographica Section A* **1975,** *31*, 514-520.

216. Harris, R. K.; Hodgkinson, P.; Pickard, C. J.; Yates, J. R.; Zorin, V., Chemical shift computations on a crystallographic basis: Some reflections and comments. *Magnetic Resonance in Chemistry* **2007,** *45*, S174-S186.

217. Hill, D. E.; Vasdev, N.; Holland, J. P., Evaluating the accuracy of density functional theory for calculating 1H and 13C NMR chemical shifts in drug molecules. *Computational and Theoretical Chemistry* **2015,** *1051*, 161-172.

218. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **1996,** *118*, 11225-11236.

219. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C., GROMACS: Fast, flexible, and free. In *Journal of Computational Chemistry*, 2005; Vol. 26, pp 1701-1718.

220. Pack, J. D.; Monkhorst, H. J., special points for Brillouin-zone integrations. *Phys Rev B* **1977,** *16*, 1748-1749.

221. Betteridge, P. W.; Carruthers, J. R.; Cooper, R. I.; Prout, K.; Watkin, D. J., CRYSTALS version 12: software for guided crystal structure analysis. *J Appl Crystallogr* **2003,** *36*, 1487-1487.
222. Watkin, D. J. P., C.K. Pearce, L.J., CAMERON. *Chemical Crystallography Laboratory* **1996**.
223. Monserrat, B.; Needs, R. J.; Pickard, C. J., Temperature effects in first-principles solid state calculations of the chemical shielding tensor made simple. *J Chem Phys* **2014,** *141*.
224. Ashbrook, S. E.; McKay, D., Combining Solid-State NMR Spectroscopy with First-Principles Calculations – A Guide to NMR Crystallography. *Chem. Commun.* **2016,** *52*, 7186-7204.
225. Dračínský, M.; Bouř, P.; Hodgkinson, P., Temperature Dependence of NMR Parameters Calculated from Path Integral Molecular Dynamics Simulations. *Journal of chemical theory and computation* **2016,** *12*, 968-73.
226. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, K. M.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Foz, T.; Caldwell, J. W.; Kollman, P. A., A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids and Organic Molecules. *J. Am. Chem. Soc* **1995,** *117*, 5179.
227. MacKerell, a. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry. B* **1998,** *102*, 3586-616.
228. Martín-García, F.; Papaleo, E.; Gomez-Puertas, P.; Boomsma, W.; Lindorff-Larsen, K., Comparing molecular dynamics force fields in the essential subspace. *PLoS ONE* **2015,** *10*, 1-16.
229. Cizmeciyan, D.; Yonutas, H.; Karlen, S. D.; Garcia-Garibay, M. A., 2H NMR and X-ray diffraction studies of methyl rotation in crystals of ortho-methyldibenzocycloalkanones. *Solid State Nuclear Magnetic Resonance* **2005,** *28*, 1-8.
230. Willatt, M. J.; Musil, F.; Ceriotti, M., Atom-density representations for machine learning. *J Chem Phys* **2019,** *150* (15).
231. Chisholm, J. A.; Motherwell, S., COMPACK: a program for identifying crystal structure similarity using distances. *J Appl Crystallogr* **2005,** *38*, 228-231.
232. Karamertzanis, P. G.; Pantelides, C. C., Ab initio crystal structure prediction - I. Rigid molecules. *Journal of Computational Chemistry* **2005,** *26* (3), 304-324.
233. Becke, A. D., Density-Functional Thermochemistry .3. The Role of Exact Exchange. *J Chem Phys* **1993,** *98* (7), 5648-5652.
234. Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J., Ab-Initio Calculation of Vibrational Absorption and Circular-Dichroism Spectra Using Density-Functional Force-Fields. *J Phys Chem-Us* **1994,** *98* (45), 11623-11627.
235. 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. *Journal of the American Chemical Society* **1996,** *118* (21), 5011-5019.
236. Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M., Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. *Physical Chemistry Chemical Physics* **2010,** *12* (30), 8478-8490.
237. Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M., Role of electrostatic interactions in determining the crystal structures of polar organic molecules. A distributed multipole study. *J Phys Chem-Us* **1996,** *100* (18), 7352-7360.
238. Stone, A. J.; Alderton, M., Distributed Multipole Analysis - Methods and Applications. *Molecular Physics* **1985,** *56* (5), 1047-1064.
239. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C., Efficient Handling of Molecular Flexibility in Lattice Energy Minimization of Organic Crystals. *Journal of Chemical Theory and Computation* **2011,** *7* (6), 1998-2016.
240. Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Montgomery Jr, J.; Vreven, T.; Kudin, K.; Burant, J., Gaussian 03, revision c. 02; Gaussian. *Inc., Wallingford, CT* **2004,** *4*.
241. Tkatchenko, A.; Scheffler, M., Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **2009,** *102* (7).
242. Dal Corso, A., Pseudopotentials periodic table: From H to Pu. *Computational Materials Science* **2014,** *95*, 337-350.
243. Szlachta, W. J.; Bartok, A. P.; Csanyi, G., Accuracy and transferability of Gaussian approximation potential models for tungsten. *Phys Rev B* **2014,** *90* (10).
244. Deringer, V. L.; Csanyi, G., Machine learning based interatomic potential for amorphous carbon. *Phys Rev B* **2017,** *95* (9).
245. Eldar, Y.; Lindenbaum, M.; Porat, M.; Zeevi, Y. Y., The farthest point strategy for progressive image sampling. *Ieee T Image Process* **1997,** *6* (9), 1305-1315.
246. Case, D. H.; Campbell, J. E.; Bygrave, P. J.; Day, G. M., Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling. *Journal of Chemical Theory and Computation* **2016,** *12* (2), 910-924.
247. Nyman, J.; Day, G. M., Static and lattice vibrational energy differences between polymorphs. *Crystengcomm* **2015,** *17* (28), 5154-5165.
248. Boles, M. O.; Girven, R. J., The structures of ampicillin: a comparison of the anhydrate and trihydrate forms. *Acta Crystallographica Section B* **1976,** *32* (8), 2279-2284.
249. Clayden, N. J.; Dobson, C. M.; Lian, L.-Y.; Twyman, J. M., A solid-state 13C nuclear magnetic resonance study of the conformational states of penicillins. *Journal of the Chemical Society, Perkin Transactions 2* **1986,** (12), 1933-1940.
250. Kolossváry, I.; Guida, W. C., Low-mode conformational search elucidated: Application to C39H80 and flexible docking of 9-deazaguanine inhibitors into PNP. *Journal of Computational Chemistry* **1999,** *20* (15), 1671-1684.
251. *MacroModel*, V9.0; Schrödinger LLC: New York, NY, 2011.

252. Harder, E.; Damm, W.; Maple, J.; Wu, C. J.; Reboul, M.; Xiang, J. Y.; Wang, L. L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A., OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *Journal of Chemical Theory and Computation* **2016,** *12* (1), 281-296.
253. Stone, A. J., Distributed multipole analysis: Stability for large basis sets. *Journal of Chemical Theory and Computation* **2005,** *1* (6), 1128-1132.
254. Charpentier, T., The PAW/GIPAW approach for computing NMR parameters: A new dimension added to NMR study of solids. *Solid State Nuclear Magnetic Resonance* **2011,** *40* (1), 1-20.
255. Harris, R. K.; Jackson, P., High-Resolution H-1 and C-13 Nmr of Solid 2-Aminobenzoic Acid. *J Phys Chem Solids* **1987,** *48* (9), 813-818.
256. Tatton, A. S.; Pham, T. N.; Vogt, F. G.; Iuga, D.; Edwards, A. J.; Brown, S. P., Probing intermolecular interactions and nitrogen protonation in pharmaceuticals by novel N-15-edited and 2D N-14-H-1 solid-state NMR. *Crystengcomm* **2012,** *14* (8), 2654-2659.
257. Ceriotti, M.; More, J.; Manolopoulos, D. E., i-PI: A Python interface for ab initio path integral molecular dynamics simulations. *Computer Physics Communications* **2014,** *185* (3), 1019-1026.
258. Ceriotti, M.; Bussi, G.; Parrinello, M., Colored-Noise Thermostats à la Carte. *Journal of Chemical Theory and Computation* **2010,** *6* (4), 1170-1180.
259. Kotsiantis, S. B., Supervised machine learning: A review of classification techniques. *Informatica* **2007,** *31*, 249-268.
260. Schmidhuber, J., Deep learning in neural networks: an overview. *Neural Netw* **2015,** *61*, 85-117.
261. Goh, G. B.; Hodas, N. O.; Vishnu, A., Deep learning for computational chemistry. 2017; Vol. 38.
262. Ryczko, K.; Mills, K.; Luchak, I.; Homenick, C.; Tamblyn, I., Convolutional neural networks for atomistic systems. *Computational Materials Science* **2018,** *149*, 134-142.
263. Le Bourhis, E., In *Glass; Mechanics and Technology*, Wiley-VCH Verlag GmbH & Co. KGaA: 2007; pp 39-51.
264. Taylor, H. F. W., *Cement Chemistry*. 1997.
265. Scrivener, K. L.; Nonat, A., Hydration of cementitious materials, present and future. *Cement and Concrete Research* **2011,** *41* (7), 651-665.
266. Sales, B. C.; Boatner, L. A., Lead-Iron Phosphate-Glass - a Stable Storage Medium for High-Level Nuclear Waste. *Science* **1984,** *226* (4670), 45-48.
267. Lutze, W.; Malow, G.; Ewing, R. C.; Jercinovic, M. J.; Keil, K., ALTERATION OF BASALT GLASSES - IMPLICATIONS FOR MODELING THE LONG-TERM STABILITY OF NUCLEAR WASTE GLASSES. *Nature* **1985,** *314* (6008), 252-255.
268. Mackenzie, K. J. D.; Smith, M. E., *Multinuclear Solid-State NMR of Inorganic Materials*. Pergamon: Oxford, United Kingdom, 2002; p 201.
269. Bakhmutov, V. I., Solid-State NMR in Materials Science: Principles and Applications. CRC Press: 2011; pp 231-257.
270. Wasylishen, R. E.; Ashbrook, S. E.; Wimperis, S., *NMR of Quadrupolar Nuclei in Solid Materials* Wiley: Chichester, United Kingdom, 2012.
271. De Jong, B.; Schramm, C. M.; Parziale, V. E., Silicon-29 magic angle spinning NMR study on local silicon environments in amorphous and crystalline lithium silicates. *Journal of the American Chemical Society* **1984,** *106* (16), 4396-4402.
272. Feng, T.; Pinal, R.; Carvajal, M. T., Process induced disorder in crystalline materials: Differentiating defective crystals from the amorphous form of griseofulvin. *Journal of Pharmaceutical Sciences* **2008,** *97* (8), 3207-3221.
273. Bonhomme, C.; Gervais, C.; Babonneau, F.; Coelho, C.; Pourpoint, F.; Azais, T.; Ashbrook, S. E.; Griffin, J. M.; Yates, J. R.; Mauri, F.; Pickard, C. J., First-Principles Calculation of NMR Parameters Using the Gauge Including Projector Augmented Wave Method: A Chemist's Point of View. *Chemical Reviews* **2012,** *112* (11), 5733-5779.
274. Charpentier, T.; Menziani, M. C.; Pedone, A., Computational simulations of solid state NMR spectra: a new era in structure determination of oxide glasses. *Rsc Advances* **2013,** *3* (27), 10550-10578.
275. Vasconcelos, F.; Cristol, S.; Paul, J.-F.; Delevoye, L.; Mauri, F.; Charpentier, T.; Le Caer, G., Extended Czjzek model applied to NMR parameter distributions in sodium metaphosphate glass. *Journal of Physics-Condensed Matter* **2013,** *25* (25).
276. Kibalchenko, M.; Yates, J. R.; Pasquarello, A., First-principles investigation of the relation between structural and NMR parameters in vitreous GeO2. *Journal of Physics-Condensed Matter* **2010,** *22* (14).
277. Kibalchenko, M.; Yates, J. R.; Massobrio, C.; Pasquarello, A., Structural Composition of First-Neighbor Shells in GeSe2 and GeSe4 Glasses from a First-Principles Analysis of NMR Chemical Shifts. *Journal of Physical Chemistry C* **2011,** *115* (15), 7755-7759.
278. Benoit, M.; Profeta, M.; Mauri, F.; Pickard, C. J.; Tuckerman, M. E., First-principles calculation of the O-17 NMR parameters of a calcium aluminosilicate glass. *Journal of Physical Chemistry B* **2005,** *109* (13), 6052-6060.
279. Angeli, F.; Villain, O.; Schuller, S.; Ispas, S.; Charpentier, T., Insight into sodium silicate glass structural organization by multinuclear NMR combined with first-principles calculations. *Geochimica Et Cosmochimica Acta* **2011,** *75* (9), 2453-2469.
280. Lee, Y. J.; Bingoel, B.; Murakhtina, T.; Sebastiani, D.; Meyer, W. H.; Wegner, G.; Spiess, H. W., High-resolution solid-state NMR studies of poly(vinyl phosphonic acid) proton-conducting polymer: Molecular structure and proton dynamics. *Journal of Physical Chemistry B* **2007,** *111* (33), 9711-9721.
281. Kins, C. F.; Dudenko, D.; Sebastiani, D.; Brunklaus, G., Molecular Mechanisms of Additive Fortification in Model Epoxy Resins: A Solid State NMR Study. *Macromolecules* **2010,** *43* (17), 7200-7211.
282. Huff, N. T.; Demiralp, E.; Cagin, T.; Goddard, W. A., Factors affecting molecular dynamics simulated vitreous silica structures. *J Non-Cryst Solids* **1999,** *253*, 133-142.
283. Tangney, P.; Scandolo, S., An ab initio parametrized interatomic force field for silica. *J Chem Phys* **2002,** *117* (19), 8898-8904.

