

NMR Crystallography in the Big Data Era: New Methods and Applications Powered by Machine Learning

Présentée le 29 septembre 2023

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

Manuel CORDOVA

Acceptée sur proposition du jury

Prof. U. Röthlisberger, présidente du jury
Prof. D. L. Emsley, directeur de thèse
Prof. S. Price, rapporteuse
Prof. F. Hansen, rapporteur
Prof. N. Marzari, rapporteur

Make good choices.

– A.R.

Acknowledgements

First of all, I would like to thank all the members of my jury: **Sarah Price** (external expert), **Flemming Hansen** (external expert), **Nicola Marzari** (internal expert) and **Ursula Röthlisberger** (president) for taking the time to read my thesis and travel to Switzerland to attend my private defence. Your insights were extremely valuable to improve my work and gave me many new ideas for future development of my projects.

None of this work would have been possible without the continuous support and guidance from my PhD supervisor **Lyndon Emsley**. Thank you for making me discover the world of NMR crystallography, and for giving me the freedom to implement my ideas while guiding me towards their completion. I have learned so much from you, from conducting science projects and collaborating efficiently with others to presenting research in a clear and compelling manner. I will always admire your ability to always find the right thing to say when I felt like my projects were stalling, and to lead your lab with passion and integrity, allowing us all to do great science in the best environment possible.

I would like to thank **Nadia Gauljaux** for handling the administrative tasks during my PhD. I am grateful for how you made my life so much easier and allowed me to fully focus on science.

A big thank you to all my former and current labmates. Thank you, **Albert Hofstetter** and **Federico Paruzzo**, for paving the way to the beginning of my PhD and for teaching me so much in the few months where we overlapped. Thanks to my officemate **Michael Hope** for always answering my basic questions about NMR and listening to my weird computer problems. Thanks to my PhD twin **Bruno Simões de Almeida** for the time spent together figuring out NMR, travelling to conferences, and talking about random things. I will never forget our trip to Florida. Thank you **Pinelopi Moutzouri**, **Daria Torodii**, **Jacob Holmes** and **Martins Balodis** for teaching me about experimental aspects of NMR and for the very fruitful collaborations. Thanks to the former and current post-docs in our group, **Pierrick Berruyer**, **Andrea Bertarello**, **Claudia Avalos**, **Gabriele Stevanato**, **Federico de Biasi**, **Saumya Badoni**, **Amrit Venkatesh** and **Ran Wei**, as well as the PhD students of the lab, **Baptiste Busi**, **Anna Morales**, **Snædís Björgvinsdóttir**, **Aditya Mishra**, **Yu Rao**, **Gülsüm Günes** and **Ray Cowen** for your help and for very interesting and fun discussions. Finally, thank you again **Michael**, **Federico**, **Saumya** and **Jacob** for proofreading my thesis.

Thank you to all our collaborators without whom most of my projects would not have been possible. Thank you, **Michele Ceriotti**, **Edgar Engel**, **Matthias Kellner** and **Philippe Schwaller**, for making sure my machine learning practices are sound. Thanks to **Staffan Schantz**, **Sten Nilsson Lill**, **Emma Eriksson**, **Arthur Pinon** and all the AstraZeneca team for giving me a glimpse into the pharma industry through your fascinating problems.

Thank you to my bandmates **Rom**, **Kris** and **Ben** for allowing me to cool down after long days of science by playing music together, our shows are among the best moments of my life. Thank you for the very fun (and sometimes a bit too ambitious) projects we have undertaken, from writing albums to filming music videos.

Thank you to my family, my **mom**, my **dad**, and my brothers **Jonathan** and **Thomas** for being there for me and always supporting me, and for the fun adventures we had either playing or making escape rooms.

Last but not least, thank you to my girlfriend **Jenny** for always being there for me. Thank you for your support and encouragements, and for the amazing time spent together. I love you.

P.S. Thank you, reader, for having a look at my work. I hope you will find it useful.

Abstract

Structure determination of materials is key to understanding their physical properties. While single-crystal X-ray diffraction is the gold standard for structures displaying long-range order, many materials of interest are polycrystalline and/or disordered, which is a challenge for diffraction methods. On the other hand, nuclear magnetic resonance (NMR) spectroscopy probes the local environment around nuclei, and does not require long-range order. It is thus the method of choice for investigating the structure of disordered solids. In particular, NMR crystallography based on chemical shifts has proven able to determine the atomic-level structure of various materials through the combination of solid-state NMR experiments, crystal structure prediction (CSP) protocols, and density functional theory (DFT) computation of chemical shifts.

However, several drawbacks prevent the widespread use of NMR crystallography, especially for disordered materials. First, the computation of chemical shifts for candidate structures generated by CSP requires significant computational resources. In addition, CSP algorithms also require intensive computations to explore the space of possible crystal structures in order to construct a comprehensive set of candidates. Moreover, experimental measurement and assignment of chemical shifts is challenging, typically requiring time consuming, multi-dimensional experiments. These challenges are exacerbated in disordered solids, owing to the need to model these materials using large structures, typically generated using molecular dynamics (MD), which prevents the use of DFT to compute chemical shifts.

In this thesis, we use machine learning to help alleviate these drawbacks. We extend the capabilities of ShiftML, a previously introduced model of chemical shifts of molecular solids, and incorporate the model into CSP protocols, in order to drive the generation of candidate crystal structures towards the experimentally observed structure. We also predict chemical shifts using ShiftML on a large database of crystal structures, and leverage the resulting database of chemical shifts to help assign measured chemical shifts to atomic sites without prior knowledge about the three-dimensional structure of the molecule under study. The database is also used to construct chemical shift-dependent interaction maps in molecular solids. The maps generated can in turn be used to score candidate crystal structures without performing any additional DFT-level chemical shift computation, and to construct structural constraints to drive CSP protocols.

Another challenge tackled in this thesis is the resolution of ^1H NMR spectra of solids. Dipolar coupling between spins lead to broadened lineshapes, which can (partially) be removed by spinning the sample at the magic angle. However, at finite spinning rates, these interactions are not completely removed. We develop a convolutional recurrent neural network to obtain the spectra that would be obtained at infinite spinning rates from a set of spectra measures at variable spinning speeds. The model is applied both to one-dimensional ^1H spectra and two-dimensional ^1H – ^1H correlation experiments.

Finally, we investigate the structure of amorphous molecular solids by NMR crystallography, by replacing DFT chemical shift computations by ShiftML. This allows the computation of chemical shifts for ensembles of large structures generated by MD, that we compare to experimental values in order to extract preferred conformations and noncovalent interactions in amorphous compounds. A general method to determine the structure of amorphous molecular solids is introduced, which involves the simultaneous comparison of experimental and computed shifts of multiple atomic sites in the molecule studied.

Keywords

solid-state NMR, NMR crystallography, machine learning, structure determination, pharmaceutical compounds, amorphous compounds, crystal structure prediction, intermolecular interactions, chemical shift assignment

Résumé

Déterminer la structure de matériaux est crucial pour comprendre leurs propriétés physiques. Alors que la diffraction à rayons X est la méthode de référence pour les structures présentant un ordre à longue portée, plusieurs matériaux d'importance sont polycristallins et/ou désordonnés, ce qui est problématique pour les méthodes par diffraction. D'autre part, la spectroscopie par résonance magnétique nucléaire (RMN) sonde l'environnement local autour des noyaux, et ne nécessite pas d'ordre à longue portée. Elle est donc la méthode de choix pour étudier la structure de solides désordonnés. En particulier, la cristallographie par RMN basée sur les déplacements chimiques s'est montrée capable de déterminer la structure au niveau atomique de matériaux variés, au travers de la combinaison d'expériences RMN à l'état solide, de protocoles de prédiction de structure cristalline (PSC), et de calcul de déplacements chimiques par la théorie de la fonctionnelle de la densité (TFD).

Cependant, plusieurs inconvénients empêchent l'utilisation généralisée de la cristallographie par RMN, en particulier pour les matériaux désordonnés. En premier lieu, le calcul de déplacements chimiques pour des structures candidates générées par PSC nécessite d'importantes ressources de calcul. De plus, les algorithmes de PSC ont également besoin de calculs intensifs pour explorer l'espace des structures cristallines possibles afin de constituer un ensemble complet de candidats. En outre, la mesure expérimentale et l'assignement des déplacements chimiques est complexe, et requièrent généralement de longues expériences multidimensionnelles. Ces problèmes sont exacerbés pour les solides désordonnés, en raison du besoin de modéliser ces matériaux à l'aide de grandes structures, généralement générées par la dynamique moléculaire (DM), ce qui empêche l'utilisation de la TFD pour calculer les déplacements chimiques.

Dans cette thèse, nous utilisons l'apprentissage automatique afin de pallier ces inconvénients. Nous étendons les capacités de ShiftML, un modèle de déplacements chimiques pour solides moléculaires précédemment introduit, et intégrons le modèle dans des protocoles de PSC, afin d'orienter la génération de structures cristallines candidates vers la structure observée expérimentalement. Nous prédisons également les déplacements chimiques à l'aide de ShiftML pour une grande base de données de structures cristallines, et nous tirons parti de la base de données de déplacements chimiques résultante afin d'assister l'assignement des déplacements chimiques mesurés expérimentalement à des sites atomiques sans connaissance préalable de la structure tridimensionnelle de la molécule étudiée. La base de données est aussi utilisée pour construire des cartes d'interactions dépendantes du déplacement chimique dans les solides moléculaires. Les cartes générées peuvent à leur tour être utilisées pour classer des structures cristallines candidates sans effectuer aucun calcul supplémentaire de déplacement chimique par TFD, et pour établir des contraintes structurelles afin d'orienter les protocoles PSC.

Un autre problème traité dans cette thèse est la résolution des spectres RMN ^1H de solides. Les couplages dipolaires entre les spins conduisent à des lignes élargies, qui peuvent (partiellement) être éliminées par rotation de l'échantillon autour de l'angle magique. Cependant, avec des vitesses de rotations finies, ces interactions ne sont pas complètement supprimées. Nous développons un réseau neuronal récurrent convolutif afin d'obtenir les spectres qui seraient mesurés à des vitesses de rotation infinies à partir d'un ensemble de spectres mesurés à différentes vitesses de rotation. Le modèle est appliqué à la fois à des spectres ^1H unidimensionnels et à des expériences de corrélation ^1H – ^1H bidimensionnelles.

Finalement, nous étudions la structure des solides moléculaires amorphes par cristallographie RMN, en remplaçant le calcul de déplacements chimiques TFD par ShiftML. Cela permet le calcul de déplacements chimiques pour des ensembles de grandes structures générées par DM, que nous comparons aux valeurs expérimentales afin d'extraire les conformations et interactions non-covalentes préférentielles dans les composés amorphes. Une méthode générale pour déterminer la structure des solides moléculaires amorphes est introduite, qui implique la comparaison simultanée des déplacements chimiques expérimentaux et calculés de plusieurs sites atomiques dans la molécule étudiée.

Mots-clés

RMN de solides, cristallographie RMN, apprentissage automatique, détermination de structure, composés pharmaceutiques, composés amorphes, prédiction de structure cristalline, interactions intermoléculaires, assignement de déplacement chimique

Table of Contents

Acknowledgements	i
Abstract	iii
Résumé.....	v
Table of Contents.....	vii
List of publications.....	9
Chapter 1 Introduction	11
1.1 <i>Structure determination methods for solids</i>	11
1.2 <i>NMR crystallography using chemical shifts</i>	11
1.2.1 Chemical shifts as a probe of local structure	11
1.2.2 Chemical shifts from electronic structure methods	12
1.2.3 NMR crystallography	14
1.2.4 Example applications of NMR crystallography	15
1.3 <i>Machine learning in NMR</i>	21
1.3.1 Machine learning chemical shifts	21
1.3.2 Machine learning for the analysis of NMR data	22
1.4 <i>Outline of the present thesis</i>	23
Chapter 2 Accelerating NMR crystallography of microcrystalline solids	25
2.1 <i>Introduction</i>	25
2.2 <i>ShiftML2: A machine learning model of chemical shifts for chemically and structurally diverse molecular solids</i>	27
2.2.1 Introduction	27
2.2.2 Methods	27
2.2.3 Results and Discussion	30
2.2.4 Conclusion	36
2.2.5 Appendix I	36
2.3 <i>De novo crystal structure determination from machine learned chemical shifts</i>	45
2.3.1 Introduction	45
2.3.2 Methods	45
2.3.3 Results and Discussion	47
2.3.4 Conclusion	52
2.3.5 Appendix II	52
2.4 <i>Chemical shift-dependent interaction maps in molecular solids</i>	61
2.4.1 Introduction	61
2.4.2 Methods	61
2.4.3 Results and Discussion	63
2.4.4 Conclusion	69

2.4.5	Appendix III	69
Chapter 3	The assignment problem	87
3.1	<i>Introduction</i>	87
3.2	<i>Bayesian probabilistic assignment of chemical shifts in organic solids</i>	89
3.2.1	Introduction	89
3.2.2	Methods	89
3.2.3	Results and Discussion	92
3.2.4	Conclusion	98
3.2.5	Appendix IV	99
3.3	<i>Pure isotropic proton NMR spectra in solids using deep learning</i>	123
3.3.1	Introduction	123
3.3.2	Methods	124
3.3.3	Results and Discussion	129
3.3.4	Conclusion	134
3.3.5	Appendix V	135
3.4	<i>Two-dimensional pure isotropic proton solid state NMR</i>	147
3.4.1	Introduction	147
3.4.2	Methods	147
3.4.3	Results and Discussion	149
3.4.4	Conclusion	154
3.4.5	Appendix VI	155
Chapter 4	NMR crystallography of amorphous solids	165
4.1	<i>Introduction</i>	165
4.2	<i>Structure determination of an amorphous drug through large-scale NMR predictions</i>	167
4.2.1	Introduction	167
4.2.2	Methods	167
4.2.3	Results and Discussion	169
4.2.4	Conclusion	174
4.2.5	Appendix VII	174
4.3	<i>Atomic-level structure determination of amorphous molecular solids by NMR</i>	183
4.3.1	Introduction	183
4.3.2	Methods	183
4.3.3	Results and Discussion	186
4.3.4	Conclusion	191
4.3.5	Appendix VIII	191
Chapter 5	Conclusion	203
5.1	<i>Results achieved</i>	203
5.2	<i>Future development</i>	204
Bibliography		205
Curriculum Vitae		227

List of publications

The present thesis is based on the following publications:

1. Cordova, M.; Balodis, M.; Hofstetter, A.; Paruzzo, F.; Nilsson Lill, S. O.; Eriksson, E. S. E.; Berruyer, P.; Simões de Almeida, B.; Quayle, M. J.; Norberg, S. T.; Svensk Ankarberg, A.; Schantz, S.; Emsley, L., Structure determination of an amorphous drug through large-scale NMR predictions. *Nature Communications* **2021**, *12* (1), 2964.
2. Morales-Melgares, A.; Casar, Z.; Moutzouri, P.; Venkatesh, A.; Cordova, M.; Kunhi Mohamed, A.; Scrivener, K. L.; Bowen, P.; Emsley, L., Atomic-Level Structure of Zinc-Modified Cementitious Calcium Silicate Hydrate. *Journal of the American Chemical Society* **2022**, *144* (50), 22915-22924.
3. Hope, M. A.; Nakamura, T.; Ahlawat, P.; Mishra, A.; Cordova, M.; Jahanbakhshi, F.; Mladenovic, M.; Runjhun, R.; Merten, L.; Hinderhofer, A.; Carlsen, B. I.; Kubicki, D. J.; Gershoni-Poranne, R.; Schneeberger, T.; Carbone, L. C.; Liu, Y.; Zakeeruddin, S. M.; Lewinski, J.; Hagfeldt, A.; Schreiber, F.; Rothlisberger, U.; Gratzel, M.; Milic, J. V.; Emsley, L., Nanoscale Phase Segregation in Supramolecular pi-Templating for Hybrid Perovskite Photovoltaics from NMR Crystallography. *Journal of the American Chemical Society* **2021**, *143* (3), 1529-1538.
4. Cordova, M.; Engel, E. A.; Stefaniuk, A.; Paruzzo, F.; Hofstetter, A.; Ceriotti, M.; Emsley, L., A Machine Learning Model of Chemical Shifts for Chemically and Structurally Diverse Molecular Solids. *The Journal of Physical Chemistry C* **2022**, *126* (39), 16710-16720.
5. Balodis, M.; Cordova, M.; Hofstetter, A.; Day, G. M.; Emsley, L., De Novo Crystal Structure Determination from Machine Learned Chemical Shifts. *Journal of the American Chemical Society* **2022**, *144* (16), 7215-7223.
6. Cordova, M.; Emsley, L., Chemical Shift-Dependent Interaction Maps in Molecular Solids. *Journal of the American Chemical Society* **2023**, *145* (29), 16109-16117.
7. Cordova, M.; Balodis, M.; Simões de Almeida, B.; Ceriotti, M.; Emsley, L., Bayesian probabilistic assignment of chemical shifts in organic solids. *Science Advances* **2021**, *7* (48), eabk2341.
8. Cordova, M.; Moutzouri, P.; Simões de Almeida, B.; Torodii, D.; Emsley, L., Pure Isotropic Proton NMR Spectra in Solids using Deep Learning. *Angewandte Chemie International Edition* **2023**, *62* (8), e202216607.
9. Moutzouri, P.; Cordova, M.; Simões de Almeida, B.; Torodii, D.; Emsley, L., Two-dimensional Pure Isotropic Proton Solid State NMR. *Angewandte Chemie International Edition* **2023**, *62* (21), e202301963.
10. Cordova, M.; Moutzouri, P.; Nilsson Lill, S. O.; Cousen, A.; Kearns, M.; Norberg, S. T.; Svensk Ankarberg, A.; McCabe, J.; Pignon, A. C.; Schantz, S.; Emsley, L., Atomic-level Structure Determination of Amorphous Molecular Solids by NMR. *In press* **2023**.

During my PhD, I have also worked on other aspects of solid-state nuclear magnetic resonance. In particular, this includes work on the structure determination of perovskite materials by NMR, as well as the analysis of the structure of radicals used to enhance the sensitivity of NMR experiments in order to understand the structural factors contributing to the efficiency of these molecules. These results are not part of the present thesis, and can be found in the following publications:

1. Stevanato, G.; Casano, G.; Kubicki, D. J.; Rao, Y.; Esteban Hofer, L.; Menzildjian, G.; Karoui, H.; Siri, D.; Cordova, M.; Yulikov, M.; Jeschke, G.; Lelli, M.; Lesage, A.; Ouari, O.; Emsley, L., Open and Closed Radicals: Local Geometry around Unpaired Electrons Governs Magic-Angle Spinning Dynamic Nuclear Polarization Performance. *Journal of the American Chemical Society* **2020**, *142* (39), 16587-16599.
2. Datta, K.; Caiazzo, A.; Hope, M. A.; Li, J.; Mishra, A.; Cordova, M.; Chen, Z.; Emsley, L.; Wienk, M. M.; Janssen, R. A. J., Light-Induced Halide Segregation in 2D and Quasi-2D Mixed-Halide Perovskites. *ACS Energy Letters* **2023**, *8* (4), 1662-1670.

Chapter 1 Introduction

1.1 Structure determination methods for solids

Structure governs the physical properties of matter. Determining the structure of molecules and materials at the atomic level is thus paramount to understanding and optimising macroscopic properties such as the efficacy of pharmaceutical compounds,¹⁻³ the photovoltaic performances of perovskite materials,⁴⁻⁶ or the efficiency of enzymes^{7, 8} and inorganic catalysts.^{9, 10} Single-crystal X-ray diffraction (XRD) is the gold standard method to determine the atomic-level structure of crystalline materials.¹¹⁻¹⁵ However, this method requires a single crystal of the material under study with a size at least in the order of a hundred microns, which can be difficult to obtain.^{14, 15} Powder diffraction (either X-ray or neutron) can be used for microcrystalline samples, but determining the structure is challenging in all but the simplest cases.¹⁶⁻²¹ In addition, amorphous samples can be studied using X-ray and neutron total scattering experiments and pair-distribution analysis, however again interpretation is extremely challenging.²²⁻²⁵ Electron microscopy (EM) can be applied to determine the nanostructure and atomic-level structure of some inorganic materials.²⁶⁻³⁰ In particular, the development of cryo-EM in recent years has expanded the range of applicable materials to organic molecules and proteins.³¹⁻³³ Nevertheless, the atomic-level structure determination of microcrystalline and amorphous materials by diffraction methods remains challenging due to the difficulty to interpret data from samples lacking long-range order.

1.2 NMR crystallography using chemical shifts

Nuclear magnetic resonance (NMR) spectroscopy is an experimental technique that probes the magnetic properties of nuclear spins. One key advantage with respect to diffraction techniques is that it does not rely on long-range order in the material under study. NMR has been widely used to determine the structure of molecules in solution,³⁴⁻³⁶ as well as a variety of solids including organic compounds,³⁷⁻⁵⁶ proteins,⁵⁷⁻⁶⁰ zeolites,⁶¹⁻⁶⁵ cementitious materials,^{66, 67} battery materials,⁶⁸ perovskites,⁶⁹ and other inorganic materials.⁷⁰⁻⁷⁶

1.2.1 Chemical shifts as a probe of local structure

NMR spectroscopy measures the precession of magnetic moments μ in a static magnetic field \mathbf{B}_0 . Nuclei that have a non-zero spin I have a magnetic moment given by

$$\mu = \gamma \hbar I, \quad (1.1)$$

where γ is the gyromagnetic ratio of the nucleus and \hbar is the reduced Planck constant.⁷⁷⁻⁸⁰ Magnetic moments can thus be detected by NMR for such so-called NMR active nuclei, including ^1H , ^{13}C and ^{15}N , among others. The frequency of precession of a magnetic moment, which is measured by NMR, results from its atomic environment. It is determined by the energy of the interactions involving the corresponding nucleus i , described by the NMR Hamiltonian,

$$H_{\text{NMR}} = -\hbar \gamma_i \mathbf{B}_0 (\bar{\mathbf{1}} - \bar{\boldsymbol{\sigma}}) \mathbf{I}_i + \frac{1}{2} \hbar^2 \sum_{j \neq i} \gamma_i \gamma_j \mathbf{I}_i (\bar{\mathbf{D}}_{ij} + \bar{\mathbf{J}}_{ij}) \mathbf{I}_j + \mathbf{I}_i \bar{\mathbf{Q}}_i \mathbf{I}_i. \quad (1.2)$$

This Hamiltonian represents the main interactions that affect the energy of nucleus i in diamagnetic solids. The first term describes the interaction between the spin \mathbf{I}_i and the applied magnetic field \mathbf{B}_0 , called the Zeeman effect. This interaction is modulated by the magnetic shielding tensor $\bar{\boldsymbol{\sigma}}$, which arises from the presence of electrons around nucleus i . The magnetic field generates electric currents in the electron cloud by electromagnetic induction, which in turn generate an induced magnetic field \mathbf{B}_{ind} that opposes \mathbf{B}_0 . The shielding tensor relates \mathbf{B}_0 to \mathbf{B}_{ind} as

$$\mathbf{B}_{\text{ind}} = -\bar{\boldsymbol{\sigma}} \mathbf{B}_0. \quad (1.3)$$

The second term in **Equation 1.2** represents interactions between two spins. Dipolar spin–spin interactions can occur directly between nuclei and are represented by the dipolar coupling \bar{D}_{ij} , or indirectly using electrons as intermediates, which is represented by the scalar coupling \bar{J}_{ij} , also called J-coupling.

The last term in **Equation 1.2** is only present for nuclei with spins greater than $\frac{1}{2}$ and describes the interaction between the quadrupolar moment of the nucleus and the surrounding electric field gradient. This interaction is described by the quadrupolar coupling tensor \bar{Q}_i .

While all interactions in **Equation 1.2** can provide information about the local environment around nuclei, I will mainly focus on the first term of the NMR Hamiltonian in this work. In NMR experiments, the absolute shielding tensor is not directly measured. Instead, the shielding tensor $\bar{\sigma}$ is measured relative to that of a reference compound $\bar{\sigma}_{ref}$ to give the chemical shift tensor,

$$\bar{\delta} = \bar{\sigma}_{ref} - \bar{\sigma}. \quad (1.4)$$

In practice, NMR crystallography of molecular solids is mainly based on the isotropic value of the chemical shift tensor for ^1H and/or ^{13}C nuclei. The isotropic chemical shift δ is obtained as one third of the trace of the chemical shift tensor. **Equation 1.4** can be adapted for isotropic values as

$$\delta = \sigma_{ref} - \sigma, \quad (1.5)$$

where σ and σ_{ref} are the isotropic shieldings of the nucleus in the sample of interest and in the reference compound, respectively, obtained as one third of the trace of the shielding tensors. Since the isotropic shielding strongly depends on the local electronic density around the nucleus, which is determined by the positions of the neighbouring atoms, the chemical shift is a direct probe of the local atomic environment around the nucleus. This provides a powerful method to determine the structure of materials by NMR.⁷⁷⁻⁸⁰

In practice, obtaining isotropic chemical shifts for solid compounds presents numerous challenges. In particular, the chemical shifts and dipolar couplings depend on the relative orientation of the sample with respect to the main magnetic field, which leads to severe broadening of the NMR spectra of powdered samples. In liquid samples, the dipolar interactions are averaged out by molecular tumbling, but solid samples require coherent averaging schemes to remove these interactions. Rotating the sample at the “magic angle” (54.74°) with respect to the main magnetic field, a method called magic angle spinning (MAS), leads to the removal of 2nd-rank anisotropic interactions in solids, including dipolar interactions.⁸¹⁻⁸⁵ However, often these interactions are not yet fully removed even at the highest rotating speeds currently available.^{84, 85}

1.2.2 Chemical shifts from electronic structure methods

Chemical shifts encode the local atomic environments around nuclei. They are thus direct probes of the structure of materials. However, decoding chemical shifts into atomic-level structure is not (currently) directly possible. Instead, model structures of the materials studied are typically constructed, and their associated expected chemical shifts are compared to the experimental values. This requires accurate methods to obtain chemical shifts for these model structures.

Many quantum-mechanical properties of materials can be obtained from their ground-state wavefunction. The development of first principles (*ab initio*) methods to determine the wavefunction Ψ and its related properties by solving the Schrödinger equation⁸⁶ (**Equation 1.6**) has thus been a highly active field of research since the establishment of the Hartree-Fock (HF) method that provides an approximate solution to the Schrödinger equation.⁸⁷⁻⁸⁹ The Schrödinger equation expresses the energy E of a system described by a wavefunction Ψ as the application of the Hamiltonian operator formed by the kinetic ($-\frac{1}{2}\nabla^2$) and potential ($V(\mathbf{r})$) energy operators to the wavefunction. The resulting eigenvalues and eigenvectors give the energy levels and associated wavefunctions of the system, respectively.

$$\left(-\frac{1}{2}\nabla^2 + V(\mathbf{r})\right)\Psi = E\Psi \quad (1.6)$$

The high computational cost of post-HF methods⁹⁰⁻⁹² has driven the need for accurate methods with computational requirements similar to the HF method. Hohenberg and Kohn have shown the equivalence of obtaining observables from the wavefunction or the electron density of atomic systems.⁹³ Based on this, density functional theory (DFT) has been the leading first-principles framework to compute physical properties of atomic systems with reasonable accuracy and at relatively low computational cost.⁹⁴ In DFT, the electronic wavefunction Ψ is replaced by a set of non-interacting pseudoelectrons, each with a wavefunction ϕ_i and energy ϵ_i experiencing a potential V_{eff} ,

$$\left(-\frac{1}{2}\nabla^2 + V_{\text{eff}}(\mathbf{r})\right)\phi_i(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r}) \quad (1.7)$$

with V_{eff} defined such that the electronic density of the system reproduces the electronic density from the true wavefunction,

$$\rho(\mathbf{r}) = |\Psi|^2 = \sum_i |\phi_i(\mathbf{r})|^2. \quad (1.8)$$

V_{eff} contains the classical electrostatic interaction between each nucleus I with charge Z_I and position \mathbf{R}_I and the electron density $\rho(\mathbf{r})$, as well as the classical electron–electron electrostatic interaction. A correction term for quantum mechanical exchange, the removal of self-interaction and the correlation of the motion of electrons is introduced as the exchange–correlation functional, E_{xc} .

$$V_{\text{eff}} = -\sum_I \int \frac{Z_I \cdot \rho(\mathbf{r})}{|\mathbf{R}_I - \mathbf{r}|} d\mathbf{r} + \iint \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}d\mathbf{r}' + E_{\text{xc}} \quad (1.9)$$

Different levels of approximation for E_{xc} make up the different so-called levels of theory which can be categorised into local density approximation (LDA) functionals that depends on $\rho(\mathbf{r})$ such as the VWN and PW92 functionals,^{95, 96} general gradient approximation (GGA) functionals that also include the gradient of $\rho(\mathbf{r})$ such as the PBE and PB86 functionals,⁹⁷⁻⁹⁹ meta-GGA functionals that incorporate higher-order derivatives of $\rho(\mathbf{r})$ such as the M06-L and TPSS functionals,¹⁰⁰⁻¹⁰³ and hybrid functionals that include Hartree-Fock exchange in addition to $\rho(\mathbf{r})$ and its gradient in the description of E_{xc} , such as the B3LYP and PBE0 functionals.^{95, 99, 104-106} While DFT is exact in its principle (with an exact E_{xc}), the choice of the functional determines the accuracy of the property computed.