284. Davila, L. P.; Caturla, M. J.; Kubota, A.; Sadigh, B.; de la Rubia, T. D.; Shackelford, J. F.; Risbud, S. H.; Garofalini, S. H., Transformations in the medium-range order of fused silica under high pressure. *Phys. Rev. Lett.* **2003,** *91* (20).
285. Trachenko, K.; Dove, M. T., Compressibility, kinetics, and phase transition in pressurized amorphous silica. *Phys Rev B* **2003,** *67* (6).
286. Richardson, I. G., Model structures for C-(A)-S-H(I). *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2014,** *70* (6), 903-923.
287. Lothenbach, B.; Nonat, A., Calcium silicate hydrates: Solid and liquid phase composition. *Cement and Concrete Research* **2015,** *78, Part A*, 57-70.
288. Prati, C.; Gandolfi, M. G., Calcium silicate bioactive cements: Biological perspectives and clinical applications. *Dental Materials* **2015,** *31* (4), 351–370.
289. Ho, C.-C.; Wei, C.-K.; Lin, S.-Y.; Ding, S.-J., Calcium silicate cements prepared by hydrothermal synthesis for bone repair. *Ceramics International* **2016,** *42* (7), 9183–9189.
290. Zhao, J.; Zhu, Y.-J.; Wu, J.; Zheng, J.-Q.; Zhao, X.-Y.; Lu, B.-Q.; Chen, F., Chitosan-coated mesoporous microspheres of calcium silicate hydrate: Environmentally friendly synthesis and application as a highly efficient adsorbent for heavy metal ions. *Journal of Colloid and Interface Science* **2014,** *418*, 208–215.
291. Okano, K.; Miyamaru, S.; Kitao, A.; Takano, H.; Aketo, T.; Toda, M.; Honda, K.; Ohtake, H., Amorphous calcium silicate hydrates and their possible mechanism for recovering phosphate from wastewater. *Separation and Purification Technology* **2015,** *144*, 63–69.
292. Dezerald, L.; Kohanoff, J. J.; Correa, A. A.; Caro, A.; Pellenq, R. J. M.; Ulm, F. J.; Saúl, A., Cement As a Waste Form for Nuclear Fission Products: The Case of 90Sr and Its Daughters. *Environmental Science & Technology* **2015,** *49* (22), 13676–13683.
293. Richardson, I. G.; Groves, G. W., Microstructure and microanalysis of hardened ordinary Portland cement pastes. *JOURNAL OF MATERIALS SCIENCE* **1993,** *28* (1), 265-277.
294. Garbev, K.; Beuchle, G.; Bornefeld, M.; Black, L.; Stemmermann, P., Cell Dimensions and Composition of Nanocrystalline Calcium Silicate Hydrate Solid Solutions. Part 1: Synchrotron-Based X-Ray Diffraction. *Journal of the American Ceramic Society* **2008,** *91* (9), 3005–3014.
295. Chen, J. J.; Sorelli, L.; Vandamme, M.; Ulm, F.-J.; Chanvillard, G., A Coupled Nanoindentation/SEM-EDS Study on Low Water/Cement Ratio Portland Cement Paste: Evidence for C–S–H/Ca(OH)2 Nanocomposites. *Journal of the American Ceramic Society* **2010,** *93* (5), 1484–1493.
296. Grangeon, S.; Claret, F.; Linard, Y.; Chiaberge, C., X-ray diffraction: a powerful tool to probe and understand the structure of nanocrystalline calcium silicate hydrates. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2013,** *69* (5), 465–473.
297. Walker, C. S.; Sutou, S.; Oda, C.; Mihara, M.; Honda, A., Calcium silicate hydrate (C-S-H) gel solubility data and a discrete solid phase model at 25 °C based on two binary non-ideal solid solutions. *Cement and Concrete Research* **2016,** *79*, 1-30.
298. Grangeon, S.; Fernandez-Martinez, A.; Baronnet, A.; Marty, N.; Poulain, A.; Elkaïm, E.; Roosz, C.; Gaboreau, S.; Henocq, P.; Claret, F., Quantitative X-ray pair distribution function analysis of nanocrystalline calcium silicate hydrates: a contribution to the understanding of cement chemistry. *J Appl Crystallogr* **2017,** *50* (1), 14-21.
299. Hirljac, J.; Wu, Z. Q.; Young, J. F., Silicate polymerization during the hydration of alite. *Cement and Concrete Research* **1983,** *13* (6), 877-886.
300. Brown, P. W.; Franz, E.; Frohnsdorff, G.; Taylor, H. F. W., Analyses of the aqueous phase during early C3S hydration. *Cement and Concrete Research* **1984,** *14* (2), 257-262.
301. Brown, P. W.; Pommersheim, J.; Frohnsdorff, G., A kinetic model for the hydration of tricalcium silicate. *Cement and Concrete Research* **1985,** *15* (1), 35-41.
302. Jennings, H. M., Aqueous Solubility Relationships for Two Types of Calcium Silicate Hydrate. *Journal of the American Ceramic Society* **1986,** *69* (8), 614-618.
303. Nicoleau, L.; Bertolim, M. A., Analytical Model for the Alite (C3S) Dissolution Topography. *Journal of the American Ceramic Society* **2016,** *99* (3), 773-786.
304. Chen, J. J.; Thomas, J. J.; Taylor, H. F. W.; Jennings, H. M., Solubility and structure of calcium silicate hydrate. *Cement and Concrete Research* **2004,** *34* (9), 1499-1519.
305. García Lodeiro, I.; Macphee, D. E.; Palomo, A.; Fernández-Jiménez, A., Effect of alkalis on fresh C–S–H gels. FTIR analysis. *Cement and Concrete Research* **2009,** *39* (3), 147-153.
306. Hall, D. A.; Maus, D. C.; Gerfen, G. J.; Inati, S. J.; Becerra, L. R.; Dahlquist, F. W.; Griffin, R. G., Polarization-enhanced NMR spectroscopy of biomolecules in frozen solution. *Science* **1997,** *276* (5314), 930-932.
307. Maly, T.; Debelouchina, G. T.; Bajaj, V. S.; Hu, K. N.; Joo, C. G.; Mak-Jurkauskas, M. L.; Sirigiri, J. R.; Van Der Wel, P. C. A.; Herzfeld, J.; Temkin, R. J.; Griffin, R. G., Dynamic nuclear polarization at high magnetic fields. *J Chem Phys* **2008,** *128* (5), 052211.
308. Ni, Q. Z.; Daviso, E.; Can, T. V.; Markhasin, E.; Jawla, S. K.; Swager, T. M.; Temkin, R. J.; Herzfeld, J.; Griffin, R. G., High frequency dynamic nuclear polarization. *Acc. Chem. Res.* **2013,** *46* (9), 1933-1941.
309. Sangodkar, R. P.; Smith, B. J.; Gajan, D.; Rossini, A. J.; Roberts, L. R.; Funkhouser, G. P.; Lesage, A.; Emsley, L.; Chmelka, B. F., Influences of Dilute Organic Adsorbates on the Hydration of Low-Surface-Area Silicates. *Journal of the American Chemical Society* **2015,** *137* (25), 8096-8112.
310. Lothenbach, B.; Winnefeld, F., Thermodynamic modelling of the hydration of Portland cement. *Cement and Concrete Research* **2006,** *36* (2), 209-226.

311. Kulik, D. A., Improving the structural consistency of C-S-H solid solution thermodynamic models. *Cement and Concrete Research* **2011,** *41* (5), 477-495.

312. Kulik, D. A.; Wagner, T.; Dmytrieva, S. V.; Kosakowski, G.; Hingerl, F. F.; Chudnenko, K. V.; Berner, U. R., GEM-Selektor geochemical modeling package: revised algorithm and GEMS3K numerical kernel for coupled simulation codes. *Comput Geosci* **2012,** *17* (1), 1-24.

313. Rosay, M.; Tometich, L.; Pawsey, S.; Bader, R.; Schauwecker, R.; Blank, M.; Borchard, P. M.; Cauffman, S. R.; Felch, K. L.; Weber, R. T.; Temkin, R. J.; Griffin, R. G.; Maas, W. E., Solid-state dynamic nuclear polarization at 263 GHz: spectrometer design and experimental results. *Phys. Chem. Chem. Phys.* **2010,** *12* (22), 5850–5860.

314. Galmarini, S.; Bowen, P., Atomistic simulation of the adsorption of calcium and hydroxyl ions onto portlandite surfaces — towards crystal growth mechanisms. *Cement and Concrete Research* **2016,** *81*, 16-23.

315. Todorov, I. T.; Smith, W.; Trachenko, K.; Dove, M. T., DL_POLY_3: new dimensions in molecular dynamics simulations via massive parallelism. *J. Mater. Chem.* **2006,** *16* (20), 1911-1918.

316. Nicoleau, L.; Nonat, A.; Perrey, D., The di- and tricalcium silicate dissolutions. *Cement and Concrete Research* **2013,** *47*, 14-30.

317. Richardson, I. g.; Skibsted, J.; Black, L.; Kirkpatrick, R. j., Characterisation of cement hydrate phases by TEM, NMR and Raman spectroscopy. *Advances in Cement Research* **2010,** *22* (4), 233–248.

318. Brough, A. R.; Dobson, C. M.; Richardson, I. G.; Groves, G. W., Application of Selective 29Si Isotopic Enrichment to Studies of the Structure of Calcium Silicate Hydrate (C-S-H) Gels. *Journal of the American Ceramic Society* **1994,** *77* (2), 593-596.

319. Cong, X.; Kirkpatrick, R. J., 29Si MAS NMR study of the structure of calcium silicate hydrate. *Advanced Cement Based Materials* **1996,** *3* (3–4), 144-156.

320. Skibsted, J.; Hall, C., Characterization of cement minerals, cements and their reaction products at the atomic and nano scale. *Cement and Concrete Research* **2008,** *38* (2), 205-225.

321. Alizadeh, R.; Raki, L.; Makar, J. M.; Beaudoin, J. J.; Moudrakovski, I., Hydration of tricalcium silicate in the presence of synthetic calcium–silicate–hydrate. *J. Mater. Chem.* **2009,** *19* (42), 7937.

322. Foley, E. M.; Kim, J. J.; Reda Taha, M. M., Synthesis and nano-mechanical characterization of calcium-silicate-hydrate (C-S-H) made with 1.5 CaO/SiO2 mixture. *Cement and Concrete Research* **2012,** *42* (9), 1225-1232.

323. Pustovgar, E.; Sangodkar, R. P.; Andreev, A. S.; Palacios, M.; Chmelka, B. F.; Flatt, R. J.; Lacaillerie, d. E. d. J.-B., Understanding silicate hydration from quantitative analyses of hydrating tricalcium silicates. *Nat Commun* **2016,** *7*, 10952.

324. Sauvée, C.; Rosay, M.; Casano, G.; Aussenac, F.; Weber, R. T.; Ouari, O.; Tordo, P., Highly Efficient, Water-Soluble Polarizing Agents for Dynamic Nuclear Polarization at High Frequency. *Angewandte Chemie International Edition* **2013,** *52* (41), 10858–10861.

325. Becerra, L. R.; Gerfen, G. J.; Temkin, R. J.; Singel, D. J.; Griffin, R. G., Dynamic nuclear polarization with a cyclotron resonance maser at 5 T. *Phys. Rev. Lett.* **1993,** *71* (21), 3561-3564.

326. Lesage, A.; Lelli, M.; Gajan, D.; Caporini, M. A.; Vitzthum, V.; Miéville, P.; Alauzun, J.; Roussey, A.; Thieuleux, C.; Mehdi, A.; Bodenhausen, G.; Copéret, C.; Emsley, L., Surface enhanced NMR spectroscopy by dynamic nuclear polarization. *Journal of the American Chemical Society* **2010,** *132* (44), 15459-15461.

327. Rossini, A. J.; Zagdoun, A.; Lelli, M.; Lesage, A.; Copéret, C.; Emsley, L., Dynamic Nuclear Polarization Surface Enhanced NMR Spectroscopy. *Acc. Chem. Res.* **2013,** *46* (9), 1942-1951.

328. Thomas, J. J.; Jennings, H. M.; Allen, A. J., Determination of the Neutron Scattering Contrast of Hydrated Portland Cement Paste using H2O/D2O Exchange. *Advanced Cement Based Materials* **1998,** *7* (3–4), 119-122.

329. Gajan, D.; Schwarzwalder, M.; Conley, M. P.; Gruning, W. R.; Rossini, A. J.; Zagdoun, A.; Lelli, M.; Yulikov, M.; Jeschke, G.; Sauvee, C.; Ouari, O.; Tordo, P.; Veyre, L.; Lesage, A.; Thieuleux, C.; Emsley, L.; Coperet, C., Solid-Phase Polarization Matrixes for Dynamic Nuclear Polarization from Homogeneously Distributed Radicals in Mesostructured Hybrid Silica Materials. *Journal of the American Chemical Society* **2013,** *135* (41), 15459-15466.

330. Alemany, L. B.; Grant, D. M.; Pugmire, R. J.; Alger, T. D.; Zilm, K. W., Cross polarization and magic angle sample spinning NMR spectra of model organic compounds. 2. Molecules of low or remote protonation. *J. Am. Chem. Soc.* **1983,** *105* (8), 2142–2147.

331. Lesage, A.; Bardet, M.; Emsley, L., Through-Bond Carbon−Carbon Connectivities in Disordered Solids by NMR. *Journal of the American Chemical Society* **1999,** *121* (47), 10987-10993.

332. Xue, X.; Kanzaki, M., Proton Distributions and Hydrogen Bonding in Crystalline and Glassy Hydrous Silicates and Related Inorganic Materials: Insights from High-Resolution Solid-State Nuclear Magnetic Resonance Spectroscopy. *Journal of the American Ceramic Society* **2009,** *92* (12), 2803–2830.

333. Richardson, I. G., The nature of C-S-H in hardened cements. *Cement and Concrete Research* **1999,** *29* (8), 1131-1147.

334. Bonaccorsi, E.; Merlino, S.; Kampf, A. R., The Crystal Structure of Tobermorite 14 Å (Plombierite), a C–S–H Phase. *Journal of the American Ceramic Society* **2005,** *88* (3), 505-512.

335. Merlino, S.; Bonaccorsi, E.; Armbruster, T., The real structures of clinotobermorite and tobermorite 9 Å: OD character, polytypes, and structural relationships. *European Journal of Mineralogy* **2000**, 411-429.

336. Merlino, S.; Bonaccorsi, E.; Armbruster, T., The real structure of tobermorite 11Å. *European Journal of Mineralogy* **2001,** *13* (3), 577-590.

337. Renaudin, G.; Russias, J.; Leroux, F.; Frizon, F.; Cau-dit-Coumes, C., Structural characterization of C-S-H and C-A-S-H samples-Part I: Long-range order investigated by Rietveld analyses. *Journal of Solid State Chemistry* **2009,** *182*, 3312–3319.

338. Roosz, C.; Gaboreau, S.; Grangeon, S.; Prêt, D.; Montouillout, V.; Maubec, N.; Ory, S.; Blanc, P.; Vieillard, P.; Henocq, P., Distribution of Water in Synthetic Calcium Silicate Hydrates. *Langmuir* **2016,** *32* (27), 6794-6805.

339. Nonat, A.; Lecoq, X., The Structure, Stoichiometry and Properties of C-S-H Prepared by C3S Hydration Under Controlled Condition. In *Nuclear Magnetic Resonance Spectroscopy of Cement-Based Materials*, Colombet, D. P.; Zanni, P. H.; Grimmer, D. A.-R.; Sozzani, P. P.; Colombet, D. P.; Zanni, P. H.; Grimmer, D. A.-R.; Sozzani, P. P., Eds. 1998; pp 197-207.

340. Richardson, I. G., Tobermorite/jennite- and tobermorite/calcium hydroxide-based models for the structure of C-S-H: applicability to hardened pastes of tricalcium silicate, β-dicalcium silicate, Portland cement, and blends of Portland cement with blast-furnace slag, metakaolin, or silica fume. *Cement and Concrete Research* **2004,** *34* (9), 1733-1777.

341. Iller, R., The Chemistry of Silica: Solubility, Polymerization, Colloid and Surface Properties and Biochemistry of Silica. 1979.

342. Brinker, C. J.; Scherer, G. W., Sol-gel Science: The Physics and Chemistry of Sol-gel Processing. **1990**.

343. Galmarini, S. Atomistic Simulation of Cementitious Systems. lausanne, 2013.

344. mboxPhySy, L., *RMN 1.1*. 2016.

345. States, D. J.; Haberkorn, R. A.; Ruben, D. J., A Two-Dimensional Nuclear Overhauser Experiment with Pure Absorption Phase in Four Quadrants. *J. Magn. Reson.* **1982,** *48*, 286–292.

346. Ernst, R. R.; Bodenhausen, G.; Wokaun, A., *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. 1987.

347. van Rossum, B. J.; Förster, H.; de Groot, H. J. M., High-Field and High-Speed CP-MAS13C NMR Heteronuclear Dipolar-Correlation Spectroscopy of Solids with Frequency-Switched Lee–Goldburg Homonuclear Decoupling. *J. Magn. Reson.* **1997,** *124* (2), 516–519.

348. Elena, B.; de Paëpe, G.; Emsley, L., Direct spectral optimisation of proton–proton homonuclear dipolar decoupling in solid-state NMR. *Chem. Phys. Lett.* **2004,** *398* (4–6), 532–538.

349. Rossini, A. J.; Zagdoun, A.; Lelli, M.; Gajan, D.; Rascón, F.; Rosay, M.; Maas, W. E.; Copéret, C.; Lesage, A.; Emsley, L., One hundred fold overall sensitivity enhancements for Silicon-29 NMR spectroscopy of surfaces by dynamic nuclear polarization with CPMG acquisition. *Chem. Sci.* **2012,** *3* (1), 108–115.

350. Watson, G. W.; Kelsey, E. T.; Leeuw, d. N. H.; Harris, D. J.; Parker, S. C., Atomistic simulation of dislocations, surfaces and interfaces in MgO. **1996,** *92* (3), 433-438.

351. Hartman, J. D.; Beran, G. J. O., Fragment-Based Electronic Structure Approach for Computing Nuclear Magnetic Resonance Chemical Shifts in Molecular Crystals. *Journal of Chemical Theory and Computation* **2014,** *10* (11), 4862-4872.

352. Hartman, J. D.; Neubauer, T. J.; Caulkins, B. G.; Mueller, L. J.; Beran, G. J., Converging nuclear magnetic shielding calculations with respect to basis and system size in protein systems. *J Biomol NMR* **2015,** *62* (3), 327-40.