Cost-effective DFT computations rely on the use of an efficient basis to describe the electron density. While Slater functions¹⁰⁷ and spherical harmonics form an effective basis set to describe atomic systems, the periodicity of crystals makes plane waves a natural choice to describe $\rho(\mathbf{r})$ for crystalline solids. A major drawback of plane wave basis sets is the need for high frequency waves to describe the core electrons, significantly increasing the computational cost compared to atomic orbital-like basis sets. To circumvent this issue, core electrons can be replaced by pseudopotentials¹⁰⁸⁻¹¹⁰ and the all-electron wavefunction (or density) can be reconstructed using the projector augmented-wave (PAW) method.¹¹¹

The magnetic shielding tensor $\bar{\sigma}$ for a nucleus I in an atomic system can be obtained from the ground-state electron density by computing the second derivative of the energy E with respect to the magnetic moment of the nucleus μ^I and the applied magnetic field \mathbf{B}_0 . In a given frame, the components of the shielding tensor are given by

$$\sigma_{ij}^I = \frac{\partial^2 E}{\partial B_j \partial \mu_i^I}, \quad (1.10)$$

where B_j is the j -th component of the magnetic field and μ_i^I is the i -th component of the magnetic moment of nucleus I . Perturbation theory is typically used to obtain $\bar{\sigma}$ from first principles. However, with finite basis sets the results depend on the absolute position of the atomic system in space. Solving this so-called gauge problem has led to the rise of the gauge-including atomic orbital (GIAO)¹¹²⁻¹¹⁶ and gauge-including projector-augmented wave (GIPAW)^{117, 118} formalisms as popular methods to compute chemical shifts for atomic systems using atomic orbital and plane wave basis sets, respectively. While the inherent periodicity of the GIPAW method renders it directly suitable to obtain chemical shifts of periodic systems, GIAO chemical shifts can be obtained in crystalline solids by constructing clusters to incorporate interactions with neighbouring molecules in the crystal packing,¹¹⁹⁻¹²² using charge embedding to model long-range electrostatic interactions,^{123, 124} or decomposing crystals into fragments and summing the many-body contributions to the chemical shift to model the global effect of packing interactions.¹²⁴⁻¹²⁸

1.2.3 NMR crystallography

NMR parameters are sensitive probes of the local electronic density, which encodes the local atomic environment. Thus, NMR has allowed the structure determination of various compounds by comparison of experimental parameters with values computed for model systems, in a process called NMR crystallography. In this thesis, I will focus on (isotropic) chemical shift-based NMR crystallography, although several examples of structure determination have been reported using other NMR parameters such as dipolar couplings,⁶¹ full chemical shift tensors,^{63, 65} quadrupolar couplings,^{129, 130} and spin diffusion rates.^{42, 44, 47, 50}

In 1993, De Dios, Pearson and Oldfield established the importance of torsion angles, hydrogen bonding and electrostatic environment to describe ¹³C, ¹⁵N and ¹⁹F chemical shifts in proteins.¹³¹ The same year, Facelli and Grant demonstrated the high sensitivity of ¹³C chemical shifts to molecular structure.¹³² Subsequently, ¹H chemical shifts have been established as sensitive probes of intermolecular interactions in solids.^{38, 39, 133} This provides useful handles to determine the structure of molecular solids by NMR. The accuracy of DFT chemical shift computations was found to be sufficient to assign chemical shifts measured experimentally by comparison with computations performed on structures obtained using single-crystal X-ray diffraction.¹³⁴⁻¹⁴⁰ In addition, combining solid-state NMR and DFT chemical shift computations has allowed the validation of X-ray structures as well as accurate determination of hydrogen positions,¹⁴¹⁻¹⁴⁴ which have historically been difficult to obtain from X-ray diffraction patterns.¹⁴⁵ In particular, NMR crystallography is a powerful method to determine the tautomeric form¹⁴⁶⁻¹⁵¹ and (zwitter)ionic character¹⁵²⁻¹⁵⁵ of molecular solids and co-crystals, in addition to being able to determine the number of molecules in the asymmetric unit (Z').^{156, 157}

Performing structure determination by chemical shift-based NMR crystallography involves measuring experimental chemical shifts for the material under study, generating a set of candidate crystal structures through chemical modelling methods such as crystal structure prediction (CSP) protocols^{158, 159} or molecular dynamics (MD) simulations,¹⁶⁰⁻¹⁶³ and comparing the chemical shifts computed for these model structures to the corresponding experimental values. If a model structure yields an error between experimental and computed shifts, (e.g., the root-mean-square error (RMSE)) below the expected error for the DFT method used, then it is considered to be correct. This process thus relies on the generation of a comprehensive set of candidate crystal structures, as well as on the accuracy of the DFT method used to compute chemical shifts. PBE⁹⁷ is a popular level of theory for GIPAW computations, while hybrid exchange-correlation functionals such as B3LYP^{95, 99, 104} or PBE0^{105, 106} are typically used within the GIAO approach. While the computed isotropic shieldings can in principle be converted to isotropic chemical shifts using the shielding value computed for a reference compound as described in **Equation 1.5**, a direct linear regression between computed shieldings and experimental chemical shifts is generally performed, either directly between the shieldings computed for the candidate crystal under consideration and the experimental shifts, or using an external set of chemically and/or structurally similar compounds with known crystal structures and chemical shifts,

$$\delta = a\sigma + b. \quad (1.11)$$

Allowing the slope a in the regression to deviate from the theoretical value of -1 allows the removal of systematic error incurred by the DFT method. The offset b is equal to σ_{ref} in **Equation 1.5** if the slope is -1.

NMR crystallography has been combined with diffraction experiments to perform structure determination of powdered molecular solids.¹⁶⁴⁻¹⁶⁹ In addition, in recent years numerous successful structure determination based solely on NMR data have been performed. In 2010, Salager *et al.* introduced a method purely based on the combination of CSP protocols and ¹H chemical shifts to determine the structure of powdered molecular solids.⁴⁹ The method was validated by obtaining the correct crystal structure of thymol. Baías *et al.* further demonstrated the generality of the method by demonstrating its ability to determine the crystal structure of cocaine, flutamide, flufenamic acid and theophylline.⁵² The method was validated by comparison of the structures determined by NMR with those obtained using single-crystal X-ray diffraction. Baías *et al.* then used this method to perform *de novo* structure determination of form 4 of a large polymorphic drug molecule named AZD8329, for which no X-ray structure was previously available.⁵³ Following these important milestones in the development of NMR crystallography, the structures of several molecular and macromolecular crystals were determined using similar methods.¹⁷⁰⁻¹⁷⁴

Although NMR crystallography is a powerful method to determine the structure of molecular solids, several bottlenecks hinder its widespread use. In particular, the DFT computation of chemical shifts for candidate structures is computationally expensive, typically requiring the use of high-performance computing facilities and restricting the number of CSP candidates that can be evaluated. Acceleration of this process is discussed in **Section 1.3**. Another bottleneck results from the high-dimensional energy landscape to explore during the CSP procedure in order to include the correct crystal structure in the CSP set of candidates. In practice, a comprehensive sampling of the energy landscape would be prohibitively expensive computationally. Thus, CSP procedures typically incorporate restrictions of the conformational space explored based on the conformational energy of the molecule in the gas phase. However, intermolecular interactions in the crystal structure can stabilise unfavourable gas-phase conformations, which may not be selected by the CSP procedure. To address this issue, Hofstetter *et al.* introduced a method to obtain experimental constraints for the generation of conformations, based on two-dimensional ^1H - ^{13}C heteronuclear correlation (HETCOR) NMR experiments.⁵⁵ They showed that this method was successful in determining the structure of ampicillin through NMR crystallography, which would have failed using fully energy-based CSP procedures due to the high energy of the gas-phase conformation of the molecule. Another advantage of this method is that it reduces the number of generated CSP candidates compatible with experiments, ultimately reducing the overall cost of subsequent DFT computations of chemical shifts to determine the crystal structure.

In recent years, several improvements to chemical shift-based NMR crystallography have been introduced. In 2017, Hofstetter *et al.* proposed a method to obtain the positional uncertainty of structures determined by NMR crystallography.¹⁷⁵ Based on the expected error of DFT-based chemical shift computations and experimental errors, the method correlates perturbations in the atomic positions within the crystal structure to the error between computed and experimental shifts, and provides the uncertainty of atomic positions by selecting structures within the expected error and analysing the corresponding displacements for each atomic site.

Identifying the correct structure among a CSP set of candidates is not always straightforward and heavily depends on the accuracy of the method used to compute chemical shifts. In addition, the simultaneous evaluation of errors between experimental and computed shifts for different nuclei may be challenging to unambiguously identify the best matching candidate structure. A Bayesian probabilistic framework was introduced in 2019 by Engel *et al.* that allows the critical evaluation of shifts from multiple elements, incorporates the expected error of the DFT method in the analysis, and provides a quantified probability for each candidate structure to match experiments.¹⁷⁶ This method provides a quantified confidence in the identification of the experimental crystal structure, as well as an indication of whether the experimental structure is present in the CSP set or not.

In this context, I have been involved in structure determination of different materials by combined solid-state NMR and DFT calculations of chemical shifts of model structures during my PhD. Rather than go into further details here to review current NMR crystallography protocols, in the following I exemplify the approaches through three application examples that I carried out during my PhD.

1.2.4 Example applications of NMR crystallography

The following three examples have been adapted with permission from:

- Cordova, M.; Balodis, M.; Hofstetter, A.; Paruzzo, F.; Nilsson Lill, S. O.; Eriksson, E. S. E.; Berruyer, P.; Simões de Almeida, B.; Quayle, M. J.; Norberg, S. T.; Svensk Ankarberg, A.; Schantz, S.; Emsley, L., Structure determination of an amorphous drug through large-scale NMR predictions. *Nature Communications* **2021**, *12* (1), 2964. (post-print)
- Morales-Melgares, A.; Casar, Z.; Moutzouri, P.; Venkatesh, A.; Cordova, M.; Kunhi Mohamed, A.; Scrivener, K. L.; Bowen, P.; Emsley, L., Atomic-Level Structure of Zinc-Modified Cementitious Calcium Silicate Hydrate. *Journal of the American Chemical Society* **2022**, *144* (50), 22915-22924. (post-print)
- Hope, M. A.; Nakamura, T.; Ahlawat, P.; Mishra, A.; Cordova, M.; Jahanbakhshi, F.; Mladenovic, M.; Runjhun, R.; Merten, L.; Hinderhofer, A.; Carlsen, B. I.; Kubicki, D. J.; Gershoni-Poranne, R.; Schneeberger, T.; Carbone, L. C.; Liu, Y.; Zakeeruddin, S. M.; Lewinski, J.; Hagfeldt, A.; Schreiber, F.; Rothlisberger, U.; Gratzel, M.; Milic, J. V.; Emsley, L., Nanoscale Phase Segregation in Supramolecular pi-Templating for Hybrid Perovskite Photovoltaics from NMR Crystallography. *Journal of the American Chemical Society* **2021**, *143* (3), 1529-1538. (post-print)

The first example is the determination of the crystal structure of the crystalline form of the drug AZD5718 from a powder sample using the most state of the art approaches at the time,¹⁷⁷ for which my contribution was to compute the chemical shifts of the candidate crystal structures and compare them to experimental values to determine the structure of the drug, as well as to determine the positional uncertainty of the atoms in the structure. This example is part of the work presented in **Section 4.2**, and is briefly described below. A more detailed description is given in **Section 4.2.3**.

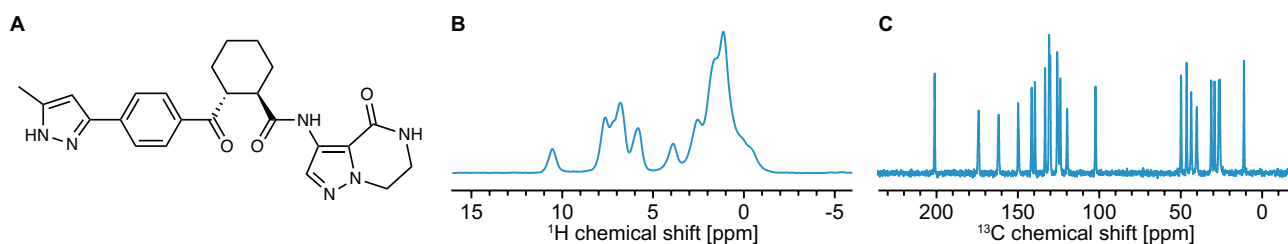


Figure 1.1. (A) Molecular structure of AZD5718. (B) ^1H and (C) ^{13}C NMR spectra of crystalline AZD5718.

Figure 1.1 shows the molecular structure of the drug along with the ^1H and ^{13}C magic angle spinning (MAS) NMR spectra of the powdered sample of AZD5718. Multidimensional ^1H - ^{13}C and ^{13}C - ^{13}C correlation spectra (see **Section 4.2.2** and **Appendix VII**) allowed the measurement and assignment of individual chemical shifts for all proton and carbon sites in the molecule. We generated a set of candidate crystal structures using a rapid CSP protocol, and computed chemical shifts for the candidates with the ten lowest predicted energy generated using a fragment- and cluster-based approach, the PBE0 density functional and the GIAO method.^{106, 115, 126-128} **Figure 1.2A** shows the RMSE between ^1H and ^{13}C chemical shifts for the candidates considered, as well as for the structure obtained using single-crystal X-ray diffraction. We note that while the candidate #1 and X-ray structures are similar, with a RMSD_{15} (root-mean-square deviation of the atomic positions in 15 molecules, ignoring hydrogen positions) of 0.42\AA , the bicyclo ring displays a different conformation in the two structures (see **Appendix VII**). The RMSEs obtained for the ^1H shifts suggested that candidate #1, the lowest energy candidate, best matches the experiment, while ^{13}C chemical shifts results identified the X-ray structure as the best match.

In order to quantitatively determine whether the candidate #1 or XRD structure best matches the NMR experiments, we applied the Bayesian analysis introduced by Engel *et al.*¹⁷⁶ to the CSP set and the XRD structure and obtained a 99.7% confidence that candidate #1 best matches experiment (**Figure 1.2B**). Although the computed shifts for the XRD structure appear closer to the experimental result (red cross) in the first two chemical shift principal components in **Figure 1.2B**, including the complete chemical shift space identifies candidate #1 as the structure that best matches experiment, as indicated by its associated probability.

The unit cell of the crystal structure of AZD5718 determined by NMR powder crystallography is shown in **Figure 1.3A**. By perturbing the structure through MD simulations and evaluating the associated extent of increase in ^1H chemical shift RMSE with respect to experiment, we obtained the positional uncertainty of the atoms in the molecule, as introduced by Hofstetter *et al.*¹⁷⁵ **Figure 1.3B** shows the ORTEP¹⁷⁸ plot of the atomic displacement parameter¹⁷⁹ (ADP) tensors corresponding to a ^1H chemical shift RMSE of 0.34 ppm . This value corresponds to the estimated error of ^1H chemical shifts computed with the fragment- and cluster-based approach.¹²⁷ The average value of the ADPs is 0.00025 \AA^2 .

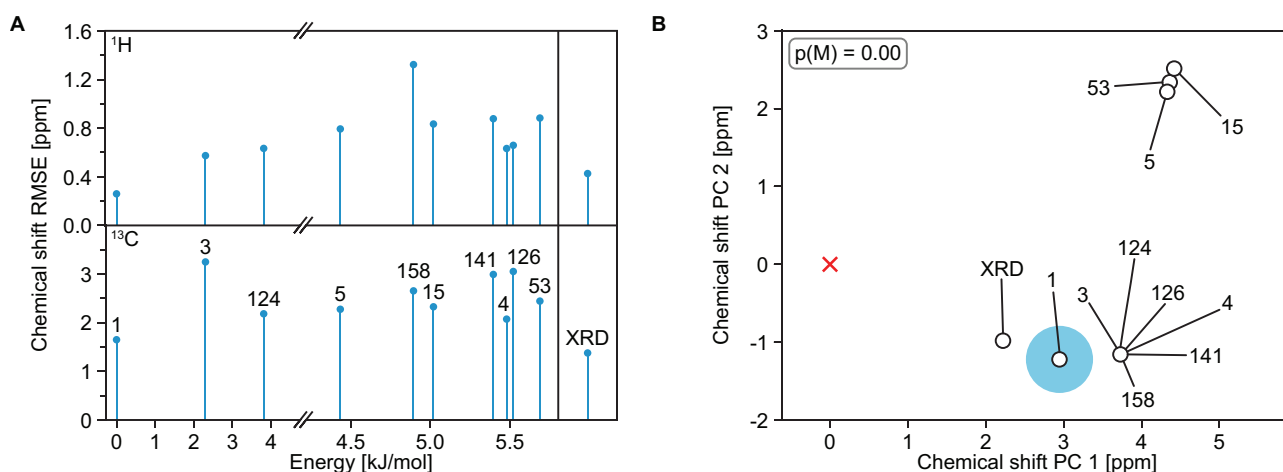


Figure 1.2. (A) ^1H (top) and ^{13}C (bottom) chemical shift RMSEs of the ten lowest energy candidate crystal structures and the single-crystal X-ray structure of AZD5718. (B) Two-dimensional projection of the similarity of the computed ^1H and ^{13}C chemical shifts of the candidate structures to the experimental data (red cross). The probability of each candidate matching experiment is represented by the area of the blue disk. $p(M)$ represents the probability that a virtual candidate, which represents structures potentially missing from the CSP candidate pool, matches experiment. A large value of $p(M)$ would indicate that the correct structure may not be present in the set of candidates considered, which is not the case here.

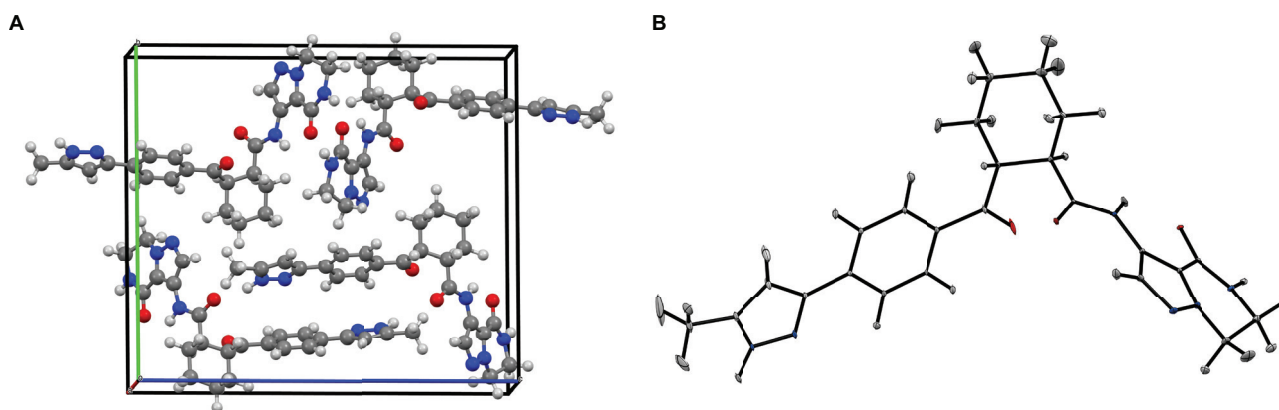


Figure 1.3. (A) Three-dimensional crystal structure of AZD5718 determined by NMR powder crystallography. (B) ORTEP plot of the ADP tensors for the NMR structure of AZD5819 drawn at the 90% probability level.

This example highlights the ability of NMR crystallography to determine the structure of microcrystalline molecular solids from powder samples. The obtained crystal structure displays very low positional uncertainty. A more detailed description of this system is described in **Section 4.2.2**.

AZD5718 provides an illustrative example of how NMR can be used to solve structures in crystalline powders. The second and third examples of the approach show how NMR can also solve complete atomic-level structures which are not crystalline in the usual sense. Indeed, due to the high sensitivity of chemical shifts to local structure, NMR crystallography can also readily be used to determine the structure of disordered materials through the comparison of experimental shifts with values computed from model structures representing local environments. An example is the determination of the structure of zinc-modified calcium silicate hydrate (CSH),¹⁸⁰ where my contribution was to compute chemical shifts for model structures and compare the results obtained with the experimental spectra.

Concrete is one of the most used substances on earth, and accounts for around 8% of anthropogenic CO₂ emissions.¹⁸¹ Lowering its carbon footprint is therefore paramount, and a promising approach involves substituting the clinker (the main ingredient used in the manufacture of Portland cement) by supplementary cementitious materials (SCMs) that have much lower associated CO₂ emissions, but have a tendency to lower the early-age strength of the resulting concrete.^{182, 183} In contrast, the addition of zinc to the main phase in clinker was found to enhance the early-age mechanical strength of the hydrated paste.¹⁸⁴⁻¹⁸⁶ This observation was associated with the growth of longer CSH particles.¹⁸⁷ Determining the incorporation of zinc into the CSH structure at the atomic level is thus important to understand the role of zinc in C-S-H growth and kinetics and would open pathways to synthetic tunability of the rate of reaction of lower-CO₂ materials.

Figure 1.4A shows the so-called dreierketten chains making up the main structure of CSH. Silicates ($\text{SiO}_{4-x}\text{H}_y^{(4-2x+y)}$, $0 \leq x < 2$, $0 \leq y \leq 4$) are found in three sites, $Q^{(1)}$, $Q^{(2b)}$ and $Q^{(2p)}$. The incorporation of zinc in the structure leads to different possible new silicate sites, $Q^{(1,Zn)}$, $Q^{(2p,Zn)}$, $Q^{(2p,2Zn)}$, $Q^{(2b,Zn)}$ and $Q^{(1,Zn_int)}$, resulting from the substitution of $Q^{(1)}$, $Q^{(2b)}$, or $Q^{(2p)}$ sites by zinc polyhedra (**Figure 1.4B**). In addition, zinc can also be present on top of the silicate chains, facing into the interlayer, where it could coordinate to one or both $Q^{(1)}$ species of a silicate dimer.

The experimental dynamic nuclear polarisation (DNP)-enhanced ^{29}Si spectra of samples with different target zinc to silica ratios ($\text{Zn}:\text{Si}$) are shown in **Figure 1.4C**. Increasing the amount of zinc in the sample leads to an enhanced signal around -72 ppm and to a decrease of the signal at -78.9 ppm. To understand the atomic-level structures making up these signals, we constructed 98 different zinc-modified CSH structural units via “brick” models.¹⁸⁸ For each structure, we computed DFT chemical shifts using the GIPAW formalism. The calculated shieldings were converted to chemical shifts using an external reference set composed of the structures of α -quartz, foshagite, hemimorphite and willemite, for which the experimental ^{29}Si shifts and crystal structures are known.¹⁸⁹⁻¹⁹⁴ The shielding-to-shift regression from DFT-computed shifts was found to have a slope of -1.05 and an offset of 345.32 ppm, with a RMSE of 0.52 ppm.

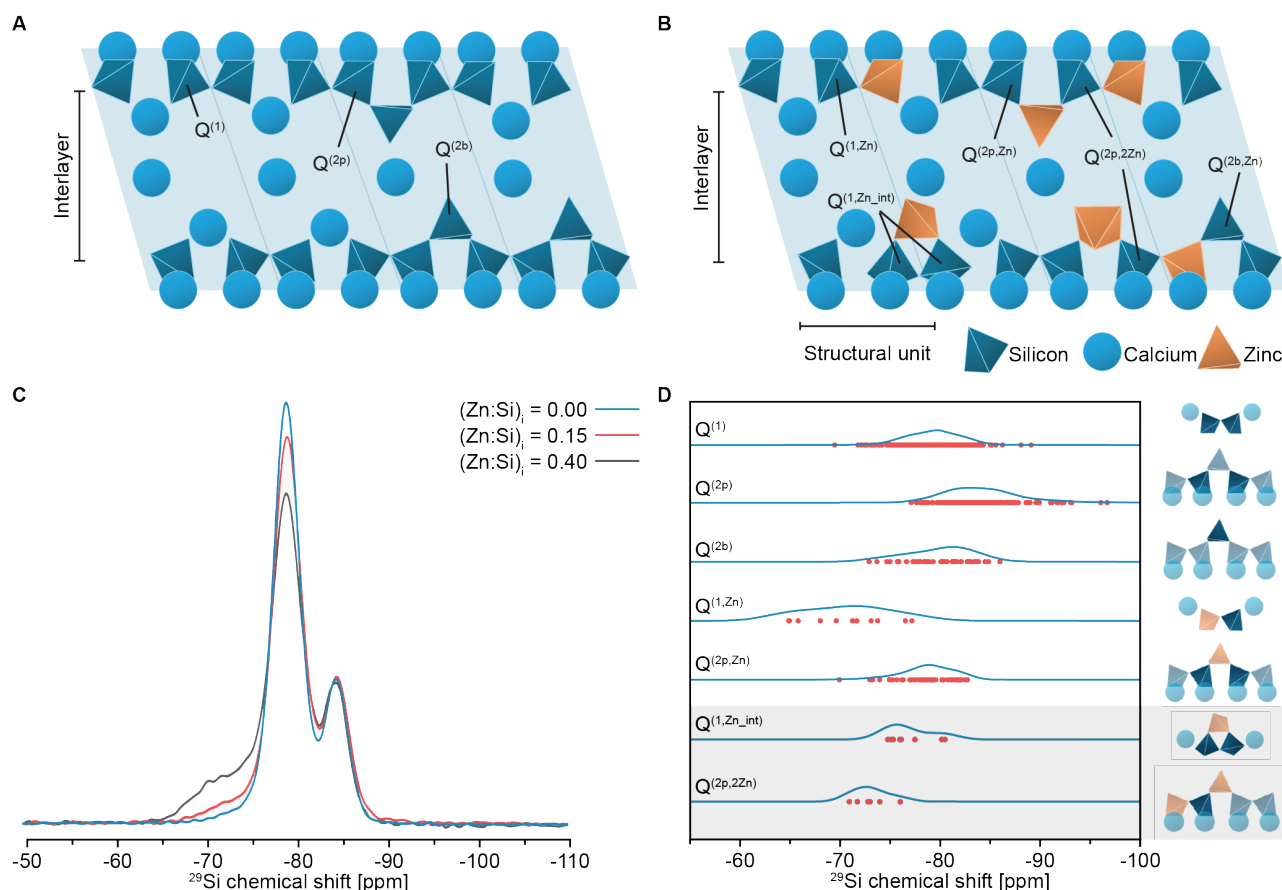


Figure 1.4. (A) Schematic of the dreierketten chains in conventional CSH showing all of the silicate species present: $Q^{(1)}$, $Q^{(2b)}$ and $Q^{(2p)}$. (B) Schematic of zinc-modified CSH showing all of the new silicate sites that could potentially be present: $Q^{(1,Zn)}$, $Q^{(2p,Zn)}$, $Q^{(2p,2Zn)}$, $Q^{(2b,Zn)}$ and $Q^{(1,Zn_int)}$. (C) DNP-enhanced experimental ^{29}Si spectra of samples with $(\text{Zn:Si})_i$ of 0.00, 0.15 and 0.40. (D) DFT-computed ^{29}Si shifts from the silicate species obtained from brick models for zinc-modified CSH and their respective schematic structures.

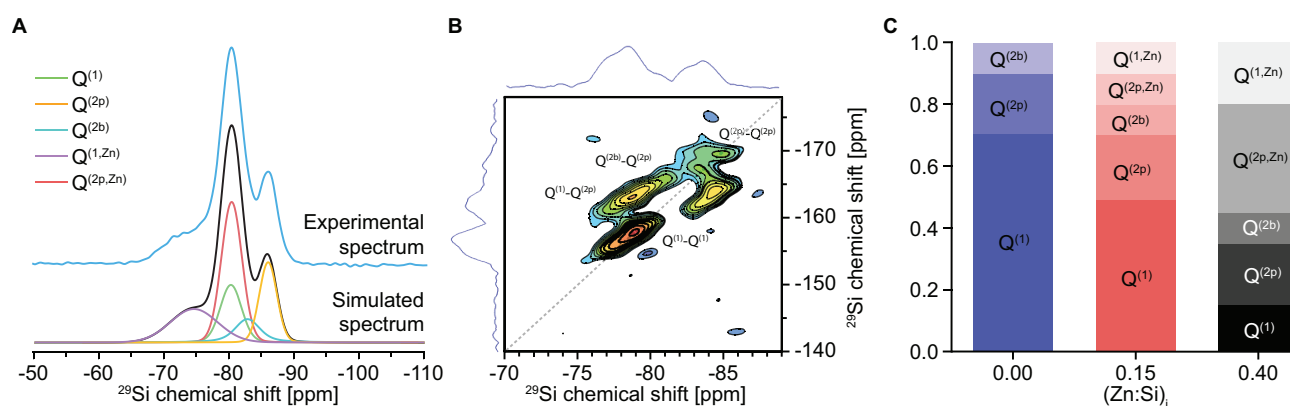


Figure 1.5. (A) 1D multi-CPMAS spectrum of the sample with $(\text{Zn:Si})_i$ of 0.40 (top) showing deconvolution into the different Q sites (bottom). (B) 2D ^{29}Si - ^{29}Si INADEQUATE spectrum of a zinc-modified C-S-H sample with a $(\text{Zn:Si})_i$ ratio of 0.40. (C) Results of the quantitative population analysis in the three samples with $(\text{Zn:Si})_i$ of 0.00, 0.15, and 0.40.

Figure 1.4D shows the chemical shifts obtained for the different silicate species in the 98 model structures. By combining the obtained shifts with DFT energies of the different model structures and experimental spectra, we could conclude that two new silicate species are formed upon incorporation of zinc in the CSH structure: $Q^{(1,Zn)}$ and $Q^{(2p,Zn)}$, the former being consistent with the signal observed around -72 ppm, and the latter being the overall most energetically favourable structure and consistent with the signal at -78.9 ppm (**Figure 1.5A**). Other possible structures were ruled out on the basis of their high energies, or from the absence of expected correlations in the 2D ^{29}Si - ^{29}Si INADEQUATE^{195, 196} spectrum that correlates shifts from linked silicates (**Figure 1.5B**). Finally, A quantitative Q species analysis of the three samples with different $(\text{Zn}:\text{Si})_i$ ratios (**Figure 1.5C**) clearly indicates a decrease in the $Q^{(1)}$ species upon zinc incorporation, as well as an increase in $Q^{(2)}$ species, which indicates the formation of longer silicate chains with higher zinc contents, providing a rationale for the observed enhanced early-age strength of concrete formed from zinc-containing CSH.

This example shows how the structure of disordered solids can be determined by NMR. Another example of structure determination in solids disordered at the nanoscale by NMR is highlighted in the following example, where we determined the structure of a hybrid layered perovskite.¹⁹⁷ My contribution in this project was to compute chemical shifts for model structures and compare them to experimental values.