353. Pickard, C. J.; Needs, R. J.; Search, H.; Journals, C.; Contact, A.; Iopscience, M., Ab initio random structure searching. *Journal of physics. Condensed matter : an Institute of Physics journal* **2011,** *23* (5), 053201-053201.

354. Demichelis, F.; Pirri, C. F.; Tresso, E., Influence of doping on the structural and optoelectronic properties of amorphous and microcrystalline silicon carbide. *Journal of Applied Physics* **1992,** *72* (4), 1327-1333.

355. Matsumoto, S., Silicon: Diffusion. In *Encyclopedia of Materials: Science and Technology*, Buschow, K. H. J.; Cahn, R. W.; Flemings, M. C.; Ilschner, B.; Kramer, E. J.; Mahajan, S.; Veyssière, P., Eds. Elsevier: Oxford, 2001; pp 8543-8549.

356. Gaul, C.; Hutsch, S.; Schwarze, M.; Schellhammer, K. S.; Bussolotti, F.; Kera, S.; Cuniberti, G.; Leo, K.; Ortmann, F., Insight into doping efficiency of organic semiconductors from the analysis of the density of states in n-doped C60 and ZnPc. *Nat Mater* **2018,** *17* (5), 439-444.

357. Zaleska, A., Doped-TiO2: A Review. *Recent Patents on Engineering* **2008,** *2* (3), 157-164.

358. Geetha, N.; Sivaranjani, S.; Ayeshamariam, A.; Kissinger, J. S.; Valan Arasu, M.; Jayachandran, M., ZnO doped Oxide Materials: Mini Review. *Fluid Mechanics: Open Access* **2016,** *03* (03).

359. Kalish, R., Doping of diamond. *Carbon* **1999,** *37* (5), 781-785.

360. Ekimov, E. A.; Sidorov, V. A.; Bauer, E. D.; Mel'nik, N. N.; Curro, N. J.; Thompson, J. D.; Stishov, S. M., Superconductivity in diamond. *Nature* **2004,** *428* (6982), 542-545.

361. Acosta, V. M.; Bauch, E.; Ledbetter, M. P.; Santori, C.; Fu, K. M. C.; Barclay, P. E.; Beausoleil, R. G.; Linget, H.; Roch, J. F.; Treussart, F.; Chemerisov, S.; Gawlik, W.; Budker, D., Diamonds with a high density of nitrogen-vacancy centers for magnetometry applications. *Phys Rev B* **2009,** *80* (11).

362. Wang, H.; Maiyalagan, T.; Wang, X., Review on Recent Progress in Nitrogen-Doped Graphene: Synthesis, Characterization, and Its Potential Applications. *ACS Catalysis* **2012,** *2* (5), 781-794.

363. Pellet, N.; Gao, P.; Gregori, G.; Yang, T. Y.; Nazeeruddin, M. K.; Maier, J.; Gratzel, M., Mixed-Organic-Cation Perovskite Photovoltaics for Enhanced Solar-Light Harvesting. *Angewandte Chemie International Edition* **2014,** *53* (12), 3151-3157.

364. Jeon, N. J.; Noh, J. H.; Yang, W. S.; Kim, Y. C.; Ryu, S.; Seo, J.; Seok, S. I., Compositional engineering of perovskite materials for high-performance solar cells. *Nature* **2015,** *517* (7535), 476-480.

365. Lee, J. W.; Kim, D. H.; Kim, H. S.; Seo, S. W.; Cho, S. M.; Park, N. G., Formamidinium and Cesium Hybridization for Photo- and Moisture-Stable Perovskite Solar Cell. *Advanced Energy Materials* **2015,** *5* (20), 1501310-1501318.

366. Li, X.; Bi, D. Q.; Yi, C. Y.; Decoppet, J. D.; Luo, J. S.; Zakeeruddin, S. M.; Hagfeldt, A.; Gratzel, M., A vacuum flash-assisted solution process for high-efficiency large-area perovskite solar cells. *Science* **2016,** *353* (6294), 58-62.

367. Xia, X.; Wu, W. Y.; Li, H. C.; Zheng, B.; Xue, Y. B.; Xu, J.; Zhang, D. W.; Gao, C. X.; Liu, X. Z., Spray reaction prepared FA(1-x)Cs(x)PbI(3) solid solution as a light harvester for perovskite solar cells with improved humidity stability. *Rsc Advances* **2016,** *6* (18), 14792-14798.

368. Yi, C. Y.; Luo, J. S.; Meloni, S.; Boziki, A.; Ashari-Astani, N.; Gratzel, C.; Zakeeruddin, S. M.; Rothlisberger, U.; Gratzel, M., Entropic stabilization of mixed A-cation ABX(3) metal halide perovskites for high performance perovskite solar cells. *Energy & Environmental Science* **2016,** *9* (2), 656-662.

369. Saliba, M.; Matsui, T.; Seo, J. Y.; Domanski, K.; Correa-Baena, J. P.; Nazeeruddin, M. K.; Zakeeruddin, S. M.; Tress, W.; Abate, A.; Hagfeldt, A.; Gratzel, M., Cesium-containing triple cation perovskite solar cells: improved stability, reproducibility and high efficiency. *Energy & Environmental Science* **2016,** *9* (6), 1989-1997.

370. Duong, T.; Mulmudi, H. K.; Shen, H. P.; Wu, Y. L.; Barugkin, C.; Mayon, Y. O.; Nguyen, H. T.; Macdonald, D.; Peng, J.; Lockrey, M.; Li, W.; Cheng, Y. B.; White, T. P.; Weber, K.; Catchpole, K., Structural engineering using rubidium iodide as a dopant under excess lead iodide conditions for high efficiency and stable perovskites. *Nano Energy* **2016,** *30*, 330-340.

371. Saliba, M.; Matsui, T.; Domanski, K.; Seo, J. Y.; Ummadisingu, A.; Zakeeruddin, S. M.; Correa-Baena, J. P.; Tress, W. R.; Abate, A.; Hagfeldt, A.; Gratzel, M., Incorporation of rubidium cations into perovskite solar cells improves photovoltaic performance. *Science* **2016,** *354* (6309), 206-209.

372. Park, Y. H.; Jeong, I.; Bae, S.; Son, H. J.; Lee, P.; Lee, J.; Lee, C. H.; Ko, M. J., Inorganic Rubidium Cation as an Enhancer for Photovoltaic Performance and Moisture Stability of HC(NH2)(2)PbI3 Perovskite Solar Cells. *Advanced Functional Materials* **2017,** *27* (16), 1605988-16059815.

373. Zhang, M.; Yun, J. S.; Ma, Q. S.; Zheng, J. H.; Lau, C. F. J.; Deng, X. F.; Kim, J.; Kim, D.; Seidel, J.; Green, M. A.; Huang, S. J.; Ho-Baillie, A. W. Y., High-Efficiency Rubidium-Incorporated Perovskite Solar Cells by Gas Quenching. *Acs Energy Letters* **2017,** *2* (2), 438-444.

374. Nam, J. K.; Chai, S. U.; Cha, W.; Choi, Y. J.; Kim, W.; Jung, M. S.; Kwon, J.; Kim, D.; Park, J. H., Potassium Incorporation for Enhanced Performance and Stability of Fully Inorganic Cesium Lead Halide Perovskite Solar Cells. *Nano Letters* **2017,** *17* (3), 2028-2033.

375. Zhao, P.; Yin, W.; Kim, M.; Han, M.; Song, Y. J.; Ahn, T. K.; Jung, H. S., Improved carriers injection capacity in perovskite solar cells by introducing A-site interstitial defects. *Journal of Materials Chemistry A* **2017,** *5* (17), 7905-7911.

376. Zhao, W.; Yao, Z.; Yu, F.; Yang, D.; Liu, S., Alkali Metal Doping for Improved CH3NH3PbI3 Perovskite Solar Cells. *Advanced Science* **2018,** *5* (2), 1700131-1700138.

377. Jacobsson, T. J.; Svanström, S.; Andrei, V.; Rivett, J. P. H.; Kornienko, N.; Philippe, B.; Cappel, U. B.; Rensmo, H.; Deschler, F.; Boschloo, G., Extending the Compositional Space of Mixed Lead Halide Perovskites by Cs, Rb, K, and Na Doping. *The Journal of Physical Chemistry C* **2018**, DOI: 10.1021/acs.jpcc.7b12464.

378. Tang, Z.; Uchida, S.; Bessho, T.; Kinoshita, T.; Wang, H.; Awai, F.; Jono, R.; Maitani, M. M.; Nakazaki, J.; Kubo, T.; Segawa, H., Modulations of various alkali metal cations on organometal halide perovskites and their influence on photovoltaic performance. *Nano Energy* **2018,** *45*, 184-192.

379. Son, D.-Y.; Kim, S.-G.; Seo, J.-Y.; Lee, S.-H.; Shin, H.; Lee, D.; Park, N.-G., Universal Approach toward Hysteresis-Free Perovskite Solar Cell via Defect Engineering. *Journal of the American Chemical Society* **2018,** *140* (4), 1358-1364.

380. Autschbach, J., Calculation of Heavy-Nucleus Chemical Shifts. Relativistic All-Electron Methods. In *Calculation of NMR and EPR Parameters*, 2004; pp 227-247.

381. Edlund, U.; Lejon, T.; Pyykko, P.; Venkatachalam, T. K.; Buncel, E., Lithium-7, silicon-29, tin-119, and lead-207 NMR studies of phenyl-substituted Group 4 anions. *Journal of the American Chemical Society* **1987,** *109* (20), 5982-5985.

382. Pyykkö, P.; Görling, A.; Rösch, N., A transparent interpretation of the relativistic contribution to the N.M.R. 'heavy atom chemical shift'. *Molecular Physics* **1987,** *61* (1), 195-205.

383. Guerra, C. F.; Snijders, J. G.; te Velde, G.; Baerends, E. J., Towards an order-N DFT method. *Theoretical Chemistry Accounts* **1998,** *99* (6), 391-403.

384. te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Guerra, C. F.; Van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T., Chemistry with ADF. *Journal of Computational Chemistry* **2001,** *22* (9), 931-967.

385. Kojima, A.; Teshima, K.; Shirai, Y.; Miyasaka, T., Organometal Halide Perovskites as Visible-Light Sensitizers for Photovoltaic Cells. *Journal of the American Chemical Society* **2009,** *131* (17), 6050-6051.

386. Era, M.; Hattori, T.; Taira, T.; Tsutsui, T., Self-organized growth of PbI-based layered perovskite quantum well by dual-source vapor deposition. *Chemistry of Materials* **1997,** *9* (1), 8-10.

387. Liang, K. N.; Mitzi, D. B.; Prikas, M. T., Synthesis and characterization of organic-inorganic perovskite thin films prepared using a versatile two-step dipping technique. *Chemistry of Materials* **1998,** *10* (1), 403-411.