Hybrid perovskite materials display high photovoltaic performances. These systems are based on the AMX_3 composition that defines a corner-sharing crystal structure consisting of A cations (e.g., Cs^+ , methylammonium or formamidinium), as well as their mixtures, along with divalent M cations (e.g., Pb^{2+} , Sn^{2+}) and halide anions X (e.g., I^- , Br^- , Cl^-).¹⁹⁸⁻²⁰⁰ The major challenge preventing the widespread application of these materials in photovoltaic systems is their limited stability due to reactivity with oxygen and water, or ion migration under operating conditions of voltage bias and light irradiation.^{198, 201, 202} Incorporating layers of hydrophobic organic cations between the hybrid perovskite slabs to form layered two-dimensional perovskites was found to improve the stability of these materials. 2-phenylethylammonium (PEA^+ , **Figure 1.6A**) is a popular organic spacer cation for layered perovskites. Mixing this ligand with 2-(perfluorophenyl)ethylammonium (FEA^+ , **Figure 1.6A**) increases the stabilisation of the layered structure,²⁰² however an atomic-level understanding of the interactions leading to higher stability is required to establish rational structure–activity-based design strategies.

The simplest model systems of layered 2D perovskites, considered here, have a S_2PbI_4 composition ($\text{S}^+ = \text{PEA}^+$ and/or FEA^+). The samples of $(\text{PEA})_2\text{PbI}_4$, $(\text{FEA})_2\text{PbI}_4$ and $(\text{PF})_2\text{PbI}_4$ (where PF denotes a 1:1 $\text{PEA}^+:\text{FEA}^+$ mixture) were analysed by measuring the $^1\text{H} \rightarrow ^{13}\text{C}$, $^{19}\text{F} \rightarrow ^{13}\text{C}$, and ^{19}F NMR spectra of the aromatic regions of the spacer cations, as shown in **Figure 1.6**. In the mixed halide perovskite structure, a weak signal intensity corresponding to the PEA^+ carbons labelled a, b, and c observed in the $^{19}\text{F} \rightarrow ^{13}\text{C}$ cross-polarisation (CP)MAS NMR spectrum indicates atomic-scale mixing of the spacer cations, since CP transfer relies on through-space dipole-dipole interactions at the sub-nanometer length scale. However, the layered perovskites containing only a single type of spacer cation, namely $(\text{PEA})_2\text{PbI}_4$ and $(\text{FEA})_2\text{PbI}_4$, exhibit very similar spectra to the samples with mixed spacers. These observations can be explained by nanoscale segregation due to self-recognition or “narcissistic” self-sorting, which would result in the local environments remaining similar to the individual spacer structures, while still affording the atomic-level contact observed by $^{19}\text{F} \rightarrow ^{13}\text{C}$ CP.

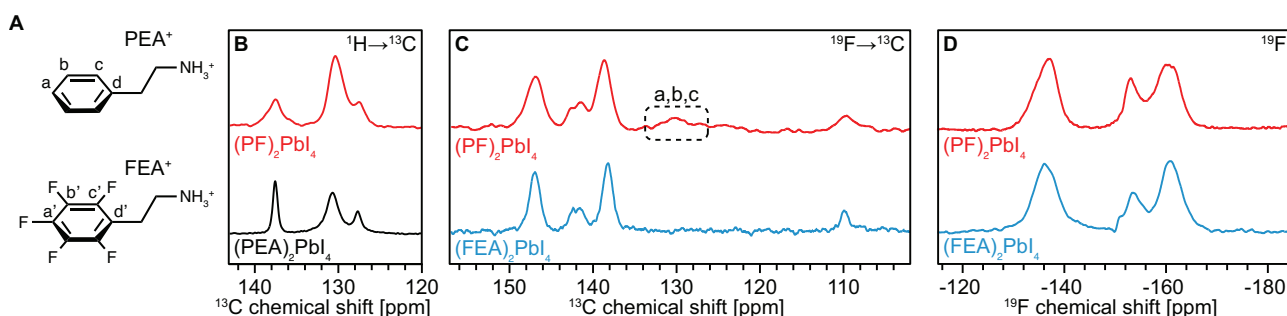


Figure 1.6. (A) Structure of PEA^+ and FEA^+ cations with the corresponding ^{13}C and ^{19}F sites labelled. (B) $^1\text{H} \rightarrow ^{13}\text{C}$ CP, (C) $^{19}\text{F} \rightarrow ^{13}\text{C}$ CP, and (D) direct ^{19}F MAS NMR spectra of the layered hybrid perovskites. PF = 1:1 $\text{PEA}^+:\text{FEA}^+$.

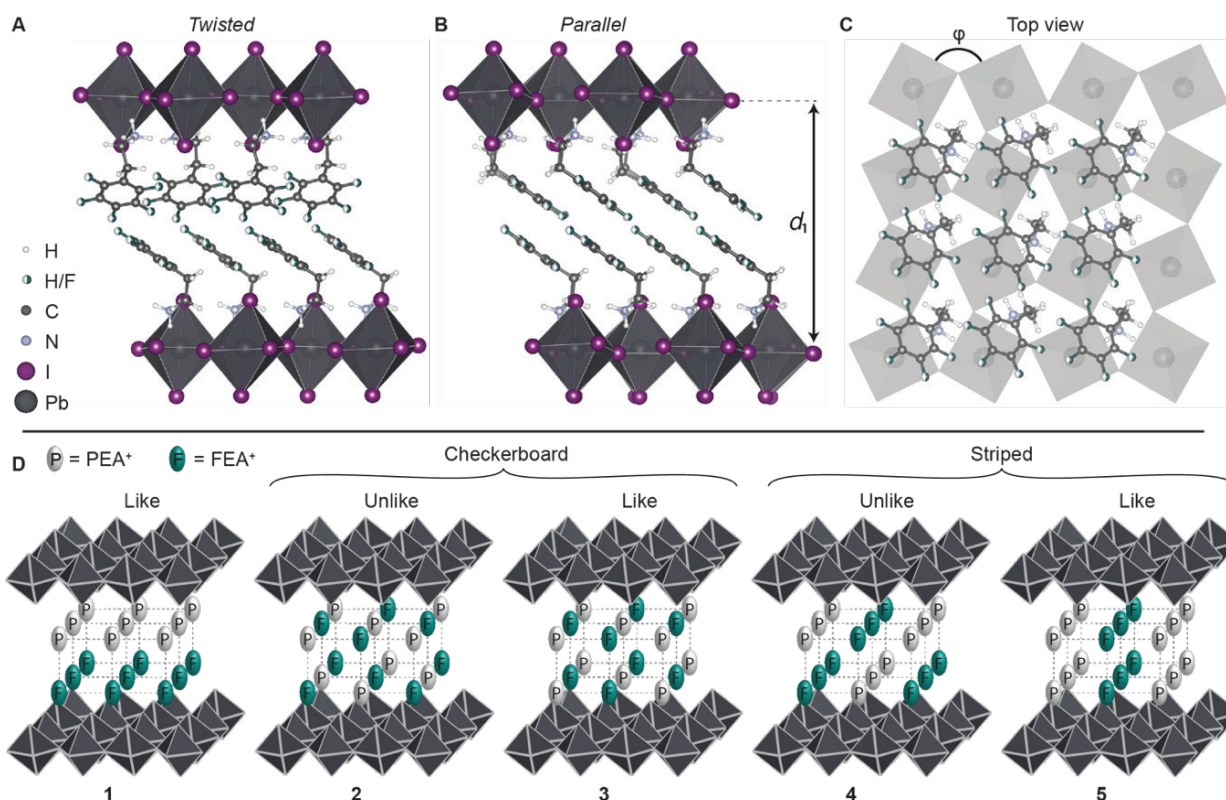


Figure 1.7. (A, B) Twisted and parallel relative orientations of the aromatic rings in adjacent layers and (C) top view of the spacer layer. (D) Schematic of different possible arrangements of PEA⁺ (P) and FEA⁺ (F) moieties ion the two opposite lattices representing the spacer bilayer within the layered perovskite.

To investigate the atomic-level structure of the layered hybrid perovskite, we performed GIPAW DFT chemical shift computations of different trial structures. The shieldings obtained were converted to chemical shifts using an external set of structures with known chemical shifts and structure,²⁰³⁻²⁰⁶ for which we computed DFT shifts to obtain the regression parameters. The trial structures were generated by selecting low-energy structures from MD simulations, followed by structure optimisation by DFT. For this analysis, only the aromatic carbons in the spacer cations were considered because the aliphatic carbons are close to the heavy Pb and I atoms and may require full relativistic treatment to obtain accurate shieldings.²⁰⁷⁻²⁰⁹

Structures with two different relative orientations of the spacer cation aromatic rings were considered: the “*twisted*” structure (Figure 1.7A), with a twist between the aromatic rings in the two opposing layers, and the “*parallel*” structure (Figure 1.7B), with aromatic rings from opposite layers aligned in parallel planes at 180° between the layers. For (PEA)₂PbI₄, the experimental ¹³C shifts agree with the calculated shifts for the *twisted* structure better than for the *parallel* structure, in agreement with the previously reported single crystal structure. In contrast, for (FEA)₂PbI₄, the calculated ¹³C and ¹⁹F shifts for the *parallel* structure are in better agreement with experiment, in accordance with the fact that the DFT energy is lower for the *parallel* structure.

Five possible arrangements of PEA⁺ and FEA⁺ spacers were investigated, as shown in Figure 1.7D. In addition to these five model structures, we also considered the possibility of phase segregation, where the shifts are computed for the separate pure *twisted* (PEA)₂PbI₄ and *parallel* (FEA)₂PbI₄ structures. Such structures would form as a result of predominantly narcissistic self-sorting. Figure 1.8 shows the comparison between experimental and computed ¹³C and ¹⁹F chemical shifts for the five mixed (PF)₂PbI₄ structures (1-5) and the phase segregated model. Considering both ¹³C and ¹⁹F chemical shifts, only the phase segregated model is in agreement with the experimental data. We therefore conclude that the layered hybrid perovskite structure formed by mixed PEA⁺ and FEA⁺ spacers comprises segregated domains of the two spacer moieties; however, since the PEA⁺ ¹³C signals are observed in the ¹⁹F→¹³C CP spectrum (Figure 1.6C), the domains must be limited to the nanoscale.

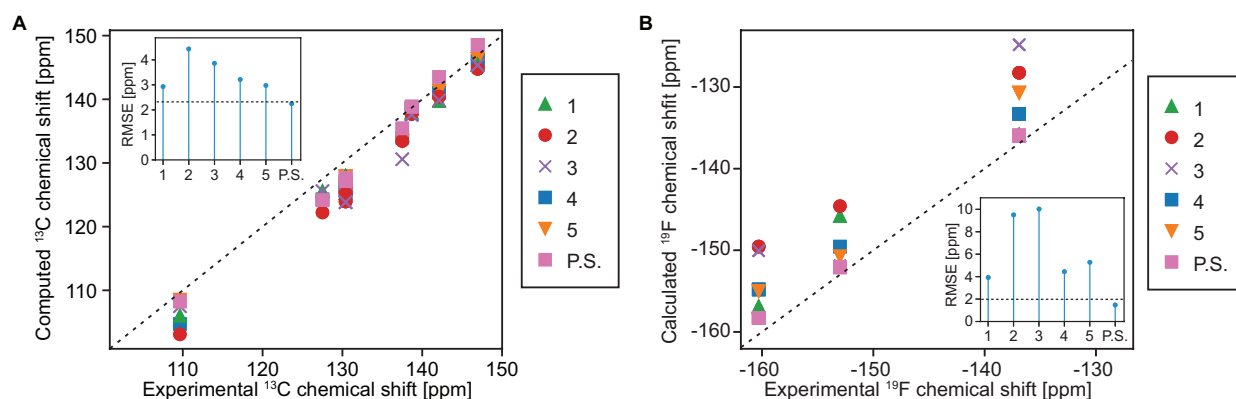


Figure 1.8. Comparison of experimental and computed (A) ^{13}C and (B) ^{19}F chemical shifts for structures 1-5 and the phase segregated model (P.S.). The dashed diagonal lines indicate perfect agreement. The insets show the RMSE between the computed and experimental chemical shifts for each structure and the horizontal lines show the expected error in DFT calculated shifts given by the RMSE in the reference set of structures.

Taken together, these three examples illustrate how chemical shifts can be used to determine *de novo* atomic-level structures, and highlight the ability of NMR crystallography to determine the structure of a wide range of crystalline and disordered materials.

1.3 Machine learning in NMR

Over the last decades, machine learning (ML) has tremendously improved many areas of science and technology. In particular, many ML models have been developed to replace resource-intensive quantum chemical computations,²¹⁰⁻²¹⁶ or improve their accuracy at a negligible additional cost,²¹⁷⁻²²⁰ in addition to other applications in chemistry and material sciences.²²¹⁻²²⁵ ML encompasses several different types of models and methods, including kernel ridge regression,²²⁶ support vector machines,²²⁷ decision trees,^{228, 229} gradient boosting,^{230, 231} and artificial neural networks.^{232, 233}

ML models statistically learn the relationship between inputs and outputs. For atomic properties, outputs are set to the desired property to predict, and the inputs should be simple to compute and capture the desired property. Thus, to calculate atomic properties, the inputs are typically atom-centered representations of local atomic environments such as the smooth overlap of atomic position (SOAP)²³⁴ or spectrum of London and Axilrod-Teller-Muto (SLATM)²³⁵ representations of local atomic environments. In particular, the SOAP descriptor expresses local atomic densities (3D Gaussian functions centered on the atomic positions) around an atom on a basis formed by spherical harmonics and radial functions. This description is invariant to translations, rotations and symmetry operations, and is particularly useful to predict properties that depend on local atomic environments and display similar invariance to such transformations, such as the (isotropic) chemical shift.

In order to train machine learning models of atomic properties, a large dataset of input structures with associated target properties must be available. While experimental databases of sufficient size are sometimes available, consistency is paramount in order to reduce the noise in the training data. Thus, ML models have also been trained on DFT-computed properties, which are generally more consistent than experimental data, and can be obtained for a large number of systems using high performance computing facilities.

1.3.1 Machine learning chemical shifts

Computing chemical shifts using DFT requires significant memory and CPU time. This prevents the use of NMR crystallography for both large systems and for large-scale screening of candidate structures. To overcome this challenge, numerous ML models of chemical shifts have been developed for small molecules in solution,²³⁶⁻²⁴⁵ proteins,²⁴⁶⁻²⁵³ and solids.²⁵⁴⁻²⁶²

Early examples of chemical shift prediction of solvated molecules without explicit quantum mechanical computation were performed by separating chemical shifts into additive contributions from atoms in their covalent environment. Dailey and Shoolery²⁶³ established a relationship between C-H proton chemical shifts of methylene groups and substituent electronegativity. Similarly, Paul and Grant²⁶⁴ expressed ^{13}C chemical shifts of linear alkanes as a sum of contributions from substituents. Such relationships were then extended to construct tables of chemical shifts for various chemical groups depending on their substituents.²⁶⁵ Chemical shifts have also been obtained from “hierarchically ordered spherical description of environment” (HOSE) codes,²⁶⁶ which encode the bonding structure around atoms in a molecule. The desired shift is obtained by averaging experimental shifts of atoms from a database that have the same HOSE code. Further development of chemical shift prediction systems led to the implementation of machine learning methods such as decision trees, support vector machines, and neural networks.²³⁶⁻²⁴³ In addition, more accurate chemical shifts can be obtained by combining DFT chemical shifts with a ML correction to reproduce the shifts obtained at higher levels of theory while reducing the computational cost.^{244, 245}

The chemical shifts in proteins heavily depend on their primary, and secondary structure.²⁶⁷ This has fuelled several methods to obtain shifts directly from the structure of proteins, either through machine learning or by constructing empirical relationships between chemical shifts and backbone torsional angles.²⁴⁶⁻²⁵³ These methods have relied on large experimental databases of assigned solution-state chemical shifts in proteins with known structures.²⁶⁸⁻²⁷⁰

In contrast, predicting chemical shifts directly from the structure of solid compounds is hindered by the lack of large databases of experimental chemical shifts. To circumvent this issue, databases of DFT-computed chemical shifts were constructed to train machine learning models of shifts for a variety of solids including silicates,^{254, 255} clay minerals,²⁵⁶ zeolites,²⁵⁷ aluminophosphates,²⁵⁸ and molecular solids.²⁵⁹⁻²⁶²

In 2019, Paruzzo *et al.*^{176, 261} introduced ShiftML, a kernel ridge regression model of chemical shifts for molecular solids containing C, H, N, O and S atoms. The model is trained on GIPAW DFT shifts computed for 3,546 diverse crystal structures. Predictions of isotropic shifts are performed based on the atom-centered SOAP representation,²³⁴ and yields an RMSE of 0.48 ppm for ^1H , 4.13 ppm for ^{13}C , 13.70 ppm for ^{15}N and 17.05 ppm for ^{17}O shifts against DFT on a test set of 500 crystal structures. Importantly, the accuracy of the model was found to be sufficient to correctly identify the crystal structure of molecular solids among sets of candidates for multiple molecular solids by comparison with experimental shifts. Overall, ShiftML reduces the computational cost of obtaining chemical shifts from hours/days to seconds for small and medium-size crystals, and allows predictions on large systems for which shift computations would be unfeasible using DFT with the currently available resources.

1.3.2 Machine learning for the analysis of NMR data

Several deep learning models have also been implemented to process NMR spectra. For example, convolutional neural network (CNN) architectures, originally developed for computer vision,²⁷¹⁻²⁷⁵ are particularly well suited to process spectral data. This has fuelled several applications of deep learning for denoising low signal-to-noise spectra,²⁷⁶⁻²⁷⁸ performing deconvolution and pick peaking,²⁷⁹⁻²⁸² applying virtual decoupling,^{283, 284} and reconstructing two-dimensional spectra from undersampled data.^{283, 285, 286} CNNs process inputs by sequentially applying non-linear convolutional filters, allowing the recognition of local features in the input data, while incorporating location invariance. This architecture was designed to approximate the inner workings of the visual cortex in mammals.^{275, 287-289}

Another class of neural networks that has been used to process NMR data is recurrent neural networks (RNNs),²⁷⁵ which are designed to process sequential data. Among particular RNN architectures, the long short-term memory (LSTM) neural network architecture²⁹⁰ has been shown to be able to reconstruct two-dimensional spectra from undersampled data by processing time-domain NMR data,²⁹¹ and to accelerate shimming algorithms that aim at correcting magnetic field inhomogeneities that impair spectral resolution.²⁹² Recurrent neural networks process input sequences one element at a time, and contain a “state vector” that encodes a memory of the past elements. In particular, RNNs have been widely used for language processing tasks.^{275, 293, 294}

Due to the large amounts of data necessary to train deep learning models, most applications presented above have required the generation of synthetic spectra for training. This relies on theoretical descriptions of the observed signals, and particular care should be taken to generate synthetic data that encompass the expected breadth of experimental variations, arising both from the expected diversity of systems studied and from experimental noise and errors, while producing realistic inputs. Coupling accurate theoretical models and realistic sources of noise and artifacts allows the generation of virtually infinite synthetic databases for training deep learning models that can then be used to process experimental data.^{295, 296}

1.4 Outline of the present thesis

In this chapter, I have presented how chemical shifts are probes of local atomic environments and how they can be obtained for model systems using *ab initio* methods, enabling the combined use of solid-state NMR and density functional theory computations to determine the atomic-level structure of materials. This provides an alternative to diffraction-based methods. In addition, I have presented selected examples of structure determination of molecular and disordered solids through NMR crystallography, and I have briefly discussed machine learning approaches to computing chemical shifts, enabling large-scale screening and investigations of large systems, as well as recent uses of deep learning for NMR data processing. During my PhD, I have focused on the further development and applications of machine learning models of chemical shifts to accelerate the structure determination of microcrystalline molecular solids, as well as amorphous materials, through NMR crystallography.

Chapter 2 focuses on the development and applications of machine learning to accelerate chemical shift-based NMR crystallography of crystalline molecular solids. We present an updated version of ShiftML that improves the accuracy and extends the capabilities of the model to predict chemical shifts. In addition, we show how experimental and machine learned chemical shifts can be incorporated in CSP procedures to drive the generation of candidate crystal structures towards the experimentally observed structures. Finally, we introduce a method to identify intermolecular interactions in crystal structures directly from experimental shifts and without any prior knowledge of the three-dimensional structure of the molecule, using a database of crystal structures with ShiftML-computed chemical shifts.

Chapter 3 discusses computational methods to accurately obtain and assign chemical shifts of molecular solids. We introduce a Bayesian framework to automatically assign chemical shifts to atomic sites of solid compounds without any prior knowledge of the three-dimensional structure of the molecule, and in a probabilistic manner. We also introduce a deep learning model to obtain pure isotropic proton solid-state NMR spectra, i.e., the spectra that would be obtained at infinite MAS rates, from datasets of experimental spectra acquired at different MAS rates. We apply the model to one-dimensional ^1H NMR spectra of various molecular solids, as well as two-dimensional ^1H - ^1H correlation spectra.

Chapter 4 presents the application of NMR crystallography to determine the structure of amorphous molecular solids. By combining solid-state NMR experiments with molecular dynamics simulations for which we predict chemical shifts using ShiftML, we determine the hydrogen bonding structure of the amorphous form of a drug molecule. We then introduce a general method to determine the structure of amorphous molecular solids through the combination of solid-state NMR, molecular dynamics and machine-learned chemical shifts.

Chapter 5 summarises the results achieved and provides an outlook of future development and applications of machine learning in NMR crystallography.

Chapter 2 Accelerating NMR crystallography of microcrystalline solids

2.1 Introduction

Atomic-level structure determination of molecular solids is a critical step in the rationalisation of their physical properties.^{297, 298} This is for example particularly important for pharmaceutical compounds, where the three-dimensional structure determines key properties of drugs delivered in either crystalline or amorphous form, such as solubility and bioavailability.²⁹⁸⁻³⁰⁰ While X-ray diffraction (XRD) is the most well-established method for determining the structure of crystalline compounds, many materials lack the long-range order required to perform single-crystal XRD. Solid-state nuclear magnetic resonance (NMR) directly probes local atomic environments, and so does not require long-range order, making it a popular method for studying the structure of microcrystalline and disordered solids from powder samples.

However, crystal structure determination by NMR is still a challenging process, in part due to the large space of candidate crystal structures to explore and the cost of computing chemical shifts for these structures using DFT. Accelerating the computation of DFT chemical shifts and incorporating experimental constraints to generate more accurate candidate structures would thus significantly accelerate NMR crystallography.

A key step in NMR crystallography is the computation of chemical shifts for candidate structures. Here, high accuracy is required in order to capture the effect of the particular conformation and packing of the molecular building blocks on the chemical shifts, and to allow the identification of the correct structure among a set of potential candidates based on a comparison between computed and measured chemical shifts.^{127, 176, 301-303} With the current best calculations, the root-mean-square error (RMSE) between experiment and calculation can be as low as 1.5 ppm for ^{13}C and 0.2 ppm for ^1H .^{127, 151, 304-306}

Plane-wave density functional theory (DFT) methods using the gauge including projected augmented wave (GIPAW) formalism^{117, 118, 307} generally offer a good trade-off between accuracy and computational cost for computing chemical shifts in small periodic structures. Consequently, DFT has been widely used in NMR crystallography to determine the structure of powdered solids.^{52, 53, 151, 308} However, the computational cost of DFT methods severely limits the size of systems accessible, preventing the study of large or disordered systems.

In recent years machine learning models have proven a powerful tool for supplementing and bypassing intensive quantum-mechanical calculations of molecular and atomic properties. In particular, NMR chemical shifts have been modelled using kernel methods^{259, 309, 310} and neural networks.^{237, 245, 247, 253, 260, 262, 311} Such approaches have proven able to yield chemical shifts to within DFT accuracy at a fraction of the computational cost, allowing applications to large ensembles of large systems.

In this context, ShiftML^{176, 261} is a machine learning model of chemical shifts of molecular solids trained on GIPAW DFT data for 3,546 structures from the Cambridge structural database (CSD),³¹² allowing fast and accurate predictions of chemical shifts for any molecular solid containing C, H, N, O, S atoms. However, two important limitations prevent its more widespread use. First, the model is limited to compounds containing only the five elements present in its training set. Second, ShiftML is trained only on structures that were geometry optimised using DFT, resulting in lower accuracy for predictions on finite temperature or distorted structures.

In **Section 2.2**, we extend the capabilities of ShiftML to predict chemical shifts for both finite temperature structures and more chemically diverse compounds, while retaining the same speed and accuracy. For a benchmark set of 13 molecular solids, we find a root-mean-squared error of 0.47 ppm with respect to experiment for ^1H shift predictions (compared to 0.35 ppm for explicit GIPAW DFT calculations using the PBE density functional), while reducing the computational cost by over four orders of magnitude.

Established approaches to *de novo* structure determination, for example by single-crystal X-ray diffraction of large molecules or by solution NMR, usually involve an iterative process where a (often random) starting structure is optimised under the combined effect of an (usually empirical) energetic potential and a penalty term that compares the computed observables with the measured values at every step of the optimisation.⁵⁷ This is a very powerful approach to finding the correct structure, and is enabled by the fact that the calculation of observables from any trial structure is very rapid. So far, this has not been possible in chemical shift-based NMR crystallography, with a few notable exceptions where chemical shifts were incorporated and derived from parametrised force-fields.^{313, 314} To make this approach general the calculation of chemical shifts so far would have required the highly accurate but very time consuming electronic structure calculations described above.^{117, 315-318} This results in *de novo* structure determination currently requiring first the generation of a large ensemble of credible candidate structures, usually done with some form of computational crystal structure prediction (CSP) protocol,^{159, 319-323} followed by DFT chemical shift calculations for the set of candidates, and only at the end of this process is there a comparison with the experimental shifts to determine which is the correct structure. This is the approach used in the example cases discussed in **Chapter 1**. While powerful, this is a time consuming and laborious approach whose efficiency could be greatly improved by making use of chemical shift data at an earlier stage of the process. Additionally, if the set of candidates does not contain the correct structure, then the whole process fails.

In **Section 2.3**, we successfully determine the crystal structures of ampicillin, piroxicam, cocaine, and two polymorphs of the drug molecule AZD8329 using on-the-fly generated machine-learned isotropic chemical shifts to directly guide a Monte Carlo-based structure determination process starting from a random gas-phase conformation.

In crystalline molecular solids, preferential interactions have previously been identified using full interaction maps (FIMs),³²⁴ where the propensity for interactions between pairs of functional groups are probed based on statistics extracted from the Cambridge structural database (CSD).³¹² This allows the identification of potential intermolecular interactions in crystalline materials, which can qualitatively inform on the intermolecular packing and be used to evaluate the relative stability of different polymorphic forms. While FIMs are useful to predict preferred non-covalent interactions in molecular solids, their usefulness in the validation of potential crystal structures based on experimental data is limited. The construction of such maps driven by experimental properties could thus help validate potential candidates in crystal structure determination and establish experimental constraints to drive candidate structure generation schemes.

In **Section 2.4**, we use a database of crystal structures with associated chemical shifts to construct three-dimensional interaction maps in molecular crystals directly derived from a molecular structure and the associated set of experimentally measured chemical shifts. We show how the maps obtained can be used to identify structural constraints for accelerating CSP protocols, and to evaluate the likelihood of candidate crystal structures without requiring any chemical shift computation.

Combining the approaches presented in this chapter could in the longer term greatly accelerate the structure determination of molecular solids, streamlining NMR crystallography and allowing a more widespread use of this method to confidently and rapidly obtain crystal structures from NMR data.

2.2 ShiftML2: A machine learning model of chemical shifts for chemically and structurally diverse molecular solids

This section has been adapted with permission from: Cordova, M.; Engel, E. A.; Stefaniuk, A.; Paruzzo, F.; Hofstetter, A.; Ceriotti, M.; Emsley, L., A Machine Learning Model of Chemical Shifts for Chemically and Structurally Diverse Molecular Solids. *Journal of Physical Chemistry C* **2022**, 126 (39), 16710-16720. (post-print)

My contribution was to select the data used to train the model from the complete dataset and to identify outliers, as well as to optimise and train the model, and test it against DFT-computed and experimental chemical shifts. I also wrote the manuscript, with the contribution of all other authors.

2.2.1 Introduction

As mentioned above, ShiftML^{176, 261} is a machine learning model of chemical shifts trained on GIPAW DFT data for 3,546 structures from the Cambridge structural database (CSD),³¹² allowing fast and accurate predictions of chemical shifts for any molecular solid containing C, H, N, O, S atoms. Although it constitutes a powerful method for computing chemical shifts with high accuracy and at a low computational cost, two important limitations prevent its more widespread use. First, the model is currently limited to compounds containing only C, H, N, O, S atoms. While these elements are among the most prevalent in the CSD, numerous organic crystals contain elements outside of this set, leaving them beyond the scope of ShiftML. Second, the training set of ShiftML only contains structures that were geometry optimised using DFT, resulting in lower accuracy for predictions on finite temperature or distorted structures, or for structures that are geometry optimised using other methods (such as semi-empirical electronic structure calculations^{325, 326}).

Here, we present ShiftML2, an updated version of ShiftML, trained on GIPAW DFT chemical shifts for an extended set of over 14,000 structures containing any of 12 common elements (H, C, N, O, S, F, P, Cl, Na, Ca, Mg and K), and composed of roughly equal amounts of relaxed and thermally perturbed structures of crystals extracted from the CSD. ShiftML2 shows slight improvements over the previous versions of ShiftML on DFT-relaxed structures (¹H RMSE of 0.47 ppm against 0.51 ppm for the ShiftML model described in Ref. 176, which we refer to as ShiftML1 here). More importantly, it effectively retains this accuracy for distorted (thermalised) structures, for which the performance of ShiftML1 degrades dramatically, while additionally allowing chemical shift computations for more chemically diverse structures.

2.2.2 Methods

Configurational sampling. In order to construct suitable reference data for an accurate and robust ShiftML2 model, we first extracted all crystal structures from the CSD with unit cells containing no more than 200 atoms (for which high-throughput first-principles calculations are comparatively affordable) and including H and C, but no additional elements other than N, O, S, F, P, Cl, Na, Ca, Mg and K. We note that we initially allowed the presence of Br and I atoms, but later discarded the structures containing these atoms due to the need for relativistic corrections to obtain accurate shieldings for atoms in their vicinity. After extracting a random selection of 1,000 molecular crystals as a test set, the selection of the training set was performed by farthest point sampling (FPS)³²⁷ of the remaining 140,373 structures based on the kernel-induced pairwise distances

$$D(X_i, X_j) = k(X_i, X_i) + k(X_j, X_j) - 2k(X_i, X_j). \quad (2.1)$$

Here, the kernel function $k(\cdot, \cdot) = (X_i \cdot X_j)^2$ measures the similarity of the average smooth overlap of atomic positions (SOAP) power spectra²³⁴ of the constituent atoms within a crystal structure, X_i , computed using the hyperparameters specified in **Table 2.3**. The first 10,000 FPS-sorted (most structurally diverse) structures were selected as the training set.