388. Kitazawa, N.; Enomoto, K.; Aono, M.; Watanabe, Y., Optical properties of (C6H5C2H4NH3)(2)Pbl(4-x)Br(x) (x=0-4) mixed-crystal doped PMMA films. *JOURNAL OF MATERIALS SCIENCE* **2004,** *39* (2), 749-751.

389. Pradeesh, K.; Baumberg, J. J.; Prakash, G. V., In situ intercalation strategies for device-quality hybrid inorganic-organic self-assembled quantum wells. *Applied Physics Letters* **2009,** *95* (3), 033309-033311.

390. Weller, M. T.; Weber, O. J.; Frost, J. M.; Walsh, A., Cubic Perovskite Structure of Black Formamidinium Lead Iodide, α-[HC(NH2)2]PbI3, at 298 K. *J Phys Chem Lett* **2015,** *6* (Copyright (C) 2018 American Chemical Society (ACS). All Rights Reserved.), 3209-3212.

391. Knop, O.; Wasylishen, R. E.; White, M. A.; Cameron, T. S.; Van Oort, M. J. M., Alkylammonium lead halides. Part 2. CH3NH3PbX3 (X = chlorine, bromine, iodine) perovskites: cuboctahedral halide cages with isotropic cation reorientation. *Canadian Journal of Chemistry* **1990,** *68* (Copyright (C) 2018 American Chemical Society (ACS). All Rights Reserved.), 412-22.

392. Roiland, C.; Trippe-Allard, G.; Jemli, K.; Alonso, B.; Ameline, J.-C.; Gautier, R.; Bataille, T.; Le Polles, L.; Deleporte, E.; Even, J.; Katan, C., Multinuclear NMR as a tool for studying local order and dynamics in CH3NH3PbX3 (X = Cl, Br, I) hybrid perovskites. *Physical Chemistry Chemical Physics* **2016,** *18* (Copyright (C) 2018 American Chemical Society (ACS). All Rights Reserved.), 27133-27142.

393. Rosales, B. A.; Men, L.; Cady, S. D.; Hanrahan, M. P.; Rossini, A. J.; Vela, J., Persistent Dopants and Phase Segregation in Organolead Mixed-Halide Perovskites. *Chemistry of Materials* **2016,** *28* (19), 6848-6859.

394. Rosales, B. A.; Hanrahan, M. P.; Boote, B. W.; Rossini, A. J.; Smith, E. A.; Vela, J., Lead Halide Perovskites: Challenges and Opportunities in Advanced Synthesis and Spectroscopy. *Acs Energy Letters* **2017,** *2* (4), 906-914.

395. Franssen, W. M. J.; van Es, S. G. D.; Dervisoglu, R.; de Wijs, G. A.; Kentgens, A. P. M., Symmetry, Dynamics, and Defects in Methylammonium Lead Halide Perovskites. *J Phys Chem Lett* **2017,** *8* (1), 61-66.

396. Senocrate, A.; Moudrakovski, I.; Kim, G. Y.; Yang, T.-Y.; Gregori, G.; Grätzel, M.; Maier, J., The Nature of Ion Conduction in Methylammonium Lead Iodide: A Multimethod Approach. *Angew. Chem., Int. Ed.* **2017,** *56* (Copyright (C) 2018 American Chemical Society (ACS). All Rights Reserved.), 7755-7759.

397. Kubicki, D. J.; Prochowicz, D.; Hofstetter, A.; Péchy, P.; Zakeeruddin, S. M.; Grätzel, M.; Emsley, L., Cation Dynamics in Mixed-Cation (MA)x(FA)1–xPbI3 Hybrid Perovskites from Solid-State NMR. *Journal of the American Chemical Society* **2017,** *139* (29), 10055-10061.

398. Prochowicz, D.; Franckevicius, M.; Cieslak, A. M.; Zakeeruddin, S. M.; Gratzel, M.; Lewinski, J., Mechanosynthesis of the hybrid perovskite CH3NH3PbI3: characterization and the corresponding solar cell efficiency. *Journal of Materials Chemistry A* **2015,** *3* (41), 20772-20777.

399. Zhu, Z.-Y.; Yang, Q.-Q.; Gao, L.-F.; Zhang, L.; Shi, A.-Y.; Sun, C.-L.; Wang, Q.; Zhang, H.-L., Solvent-Free Mechanosynthesis of Composition-Tunable Cesium Lead Halide Perovskite Quantum Dots. *J. Phys. Chem. Lett.* **2017,** *8* (7), 1610-1614.

400. Prochowicz, D.; Yadav, P.; Saliba, M.; Saski, M.; Zakeeruddin, S. M.; Lewinski, J.; Gratzel, M., Reduction in the Interfacial Trap Density of Mechanochemically Synthesized MAPbI3. *ACS Appl. Mater. Interfaces* **2017,** *9* (Copyright (C) 2018 American Chemical Society (ACS). All Rights Reserved.), 28418-28425.

401. Breternitz, J.; Levcenko, S.; Hempel, H.; Gurieva, G.; Franz, A.; Hoser, A.; Schorr, S. Mechanochemical Synthesis of the Lead-Free Double Perovskite Cs2[AgIn]Br6 and its Optical Properties *arXiv e-prints* [Online], 2018. https://ui.adsabs.harvard.edu/abs/2018arXiv181011330B (accessed October 01, 2018).

402. Beck, H. P.; Milius, W., Study on A4BX6 compounds .2. Refinement of the structure of Rb4PbBr6 and a note on the excistence of Rb4HgL6 and K4CDL6. *Zeitschrift Fur Anorganische Und Allgemeine Chemie* **1988,** *562* (7), 102-104.

403. Prochowicz, D.; Yadav, P.; Saliba, M.; Saski, S. M.; Zakeeruddin, S. M.; Lewinski, J.; Gratzel, M., Mechanosynthesis of pure phase mixed-cation MAxFA1−xPbI3 hybrid perovskites: photovoltaic performance and electrochemical properties. *Sustainable Energy and Fuels* **2017,** *1*, 689-693.

404. Hayashi, S.; Hayamizu, K., Accurate determination of NMR chemical shifts in alkali halides and their correlation with structural factors. *Bulletin of the Chemical Society of Japan* **1990,** *63* (Copyright (C) 2018 American Chemical Society (ACS). All Rights Reserved.), 913-19.

405. Perdew, J. P., Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys Rev B* **1986,** *33* (12), 8822-8824.

406. Becke, A. D., Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A* **1988,** *38* (6), 3098-3100.

407. van Lenthe, E.; Baerends, E. J.; Snijders, J. G., Relativistic regular 2-component hamiltonians. *J Chem Phys* **1993,** *99* (6), 4597-4610.

408. van Lenthe, E.; Baerends, E. J.; Snijders, J. G., Relativistic total energy using regular approximations. *J Chem Phys* **1994,** *101* (Copyright (C) 2018 American Chemical Society (ACS). All Rights Reserved.), 9783-92.

409. van Lenthe, E.; Ehlers, A.; Baerends, E.-J., Geometry optimizations in the zero order regular approximation for relativistic effects. *J Chem Phys* **1999,** *110* (Copyright (C) 2018 American Chemical Society (ACS). All Rights Reserved.), 8943-8953.

410. Xie, L. Q.; Chen, L.; Nan, Z. A.; Lin, H. X.; Wang, T.; Zhan, D. P.; Yan, J. W.; Mao, B. W.; Tian, Z. Q., Understanding the Cubic Phase Stabilization and Crystallization Kinetics in Mixed Cations and. Halides Perovskite Single Crystals. *Journal of the American Chemical Society* **2017,** *139* (9), 3320-3323.

411. Nazarenko, O.; Yakunin, S.; Morad, V.; Cherniukh, I.; Kovalenko, M. V., Single crystals of caesium formamidinium lead halide perovskites: solution growth and gamma dosimetry. *Npg Asia Materials* **2017,** *9*, e373.

412. Hirotsu, S.; Kunii, Y., Brillouin-scattering study of cubic CsPbC13. *Journal of the Physical Society of Japan* **1981,** *50* (4), 1249-1254.

413. Stoumpos, C. C.; Malliakas, C. D.; Kanatzidis, M. G., Semiconducting Tin and Lead Iodide Perovskites with Organic Cations: Phase Transitions, High Mobilities, and Near-Infrared Photoluminescent Properties. *Inorganic Chemistry* **2013,** *52* (15), 9019-9038.

414. Liu, F. Z.; Dong, Q.; Wong, M. K.; Djurisic, A. B.; Ng, A. N.; Ren, Z. W.; Shen, Q.; Surya, C.; Chan, W. K.; Wang, J.; Ng, A. M. C.; Liao, C. Z.; Li, H. K.; Shih, K. M.; Wei, C. R.; Su, H. M.; Dai, J. F., Is Excess PbI2 Beneficial for Perovskite Solar Cell Performance? *Advanced Energy Materials* **2016,** *6* (7), 1502206-1502215.

415. Hu, Y.; Aygüler, M.; Petrus, M. L.; Bein, T.; Docampo, P., The Impact of Rubidium and Cesium Cations on the Moisture Stability of Multiple-Cation Mixed-Halide Perovskites. *ACS Energy Letters* **2017**, 10.1021/acsenergylett.7b00731

416. Skibsted, J.; Jakobsen, H. J., Variable-temperature Rb-87 magic-angle spinning NMR spectroscopy of inorganic rubidium salts. *Journal of Physical Chemistry A* **1999,** *103* (40), 7958-7971.

417. Swamy, T. K.; Subhadra, K. G.; Sirdeshmukh, D. B., X-RAY-DIFFRACTION STUDIES OF RBBR-RBI MIXED-CRYSTALS. *Pramana-Journal of Physics* **1994,** *43* (1), 33-39.

418. Philippe, B.; Saliba, M.; Correa-Baena, J. P.; Cappel, U. B.; Turren-Cruz, S. H.; Gratzel, M.; Hagfeldt, A.; Rensmo, H., Chemical Distribution of Multiple Cation (Rb+, Cs+, MA(+), and FA(+)) Perovskite Materials by Photoelectron Spectroscopy. *Chemistry of Materials* **2017,** *29* (8), 3589-3596.

419. Huang, T. L.; Ruoff, A. L., Equation of state and high-pressure phase-transition oc CsI. *Phys Rev B* **1984,** *29* (2), 1112-1114.

420. Van den Bosch, A.; Dresselaers, J.; Vansummeren, J.; Hovi, M., Susceptibility and lattice parameter of single K1−xRbxI crystals. *physica status solidi (a)* **1972,** *11* (2), 479-482.

421. Weber, D., CH3NH3PBX3, A PB(II)-SYSTEM WITH CUBIC PEROVSKITE STRUCTURE. *Zeitschrift Fur Naturforschung Section B-a Journal of Chemical Sciences* **1978,** *33* (12), 1443-1445.

422. Trots, D. M.; Myagkota, S. V., High-temperature structural evolution of caesium and rubidium triiodoplumbates. *J Phys Chem Solids* **2008,** *69* (10), 2520-2526.

423. Pack, J. D.; Monkhorst, H. J., Special points for Brillouin-zone integrations - reply. *Phys Rev B* **1977,** *16* (4), 1748-1749.

424. Dmitrenko, O.; Bai, S.; Beckmann, P. A.; van Bramer, S.; Vega, A. J.; Dybowski, C., The Relationship between 207Pb NMR Chemical Shift and Solid-State Structure in Pb(II) Compounds. *The Journal of Physical Chemistry A* **2008,** *112* (14), 3046-3052.

425. Even, J.; Pedesseau, L.; Jancu, J.-M.; Katan, C., Importance of Spin–Orbit Coupling in Hybrid Organic/Inorganic Perovskites for Photovoltaic Applications. *J. Phys. Chem. Lett.* **2013,** *4* (17), 2999-3005.

426. Alkan, F.; Dybowski, C., Effect of Co-Ordination Chemistry and Oxidation State on the Pb-207 Magnetic-Shielding Tensor: A DFT/ZORA Investigation. *Journal of Physical Chemistry A* **2016,** *120* (1), 161-168.

427. Giorgi, G.; Yoshihara, T.; Yamashita, K., Structural and electronic features of small hybrid organic-inorganic halide perovskite clusters: a theoretical analysis. *Physical Chemistry Chemical Physics* **2016,** *18* (39), 27124-27132.

428. Li, W.; Wang, Z.; Deschler, F.; Gao, S.; Friend, R. H.; Cheetham, A. K., Chemically diverse and multifunctional hybrid organic–inorganic perovskites. *Nature Reviews Materials* **2017,** *2*, 16099.

429. Yang, W. S.; Park, B.-W.; Jung, E. H.; Jeon, N. J.; Kim, Y. C.; Lee, D. U.; Shin, S. S.; Seo, J.; Kim, E. K.; Noh, J. H.; Seok, S. I., Iodide management in formamidinium-lead-halide–based perovskite layers for efficient solar cells. *Science* **2017,** *356* (6345), 1376-1379.

430. Snaith, H. J.; Abate, A.; Ball, J. M.; Eperon, G. E.; Leijtens, T.; Noel, N. K.; Stranks, S. D.; Wang, J. T.-W.; Wojciechowski, K.; Zhang, W., Anomalous Hysteresis in Perovskite Solar Cells. *J. Phys. Chem. Lett.* **2014,** *5* (9), 1511-1515.

431. Eames, C.; Frost, J. M.; Barnes, P. R. F.; O'Regan, B. C.; Walsh, A.; Islam, M. S., Ionic transport in hybrid lead iodide perovskite solar cells. *Nat Commun* **2015,** *6*, 7497.

432. Yuan, Y.; Huang, J., Ion Migration in Organometal Trihalide Perovskite and Its Impact on Photovoltaic Efficiency and Stability. *Acc. Chem. Res.* **2016,** *49* (Copyright (C) 2018 U.S. National Library of Medicine.), 286-93.

433. Levine, I.; Nayak, P. K.; Wang, J. T.-W.; Sakai, N.; Van Reenen, S.; Brenner, T. M.; Mukhopadhyay, S.; Snaith, H. J.; Hodes, G.; Cahen, D., Interface-Dependent Ion Migration/Accumulation Controls Hysteresis in MAPbI3 Solar Cells. *The Journal of Physical Chemistry C* **2016,** *120* (30), 16399-16411.

434. Jacobs, D. A.; Wu, Y.; Shen, H.; Barugkin, C.; Beck, F. J.; White, T. P.; Weber, K.; Catchpole, K. R., Hysteresis phenomena in perovskite solar cells: the many and varied effects of ionic accumulation. *Physical Chemistry Chemical Physics* **2017,** *19* (4), 3094-3103.

435. Shao, Y.; Xiao, Z.; Bi, C.; Yuan, Y.; Huang, J., Origin and elimination of photocurrent hysteresis by fullerene passivation in CH3NH3PbI3 planar heterojunction solar cells. *Nat Commun* **2014,** *5* (Copyright (C) 2018 U.S. National Library of Medicine.), 5784.

436. Jiang, Q.; Zhang, L.; Wang, H.; Yang, X.; Meng, J.; Liu, H.; Yin, Z.; Wu, J.; Zhang, X.; You, J., Enhanced electron extraction using SnO2 for high-efficiency planar-structure HC(NH2)2PbI3-based perovskite solar cells. *Nature Energy* **2016,** *2*, 16177.

437. Tan, H.; Jain, A.; Voznyy, O.; Lan, X.; Garcia, d. A. F. P.; Fan, J. Z.; Quintero-Bermudez, R.; Yuan, M.; Zhang, B.; Zhao, Y.; Fan, F.; Quan, L. N.; Yang, Z.; Hoogland, S.; Sargent, E. H.; Li, P.; Zhao, Y.; Lu, Z.-H., Efficient and stable solution-processed planar perovskite solar cells via contact passivation. *Science (New York, N.Y.)* **2017,** *355* (Copyright (C) 2018 U.S. National Library of Medicine.), 722-726.

438. Giordano, F.; Abate, A.; Correa Baena, J. P.; Saliba, M.; Matsui, T.; Im, S. H.; Zakeeruddin, S. M.; Nazeeruddin, M. K.; Hagfeldt, A.; Graetzel, M., Enhanced electronic properties in mesoporous TiO2 via lithium doping for high-efficiency perovskite solar cells. *Nat Commun* **2016,** *7*, 10379.

439. Tang, Z.; Bessho, T.; Awai, F.; Kinoshita, T.; Maitani, M. M.; Jono, R.; Murakami, T. N.; Wang, H.; Kubo, T.; Uchida, S.; Segawa, H., Hysteresis-free perovskite solar cells made of potassium-doped organometal halide perovskite. *Sci Rep-Uk* **2017,** *7* (1), 12183.

440. Pecharsky, V.; Zavalij, P., *Fundamentals of Powder Diffraction and Structural Characterization of Materials, Second Edition*. 2009.

441. Van Gompel, W. T. M.; Herckens, R.; Reekmans, G.; Ruttens, B.; D'Haen, J.; Adriaensens, P.; Lutsen, L.; Vanderzande, D., Degradation of the Formamidinium Cation and the Quantification of the Formamidinium-Methylammonium Ratio in Lead Iodide Hybrid Perovskites by Nuclear Magnetic Resonance Spectroscopy. *Journal of Physical Chemistry C* **2018,** *122* (Copyright (C) 2018 American Chemical Society (ACS). All Rights Reserved.), 4117-4124.

442. Karmakar, A.; Askar, A. M.; Bernard, G. M.; Terskikh, V. V.; Ha, M.; Patel, S.; Shankar, K.; Michaelis, V. K., Mechanochemical Synthesis of Methylammonium Lead Mixed–Halide Perovskites: Unraveling the Solid-Solution Behavior using Solid-State NMR. *Chemistry of Materials* **2018,** *30*, 2309-2321.

443. Kubicki, D. J.; Prochowicz, D.; Hofstetter, A.; Zakeeruddin, S. M.; Grätzel, M.; Emsley, L., Phase Segregation in Cs-, Rb- and K-Doped Mixed-Cation (MA)x(FA)1–xPbI3 Hybrid Perovskites from Solid-State NMR. *Journal of the American Chemical Society* **2017,** *139* (40), 14173-14180.

444. Kubicki, D. J.; Prochowicz, D.; Hofstetter, A.; Saski, M.; Yadav, P.; Bi, D.; Pellet, N.; Lewiński, J.; Zakeeruddin, S. M.; Grätzel, M.; Emsley, L., Formation of Stable Mixed Guanidinium–Methylammonium Phases with Exceptionally Long Carrier Lifetimes for High-Efficiency Lead Iodide-Based Perovskite Photovoltaics. *Journal of the American Chemical Society* **2018,** *140* (9), 3345-3351.