All training and test structures were relaxed using DFT fixed cell geometry optimisations using the Quantum ESPRESSO (QE) electronic structure package^{328, 329} with the PBE density functional,⁹⁷ a Grimme D2 dispersion correction,^{330, 331} wavefunction and charge density energy cut-offs of 60 Ry and 240 Ry, respectively, and ultrasoft pseudopotentials with GIPAW reconstruction.^{332, 333} To render this computation efficient, only the Gamma-point was accounted for. Further details may be found in **Appendix I**.

Subsequently, short constant-volume molecular dynamics (MD) simulations of 500 fs were performed using i-PI^{334, 335} to drive the dynamics, and the above QE setup to evaluate energies and forces. We used a timestep of 1 fs and a Generalised Langevin Equation thermostat^{336, 337} to equilibrate the system at 300 K.

Finally, we collected two structures for each molecular crystal in the training and test sets, the relaxed structure and a thermalised MD structure (the last in the trajectory), and proceeded to compute the associated GIPAW-DFT chemical shieldings for all 22,000 resulting structures.

GIPAW-DFT chemical shieldings. The GIPAW NMR calculations were performed using the QE code with the same DFT parameters as for the structure relaxation above, but using refined plane wave and charge density energy cut-offs of 100 Ry and 400 Ry, respectively, a Monkhorst-Pack k-point grid³³⁸ with a maximum spacing of 0.06 Å⁻¹, and the ultrasoft pseudopotentials with GIPAW reconstruction from the USSP pseudopotential database v1.0.0.

Finally, all structures were discarded which displayed at least one outlier shift (defined as being outside the range of chemical shifts between the 1st and 99th percentile of all shifts of that element by at least 1.5 times that range), or where the calculation failed. Overall, 2,650 structures were discarded because the self-consistent loop did not reach the high level of convergence needed for reliable GIPAW calculations, we removed 3,313 additional structures containing Br or I atoms, and we discarded 24 structures that displayed outlier shieldings. This led to final training and test sets containing 14,254 and 1,759 structures respectively.

Machine learning model. We use kernel ridge regression (KRR)²²⁶ to predict the isotropic chemical shielding of an atom based on its local atomic environment as

$$\sigma(X) = \sum_i^N w_i k(X, X_i) = \sum_i^N w_i (X^T \cdot X_i)^\zeta, \quad (2.2)$$

where X and X_i are symmetry-adapted descriptors, which encode the local atomic environment around the atom of interest and those in the training set, respectively, and w_i denotes the regression weight associated with training sample i . $k(\cdot, \cdot)$ is the kernel function that defines the similarity between two atomic environments. Here, we measure the similarity between two environments as the scalar product between the vectors corresponding to their descriptor, raised to a power ζ . Training a KRR model involves determining the weights w_i such that **Equation 2.2** is best satisfied for the training data, with an additional regularisation term that reduces the magnitude of regression weights. Further information is available in **Appendix I**.

Uncertainty estimation. Uncertainty estimation is performed using a resampling approach to generate a committee of $M = 32$ KRR models,³³⁹ trained on random two-fold splits of the training data. The final prediction for a sample i in the test set, $\hat{\sigma}_i$, is given by the mean of the prediction for each model, and the estimated uncertainty is defined as the standard deviation s_i of the prediction of each model, rescaled by a factor α given by³³⁹

$$\alpha = -\frac{1}{M} + \frac{M-3}{M-1} \sqrt{\frac{1}{N_{\text{test}}} \sum_{i \in \text{test}} \frac{(\sigma_i - \hat{\sigma}_i)^2}{s_i^2}}, \quad (2.3)$$

where N_{test} is the size of the test set, and the sum runs over all test samples.

Local atomic environment descriptor. We describe local atomic environments using smooth overlap of atomic positions (SOAP) power-spectra²³⁴ as implemented in librascal.³⁴⁰ We use a sparse implementation of the SOAP descriptors, making use of the sparsity of elements in local environments around individual atoms.

The relevant hyperparameters were optimised by five-fold cross validation performed on the ¹H environments of a subset of 1,000 training structures, selected at random other than including all training structures containing Na, Ca, Mg or K. The latter ensures that these elements are represented during hyperparameter optimisation, despite their low abundance in the training data. The structures selected for hyperparameter optimisation contain a total of 27,802 ¹H environments. In each cross-validation fold, the training data was partitioned into three equal parts, and a KRR model was trained on each part. This was done in order to reduce the computational resources required to train the models for each split. The selection of descriptor parameter values was based on the RMSE obtained on the validation data. The explored and selected hyperparameter values can be found in **Appendix I**. We note that Ref. 176 found almost identical hyperparameters to be optimal for H, C, N, O, and S through independent optimisations for the different elements. We therefore apply the hyperparameters optimised for ¹H to the other elements without further optimisation, except for the optimal radial basis,³⁴¹ which was constructed individually with the complete final training data for each element. We note that the cutoff radius chosen here is well above the Van der Waals radius of all atoms considered (< 2.8 Å).

Farthest point sampling of training environments. The training data was sorted using FPS³²⁷ based on distances between pairs of environments X_i and X_j defined as in **Equation 2.1**. This serves two purposes: first, it permits the removal of duplicate environments arising from, e.g., equivalent atomic sites related by the crystal symmetries in relaxed structures. Second, it identifies the most structurally diverse set of training environments.

To eliminate redundant environments and distil a computationally manageable number of informative environments, we split the training data into randomly selected batches of 50,000 samples (atomic environments) (because FPS is not computationally feasible on the whole set). FPS was then used on each batch and stopped once the minimum distance between FPS-selected samples reached 10^{-2} for ^1H and 10^{-3} for all other elements. The FPS selection was then repeated after shuffling the environments, recombining them into different batches of 50,000 samples and increasing the distance threshold in each batch by steps of 10^{-3} , until a total of fewer than 100,000 environments remained.

Outlier detection and model training. When required, the FPS-selected training environments were randomly selected to a maximum of 2^{16} samples in order to limit the size of the kernel required to predict chemical shifts. Then, five-fold cross-validation was performed. For each fold, a committee of eight KRR models was trained. To this end the training split was further subsampled, training each KRR model on a random selection of half of the training split for a given fold. For each fold, the predictions and associated uncertainty estimates for the validation split were used to identify and discard outlier environments. In practice, environments were discarded if the residual error exceeded both the standard deviation of the shifts in the training data and twice the associated uncertainty estimate. After removing these outliers, 32 KRR models were trained on randomly selected environments making up half of the remaining curated data to construct the final model of shifts. The rescaling factor α for uncertainty estimation was obtained from the predictions on the test set. A summary of the number of structures and environments during the data selection and cleanup (FPS selection of environments, outlier removal) is given in **Figure 2.14**.

Atom type identification. The different atom types, defined here as hybridisation and formal charge, in the training and test structures were identified using the RDKit³⁴² Python package on the asymmetric unit of the crystals extracted using the CSD Python API.³¹² The structures where RDKit failed to identify bonds and/or formal charges were discarded from the atom type analysis. Carbon atoms identified as charged were set to a neutral charge, as well as nitrogen atoms identified with a negative charge and oxygen atoms identified as positively charged. This was done upon visual inspection of a subset of crystal structures displaying such unusual atom types, confirming that such atom types were incorrectly determined by the package. In total, atomic types of 6,960 out of the 10,593 final training structures and 1,443 out of the 1,759 test structures were identified. We note that this is a *post hoc* analysis of the atom types in the training and test sets, and that ShiftML2 does not require identification of the atom types to perform chemical shift predictions.

Comparison with experimental chemical shifts. To further test the resulting models, we performed plane-wave DFT calculations for 13 structures with assigned experimental chemical shifts with the same level of theory as for the computation of DFT shieldings of the training and test sets. Comparison between computed (or predicted) shieldings and experimental chemical shifts was performed by linear regression of the shieldings computed with the corresponding experimental shifts, using average values of chemically equivalent shifts and resolving any assignment ambiguity by selecting the assignment resulting in the minimum RMSE.

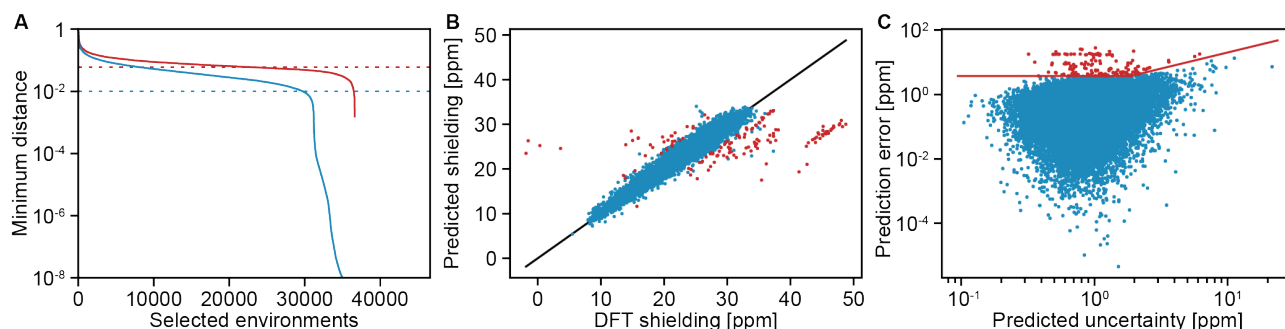


Figure 2.1. (A) First (blue) and last (red) FPS selection step for a batch of up to 50,000 ^1H environments. The blue and red dashed lines show the threshold for the minimum distance between FPS-selected samples set to select environments in the first and last FPS selection steps, respectively. (B) Comparison of DFT-computed ^1H shieldings and predictions for the training environments obtained through 5-fold cross-validation. (C) Comparison of the absolute error of the prediction and predicted uncertainty for the training environments selected by FPS. The red lines indicate the criteria used to discard outliers (red points in (B) and (C)).

2.2.3 Results and Discussion

Training set selection and model training. Due to the lack of large databases of experimental chemical shifts in molecular solids, we trained the model on shielding values computed by DFT, as was done previously for ShiftML1.^{176, 261} This ensures both consistency in the training data as well as the ability to perform high throughput computations to obtain a substantial amount of training data in reasonable time. The training structures were chosen to be as diverse as possible through FPS. Since computed shieldings are related to chemical shifts by a simple linear relationship, we use the two terms interchangeably.

High quality of the training data is key to producing an accurate machine learning model. In addition, the kernel model framework used here has a linear time and memory complexity with respect to the training set size for inference. It is thus important to reduce the amount of training data while retaining diverse atomic environments and removing outliers to obtain both fast and accurate predictions of chemical shifts. To this end, we performed an iterative, batched FPS of the chemical environments as described in **Section 2.2.2**. **Figure 2.1A** shows the first and last FPS iterations on typical batches. The significant drop in minimum distance between FPS-selected samples after selecting 30,000 of the 50,000 environments in an initial batch corresponds to symmetrically equivalent atomic sites in relaxed crystal structures. After gathering the FPS-selected environments from all batches after the final iteration, we obtained 67,535 ^1H environments.

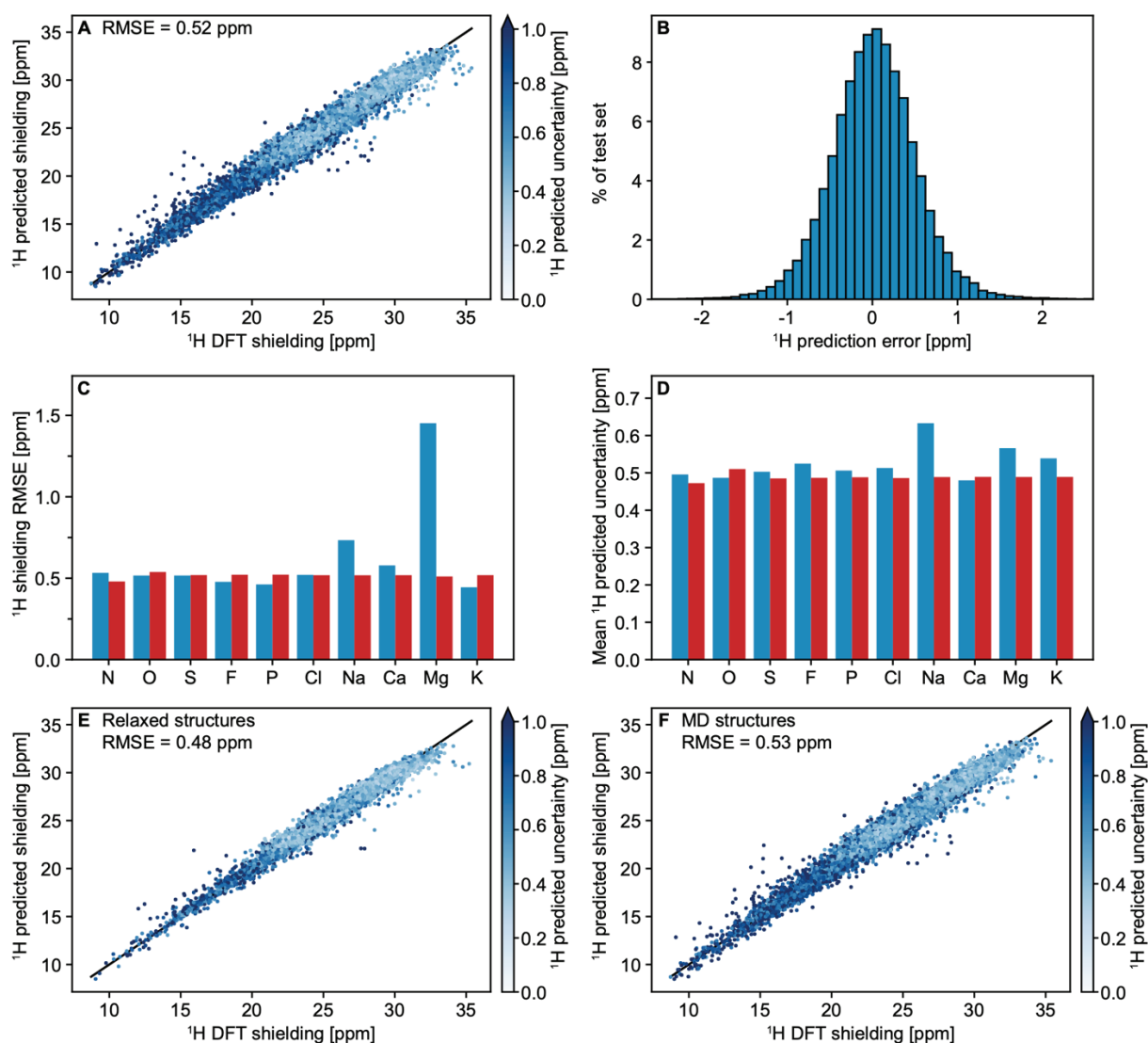


Figure 2.2. (A) Comparison of DFT-computed ^1H shieldings and ShiftML2 predictions on the test set. (B) Histogram of the of prediction error between ShiftML2 predictions and DFT-computed shieldings for ^1H environments. Comparison of ^1H (C) chemical shift RMSE and (D) average prediction uncertainties on test structures containing (blue) or lacking (red) a given element. Comparison of DFT-computed ^1H shieldings and ShiftML2 predictions on (E) relaxed and (F) MD structures in the test set. Black lines in (A), (E) and (F) show perfect correlations.

Figure 2.1B-C highlights the outliers among the selected ^1H training environments identified following the scheme described in **Section 2.2.2**. In total, 145 ^1H environments were considered as outliers because they exhibit both relatively large prediction error and comparatively small prediction uncertainty (red points and lines in **Figure 2.1C**). Among the final ^1H training environments, 86% were from distorted structures and 14% from relaxed structures. This highlights the importance of the presence of distorted structures in the training data in order to obtain a uniform sampling of the space of possible atomic environments.

The final model was constructed by training 32 models on random half splits of the remaining training environments. Prediction uncertainties were estimated as the rescaled standard deviation of the 32 predictions to fit the error distribution, as described in Ref. 343.

Model evaluation and comparison to ShiftML1. **Figure 2.2** shows correlation plots between predicted and DFT-computed ^1H shieldings in the test set as well as the associated distribution of prediction errors. We obtain an RMSE of 0.52 ppm and an R^2 coefficient of 0.97, with 95% of the predictions having an error below 1 ppm. The RMSE was found to be slightly lower in relaxed structures (0.48 ppm) compared to MD structures (0.53 ppm). The presence of sodium or magnesium in crystal structures was found to raise both the prediction error (**Figure 2.2C**) and, to a lesser extent, uncertainty (**Figure 2.2D**). We attribute that to the relatively low number of structures containing these elements in the training set (226 structures containing Na, 65 containing Mg), coupled to the high charge density of these ions which induces a large change in the shielding on neighbouring atomic sites. Although calcium and potassium are not significantly better represented in the training set (145 structures containing Ca, 176 containing K), their reduced charge densities compared to Mg and Na induce lower perturbations of the shielding of neighbouring atomic sites, which are better captured by the kernel.

We observe a reduced prediction uncertainty and error for shieldings above 20 ppm (see **Figure 2.14**). This behaviour is expected considering that 90% of the training data have DFT shieldings computed above 20 ppm, which corresponds to typical chemical shifts of aliphatic and aromatic CH protons (< 10 ppm). The reduced density of training data at lower shieldings (corresponding to higher chemical shifts) results in increased error and uncertainty of the predictions.

To compare ShiftML1 and ShiftML2 we apply both models to the ShiftML1 test set, as well as all structures from the current test set which contain exclusively H, C, N, O and S atoms (i.e., those for which ShiftML1 is applicable). **Figure 2.3** shows the ^1H shift predictions of the two models for the ShiftML1 test set (**Figure 2.3A-B**), and for the relaxed (**Figure 2.3C-D**) and finite temperature (**Figure 2.3E-F**) structures from the ShiftML2 test set, which only contain H, C, N, O and S atoms. **Table 2.1** summarises the results obtained by both models. There are two striking conclusions that are illustrated by the figure and table. First, overall, ShiftML2 displays slight improvements over ShiftML1 for relaxed structures (0.47 ppm RMSE compared to 0.49 ppm on the ShiftML1 test set, and 0.47 ppm RMSE compared to 0.51 ppm on relaxed structures from the ShiftML2 test set), indicating that the increase in the number of training environments was sufficient to avoid deterioration of the accuracy despite the greater chemical diversity. Second, ShiftML2 is substantially more accurate for finite temperature structures (0.53 ppm RMSE for ShiftML2 compared to 0.98 ppm for ShiftML1), highlighting the greater robustness of a model trained on finite temperature structures when predicting atomic properties for distorted structures. To confirm the robustness of ShiftML2 towards distorted structures, we evaluated the error against DFT-computed ^1H shieldings for up to 50 snapshots taken every 100 fs from MD simulations of the crystal structures of cocaine, AZD5718 and form 4 of AZD8329. We found that the average RMSEs along the MD trajectories were only slightly above the RMSEs obtained for the relaxed structures (0.58 ppm against 0.55 ppm RMSE for AZD5718, 0.50 ppm against 0.45 ppm RMSE for form 4 of AZD8329, and 0.49 ppm against 0.42 ppm RMSE for cocaine, see **Figure 2.15**). This is a key improvement compared to the previous ShiftML version, since it allows accurate predictions of chemical shifts beyond relaxed structures, and yields a better description of shifts in (PI)MD snapshots, and for intermediate structures during structural optimisation.

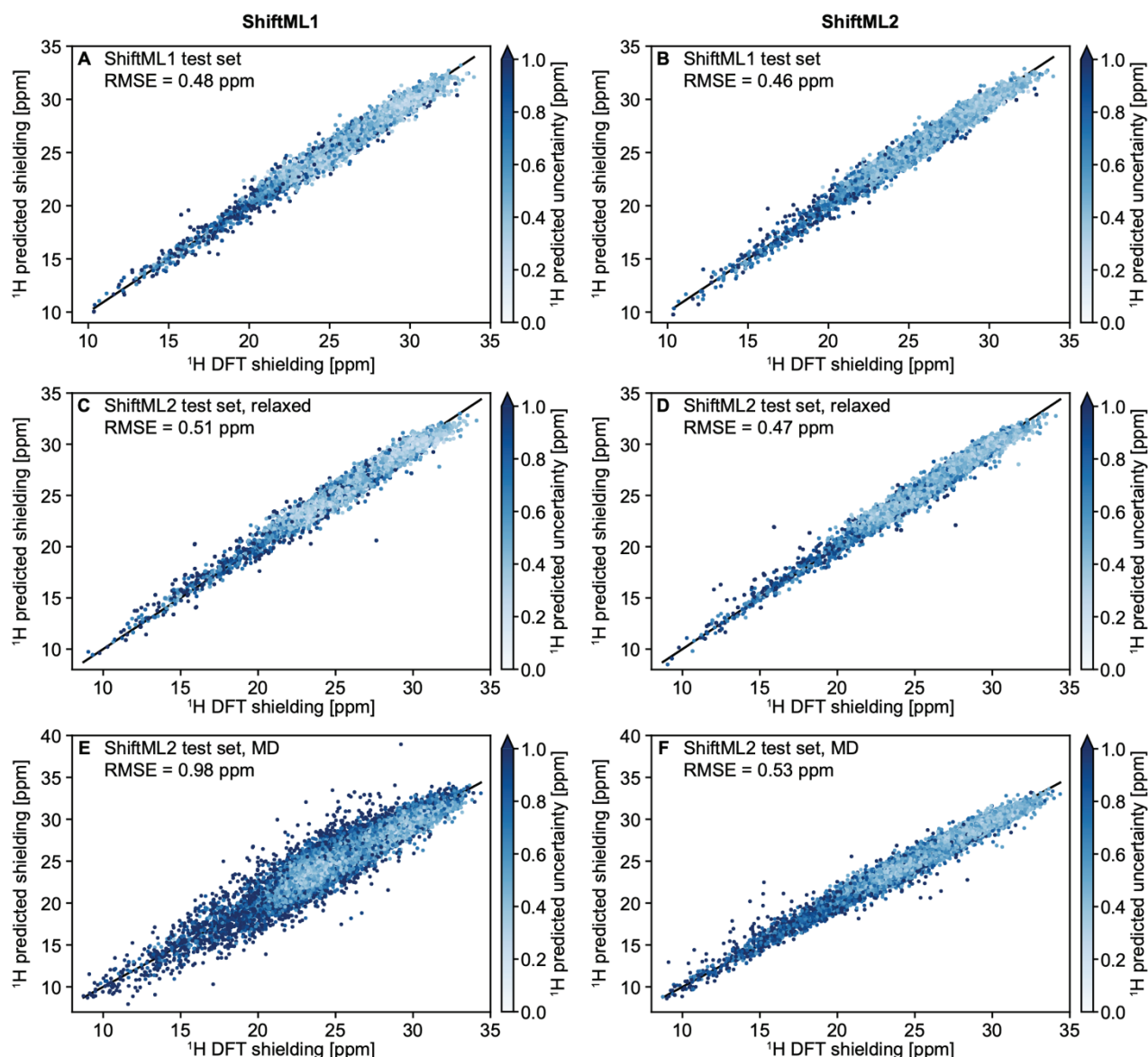


Figure 2.3. Comparison of DFT-computed ^1H shieldings and predictions using ShiftML1.1 (A, C, E) or ShiftML2 (B, D, F) on: (A, B) the ShiftML1 test set, (C, D) relaxed structures containing only H, C, N, O and S in the ShiftML2 test set, and (E, F) MD structures containing only H, C, N, O and S in the ShiftML2 test set. Black lines show perfect correlations.

Table 2.1. Chemical shift root-mean-square error (RMSE), mean absolute error (MAE) and R^2 coefficient of ShiftML1 and ShiftML2 compared to DFT-computed shieldings. The values are given for ShiftML1 and ShiftML2, separated by a slash.

Test set	RMSE [ppm]	MAE [ppm]	R^2
ShiftML1	0.48/0.46	0.37/0.35	0.98/0.98
ShiftML2, relaxed only	0.51/0.47	0.38/0.35	0.98/0.98
ShiftML2, MD only	0.98/0.53	0.71/0.40	0.91/0.97
ShiftML2, all	0.78/0.50	0.54/0.38	0.94/0.98

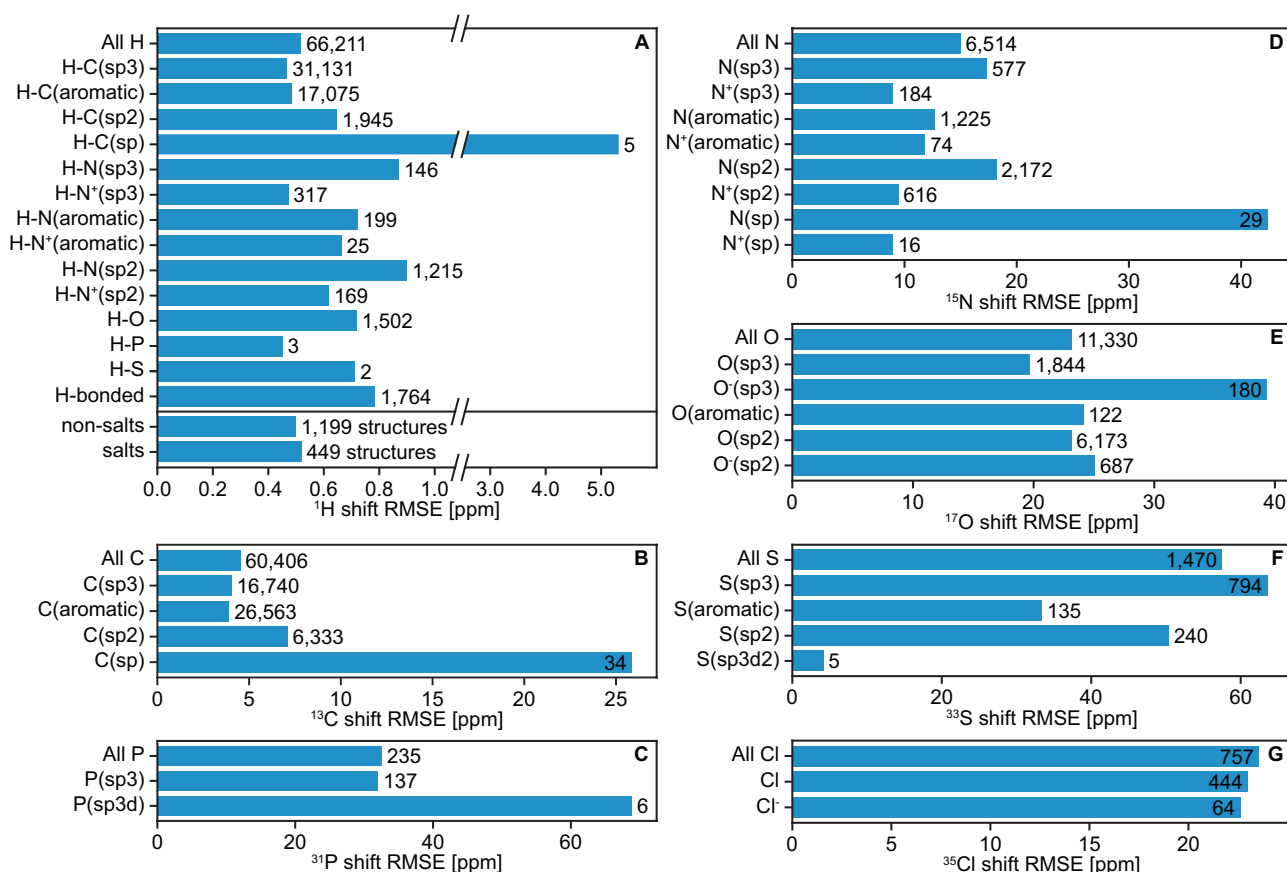


Figure 2.4. Chemical shift RMSE for different types of (A) ^1H , (B) ^{13}C , (C) ^{31}P , (D) ^{15}N , (E) ^{17}O , (F) ^{33}S and (G) ^{35}Cl in the test set. The number of environments (or structures) in the test set contributing to each bar is indicated next to it.

The ability of the model to generalise to distorted structures is key in many applications of NMR crystallography. In particular, it allows accurate computation of chemical shifts on structures that are geometry optimised with different levels of theories (e.g., force fields or DFTB), which is important for the accurate description of shifts in molecular dynamics simulations of materials.¹⁷⁷ It also enables more confident on-the-fly computations of chemical shifts of intermediate structures during the optimisation of crystal structures in chemical-shift driven structure determination protocols, resulting in a potentially more powerful driving force towards the experimental structure.³⁴⁴

Figure 2.4 shows the prediction error for different types of protons in the test set. The two most common proton types H-C(sp3) and H-C(aromatic), making up 90% of the test set, yielded chemical shift RMSEs below 0.5 ppm. All other proton types displayed chemical shift RMSEs below 0.9 ppm, with the exception of alkyne protons, for which the RMSE was found to be 5.3 ppm. Such high error is explained by the presence of only two alkyne protons identified in the final training data. Interestingly, we find that protons attached to nitrogens in charged groups display a lower error compared to their neutral counterparts. Molecular salts were found to display comparable shift RMSEs to neutral compounds. H-bonded protons yielded a chemical shift RMSE of 0.79 ppm.