445. Askar, A. M.; Karmakar, A.; Bernard, G. M.; Ha, M.; Terskikh, V. V.; Wiltshire, B. D.; Patel, S.; Fleet, J.; Shankar, K.; Michaelis, V. K., Composition-Tunable Formamidinium Lead Mixed Halide Perovskites via Solvent-Free Mechanochemical Synthesis: Decoding the Pb Environments Using Solid-State NMR Spectroscopy. *J Phys Chem Lett* **2018,** *9* (10), 2671-2677.
446. Moudrakovski, I. L.; Ripmeester, J. A., 39K NMR of Solid Potassium Salts at 21 T: Effect of Quadrupolar and Chemical Shift Tensors. *J. Phys. Chem. B* **2007,** *111* (3), 491-495.
447. Stenger, V. A.; Recchia, C.; Pennington, C. H.; Buffinger, D. R.; Ziebarth, R. P., NMR studies of alkali C60 superconductors. *Journal of Superconductivity* **1994,** *7* (6), 931-936.
448. Bowers, G. M.; Bish, D. L.; Kirkpatrick, R. J., H2O and Cation Structure and Dynamics in Expandable Clays: 2H and 39K NMR Investigations of Hectorite. *The Journal of Physical Chemistry C* **2008,** *112* (16), 6430-6438.
449. Michaelis, V. K.; Aguiar, P. M.; Kroeker, S., Probing alkali coordination environments in alkali borate glasses by multinuclear magnetic resonance. *J Non-Cryst Solids* **2007,** *353* (26), 2582-2590.
450. Xu, J.; Lucier, B. E. G.; Lin, Z.; Sutrisno, A.; Terskikh, V. V.; Huang, Y., New Insights into the Short-Range Structures of Microporous Titanosilicates As Revealed by 47/49Ti, 23Na, 39K, and 29Si Solid-State NMR Spectroscopy. *The Journal of Physical Chemistry C* **2014,** *118* (47), 27353-27365.
451. Wu, G.; Wong, A.; Gan, Z.; Davis, J. T., Direct Detection of Potassium Cations Bound to G-Quadruplex Structures by Solid-State 39K NMR at 19.6 T. *Journal of the American Chemical Society* **2003,** *125* (24), 7182-7183.
452. Wu, G.; Gan, Z.; Kwan, I. C. M.; Fettinger, J. C.; Davis, J. T., High-Resolution 39K NMR Spectroscopy of Bio-organic Solids. *Journal of the American Chemical Society* **2011,** *133* (49), 19570-19573.
453. Wong, A.; Whitehead, R. D.; Gan, Z.; Wu, G., A Solid-State NMR and Computational Study of Sodium and Potassium Tetraphenylborates: 23Na and 39K NMR Signatures for Systems Containing Cation−π Interactions. *The Journal of Physical Chemistry A* **2004,** *108* (47), 10551-10559.
454. Lee, P. K.; Chapman, R. P.; Zhang, L.; Hu, J.; Barbour, L. J.; Elliott, E. K.; Gokel, G. W.; Bryce, D. L., K-39 Quadrupolar and Chemical Shift Tensors for Organic Potassium Complexes and Diatomic Molecules. *The Journal of Physical Chemistry A* **2007,** *111* (50), 12859-12863.
455. Wu, G.; Terskikh, V., A Multinuclear Solid-State NMR Study of Alkali Metal Ions in Tetraphenylborate Salts, M[BPh4] (M = Na, K, Rb and Cs): What Is the NMR Signature of Cation−π Interactions? *The Journal of Physical Chemistry A* **2008,** *112* (41), 10359-10364.
456. Lenthe, v. E.; Baerends, E. J.; Snijders, J. G., Relativistic regular two-component Hamiltonians. *The Journal of Chemical Physics* **1993,** *99* (6), 4597-4610.
457. Dybowski, C.; Smith, M. L.; Hepp, M. A.; Gaffney, E. J.; Neue, G.; Perry, D. L., 207Pb NMR Chemical-Shift Tensors of the Lead (II) Halides and the Lead (II) Hydroxyhalides. *Appl. Spectrosc.* **1998,** *52* (3), 426-429.
458. Giorgi, G.; Fujisawa, J.-I.; Segawa, H.; Yamashita, K., Organic-Inorganic Hybrid Lead Iodide Perovskite Featuring Zero Dipole Moment Guanidinium Cations: A Theoretical Analysis. *Journal of Physical Chemistry C* **2015,** *119* (Copyright (C) 2018 American Chemical Society (ACS). All Rights Reserved.), 4694-4701.
459. Park, B.-w.; Philippe, B.; Jain, S. M.; Zhang, X.; Edvinsson, T.; Rensmo, H.; Zietz, B.; Boschloo, G., Chemical engineering of methylammonium lead iodide/bromide perovskites: tuning of opto-electronic properties and photovoltaic performance. *Journal of Materials Chemistry A* **2015,** *3* (43), 21760-21771.
460. Bu, T.; Liu, X.; Zhou, Y.; Yi, J.; Huang, X.; Luo, L.; Xiao, J.; Ku, Z.; Peng, Y.; Huang, F.; Cheng, Y.-B.; Zhong, J., A novel quadruple-cation absorber for universal hysteresis elimination for high efficiency and stable perovskite solar cells. *Energy & Environmental Science* **2017,** *10* (12), 2509-2515.
461. Isaenko, L. I.; Merkulov, A. A.; Melnikova, S. V.; Pashkov, V. M.; Tarasova, A. Y., Effect of K ↔ Rb Substitution on Structure and Phase Transition in Mixed KxRb1−xPb2Br5 Crystals. *Cryst Growth Des* **2009,** *9* (5), 2248-2251.
462. Whitfield, P. S.; Herron, N.; Guise, W. E.; Page, K.; Cheng, Y. Q.; Milas, I.; Crawford, M. K., Structures, Phase Transitions and Tricritical Behavior of the Hybrid Perovskite Methyl Ammonium Lead Iodide. *Sci Rep-Uk* **2016,** *6*, 35685.
463. Ahn, N.; Son, D.-Y.; Jang, I.-H.; Kang, S. M.; Choi, M.; Park, N.-G., Highly Reproducible Perovskite Solar Cells with Average Efficiency of 18.3% and Best Efficiency of 19.7% Fabricated via Lewis Base Adduct of Lead(II) Iodide. *Journal of the American Chemical Society* **2015,** *137* (27), 8696-8699.
464. Lee, J.-W.; Dai, Z.; Lee, C.; Lee, H. M.; Han, T.-H.; De Marco, N.; Lin, O.; Choi, C. S.; Dunn, B. S.; Koh, J.; Di Carlo, D.; Ko, J. H.; Maynard, H. D.; Yang, Y., Tuning Molecular Interactions for Highly Reproducible and Efficient Formamidinium Perovskite Solar Cells via Adduct Approach. *Journal of the American Chemical Society* **2018**, 10.1021/jacs.8b01037.
465. Abdi-Jalebi, M.; Andaji-Garmaroudi, Z.; Cacovich, S.; Stavrakas, C.; Philippe, B.; Richter, J. M.; Alsari, M.; Booker, E. P.; Hutter, E. M.; Pearson, A. J.; Lilliu, S.; Savenije, T. J.; Rensmo, H.; Divitini, G.; Ducati, C.; Friend, R. H.; Stranks, S. D., Maximizing and stabilizing luminescence from halide perovskites with potassium passivation. *Nature* **2018,** *555*, 497.
466. Price, D. L.; Rowe, J. M.; Rush, J. J.; Prince, E.; Hinks, D. G.; Susman, S., Single Crystal Neutron Diffraction Study of Potassium Cyanide. *The Journal of Chemical Physics* **1972,** *56* (7), 3697-3702.
467. Broch, E.; Oftedal, I.; Pabst, A., Neubestimmung der Gitterkonstanten von KF, CsCl und BaF2. *Zeitschrift fuer Physikalische Chemie, Abteilung B: Chemie der Elementarprozesse, Aufbau der Materie* **1929,** *3*, 209-214.
468. Ott, H., Die Strukturen von MnO, MnS, AgF, NiS, SnI4, SrCl2, BaF2, Praezisionsmessungen einiger Alkalihalogenide. *Z Kristallogr* **1926,** *63*, 222-230.
469. Shimoda, K.; Yamane, A.; Ichikawa, T.; Kojima, Y., First-Principles Calculations of Potassium Amidoborane KNH2BH3: Structure and 39K NMR Spectroscopy. *The Journal of Physical Chemistry C* **2012,** *116* (39), 20666-20672.

470. Kubicki, D. J.; Prochowicz, D.; Hofstetter, A.; Zakeeruddin, S. M.; Gratzel, M.; Emsley, L., Phase Segregation in Potassium-Doped Lead Halide Perovskites from K-39 Solid-State NMR at 21.1 T. *Journal of the American Chemical Society* **2018,** *140* (23), 7232-7238.

471. Bi, D. Q.; Li, X.; Milic, J. V.; Kubicki, D. J.; Pellet, N.; Luo, J. S.; Lagrange, T.; Mettraux, P.; Emsley, L.; Zakeeruddin, S. M.; Gratzel, M., Multifunctional molecular modulators for perovskite solar cells with over 20% efficiency and high operational stability. *Nat Commun* **2018,** *9*.

472. Franssen, W. M. J.; Bruijnaers, B. J.; Portengen, V. H. L.; Kentgens, A. P. M., Dimethylammonium Incorporation in Lead Acetate Based MAPbI(3) Perovskite Solar Cells. *Chemphyschem* **2018,** *19* (22), 3107-3115.

473. Hanrahan, M. P.; Men, L.; Rosales, B. A.; Vela, J.; Rossini, A. J., Sensitivity-Enhanced Pb-207 Solid-State NMR Spectroscopy for the Rapid, Non-Destructive Characterization of Organolead Halide Perovskites. *Chemistry of Materials* **2018,** *30* (20), 7005-7015.

474. Bernard, G. M.; Wasylishen, R. E.; Ratcliffe, C. I.; Terskikh, V.; Wu, Q. C.; Buriak, J. M.; Hauger, T., Methylammonium Cation Dynamics in Methylammonium Lead Halide Perovskites: A Solid-State NMR Perspective. *Journal of Physical Chemistry A* **2018,** *122* (6), 1560-1573.

475. Tavakoli, M. M.; Tress, W.; Milic, J. V.; Kubicki, D.; Emsley, L.; Gratzel, M., Addition of adamantylammonium iodide to hole transport layers enables highly efficient and electroluminescent perovskite solar cells. *Energy & Environmental Science* **2018,** *11* (11), 3310-3320.

476. Almeida, G.; Goldoni, L.; Akkerman, Q.; Dang, Z. Y.; Khan, A. H.; Marras, S.; Moreels, I.; Manna, L., Role of Acid-Base Equilibria in the Size, Shape, and Phase Control of Cesium Lead Bromide Nanocrystals. *Acs Nano* **2018,** *12* (2), 1704-1711.

477. Zhou, Y.; Chen, J.; Bakr, O. M.; Sun, H. T., Metal-Doped Lead Halide Perovskites: Synthesis, Properties, and Optoelectronic Applications. *Chemistry of Materials* **2018,** *30* (19), 6589-6613.

478. Milic, J. V.; Im, J. H.; Kubicki, D. J.; Ummadisingu, A.; Seo, J. Y.; Li, Y.; Ruiz-Preciado, M. A.; Dar, M. I.; Zakeeruddin, S. M.; Emsley, L.; Gratzel, M., Supramolecular Engineering for Formamidinium-Based Layered 2D Perovskite Solar Cells: Structural Complexity and Dynamics Revealed by Solid-State NMR Spectroscopy. *Advanced Energy Materials* **2019,** *9* (20).

479. Alharbi, E. A.; Alyamani, A. Y.; Kubicki, D. J.; Uhl, A. R.; Walder, B. J.; Alanazi, A. Q.; Luo, J. S.; Burgos-Caminal, A.; Albadri, A.; Albrithen, H.; Alotaibi, M. H.; Moser, J. E.; Zakeeruddin, S. M.; Giordano, F.; Emsley, L.; Gratzel, M., Atomic-level passivation mechanism of ammonium salts enabling highly efficient perovskite solar cells. *Nat Commun* **2019,** *10*.

480. Xiang, J. Y.; Yang, J. W.; Luo, N. J.; Zhu, J.; Huang, S. P.; Mao, Y., Optimized photoluminescence and electronic properties of europium doped phosphate red phosphor. *Results Phys* **2019,** *13*.

481. Senocrate, A.; Maier, J., Solid-State Ionics of Hybrid Halide Perovskites. *Journal of the American Chemical Society* **2019,** *141* (21), 8382-8396.

482. 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* **2013,** *88* (5).

483. Vreven, T.; Morokuma, K., On the application of the IMOMO (integrated molecular orbital plus molecular orbital) method. *Journal of Computational Chemistry* **2000,** *21* (16), 1419-1432.

484. Vreven, T.; Morokuma, K., The ONIOM (our own N-layered integrated molecular orbital plus molecular mechanics) method for the first singlet excited (S-1) state photoisomerization path of a retinal protonated Schiff base. *J Chem Phys* **2000,** *113* (8), 2969-2975.

485. Bartok, A. P.; Payne, M. C.; Kondor, R.; Csanyi, G., Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys Rev Lett* **2010,** *104* (13), 136403.

486. Glielmo, A.; Zeni, C.; De Vita, A., Efficient nonparametric n-body force fields from machine learning. *Phys Rev B* **2018,** *97* (18).