Experimental benchmark set and polymorphs. We evaluate the accuracy of the model with respect to experimental ^1H chemical shifts using a benchmark set of 13 molecular crystals made up of cocaine, form 4 of AZD8329, theophylline, uracil, naproxen, the co-crystal of 3,5-dimethylimidazole and 4,5-dimethylimidazole, AZD5718, furosemide, flutamide, the co-crystal of indomethacin and nicotinamide, flufenamic acid, the potassium salt of penicillin G, and phenylphosphonic acid.^{52, 139, 144, 177, 261, 345, 346} **Figure 2.5A** compares the predicted and experimentally measured shifts for this set. We obtain a RMSE of 0.47 ppm, compared to 0.35 ppm using DFT. For reference, the RMSE obtained on the experimental benchmark set for ShiftML1 (containing the six first molecular solids mentioned previously) is 0.41 ppm for ShiftML2, compared to 0.39 ppm for ShiftML1 and 0.36 for DFT. This further highlights that the accuracy of ShiftML1 for relaxed structures has been retained by ShiftML2, while extending the capabilities of the model to predict shifts for more chemically and structurally diverse structures. Notably, within the limits of the small experimental set used here, the accuracy against experimental shifts is found to decrease when including structures containing F, Cl, P or K atoms, while DFT remained at the same level of accuracy. Since no such deterioration is observed for the structures in the test set (see **Figure 2.2C**), we attribute this to the chemical environments in the experimental set, which are not well represented in the training data.

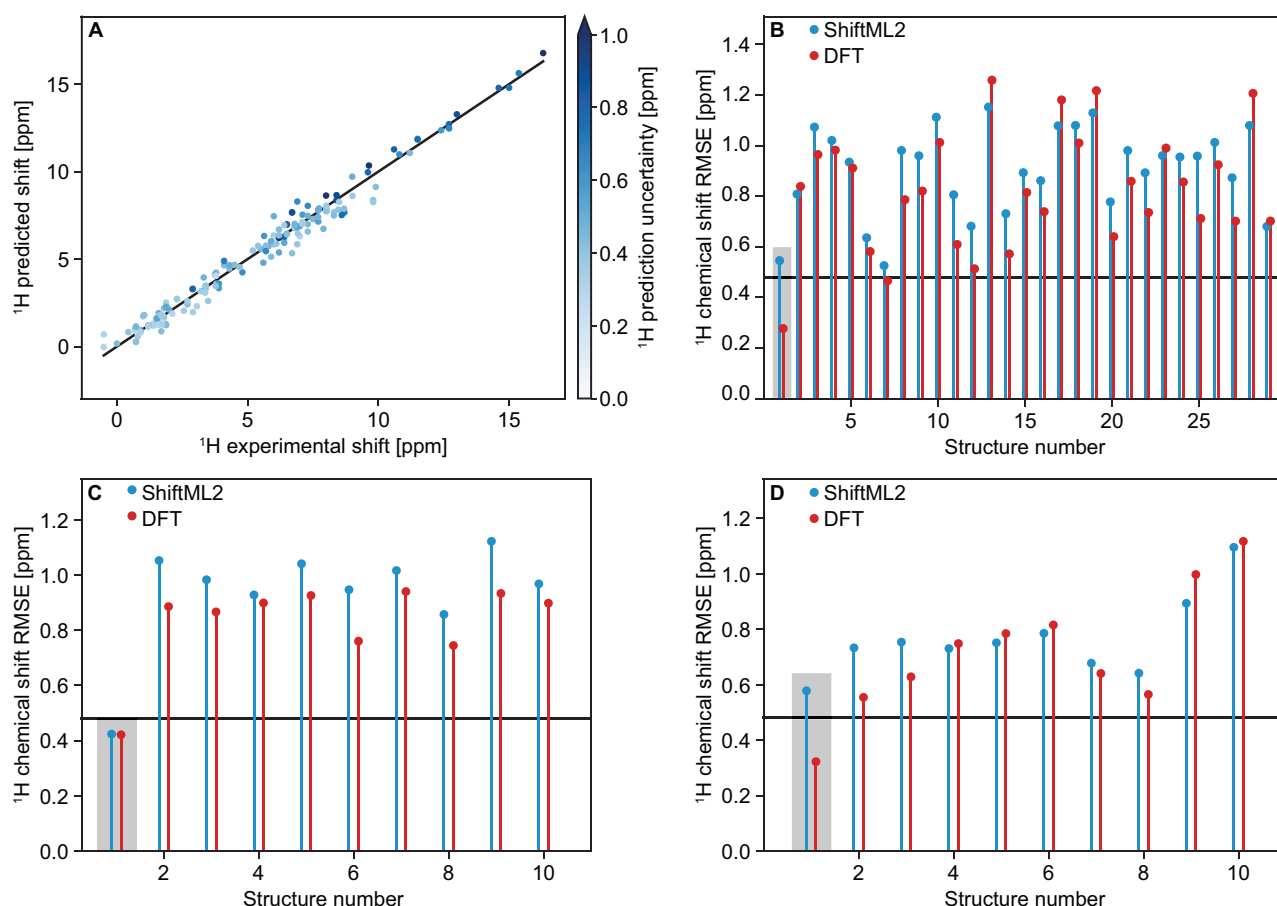


Figure 2.5. (A) Comparison between predicted and experimental ^1H shifts for 13 molecular solids. The black line shows perfect correlation. Chemical shift RMSE obtained by ShiftML2 (blue) and DFT (red) against experimental shifts for candidate structures of (B) Cocaine, (C) AZD8329 form 4, and (D) AZD5718. The correct crystal structure is indicated by the grey zone. The black horizontal lines indicate the expected RMSE between ShiftML2 predictions and experimental shifts (0.47 ppm).

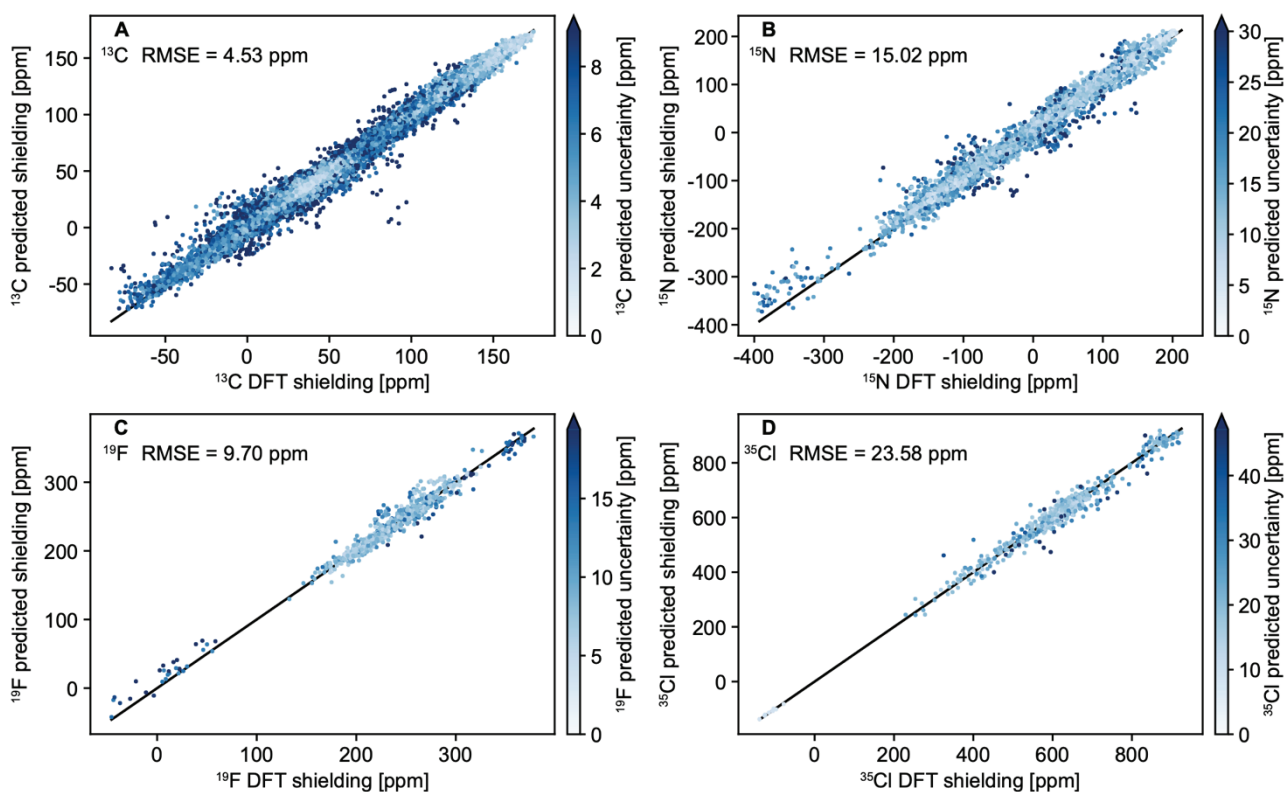
Computing DFT chemical shifts for the 13 structures required over 56 CPU days, while ShiftML2 required less than 20 CPU minutes to predict the shifts of all atoms in the structures considered. If only ^1H chemical shifts are required, this time is reduced to less than four CPU minutes, which corresponds to a more than 24,000-fold speedup compared to DFT shift computation.

The ability to determine the correct structure from among a set of candidates based on comparison between experimental and computed shifts is key to NMR crystallography. **Figure 2.5B-D** shows the RMSE between experimental and predicted ^1H shifts for different sets of candidate structures for cocaine, form 4 of AZD8329, and AZD5718. The correct candidates systematically yielded a chemical shift RMSE below 0.6 ppm, and corresponded to the lowest RMSE among the sets of candidates for form 4 of AZD8329 and AZD5718, and to the second lowest RMSE for cocaine.

Models for other nuclei. In addition to ^1H , we constructed models for all the other nuclei present in the training data. **Figure 2.6**, **Figure 2.13** and **Table 2.2** compare the resulting predictions for the nuclei beyond ^1H to GIPAW DFT shieldings. We note that although we refer to a particular nucleus (e.g., ^{15}N), the isotropic chemical shift of all NMR-active isotopes of a particular element can be predicted with the same accuracy, by adapting the offset (and slope) used to convert computed shieldings into chemical shifts. We obtain strong correlations ($R^2 > 0.95$) for ^{13}C , ^{15}N , ^{17}O , ^{19}F and ^{35}Cl . This indicates that ShiftML2 can accurately predict chemical shifts for these elements, although the absolute error is higher than for ^1H due to the larger chemical shift ranges for these nuclei (see **Table 2.2**). The lower number of training environments for ^{31}P , ^{23}Na , ^{43}Ca , ^{25}Mg and ^{39}K was found to lead to lower correlation with DFT-computed shifts. While we still provide models for these nuclei, we acknowledge that more accurate models based on more extensive training data would be required to obtain more accurate predictions for these elements. We reiterate that the main purpose of including these elements in the training data was to allow prediction of ^1H , ^{13}C or ^{15}N chemical shifts for structures containing such elements. Detailed ShiftML2 prediction accuracies for different types of ^{13}C , ^{15}N , ^{17}O , ^{31}P , ^{33}S and ^{35}Cl nuclei are shown in **Figure 2.4B-G**. As for ^1H , we observe a loss of accuracy for sp-hybridised ^{13}C and ^{15}N . The other nuclei (^{19}F , ^{23}Na , ^{43}Ca , ^{25}Mg and ^{39}K) each displayed a unique atomic type across the test set.

Table 2.2. Training and test size, chemical shift root-mean-square error (RMSE), mean absolute error (MAE) and R^2 coefficient for ShiftML2 models trained on nuclei beyond ^1H .

Nucleus	Training set size	Test set size	RMSE [ppm]	MAE [ppm]	R^2
^{13}C	65,498	60,406	4.53	3.12	0.99
^{15}N	65,506	6,514	15.02	9.99	0.98
^{17}O	65,488	11,330	23.18	16.21	0.98
^{19}F	23,958	865	9.70	6.85	0.97
^{33}S	18,509	1,470	57.53	35.12	0.87
^{31}P	5,337	235	32.61	17.64	0.70
^{35}Cl	15,780	757	23.58	17.02	0.97
^{23}Na	728	14	5.77	4.58	0.57
^{43}Ca	386	8	13.01	10.77	0.99
^{25}Mg	186	10	12.27	8.21	0.94
^{39}K	632	9	9.33	7.07	0.39

**Figure 2.6.** Comparison of DFT-computed and predicted (A) ^{13}C , (B) ^{15}N , (C) ^{19}F and (D) ^{35}Cl chemical shifts in the test set. The black lines show perfect correlation.

2.2.4 Conclusion

We have presented a machine learning model of chemical shifts that improves on the previously published model^{176, 261} in two key ways. First, the chemical diversity covered by the model has been extended from 5 to 12 elements, meaning that shifts for a much larger space of compounds can now be accessed. Second, finite temperature structures have been included in the training data, allowing reliable chemical shift predictions for distorted structures.

Compared to GIPAW DFT, we obtain R^2 correlation coefficients above 0.95 for ^1H , ^{13}C , ^{15}N , ^{17}O , ^{19}F and ^{35}Cl shifts, and a chemical shift RMSE below 0.5 ppm for ^1H . The model is able to massively accelerate the computation of shifts in molecular solids while retaining DFT level accuracy with respect to experimental shifts for ^1H (0.47 ppm RMSE). Importantly, the cases of cocaine, form 4 of AZD8329 and AZD5718 demonstrate that ShiftML2 permits fast and reliable NMR crystal structure determination for complex organic molecular crystals.

The capacity to calculate shifts for distorted structures is important for two reasons. First it allows reliable shifts to be calculated for structures that are not geometry optimised using DFT, such as structures optimised using (semi-)empirical approaches such as DFTB, and for structures from molecular dynamics simulations. Second, it means that shifts calculated for structures generated in a simulated annealing structure determination protocol³⁴⁴ will be accurate even when the trial structure is not in an energy minimum, potentially providing a much more efficient driving force towards the correct structures, and this will be the subject of future studies. The model presented here scales linearly with respect to the number of local atomic environments in a structure of interest, making shifts for large ensembles of large structures accessible. The new model will thus accelerate NMR crystallography by allowing large-scale computations for candidate structures, either from MD trajectories or in direct optimisation methods.

The models are freely available on <https://dx.doi.org/10.5281/zenodo.7097427>.

2.2.5 Appendix I

Raw data. The complete sets of training and test structures, along with the Python scripts used in this study are available from <https://dx.doi.org/10.5281/zenodo.7097427>. The model is also available via the same link. All data are made available under the CC-BY-4.0 license (Creative Commons Attribution-ShareAlike 4.0 International).

SOAP hyperparameters. We describe atomic environments using smooth overlap of atomic positions (SOAP) power spectra, which expand translational-rotational invariants of a decorated atom density using a basis of orthonormal radial basis functions (Gaussian-type orbitals) and the spherical harmonics. The resulting description depends on several hyperparameters. The cutoff radius r_c defines the distance from the central atom, beyond which any further atom density is disregarded. The cutoff smoothing width σ_c represents the distance over which the atom density is smoothed to zero. The number of radial basis functions before (n_{r0}) and after (n_r) dimensionality reduction by principal component analysis (PCA), and the maximal angular momentum number n_l define the basis for the expansion. The atomic Gaussian width σ defines the width of three-dimensional Gaussian density associated with each atomic position. To reflect the greater importance of atoms close to the central atom compared to more distant ones, the decorated atom density is radially scaled,³⁴⁷ where the rate c , the scale r_0 and the exponent m are parameters to optimise. All these hyperparameters were optimised through five-fold cross-validation. **Figure 2.7** and **Table 2.3** show the parameter values explored and the optimised hyperparameters, respectively.

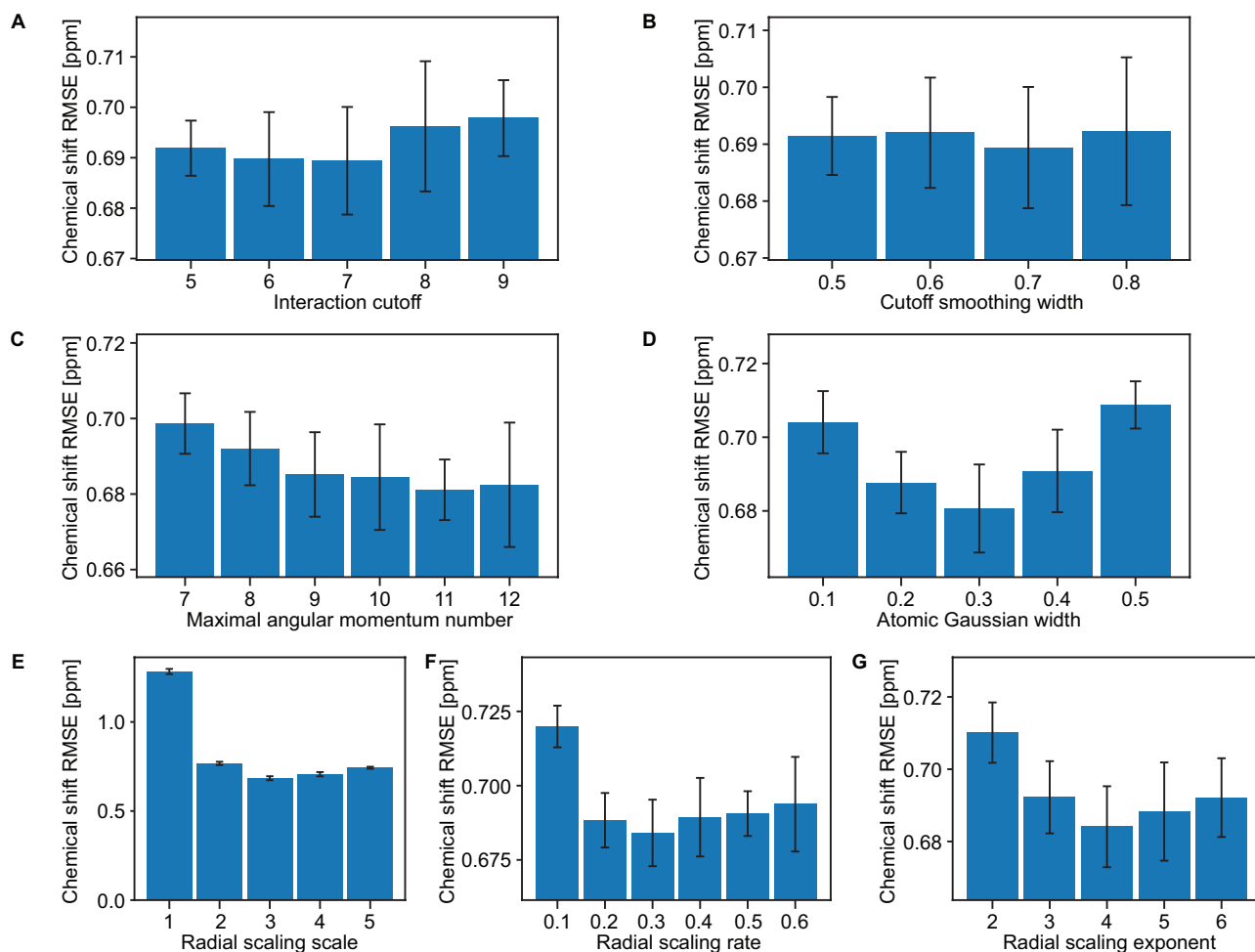


Figure 2.7. Optimisation of SOAP hyperparameters. (A) Interaction cutoff r_c , (B) cutoff smoothing width σ_c , (C) maximum angular momentum number n_l , (D) atomic Gaussian width σ , (E) radial scaling scalar r_0 , (F) radial scaling rate c , (G) radial scaling exponent m . Error bars indicate the standard deviation of the chemical shift RMSE between the five cross-validation folds.

Table 2.3. SOAP parameters used for configurational sampling and training of the model.

	r_c [Å]	σ_c [Å]	n_{r0}	n_r	n_l	σ [Å]	r_0 [Å]	c	m
Configurational sampling	4.0	0.5	9	-	4	0.4	0.0	0.0	0.0
Training	7.0	0.5	20	8	8	0.3	3.0	0.3	4.0

Pseudopotentials used for DFT computations of the training set.**Table 2.4.** Pseudopotentials used for DFT computations of the training set.

Element	Atomic mass	Pseudopotential (relaxation)	Pseudopotential (GIPAW)
H	1.0079	H.pbe-tm-new-gipaw-dc.UPF ^a	H.pbe-kjpaw_psl.1.0.0.UPF ^b
C	12.0107	C.pbe-tm-new-gipaw-dc.UPF ^a	C.pbe-n-kjpaw_psl.1.0.0.UPF ^b
N	14.0067	N.pbe-n-kjpaw_psl.1.0.0.UPF ^b	N.pbe-n-kjpaw_psl.1.0.0.UPF ^b
O	15.9994	O.pbe-n-kjpaw_psl.1.0.0.UPF ^b	O.pbe-n-kjpaw_psl.1.0.0.UPF ^b
S	32.065	S.pbe-n-kjpaw_psl.1.0.0.UPF ^b	S.pbe-n-kjpaw_psl.1.0.0.UPF ^b
F	18.998	F.pbe-n-kjpaw_psl.1.0.0.UPF ^b	F.pbe-n-kjpaw_psl.1.0.0.UPF ^b
P	30.974	P.pbe-n-kjpaw_psl.1.0.0.UPF ^b	P.pbe-n-kjpaw_psl.1.0.0.UPF ^b
Cl	35.453	Cl.pbe-n-kjpaw_psl.1.0.0.UPF ^b	Cl.pbe-n-kjpaw_psl.1.0.0.UPF ^b
Na	22.989	Na.pbe-spn-kjpaw_psl.1.0.0.UPF ^b	Na.pbe-spn-kjpaw_psl.1.0.0.UPF ^b
Ca	40.078	Ca.pbe-spn-kjpaw_psl.1.0.0.UPF ^b	Ca.pbe-spn-kjpaw_psl.1.0.0.UPF ^b
Mg	24.305	Mg.pbe-spn-kjpaw_psl.1.0.0.UPF ^b	Mg.pbe-spn-kjpaw_psl.1.0.0.UPF ^b
K	39.098	K.pbe-spn-kjpaw_psl.1.0.0.UPF ^b	K.pbe-spn-kjpaw_psl.1.0.0.UPF ^b

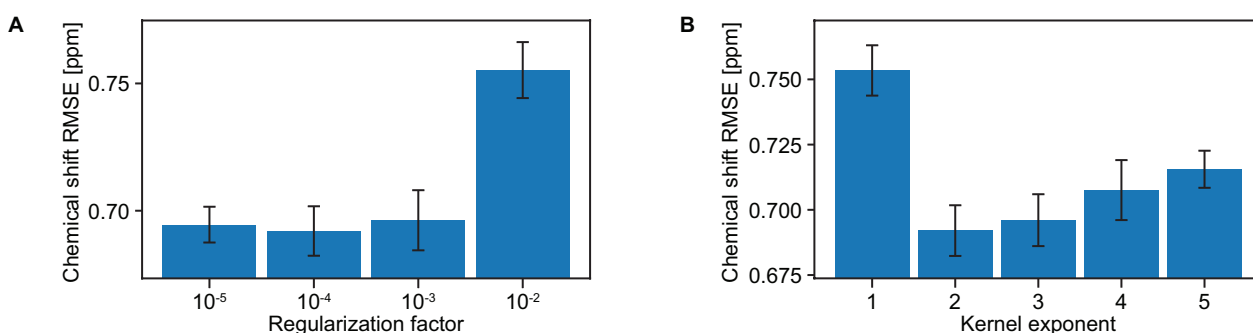
^aFrom <https://sites.google.com/site/dceresoli/pseudopotentials>. ^bFrom PSLibrary.³³²**Model training.** Predictions are performed using the following equation:

$$\sigma(X) = \sum_i^N w_i k(X, X_i) = \sum_i^N w_i (X^T \cdot X_i)^\zeta, \quad (2.4)$$

where X is the SOAP vector describing the atomic environment to compute the shift for, X_i and w_i denote the atomic environment and regression weight associated with training sample i , respectively, and $k(\cdot, \cdot)$ is the kernel function that defines the similarity between two atomic environments. In practice, training the model involves solving the equation

$$\vec{\sigma} = (K + \lambda I) \cdot W \quad (2.5)$$

for the weight matrix W given the kernel matrix between all training environments K and associated vector chemical shieldings $\vec{\sigma}$, and regularisation parameter λ . The weights were determined using least-squares regression as implemented in the Numpy Python library.³⁴⁸ Optimisation of the kernel power ζ and regularisation parameter λ are shown in **Figure 2.8**. The optimal values were found to be $\lambda = 10^{-4}$ and $\zeta = 2$.

**Figure 2.8.** Optimisation of the (A) regularisation factor and (B) kernel exponent.

ShiftML models of other nuclei.

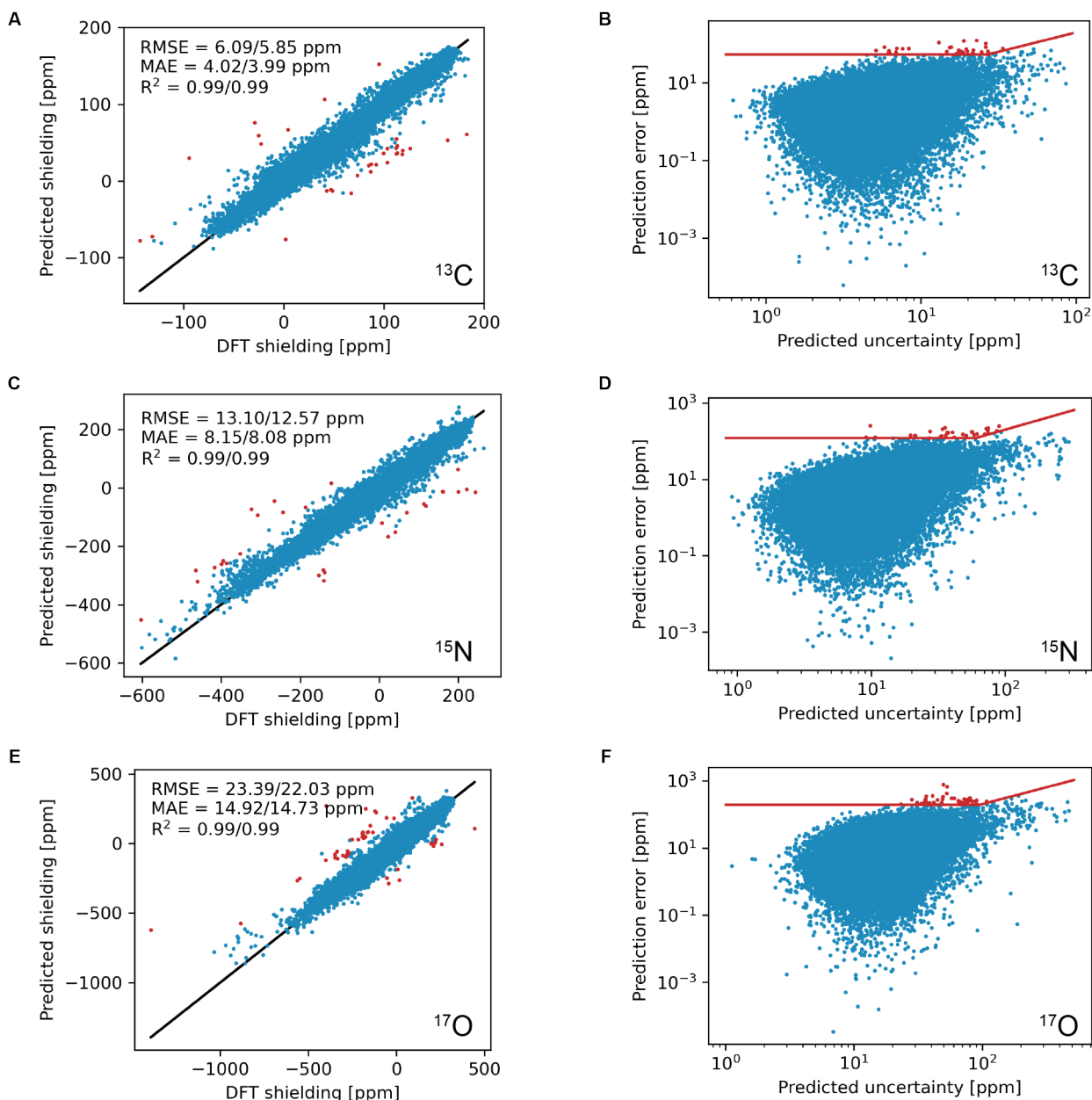


Figure 2.9. Comparison of DFT shieldings and predictions for the training environments obtained through 5-fold cross-validation for (A) ^{13}C , (C) ^{15}N , and (E) ^{17}O . Comparison of the absolute error of the prediction and predicted uncertainty for the FPS-selected training environments for (B) ^{13}C , (D) ^{15}N , and (F) ^{17}O . The red lines indicate the criteria used to discard outliers (red points).

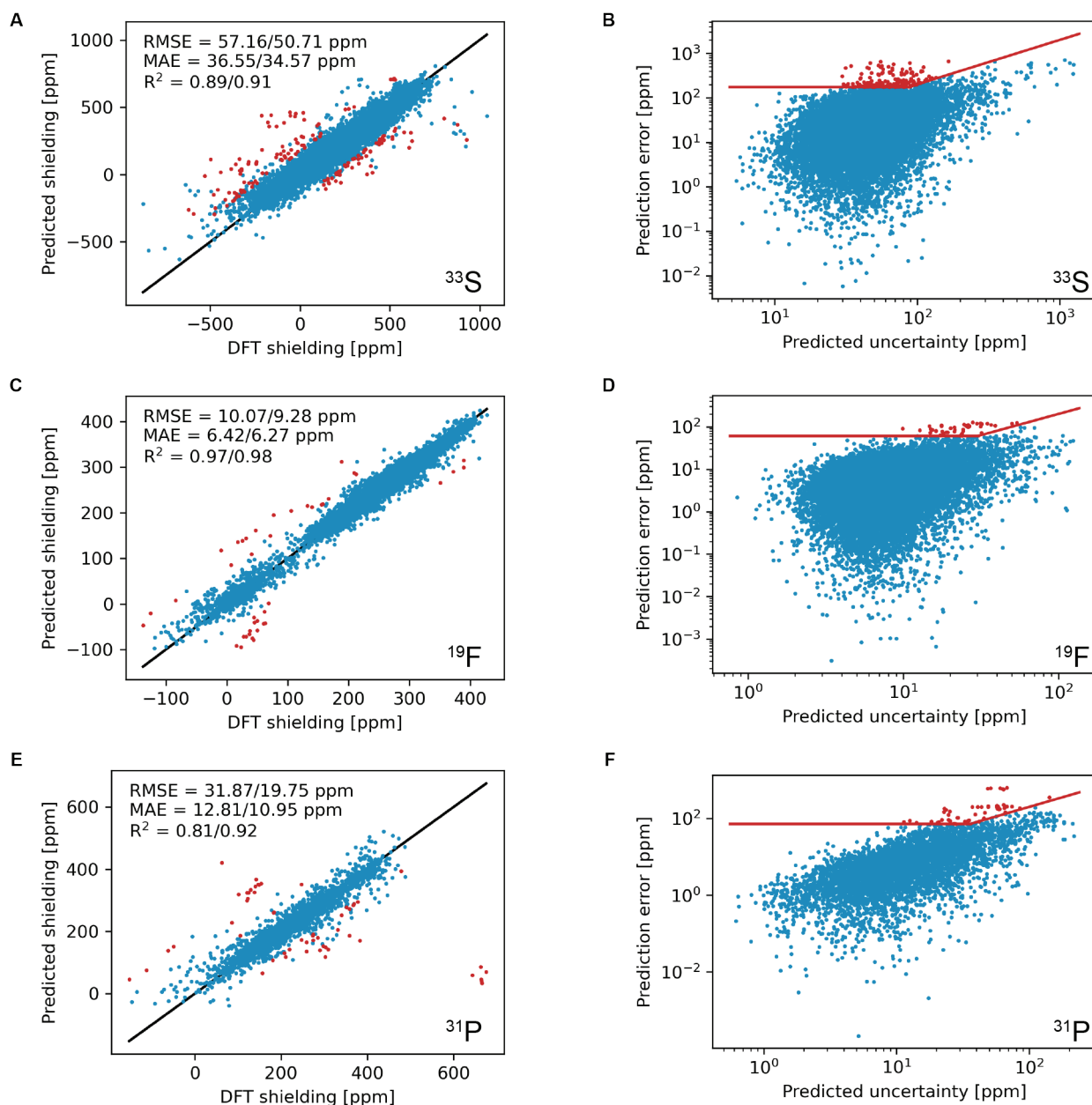


Figure 2.10. Comparison of DFT shieldings and predictions for the training environments obtained through 5-fold cross-validation for (A) ^{33}S , (C) ^{19}F , and (E) ^{31}P . Comparison of the absolute error of the prediction and predicted uncertainty for the FPS-selected training environments for (B) ^{33}S , (D) ^{19}F , and (F) ^{31}P . The red lines indicate the criteria used to discard outliers (red points).

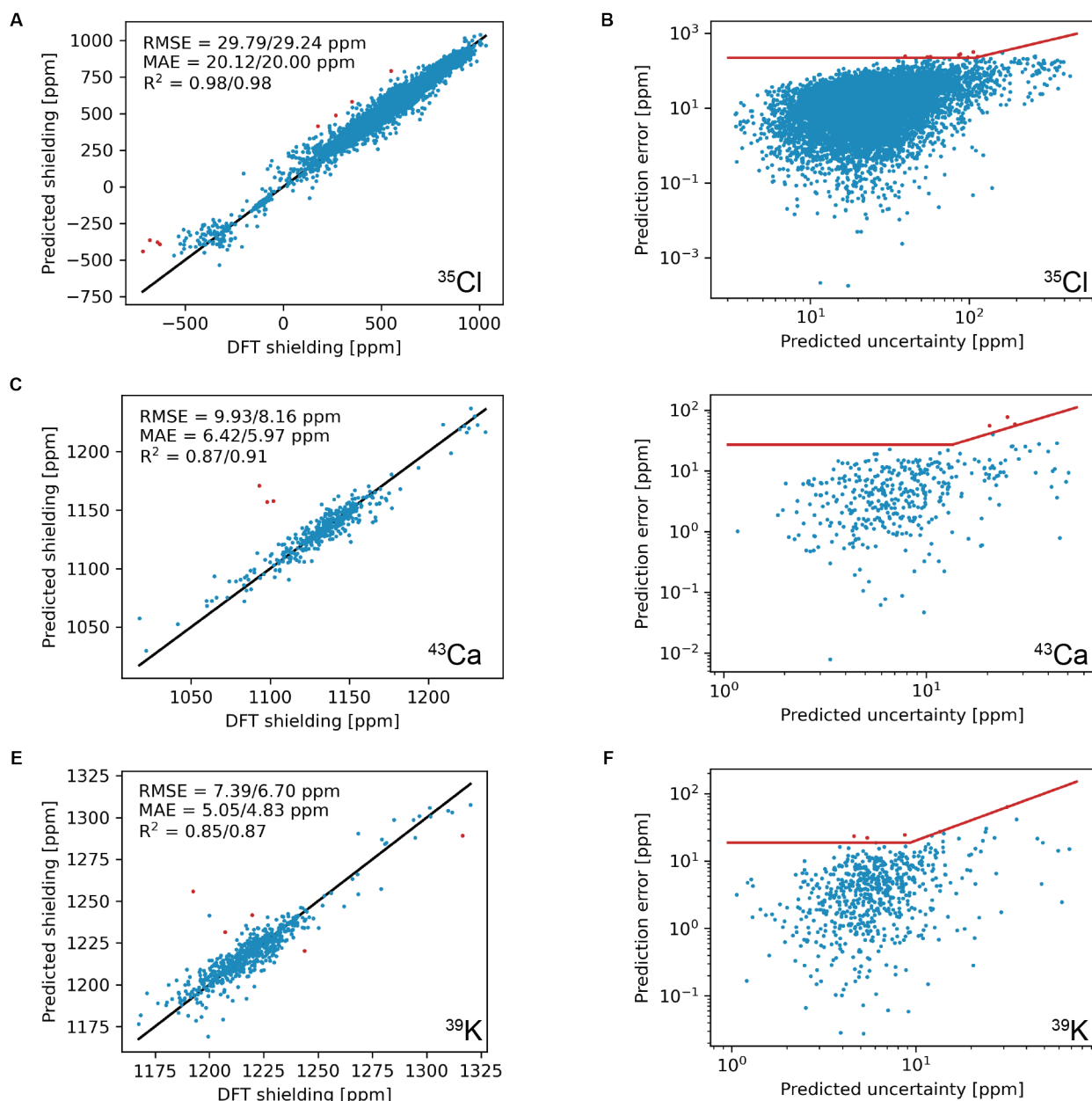


Figure 2.11. Comparison of DFT shieldings and predictions for the training environments obtained through 5-fold cross-validation for (A) ^{35}Cl , (C) ^{43}Ca , and (E) ^{39}K . Comparison of the absolute error of the prediction and predicted uncertainty for the FPS-selected training environments for (B) ^{35}Cl , (D) ^{43}Ca , and (F) ^{39}K . The red lines indicate the criteria used to discard outliers (red points).

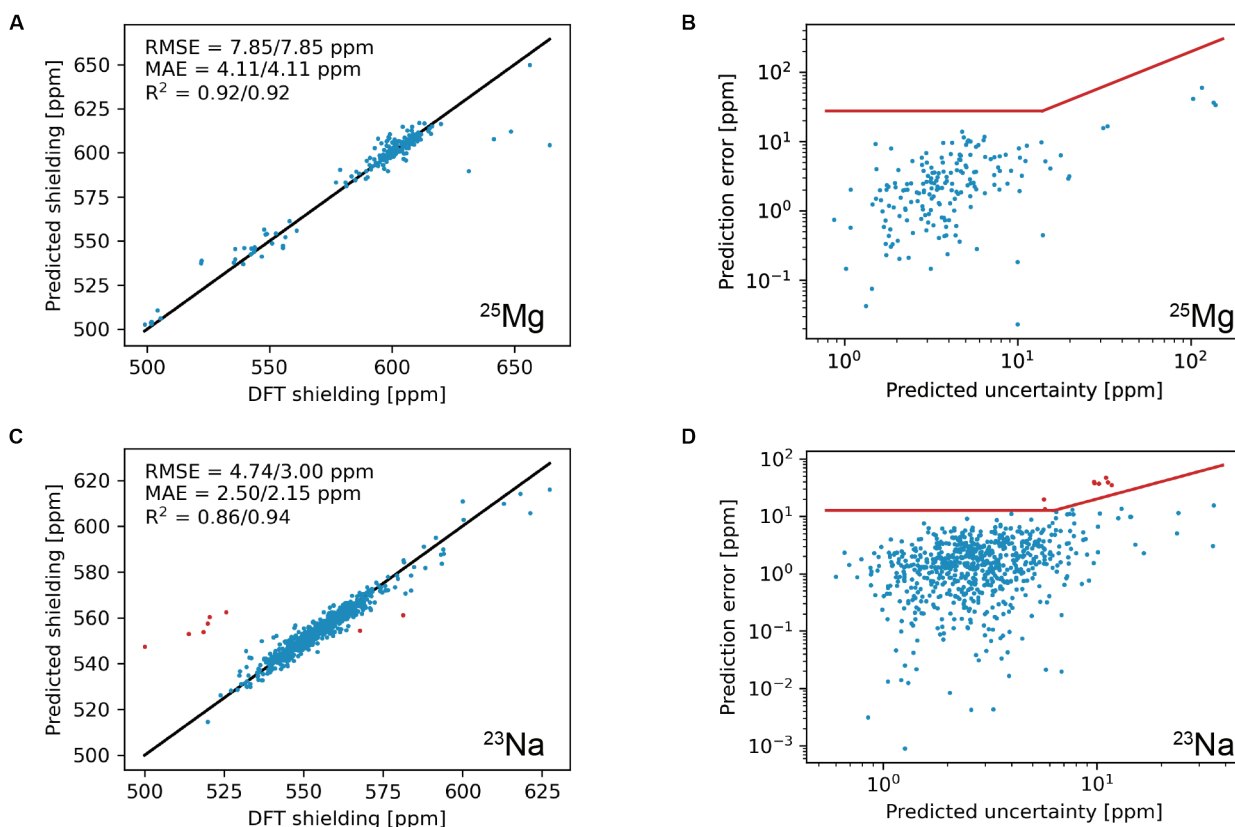


Figure 2.12. Comparison of DFT shieldings and predictions for the training environments obtained through 5-fold cross-validation for (A) ^{25}Mg , and (C) ^{23}Na . Comparison of the absolute error of the prediction and predicted uncertainty for the FPS-selected training environments for (B) ^{25}Mg , and (D) ^{23}Na . The red lines indicate the criteria used to discard outliers (red points).

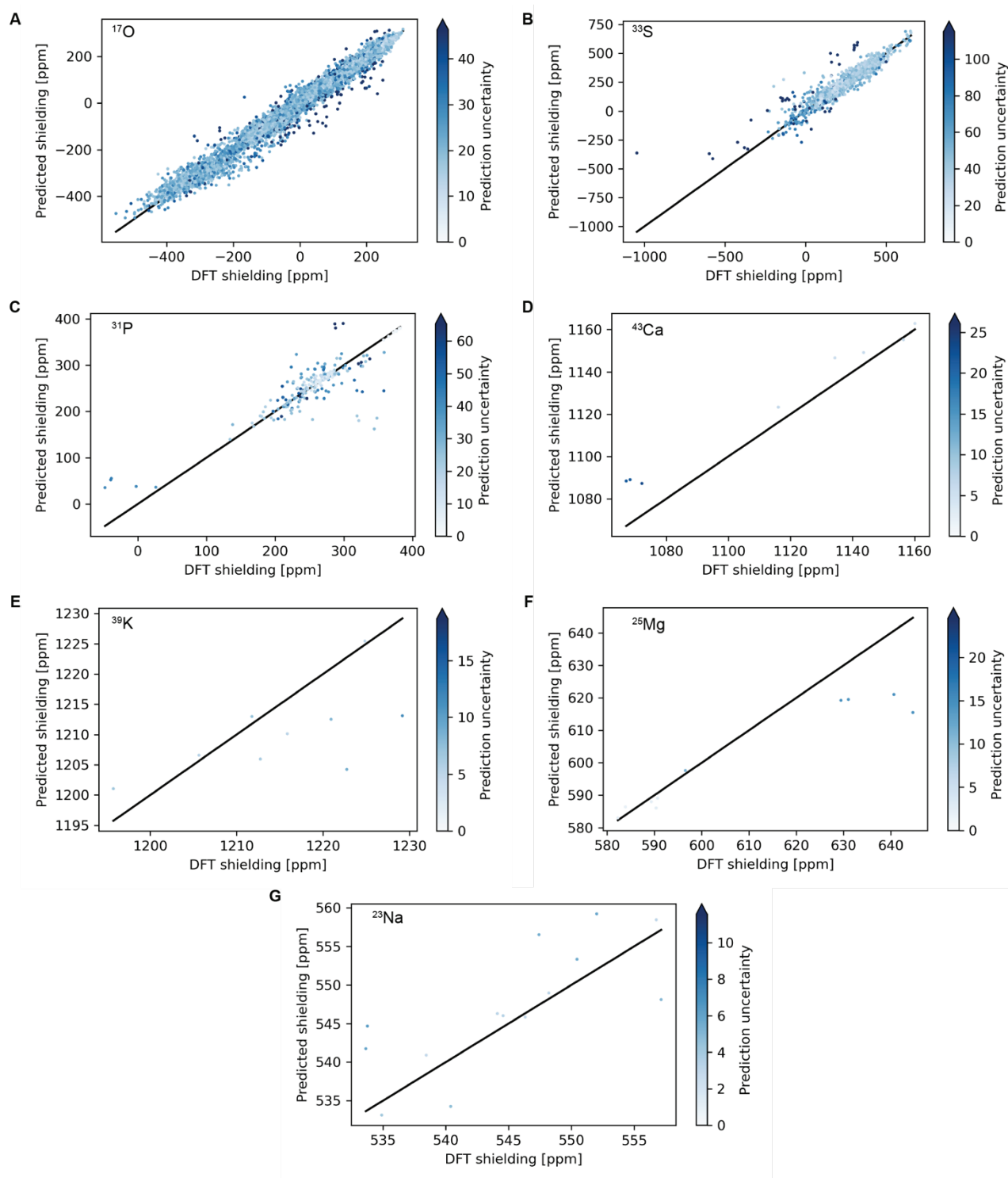


Figure 2.13. Comparison of DFT and ShiftML2 (A) ^{17}O , (B) ^{33}S , (C) ^{31}P , (D) ^{43}Ca , (E) ^{39}K , (F) ^{25}Mg and (G) ^{23}Na chemical shifts for the test set. The black lines show perfect correlation.

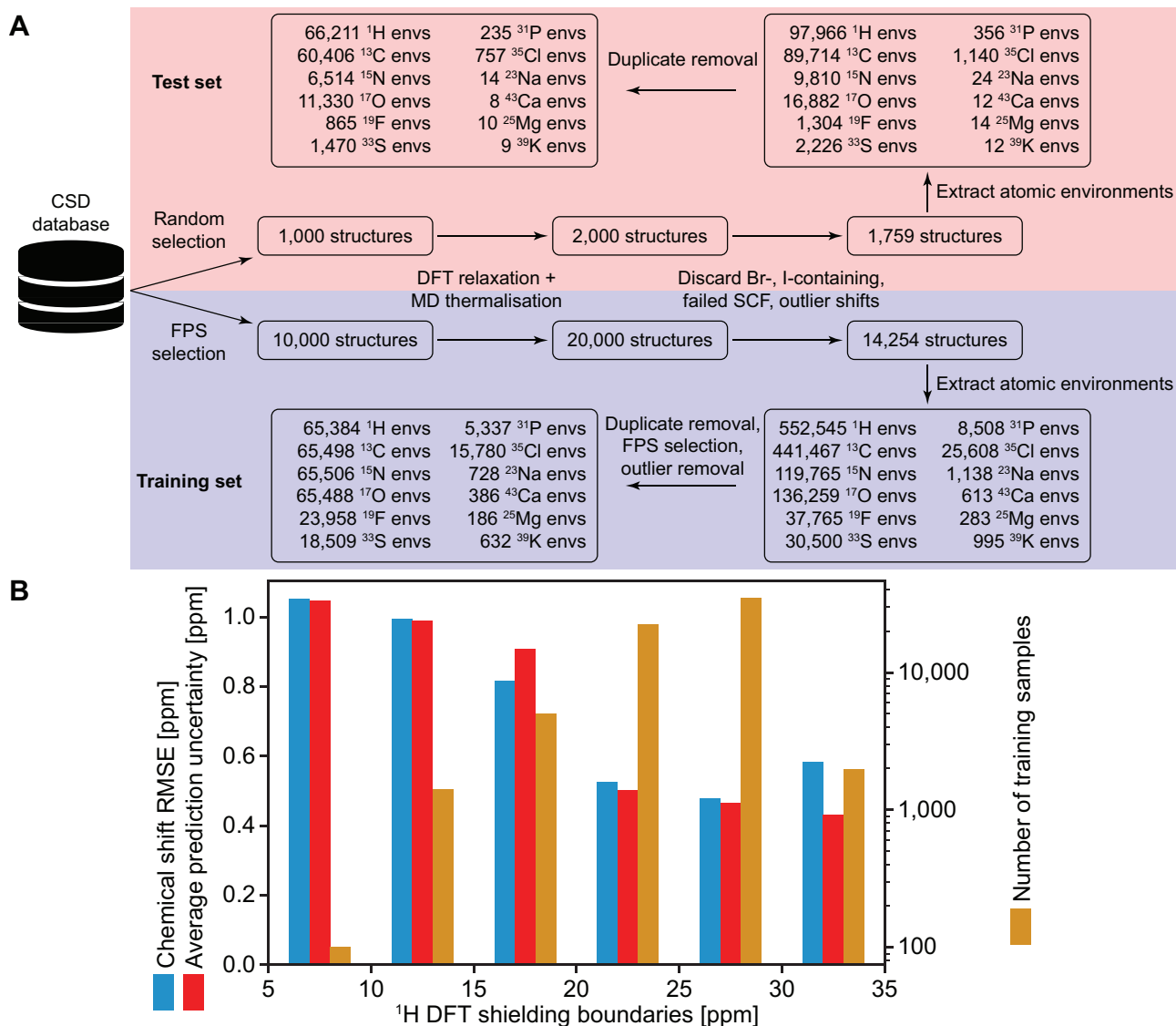


Figure 2.14. (A) Summary of the number of training and test structures and environments during the data selection and cleanup process. (B) ^1H chemical shift RMSE (blue), average prediction uncertainty (red) and number of training samples (beige) for different shielding ranges.

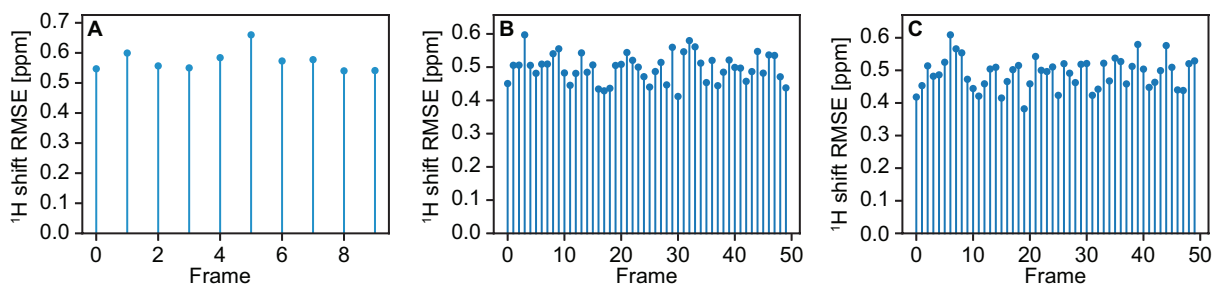


Figure 2.15. ^1H chemical shift RMSE against DFT computed shieldings along MD trajectories of the crystal structures of (A) AZD5718, (B) form 4 of AZD8329 and (C) cocaine. Frame 0 corresponds to relaxed structures.

2.3 *De novo* crystal structure determination from machine learned chemical shifts

This section has been adapted with permission from: Balodis, M.; Cordova, M.; Hofstetter, A.; Day, G. M.; Emsley, L., De Novo Crystal Structure Determination from Machine Learned Chemical Shifts. *Journal of the American Chemical Society* **2022**, *144* (16), 7215-7223. (post-print)

My contribution was to develop the method presented in this project and to analyse the results. I also contributed to the writing of the manuscript.

2.3.1 Introduction

In this section we show how by using a recently introduced machine learning model to predict chemical shifts, the structure of powdered organic solids can be determined in a manner fully analogous to the methods used in solution NMR or X-ray diffraction, by integrating on-the-fly solid-state NMR shift calculations into a Monte Carlo simulated annealing optimisation protocol. The approach does not require any structural hypothesis or knowledge of candidate structures (such as those from CSP). The approach is demonstrated to successfully determine five crystal structures, for two different polymorphs of the drug molecule AZD8329 (**1**), ampicillin (**2**), piroxicam (**3**) and cocaine (**4**) (Figure 2.16).

Among these molecules the structures of AZD8329 forms I and IV,⁵³ ampicillin⁵⁵ and cocaine⁵² have been previously found by NMR crystallography. AZD8329 form IV is notable because the structure was not found by X-ray diffraction methods prior to the original NMR crystallography study.⁵³ Having a rich polymorphic landscape, it is also an interesting example to test the ability to distinguish between different polymorphs. Ampicillin is notable because CSP methods failed to predict the correct structure until NMR constraints were included to bias the starting conformers.⁵⁵ Cocaine is one of the first examples on which it was shown that NMR chemical shifts can reliably determine the correct structure amongst a set of candidate structures.⁵² The structure of piroxicam so far has not been determined by NMR crystallography, although comparison of calculated and measured chemical shifts was used to validate a structure proposed from powder X-ray diffraction.¹⁴³

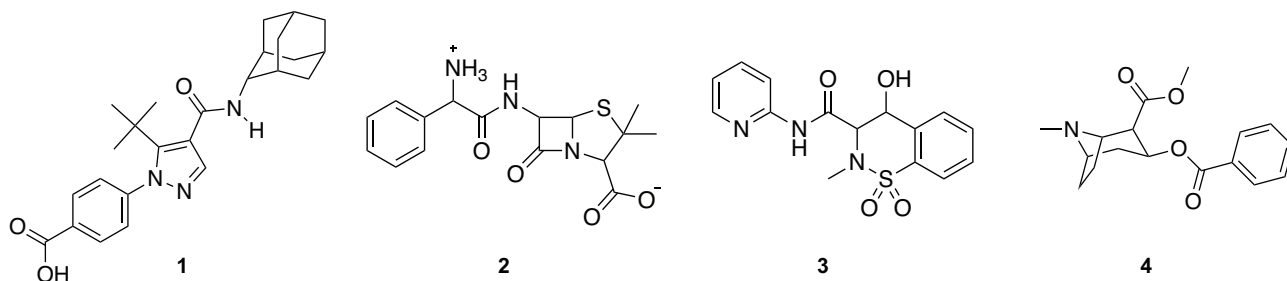


Figure 2.16. Molecular structures of AZD8329 (**1**), ampicillin (**2**), piroxicam (**3**) and cocaine (**4**).

2.3.2 Methods

Crystal structure determination. Crystal structure generation and optimisation were performed using a home-written Python script. The structure determination process follows the scheme shown in Figure 2.17, and is a version of constrained geometry optimisation that is completely analogous to the methods currently used to determine, for example, protein structures from liquid or solid-state NMR data, adapted to the case of molecular crystals. First, an initial conformation is generated with random torsional angles. The generated conformer is then placed in a randomly generated unit cell with randomly chosen position and orientation. Details of the structure generation are given in Appendix II. After the initial generation of a random crystal structure, 4,000 Monte Carlo steps are performed with a linear temperature profile between 2,500 and 50 K. The structures are generated in a given space group, and the space group symmetry is conserved during the optimisation. In each step one of the parameters defining the crystal structure (cell length or angle, conformer position or orientation, or conformer dihedral angle) is randomly selected and updated within a given maximum step size. If the change leads to better agreement (as determined by the pseudo-energies discussed below) it is accepted. Otherwise, the step is accepted with a probability $P_{acc} = e^{-\frac{\Delta E}{RT}}$, where ΔE is the change of pseudo-energy induced by the step, R is the gas constant, and T is the temperature. The step size of the updated parameter is doubled if the step is accepted, and halved otherwise (see Appendix II for detailed parameters including the step sizes). Every 500 steps hydrogen positions were optimised using tight binding DFT (DFTB3-D3H5).

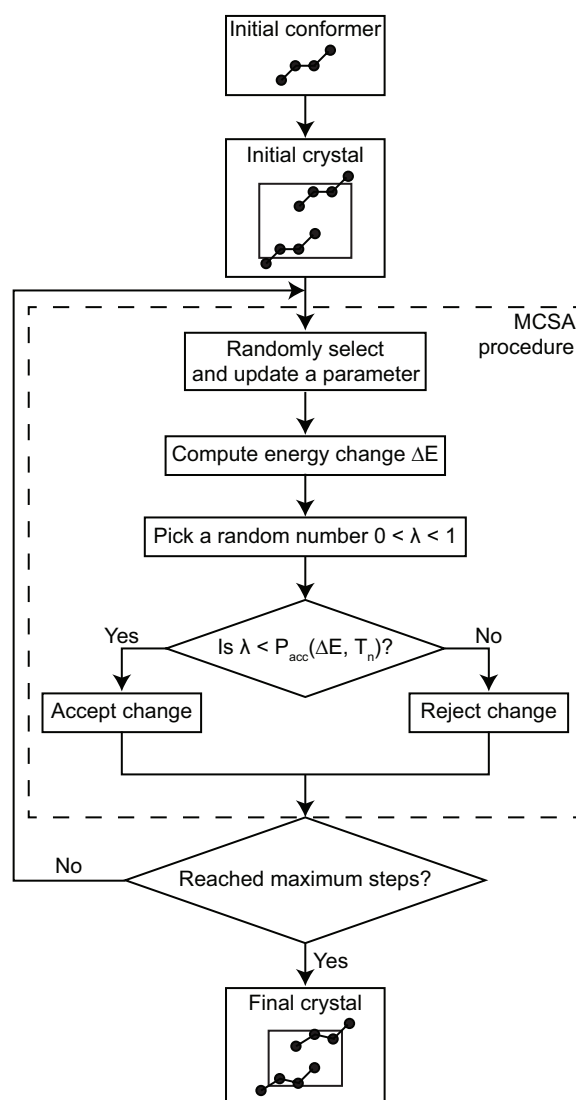


Figure 2.17. The scheme for crystal structure determination used in this study where $P_{acc} = \exp(-\Delta E/RT)$.

Energy calculations were performed at the semiempirical DFTB3-D3H5 level of theory using the 3ob-3-1 parameter set and the DFTB+ software version 20.1.^{325, 326, 349-352}

The chemical shieldings were predicted using ShiftML version 1.2 (publicly available at <https://shiftml.epfl.ch>).^{176, 261} Shieldings were converted to chemical shifts via the relation:

$$\delta = a + b\sigma, \quad (2.6)$$

where δ is the chemical shift, a and b are the experimentally determined calibration constants (see **Appendix II** for details) and σ is the calculated chemical shielding. Here we set a to 30.36 and b to -1. To account for ambiguity when comparing chemical shifts of protons for CH_2 groups, shifts were compared using the best matching criteria. Shifts which are hard or impossible to distinguish experimentally such as aromatic protons or CH_3 groups were averaged when making the comparison.

Crystal structure comparison. The optimised crystal structures were compared using the COMPACK algorithm,³⁵³ included in the commercial CSD package,³¹² which compares interatomic distances and angles within a cluster of molecules taken from the reference and comparison crystal structures. A cluster of 20 molecules was used for the comparison in this work. Before the comparison, physically unrealistic structures were removed, e.g., structures where neighbouring molecules are too close in space or where the density is unrealistically low. Most of the physically unrealistic structures are easily spotted due to their high energy or shift RMSD compared to the bulk of the structures generated.

2.3.3 Results and Discussion

The optimisation scheme introduced here is summarised in in **Figure 2.17**. In the first step, a viable conformation of the single molecule is generated, and bond-angles and lengths are optimised using, here, DFTB3-D3H5 which provides a good compromise between accuracy and computational cost (on the same timescale as ShiftML chemical shift calculations) (see **Appendix II** for details). Then, for each run, a random conformation is generated by randomising the flexible torsion angles, and a starting crystal structure is generated by randomly selecting cell parameters in a given space group (cell lengths, cell angles, position and orientation of the molecule). Between 1,000 and 10,000 trial structures were generated for each system. Each structure was then optimised by a Monte Carlo simulated annealing process described in **Section 2.3.2**, where in each step one of the parameters defining the crystal structure (i.e., a single torsion angle or cell parameter) was randomly changed and chemical shifts and the DFTB system energy were calculated following the change.

Here, to enable the possibility to calculate shifts at each step, the ShiftML prediction algorithm was used.^{176, 261} ShiftML is a fast and accurate method to compute chemical shifts in a matter of seconds even for the largest of molecular crystals. It was recently developed using DFT optimised structures derived from the Cambridge Structural Database (CSD) as a training set for a machine learning framework. The current version (at the time of this study) can predict chemical shifts for molecules containing H, C, N, O or S atoms.

The cost function used in the Monte Carlo process is:

$$E_{tot} = E_{DFTB} + cE_{cs}, \quad (2.7)$$

where

$$E_{cs} = \sqrt{\frac{\sum_{i=1}^n (\delta_{i,trg} - \delta_{i,ShiftML})^2}{n}}, \quad (2.8)$$

where $\delta_{i,trg}$ is the target chemical shift of the i th nucleus in the molecule containing n nuclei and $\delta_{i,ShiftML}$ is the corresponding shift computed using the ShiftML model. c is an empirically adjusted constant (in kJ/mol) that weights the relative contribution of the internal energy and the agreement with experiment in the cost function. (Note that the values of E_{cs} are independent of the size of the molecule, but will change from one type of nucleus to another, and that E_{DFTB} will depend on the size of the molecule. In the examples here, satisfactory results were found with vales of c such that $\Delta E_{DFTB} \sim \Delta E_{cs}$, where ΔE is the difference observed between two Monte Carlo steps at the end of the optimisation process.) In the following, for the proof of principle demonstration here, we use shifts calculated with ShiftML from the known structure as the $\delta_{i,trg}$ target set in E_{cs} . This reduces any bias due to experimental variability between compounds in the comparisons below, and makes the process fully self-consistent. We note that the estimated errors on ShiftML shifts are in any case similar to or larger than the error ranges in the experimental shifts. The other parameters in the simulated annealing process are given in **Section 2.3.2** and **Appendix II**.