487. Dracinsky, M.; Hodgkinson, P., A molecular dynamics study of the effects of fast molecular motions on solid-state NMR parameters. *Crystengcomm* **2013,** *15* (43), 8705-8712.

488. Rossano, S.; Mauri, F.; Pickard, C. J.; Farnan, I., First-principles calculation of O-17 and Mg-25 NMR shieldings in MgO at finite temperature: Rovibrational effect in solids. *Journal of Physical Chemistry B* **2005,** *109* (15), 7245-7250.

489. Schmidt, J.; Sebastiani, D., Anomalous temperature dependence of nuclear quadrupole interactions in strongly hydrogen-bonded systems from first principles. *J Chem Phys* **2005,** *123* (7).

490. Dumez, J.-N.; Pickard, C. J., Calculation of NMR chemical shifts in organic solids: Accounting for motional effects. *J Chem Phys* **2009,** *130* (10).

491. Robinson, M.; Haynes, P. D., Dynamical effects in ab initio NMR calculations: Classical force fields fitted to quantum forces. *J Chem Phys* **2010,** *133* (8).

492. Gortari, I. D.; Portella, G.; Salvatella, X.; Bajaj, V. S.; van der Wel, P. C. A.; Yates, J. R.; Segall, M. D.; Pickard, C. J.; Payne, M. C.; Vendruscolo, M., Time Averaging of NMR Chemical Shifts in the MLF Peptide in the Solid State. *Journal of the American Chemical Society* **2010,** *132* (17), 5993-6000.

493. Dracinsky, M.; Bour, P., Vibrational averaging of the chemical shift in crystalline a-glycine. *Journal of Computational Chemistry* **2012,** *33* (10), 1080-1089.

494. Foppa, L.; Yamamoto, K.; Liao, W. C.; Comas-Vives, A.; Coperet, C., Electronic Structure-Reactivity Relationship on Ruthenium Step-Edge Sites from Carbonyl (13)C Chemical Shift Analysis. *J Phys Chem Lett* **2018,** *9* (12), 3348-3353.

495. Gordon, C. P.; Shirase, S.; Yamamoto, K.; Andersen, R. A.; Eisenstein, O.; Coperet, C., NMR chemical shift analysis decodes olefin oligo- and polymerization activity of d(0) group 4 metal complexes. *P Natl Acad Sci USA* **2018,** *115* (26), E5867-E5876.

Curriculum Vitae

First Name	Albert
Last Name	Hofstetter
Date of birth	February 21, 1989
Place of birth	Basel, Switzerland
Nationality	Swiss
Address	St. Alban-Ring 185, 4052 Basel (CH)
Email	albert.hofstetter@epfl.ch
	alboeser@gmail.com
Phone number	+41 79 230 05 66

Education

2015-2019 **Ph.D. computational and physical chemistry / chemical engineering**

Ecole Polytechnique Fédérale de Lausanne, EPFL, Switzerland

Thesis Supervisor: Prof. Dr. Emsley Lyndon

Thesis Title: “Advanced Computational Methods for NMR Crystallography”

2012-2014 **M.S. computational physics, University of Basel, Basel, Switzerland**

final grade 5.5, insigni cum laude

Research Advisor: Prof. Dr. Stefan Goedecker and Prof. Dr. Ernst Meyer

Thesis Title: “Introducing charge transfer into force fields for ionic systems to obtain density functional accuracy”

final grade 6.0, summa cum laude

2009-2012 **B.S. Nanoscience, University of Basel, Basel, Switzerland**

final grade 5.0, magna cum laude

2007 **Eidgenössische Matura (Swiss High School Diploma)**

Gymnasium Bäumlihof, Basel, Switzerland

emphasis on physics and mathematics

Professional experience

2015-2019 **Doctoral Assistant, Laboratory of Magnetic Resonance, LRM**

Ecole Polytechnique Fédérale de Lausanne, EPFL, Switzerland

full-time, 80% Research in solid state nuclear magnetic resonance with main emphasis on the development and application of computational methods

20% Teaching assistant at EPFL and UNIL (University of Lausanne)

2008-2015 **Various positions at messenger business**

KurierZentrale GMBH and Velogourmet GMBH, Basel, part-time

Bicycle messenger, car driver and dispatcher, as which I was responsible for around 15 employees.

2014 **Scientific Researcher, University of Basel, Switzerland, part-time**

Supervisor: Prof. Dr. Stefan Goedecker at the Computational Physics Department.

2014 **Scientific Assistant, University of Heidelberg, Germany, part-time**

Supervisor: PD Dr. Ahmad A. Hujeirat at the Interdisciplinary Center for Scientific Computing. Responsible for setting up a homepage containing computational tools in the area of relativistic hydrodynamics.

2007-2008 **Intern in a histology laboratory**

Novartis Institute for BioMedical Research, Basel, full-time

Industrial collaborations

2018-2019 **AstraZeneca**

Languages

German	Native
English	Native
French	Intermediate

Publications

Refereed Journals

1. Engel, E.A.; Anelli, A.; Hofstetter, A.; Paruzzo, F.; Emsley, L.; Ceriotti, M., "A Bayesian approach to NMR crystal structure determination". *Submitted* **2019**.

2. Ruiz-Preciado, M.A.; Kubicki, D.J.; Hofstetter, A.; Ummadisingu, A.; Gershoni-Poranne, R.; Zakeeruddin, S.M.; Emsley, L.; Milic, J.V.; Grätzel, M., "Supramolecular Modulation of Hybrid Perovskite Solar Cells via Bifunctional Halogen Bonding Revealed by Two-Dimensional ^{19}F Solid-State NMR Spectroscopy". *Submitted* **2019**.

3. Xiang, W.; Wang, Z.; Kubicki, D.J.; Tress, W.; Luo, J.; Wang, X.; Zhang, J.; Hofstetter, A.; Zhang, L.; Emsley, L.; Grätzel, M.; Hagfeldt, A., "Ba-induced phase segregation and band gap reduction in mixed-halide $CsPbI_2Br$ for inorganic perovskite solar cells". *Nature Communications* **2019**.

4. Hofstetter, A.; Balodis, M.; Paruzzo, F.M.; Widdifield, C..M.; Stevanato, G.; Pinon, A.C.; Bygrave, P.; Day, G.M.; Emsley, L., "Rapid Structure Determination of Molecular Solids Using Chemical Shifts Directed by Unambiguous Prior Constraints". *Journal of the American Chemical Society* **2019**, XXXX, XXX

5. D.J.Kubicki, D. Prochowicz, A. Pinon, G. Stevanato, A.Hofstetter, S.M. Zakeeruddin, M. Grätzel, L. Emsley, "Doping and phase segregation in Mn^{2+}-and Co^{2+}-doped lead halide perovskites from ^{133}Cs and ^{1}H NMR relaxation enhancement". *Journal of Materials Chemistry A* **2019**, 7, 2326**.**

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

7. Kubicki, D. J.; Prochowicz, D.; Hofstetter, A.; Zakeeruddin, S. M.; Grätzel, M.; Emsley, L., "Phase Segregation in Potassium-Doped Lead Halide Perovskites from 39K solid-state NMR at 21.1 T". *Journal of the American Chemical Society* **2018**, 140 (23), 7232-7238.

8. Kubicki, D. J.; Prochowicz, D.; Hofstetter, A.; Saski, M.; Yadav, P.; Bi, D.; Pellet, N.; Lewiński, J.; Zakeeruddin, S. M.; Grätzel, M., "Formation of Stable Mixed Guanidinium–Methylammonium Phases with Exceptionally Long Carrier Lifetimes for High-Efficiency Lead Iodide-Based Perovskite Photovoltaics". *Journal of the American Chemical Society* **2018,** *140* (9), 3345-3351.

9. Busi, B.; Yarava, J. R.; Hofstetter, A.; Salvi, N.; Cala-De Paepe, D.; Lewandowski, J. R.; Blackledge, M.; Emsley, L., "Probing Protein Dynamics Using Multifield Variable Temperature NMR Relaxation and Molecular Dynamics Simulation". *J Phys Chem B* **2018,** *122* (42), 9697-9702.

10. Kumar, A.; Walder, B. J.; Kunhi Mohamed, A.; Hofstetter, A.; Srinivasan, B.; Rossini, A. J.; Scrivener, K.; Emsley, L.; Bowen, P., "The atomic-level structure of cementitious calcium silicate hydrate". *The Journal of Physical Chemistry C* **2017,** *121* (32), 17188-17196.

11. Kubicki, D. J.; Prochowicz, D.; Hofstetter, A.; Zakeeruddin, S. M.; Grätzel, M.; Emsley, L., "Phase Segregation in Cs-, Rb-and K-Doped Mixed-Cation $(MA)_x (FA)_{1-x} PbI_3$ Hybrid Perovskites from Solid-State NMR". *Journal of the American Chemical Society* **2017,** *139* (40), 14173-14180.

12. Kubicki, D. J.; Prochowicz, D.; Hofstetter, A.; Pechy, P.; Zakeeruddin, S. M.; Grätzel, M.; Emsley, L., "Cation Dynamics in Mixed-Cation $(MA)_x (FA)_{1-x} PbI_3$ Hybrid Perovskites from Solid-State NMR". *Journal of the American Chemical Society* **2017,** *139* (29), 10055-10061.

13. Hofstetter, A.; Emsley, L., "Positional variance in NMR crystallography". *Journal of the American Chemical Society* **2017,** *139* (7), 2573-2576.

14. Ghasemi, S. A.; Hofstetter, A.; Saha, S.; Goedecker, S., "Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network". *Physical Review B* **2015,** *92* (4), 045131.

Websites

1. F. M. Paruzzo, A. Hofstetter, F. Musil, D. Sandip, M. Ceriotti, L. Emsley, ***http://shiftml.epfl.ch/,*** **ShiftML:** Chemical Shifts in Molecular Solids by Machine Learning, **2018**.

News and Communications

1. ***JACS Spotlights***, E.G. Berg, Evaluating Error in NMR Crystallography. *Journal of the American Chemical Society* **2017,** *139* (9), 3301-3301.
2. ***EPFL Mediacom***, https://actu.epfl.ch/news/ai-and-nmr-spectroscopy-determine-atoms-configurat/, **2018**

Professional outreach and professional society service

2017	Scripting of a program to generate an automated conference schedule, 10th Alpine Conference on Solid-State NMR, Chamonix-Mont Blanc, France

Awards and Grants

2018	Laura Marinelli poster award at the Rocky Mountain Conference, Salt Lake City, USA
2018	"Travel funding for SMARTER6 conference 2018,” Funding from the conference committee.
2018	" Rocky Mountain Conference Travel Stipend 2018,” funding from the conference committee.
2017	" Travel funding and scholarship for the Small Molecule NMR Conference,” funding from the conference committee.
2017	"Travel funding for Gordon Research Conference and Seminar on Computational Aspects - Biomolecular NMR,” funding from EDCH doctoral program at EPFL.