Figure 2.18 shows the results for AZD8329 Form I, AZD8329 Form IV, ampicillin, piroxicam and cocaine. In order to demonstrate that the chemical shifts are indeed the driving force for the structure determination, for each case, the optimisation was performed with the penalty function that includes both the DFTB energy and chemical shift differences and, for comparison, using only the DFTB energy. **Figure 2.19** shows expansions of the regions below 100 kJ/mol and 0.5 ppm.

We expect correct structures to occur in the region of low chemical shift RMSD and low calculated energy. For ^1H shift root-mean-square deviation (RMSD) we use a cutoff of 0.5 ppm, taken from Engel *et al.* where they determined the expected error of the ShiftML model for ^1H to be 0.48 ppm.¹⁷⁶ Nyman and Day showed that with accurate calculations most polymorphs are separated by less than 7.2 kJ/mol,³⁵⁴ which can be treated as the most relevant energy range on CSP landscapes. In this study we use DFTB, whose energies are less accurate and have been shown to place observed crystal structures over a much wider energy range in CSP studies.³⁵⁵ To account for this larger spread, we use a cutoff for the accepted structures of up to 20 kJ/mol from the lowest energy structure. In-deed, the spread of predicted energies decreases significantly when the structures that are within 20 kJ/mol and 0.5 ppm RMSD are further optimised using DFT, as illustrated in **Figure 2.26** (and **Table 2.8**). Typically, after optimisation the predicted DFT energy difference between the structures is less than ~ 2 kJ/mol.

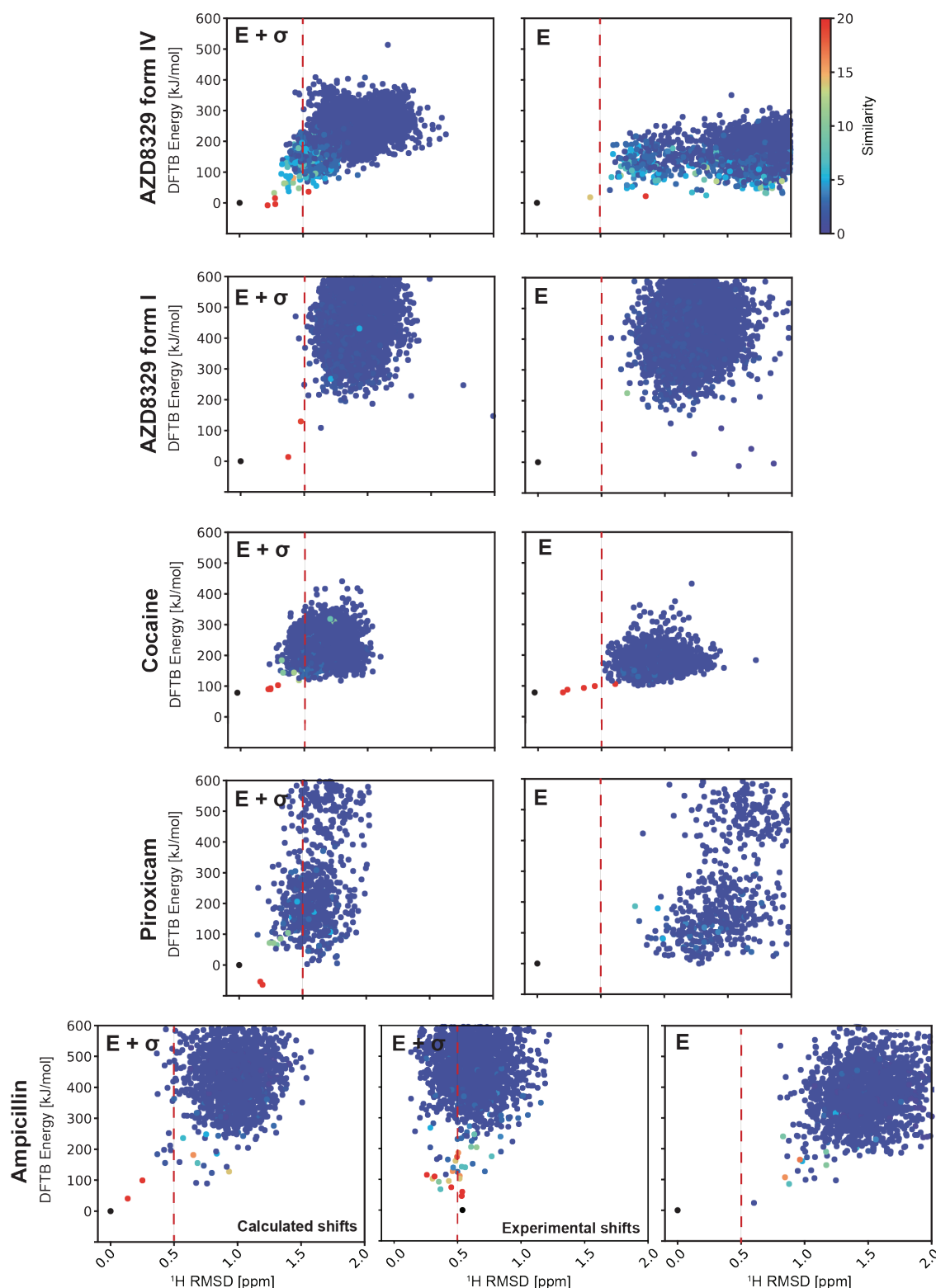


Figure 2.18. Plots of DFTB energy vs ^1H chemical shift RMSD for the results of 10,000 simulated annealing runs on AZD8329 form IV, 10,000 runs on AZD8329 form I, 2,500 runs for ampicillin, 1,000 runs for piroxicam and 2,500 runs on cocaine. The left column shows the optimisations done using both chemical shift and energy while the right column shows the optimisations done using only energy. For ampicillin results are shown for both where ^1H shifts calculated from the known reference structure were used, and where the experimental ^1H shifts were used as targets for the optimisation. Each point represents a structure optimised as described in Section 2.3.2. The vertical axis shows DFTB energies and the horizontal axis ^1H shift RMSD values with respect to the shifts calculated for the known experimental structure which is set to 0 and is coloured black. The colour of each point reflects the similarity between each of the calculated structures and the reference structure, according to the scale on the right and as described in Section 2.3.2. The red vertical dashed line shows the cutoff value of 0.5 ppm for the ^1H RMSD. For piroxicam, unconstrained optimisation of the experimental structure leads to a large deviation in the structure, so the reference energy is the energy of the experimental structure with only hydrogen atom positions optimised.

For all the compounds we note that the majority of Monte Carlo runs do not yield any results with either low DFTB energy or with a low chemical shift RMSD to experiment. Indeed, if we define a region of acceptable structures to have simultaneously a DFTB energy within 20 kJ/mol of the lowest energy structure in the Monte Carlo set and a chemical shift RMSD to experiment below 0.5 ppm, then the pure Monte Carlo approach using only DFTB energy as the driving force does not find any structures that match the RMSD₂₀ (RMSD of atomic positions of 20 molecules matched by the COMPACT algorithm) criteria for either form of AZD8329. This is completely in line with expectations since this simple semi-empirical type approach is not expected to easily find crystalline polymorphs.

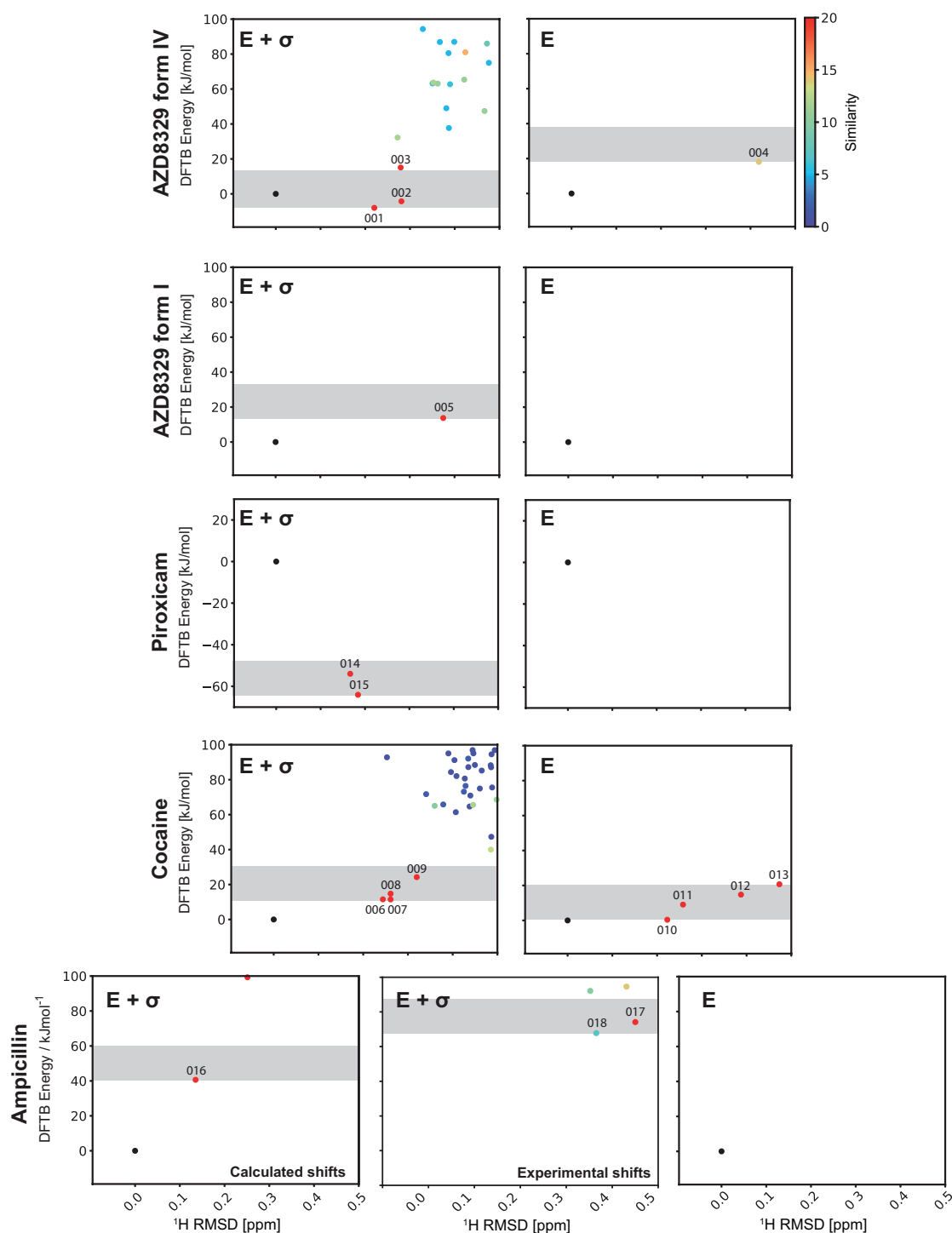


Figure 2.19. Plots of DFTB energy vs ¹H chemical shift RMSD as shown in Figure 2.18, expanded to include a range of 100 kJ/mol and up to 0.5 ppm ¹H RMSD. The grey areas represent the area within 20 kJ/mol of the lowest energy structure found in the optimisation. Labels refer to the structures as defined in Table 2.6. For piroxicam, unconstrained optimisation of the experimental structure leads to a large deviation in the structure, so the reference energy is the energy of the experimental structure with only hydrogen atom positions optimised.

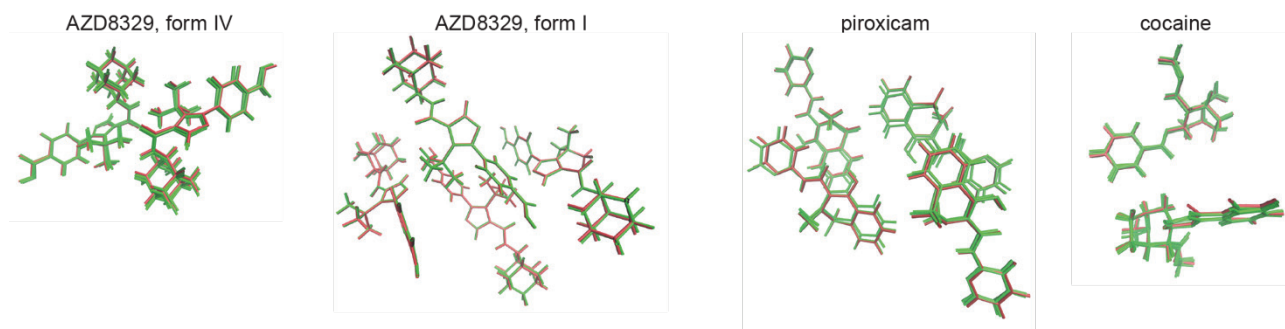


Figure 2.20. Overlay of the unit cell for the structures determined here for AZD8329 form IV, AZD8329 form I, piroxicam, and cocaine. For AZD8329 form IV, there are three structures (**Figure 2.19**), one for form I, 2 for piroxicam, and 4 for cocaine. The red structures are the known structures, and the green structures are the structures determined here that are less than 20 kJ/mol from the lowest energy determined structure and 0.5 ppm ^1H RMSD compared to the target shifts.

Including chemical shifts in the penalty function yields three structures for Form IV (001-003) within the acceptable ranges, and one structure for Form I (005). These structures for both forms are shown in **Figure 2.20**, superimposed on the known structures, and we see that they are in excellent agreement with the correct structures as previously determined by X-ray diffraction or NMR.

Ampicillin is another interesting example as noted in the introduction, since it is a case where current crystal structure prediction methods fail since the conformer present in the crystal structure has a relatively high energy in the gas phase.⁵⁵ As a result, chemical shift driven structure determination based on prior generation of candidates fails. In contrast, Monte Carlo runs for ampicillin including DFTB energy and chemical shifts produced two structures that perfectly match with the known crystal structure, with one of them (016) being selected by our criteria. The structure determined by our criteria is superimposed on the known crystal structure in **Figure 2.21**. Runs using only DFTB energy did not produce any matching structure, either in the acceptable region or outside it.

Similarly to ampicillin, runs for piroxicam produced two structures (014 and 015) matching with the known crystal structure, both of which are in the acceptable region. Again, no matching structures were found for the runs using only energy in the penalty function. Overlay of the structures determined here with the known crystal structure are shown in **Figure 2.20**. From **Figure 2.19** it is seen that both of the structures found are significantly lower in DFTB energy than the known structure. We note that to compare our determined structures and the known reference structures we systematically relaxed the atom positions and the cell parameters for the experimental reference structures using DFTB. While the results of the relaxation were fairly similar to the starting structures for most of the reference structures this was not the case for the structure of piroxicam. Full DFTB relaxation of piroxicam changed the structure to a point where its space group changed. To avoid this, we relaxed only ^1H positions with DFTB, and we suspect that this is why the energy of the reference structure appears higher than expected. When both the determined structures and the known structure were optimised with DFT the (DFT) energy difference between them was reduced to 0.4 kJ/mol for the best matching structure.

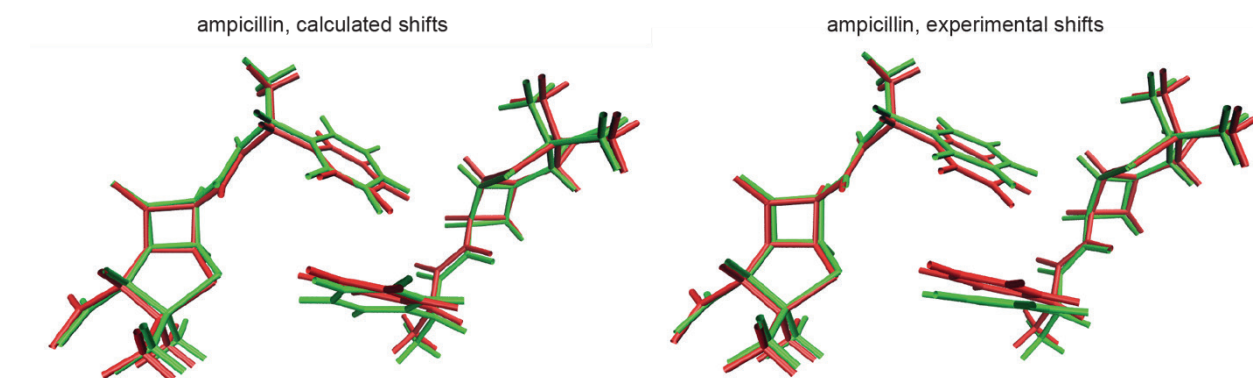


Figure 2.21. Overlay of the unit cell for the structures determined here for ampicillin with calculated (top, structure 016) or with experimental (bottom, structure 017) chemical shifts. The red structures are the known structures, and the green structures are the structures determined here.

Cocaine is an interesting example, since it is significantly less flexible than AZD8329. In this case, the Monte Carlo approach with energy alone does already produce four structures in the acceptable region (010-013). Adding chemical shifts did not improve the result, and the same number of structures were found in the acceptable region (006-009). The four structures optimised using shifts are shown in **Figure 2.20** superimposed on the known structure for cocaine, and we again see that they are in good agreement with the correct structure. We explain this as cocaine having a relatively simple energy landscape with few competing structures: the results of the Monte Carlo search using only energy directs the search efficiently towards the known crystal structure of the only known polymorph, suggesting that there are few competing, ‘false’ structures. It is in cases where there are many energetically competing structures, which is the norm, where adding the chemical shift to the fitness function is expected to increase the effectiveness of the search at locating the correct structure. The other compounds studied here, on the other hand, have much richer energy landscapes, with at least 4 anhydrous polymorphs known for AZD8329 for example,⁵³ and by using the chemical shifts of two different forms as targets, we were able to successfully determine both structures here. **Figure 2.20** and **Figure 2.21** show the overlay of the asymmetric unit of the crystal structures determined here for each compound (green) with the known reference structures (red).

All atom RMSD₂₀ values between determined structures and the known reference crystal structures are given in **Table 2.5**. The highest RMSD₂₀ value is 0.51 Å for ampicillin, meaning that all of the optimised structures correspond very well to the experimental reference crystal structure. In comparison, in the current latest crystal structure prediction blind test (6th) the highest RMSD₂₀ value was 0.81 Å which while considered high, was still considered acceptable.³⁵⁶ In the examples here, after the DFT optimisation the highest value decreased to 0.49 Å and lowest to 0.05 Å. **Table 2.5** also gives the distribution of the unit cell dimensions for the optimised structures which are very close to the experimental values. Individual RMSD₂₀ values and the cell parameters for all best matching structures are given in **Table 2.7**.

Table 2.5. The reduced unit cell parameters and atom RMSD₂₀ values for the determined structures using chemical shifts, without subsequent DFT relaxation. The number in the brackets after the name of the compound is the number of the structures found. Standard deviation is given where more than one structure is found. The number in parentheses after the mean value of the cell parameters is the value for the known experimental structure.

Name	a [Å]	b [Å]	c [Å]	α [°]	β [°]	γ [°]	RMSD ₂₀ [Å]
AZ8329, form IV (3)	9.5±0.1 (9.9)	11.0±0.1 (10.8)	11.8±0.3 (11.6)	65.3±1.7 (65.7)	75.9±2.2 (75.0)	75.5±3.4 (74.0)	0.44±0.15
AZ8329, form I (1)	11.3 (11.4)	13.2 (13.1)	15.1 (15.0)	114.2 (113.0)	90 (90)	90 (90)	0.14
Piroxicam (2)	6.9±0.1 (6.8)	13.3±0.2 (13.9)	15.12±0.1 (15.1)	90 (90)	90 (90)	93.2±1.0 (97.3)	0.40±0.13
Cocaine (4)	8.1±0.1 (8.1)	9.2±0.1 (9.0)	10.1±0.2 (10.0)	90 (90)	105.8±1.0 (106.0)	90 (90)	0.28±0.02
Ampicillin calculated (1)	5.8 (5.8)	12.3 (11.4)	12.5 (12.3)	116.4 (113.6)	90 (90)	90 (90)	0.51
Ampicillin experimental (1)	5.8 (5.8)	11.3 (11.4)	12.3 (12.3)	117.2 (113.6)	90 (90)	90 (90)	0.43

Optimisation using experimental target shifts. As noted above, we use ¹H chemical shifts calculated for the known crystal structures as the target for optimisation here. This allows us to explore the method without any biases introduced by any possible errors in chemical assignments, and to make the analysis self-consistent. Of course, it is most important that the method also works using experimental shifts. This is demonstrated in **Figure 2.18** and **Figure 2.19** where we also show the results of optimisation against experiment ¹H shifts for ampicillin. The experimental shifts were taken from Hofstetter *et al.*⁵⁵ In this case two structures (017 and 018) matched the selection criteria. One structure (017) yielded a very good RMSD₂₀ of 0.41 Å with respect to the known structure, as illustrated in **Figure 2.21**. It is interesting to note that the other structure (018) at first glance matches less well, but on further examination we see that the cell parameters match very well (see **Table 2.7**), and the main difference is a slight change in the orientation of the aromatic ring position. An overlay of the unit cell of the known structure and structure 018 is shown in **Figure 2.25**. After optimisation with DFT the relative (DFT) energy for the structures converged to -0.4 and 9.4 kJ/mol for (017) and (018) respectively with respect to the known structure (see **Table 2.8**), and the ¹H RMSD to DFT calculated shifts was 0.13 and 0.41 ppm, suggesting that the optimised structure 017 is in better agreement with the experiment.

This is the first example of a molecular crystal structure determined directly from experimentally measured chemical shifts in contrast to earlier approaches where chemical shifts were used to select from a predetermined set of predicted crystal structures.

2.3.4 Conclusion

In this section we have shown that crystal structures can be directly determined from chemical shifts, without any prior structural hypothesis and without any knowledge from candidate structures (such as from CSP), through the use of machine learned chemical shifts which enable on-the-fly calculation of shifts at each step of a simulated annealing structure determination protocol. We have illustrated this for the structures of ampicillin, piroxicam and cocaine, as well as for AZD8329 where the inclusion of machine learned chemical shifts allows the determination of the correct structures for two different polymorphic forms. We note that the AZD8329 case is a particularly important illustration, since it clearly shows how the chemical shifts can drive the optimisation towards two very different structures for the same molecule.

Here we chose to use a Monte Carlo simulated annealing algorithm due to its relatively straightforward nature, but in principle machine learned chemical shifts can be incorporated into other optimisation methods as they are easy to add as an additional pseudo energy term, and we believe there is significant room for further development and increased efficiency of this approach to chemical shift-based structure determination in molecular solids. Finally, we note that the method presented here no longer relies on a purely energy driven computational candidate crystal structure generation step. By driving the structure determination directly from chemical shifts, integrated through the entire optimisation procedure, the method is applicable even in cases where crystal structure prediction is extremely challenging, such as the example of ampicillin here.

2.3.5 Appendix II

Code availability. All code used in this study is freely available on <https://github.com/manucordova/NMRX>.

Trial crystal structure generation. A gas phase conformation was first generated by randomising the non-trivial dihedral angles in the molecule (shown in **Figure 2.16**). No energetic criterion was set to select the generated conformations. In the case of AZD8329 the OCNH angle that corresponds to the amide bond was fixed to the experimentally known value knowing that it can take only cis or trans position, and if needed the other configuration can also be explored. In the case of ampicillin, the zwitterionic form was chosen as this is easily seen from the NMR spectrum. In the case of piroxicam it was assumed that an intramolecular 6-atom aromatic system is formed via hydrogen bond, which reduces the number of non-trivial dihedral angles from 4 to 2.

The conformer was then introduced into a randomly generated crystal in the selected space group (here we take the known space group, but in principle the process can be repeated for all possible space groups). Cell lengths, cell angles, and the position and orientation of the molecule in the asymmetric unit were randomly initialised. The maximum volume of the crystal was set to be no larger than twice the sum of the van der Waals sphere volume of each atom in the unit cell. Cell lengths and angles were sampled from uniform distributions in the ranges [1 Å, 50 Å] and [45°, 135°], respectively. The position of the asymmetric unit was sampled from a uniform 3D distribution in the range [0, 1] in each dimension, corresponding to the fractional coordinates of the centre of mass of the conformer.

During both the conformation and crystal generation steps interatomic clashes were avoided by generating new conformers and crystals until no clash was detected. A clash was defined as two atoms being closer than a set factor times the sum of their covalent radii. The factor was set to 0.85 to detect clashes within a single molecule, and 1.2 to detect clashes between different molecules in the unit cell.

Figure 2.22 shows the energies and chemical shift RMSDs of the generated starting structures.

Monte Carlo run parameters. After the generation of the trial crystal structure, it was subjected to a Monte Carlo Simulated Annealing (MCSA) optimisation protocol.³⁵⁷ 4,000 Monte Carlo steps were performed with a linear temperature profile from 2,500 to 50 K. In each step one of the structure defining parameters was randomly selected (cell lengths, cell angles, position of the asymmetric unit, orientation of the asymmetric unit and torsional angles) and was randomly modified. Parameter updates were uniformly sampled in ranges initially set to [-2 Å, 2 Å], [-20°, 20°], [-0.05, 0.05] and [-40°, 40°] for cell lengths, cell angles, asymmetric unit position (in fractional coordinates), and dihedral angles, respectively. Updates to the asymmetric unit orientation were performed by first randomly selecting a direction and rotating the conformer about its centre of mass and along the selected direction with an angle uniformly sampled in a range set initially to [-30°, 30°].

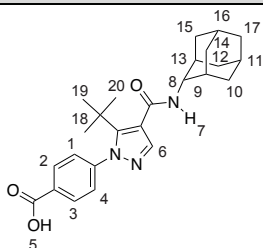
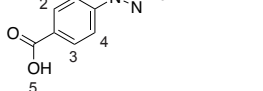
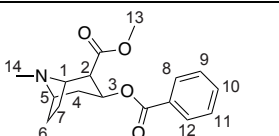
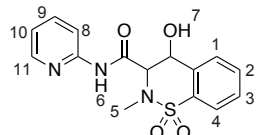
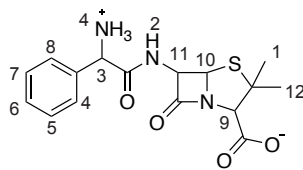
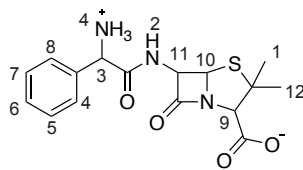
After each random change a new, updated crystal was generated and the cost function was calculated (see **Section 2.3.3**). If the step was found to lead to a lower cost function, it was accepted. Otherwise, it was accepted with a probability given by:

$$P_{acc} = e^{\frac{E_{tot,old} - E_{tot,new}}{RT}}, \quad (2.9)$$

where $E_{tot,old}$ is the old cost, $E_{tot,new}$ is the new cost, R is the gas constant and T is the temperature. When a step was accepted, the range of parameter update corresponding to the crystal parameter change was doubled, otherwise it was halved. Maximum parameter update ranges were set to ± 20 Å, $\pm 360^\circ$, ± 1 , $\pm 360^\circ$ and $\pm 360^\circ$ for cell lengths, cell angles, asymmetric unit position, asymmetric unit orientation and dihedral angles respectively. Parameter update ranges were modified individually for each cell length, cell angle and dihedral angle. Every 500 steps proton positions were optimised with DFTB.

Chemical shift referencing. To convert from chemical shieldings calculated by ShiftML to chemical shifts a set of 7 small organic molecules with known experimental chemical shifts were chosen.¹²⁷ Their shieldings were calculated and the calibration constants a and b in **Equation 2.6** were obtained by least squares regression against experimental values. The calibration constants were 30.36 and -1.0 for a and b , respectively.

Table 2.6. Target chemical shifts used.

Structure	¹ H chemical shifts	Labelling
AZD8329, form IV	1: 6.63, 2: 8.29, 3: 9.12, 4: 8.27, 5: 17.13, 6: 7.65, 7: 10.43, 8: 2.72, 9: 1.59, 10: 1.61, 2.46, 11: 1.57, 12: 0.84, 0.55, 13: 1.61, 14: 2.23, 1.8, 15: 0.83, -0.09, 16: 1.41, 17: 1.63, 0.94, 18: 0.42, 0.42, 19: 0.73, 0.73, 0.73, 20: -0.46, -0.46, -0.46	
AZD8329, form I	1: 9.2, 2: 7.62, 3: 4.69, 4: 7.89, 5: 1.41, 6: 8.05, 7: 4.13, 8: 1.8, 9: 0.91, 10: 1.59, 0.19, 11: 6.54, 12: 2.59, -0.01, 13: 1.24, 14: 1.58, 0.96, 15: 0.08, -0.11, 16: 9.04, 17: 0.53, 2.06, 18: -0.23, -0.23, -0.23, 19: 1.64, 1.64, 1.64, 20: -0.61, -0.61, -0.61	
Cocaine	1: 3.98, 2: 3.64, 3: 5.5, 4: 1.7, 2.97, 5: 3.63, 6: 3.48, 1.77, 7: 2.05, 1.69, 8: 7.87, 9: 7.46, 10: 7.58, 11: 7.57, 12: 7.35, 13: 3.76, 3.76, 3.76, 14: 1.46, 1.46, 1.46	
Piroxicam	1: 5.87, 2: 7.43, 3: 6.19, 4: 7.02, 5: 1.9, 1.9, 1.9, 6: 9.45, 7: 10.49, 8: 6.15, 9: 6.16, 10: 6.08, 11: 7.98	
Ampicillin	1: 1.05, 1.05, 1.05, 2: 7.76, 3: 5.89, NH3: 8.37, 8.37, 8.37, 4: 7.92, 5: 3.74, 6: 6.38, 7: 7.29, 8: 6.35, 9: 3.68, 10: 6.16, 11: 7.61, 12: 0.14, 0.14, 0.14	
Ampicillin, experimental ⁵⁵	1: 1.6, 1.6, 1.6, 2: 7.5, 3: 4.8, NH3: 10.0, 10.0, 10.0, 4: 7.3, 5: 7.3, 6: 7.3, 7: 7.3, 8: 7.3, 9: 4.0, 10: 5.2, 11: 6.6, 12: 0.6, 0.6, 0.6	

DFT calculations. All DFT computations performed on the structures determined were carried out using the plane-wave density functional theory (DFT) software Quantum ESPRESSO, version 6.5.^{328, 329} The PBE level of theory,⁹⁷ Grimme D2 dispersion correction³³⁰ and projector augmented wave scalar relativistic pseudopotentials obtained from PSLibrary version 1.0.0³³² were used for all computations. Wavefunction and charge density energy cutoffs were set to 160 Ry and 1280 Ry, respectively. A Monkhorst-Pack grid of k-points³³⁸ corresponding to a maximum spacing of 0.05 Å⁻¹ in reciprocal space was used. After relaxation of atomic positions and lattice parameters, a single-point computation was performed using the same parameters, and chemical shieldings were computed using the GIPAW method.^{117, 118}

Data on the structures determined.**Table 2.7.** RMSD₂₀, relative energy, ¹H shift RMSD and the reduced cell parameters for the structures that are in the region of 20 kJ/mol from the lowest energy structure and less than 0.5 ppm RMSD after the Monte Carlo optimisation and prior to DFT relaxation.

Structure	Label	RMSD ₂₀ [Å]	Relative energy [kJ/mol]	¹ H shift RMSD [ppm]	a [Å]	b [Å]	c [Å]	α [°]	β [°]	γ [°]
AZD8329, form IV, E+σ	001	0.42	-4.3	0.28	9.5	10.8	11.8	65.2	75.8	73.8
AZD8329, form IV, E+ σ	002	0.64	15.0	0.28	9.4	11.1	12.0	63.4	75.1	73.2
AZD8329, form IV, E+ σ	003	0.28	-8.0	0.22	9.6	11.0	11.3	67.6	78.9	80.4
AZD8329, form IV, E	004	-	18.2	0.42	9.3	11.7	12.2	63.0	74.7	73.9
Known structure	-	-	-	-	9.9	10.8	11.6	65.7	75.0	74.0
AZD8329, form I, E+σ	005	0.14	13.7	0.37	11.3	13.2	15.1	114.2	90	90
Known structure	-	-	-	-	11.4	13.1	15.0	113.0	90	90
Cocaine, E+σ	006	0.30	11.5	0.24	8.1	9.3	9.9	90	106.4	90
Cocaine, E+σ	007	0.25	11.5	0.26	8.1	9.2	9.9	90	104.0	90
Cocaine, E+σ	008	0.29	14.8	0.26	8.0	9.2	10.2	90	106.4	90
Cocaine, E+σ	009	0.31	24.2	0.32	8.0	9.2	10.2	90	106.4	90
Cocaine, E	010	0.17	0.38	0.22	8.2	9.4	10	90	106.3	90
Cocaine, E	011	0.60	9.10	0.26	8.3	9.1	10.0	90	106.5	90
Cocaine, E	012	0.21	14.8	0.39	8.2	9.8	9.7	90	108.6	90
Cocaine, E	013	0.35	20.7	0.47	8.3	9.3	9.8	90	105.1	90
Known structure	-	-	-	-	8.1	9.0	10.0	90	105.0	90
Piroxicam, E+σ	014	0.28	-54.0	0.17	6.9	13.1	15.2	90	90	92.3
Piroxicam, E+σ	015	0.53	-64.1	0.18	6.8	13.4	15.1	90	90	94
Known structure	-	-	-	-	6.8	13.9	15.1	90	90	97.3
Ampicillin, calculated, E+σ	016	0.51	40.6	0.14	5.8	12.3	12.5	116.4	90	90
Ampicillin, experimental, E+σ	017	0.41	74.1	0.45	5.8	12.1	12.4	116.4	90	90
Ampicillin, experimental, E+σ	018	-	67.8	0.37	5.8	11.3	12.3	117.2	90	90
Known structure	-	-	-	-	5.8	11.4	12.3	113.6	90	90

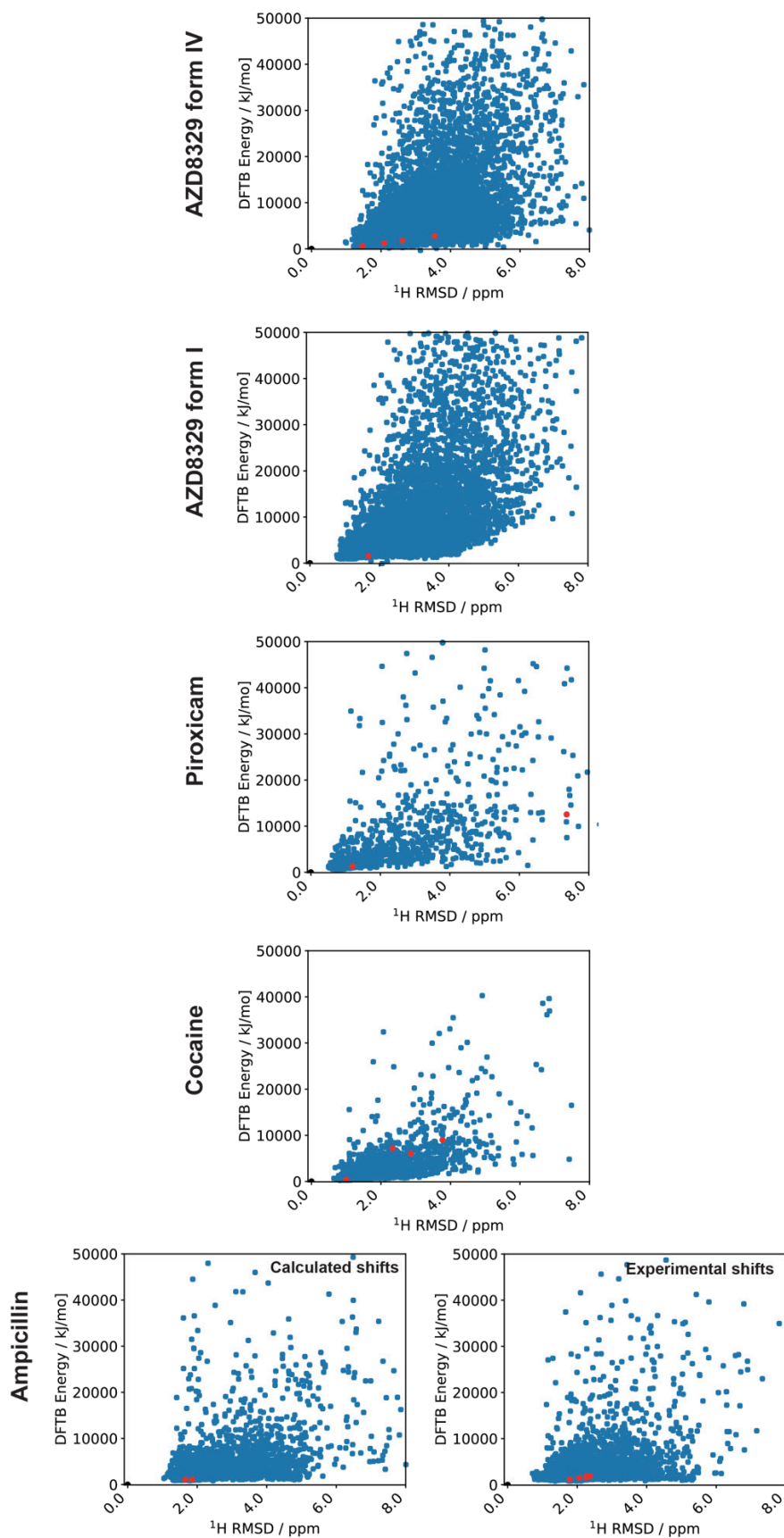


Figure 2.22. Plots of DFTB energy vs ^1H chemical shift RMSD for the initially generated structures. The vertical axis shows DFTB energies and the horizontal axis ^1H shift RMSD values with respect to the shifts calculated for the known experimental structure which is set to 0 and is coloured black. The structures coloured red are the ones that lead to the structures matching the structure previously determined by XRD or NMR, following optimisation with chemical shifts.

Energy-density plots. Figure 2.23 shows results of plotting the density vs DFTB energy. As expected, the structures found to correspond to the experimental crystal structure lie in the high density-low energy region.

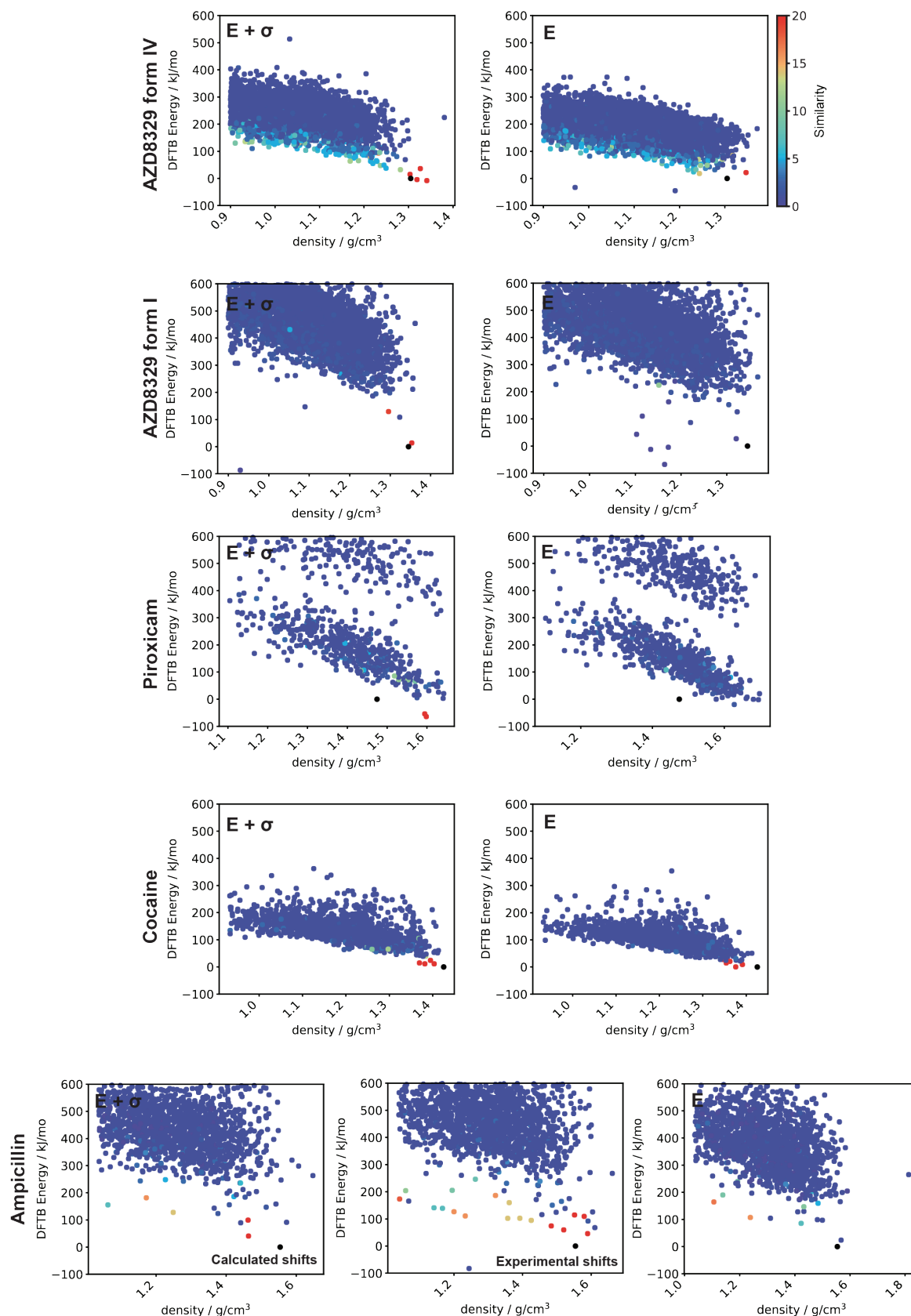


Figure 2.23. Energy-density plots for the optimised structures. The vertical axis shows DFTB energies and the horizontal axis the density of the determined structures. The black dot corresponds to the previously determined known structures.

Total energy during optimisation. To illustrate how relative energy contributions change over the course of the optimisation, **Figure 2.24** shows how total energy, DFTB energy and the energy coming from chemical shift contribution changes over time for the four AZD8329 form IV structures that were found to match the known crystal structure.

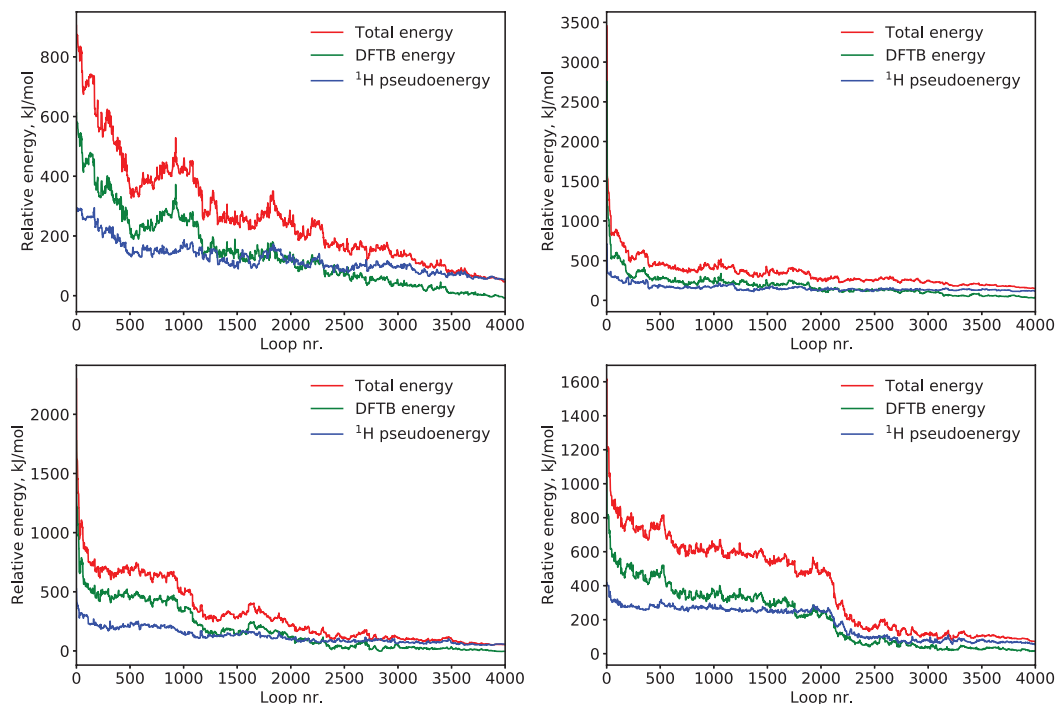


Figure 2.24. Plots of total energy, DFTB energy, and chemical shift pseudo-energy during optimisation, shown for the four AZD8329 form IV structures that were found to match the known crystal structure.

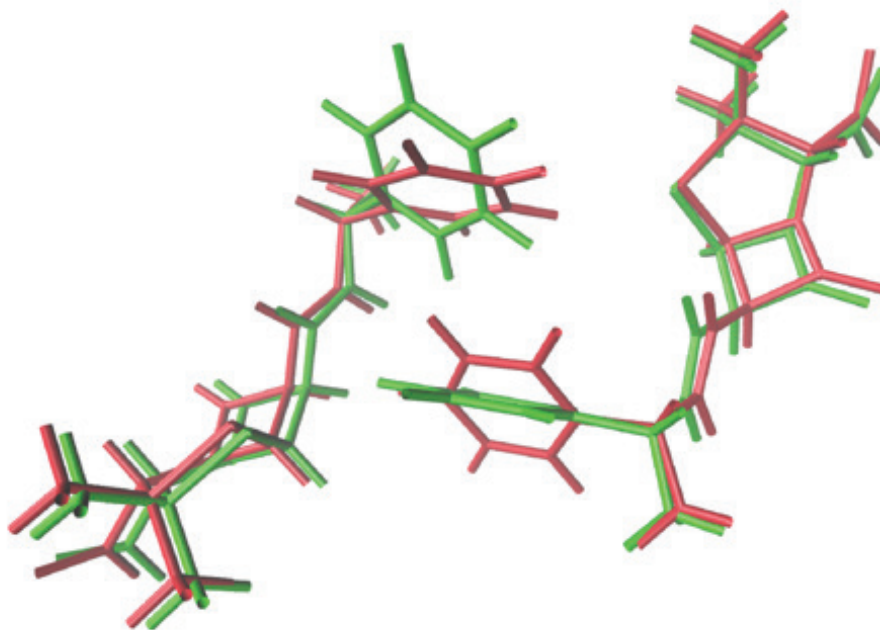


Figure 2.25. Overlay of the unit cell for structure 018 determined here for ampicillin using experimental shifts. The red structure is the known structure and 018 is in green.

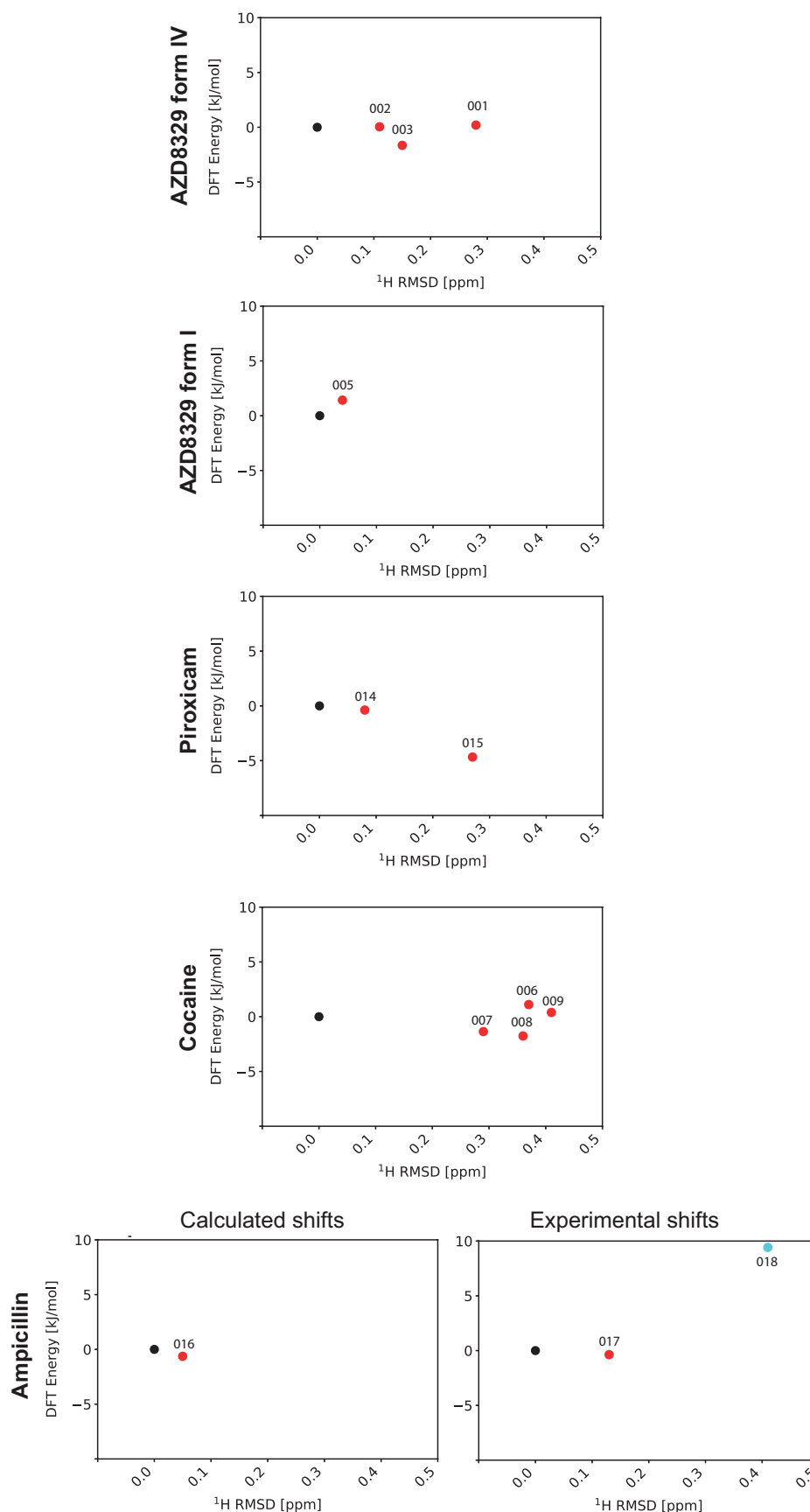


Figure 2.26. Plots of DFT energy vs ^1H chemical shift RMSD for the optimised structure fulfilling the initial selection criteria after full geometry optimisation with DFT. The black dot is the (DFT-relaxed) known structure used as a reference. Labels refer to the initial structures in Table 2.7.

Table 2.8. RMSD₂₀, relative energy and ¹H shift RMSD before and after the optimisation with DFT on the optimised structures matching the selection criteria.

Structure	Label	RMSD ₂₀ before [Å]	RMSD ₂₀ after[Å]	Relative energy before [kJ/mol]	Relative energy after [kJ/mol]	¹ H shift RMSD before [ppm]	¹ H shift RMSD after[ppm]
AZD8329, form IV, E+σ	001	0.64	0.39	15.0	0.2	0.28	0.28
AZD8329, form IV, E+σ	002	0.42	0.19	-4.3	-1.6	0.28	0.15
AZD8329, form IV, E+σ	003	0.28	0.23	-8.0	0.04	0.22	0.11
AZD8329, form IV, E	004	-	0.31	18.2	-0.6	0.42	0.15
AZD8329, form I, E+σ	005	0.14	0.06	13.7	1.4	0.37	0.04
Cocaine, E+σ	006	0.30	0.20	11.5	1.1	0.24	0.37
Cocaine, E+σ	007	0.25	0.19	11.5	-1.4	0.26	0.29
Cocaine, E+σ	008	0.29	0.21	14.8	-1.8	0.26	0.36
Cocaine, E+σ	009	0.31	0.20	24.2	0.4	0.32	0.41
Cocaine, E	010	0.17	0.25	0.40	1.9	0.22	0.37
Cocaine, E	011	0.60	0.26	9.10	-1.9	0.26	0.35
Cocaine, E	012	0.21	0.26	14.8	1.0	0.39	0.37
Cocaine, E	013	0.35	0.20	20.7	0.5	0.47	0.37
Piroxicam, E+σ	014	0.28	0.28	-54.0	-0.3	0.17	0.09
Piroxicam, E+σ	015	0.53	0.49	-64.1	-4.7	0.18	0.27
Ampicillin, calculated, E+σ	016	0.51	0.08	40.6	-0.6	0.14	0.05
Ampicillin, experimental, E+σ	017	0.41	0.05	74.1	-0.4	0.45	0.13
Ampicillin, experimental, E+σ	018	-	-	67.8	9.4	0.37	0.41

2.4 Chemical shift-dependent interaction maps in molecular solids

This section has been adapted with permission from: Cordova, M.; Emsley, L., Chemical Shift-Dependent Interaction Maps in Molecular Solids. *Journal of the American Chemical Society* **2023**, 145 (29), 16109-16117. (post-print)

My contribution was to develop and apply the method and to analyse the results obtained. I also wrote the manuscript, with contribution of the other author.

2.4.1 Introduction

In this section we construct three-dimensional atomic density maps similar to the previously reported full interaction maps (FIMs), constructed from local atomic environments from the CSD database and associated predicted chemical shifts.³⁵⁸ The atomic density maps can be considered as three-dimensional probability functions to find an atom of a given element at a given point in space in the selected environments. By selecting only environments with predicted shifts matching the experimental value, we show how the resulting chemical shift-dependent interaction maps (SIMs) predict key interactions present in the crystal structures of the samples of AZD8329 (form 1 and form 4), decitabine and lisinopril dihydrate studied here. The SIMs obtained are compared to chemical shift-independent interaction maps (IIMs), constructed analogously from local atomic environments selected without targeting a particular chemical shift. The differences between these maps enables the identification of noncovalent interactions either promoted or reduced by applying the chemical shift constraint in the construction of the atomic density maps.

The SIMs presented here are particularly sensitive to hydrogen bonding and to the proximity of aromatic rings in the crystal packing, the latter being related to aromatic ring currents. While nucleus-independent chemical shift (NICS) maps can explain the shifts observed for nuclei in the vicinity of aromatic rings,³⁵⁹⁻³⁶³ the SIMs do not require the three-dimensional structure of the material to predict the presence of neighbouring aromatic rings directly from experimental shifts.

2.4.2 Methods

The method presented here was applied to AZD8329 (form 1 and form 4), decitabine, lisinopril dihydrate and AZD5718. All experimental chemical shifts and crystal structures of the organic crystals studied here have been previously reported.^{53, 172, 177, 364} The database of crystal structures and associated chemical shifts is a subset of the Cambridge Structural Database (CSD)³¹² for which chemical shifts predictions were previously performed using ShiftML,^{176, 261} in order to assign chemical shifts in a probabilistic manner (see **Section 3.2**).³⁵⁸ Here, we recomputed the chemical shifts using the updated model ShiftML2³⁶⁵ and extended the database to all structures available for chemical shift prediction using ShiftML2 as described in Ref. 358. The database now comprises over 338,000 crystal structures.

The construction of the SIM and IIM for a given covalent environment and associated shift involves identifying local atomic environments in the database that match the covalent environment, selecting 1,000 environments either randomly or using the chemical shift as a constraint in the selection process to construct the IIM and SIM, respectively, aligning the selected environments on defined atoms in the covalent environment and extracting the three-dimensional atomic density maps by summing 3D Gaussians placed at each atomic position for each element found in the local atomic environments. The complete procedure is described step-by-step in more detail below. With the current database, the method can in principle be applied to compounds containing any subset of the 12 elements present in the database (H, C, N, O, S, F, P, Cl, Na, Ca, Mg, K).

For each ¹H and ¹³C site, as well as bonded ¹³C-¹H sites in each molecule, corresponding local atomic environments in the database were obtained by identifying covalent environment descriptors matching that of the atomic site. The descriptor is a graph representing atomic species as nodes and covalent bonds as edges for all atoms within *w* bonds away from the central atomic site (detailed in **Section 3.2.2**), as illustrated in **Figure 2.27**. A match is identified by isomorphism between the compared graphs. Importantly, this descriptor does not contain any information about the three-dimensional structure of the molecule nor intermolecular interactions, allowing for searches directly from the molecular (two-dimensional) structure, without requiring knowledge of the geometry of the molecule nor packing in the crystal structure. For each atomic site, we initially set *w* to a value of six, and reduced it until the number of matches was found to be higher than 3,000.

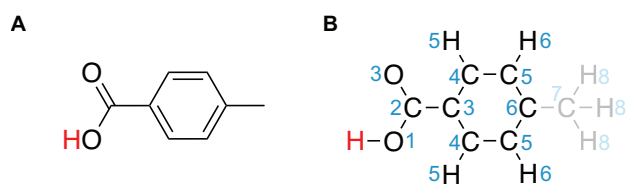


Figure 2.27. (A) molecular structure of 4-methylbenzoic acid and (B) its associated graph representation around the carboxylic acid proton (red) up to $w = 6$ bonds away. Blue numbers indicate the number of bonds away from the red proton for each node (atom) in the graph representation. Atoms further than 6 bonds away are greyed out in (B).

Each database instance contains the crystal structure and atomic site corresponding to the covalent environment descriptor searched for, as well as its associated ShiftML2-predicted chemical shift and predicted uncertainty. The local atomic environment corresponding to each database instance was defined as all atoms within a sphere with a radius of 7 Å centered at the atomic site. When required, the unit cell of crystal structures from the database were repeated in order to completely fill the defined sphere with atoms from the selected structures.

The local atomic environments corresponding to each atomic site were then aligned to a chosen conformation of the molecule under study by minimising the root-mean-square displacement (RMSD) between the positions of selected atoms in the environments through rotation and translation of the whole environments. Although we aligned all environments to the molecular conformation found in the experimental crystal structure of each compound, we note that this alignment can be performed on any conformation without loss of generality, provided that the geometry of the atoms selected for the alignment does not change upon conformational changes. To ensure that, we aligned between three and four atoms, all within at most two bonds of each other, except for rigid molecular motifs such as phenyl rings and carboxylic acids, where we allowed more distant atoms to be aligned. The set of atoms selected for alignment around each atomic site is described in **Tables 2.9-2.20**.

For each atomic site, we randomly selected 1,000 environments to obtain the average environment around the selected atomic site regardless of its chemical shift, and then another 1,000 environments were selected by drawing numbers from a Gaussian distribution centered at the experimental chemical shift and with a width given by the expected uncertainty of the ShiftML2 prediction,³⁶⁵ which corresponds to 0.5 ppm for ^1H and 5 ppm for ^{13}C . For each number drawn, the environment with the closest chemical shift was selected. The environments for bonded ^{13}C - ^1H sites were selected similarly by drawing numbers from a two-dimensional Gaussian distribution centered at the experimental ^{13}C ($\delta_{^{13}\text{C}}^{\text{exp}}$) and ^1H ($\delta_{^1\text{H}}^{\text{exp}}$) shifts and with a width of $\sigma_{^{13}\text{C}} = 5$ ppm and $\sigma_{^1\text{H}} = 0.5$ ppm in the first (^{13}C) and second (^1H) dimensions, respectively. The environment with the closest correlated chemical shift was identified by defining the distance d from the experimental chemical shift as

$$d = \frac{(\delta_{^{13}\text{C}}^{\text{exp}} - \delta_{^{13}\text{C}}^{\text{env}})^2}{\sigma_{^{13}\text{C}}^2} + \frac{(\delta_{^1\text{H}}^{\text{exp}} - \delta_{^1\text{H}}^{\text{env}})^2}{\sigma_{^1\text{H}}^2}, \quad (2.10)$$

where $\delta_{^{13}\text{C}}^{\text{env}}$ and $\delta_{^1\text{H}}^{\text{env}}$ are the ^{13}C and ^1H chemical shifts of the bonded pair of atoms in the environment, respectively.

Three-dimensional atomic density maps were generated by summing three-dimensional Gaussian functions with a width $\sigma = 0.5$ Å placed at the atomic positions \vec{r}_{a_i} of the aligned local environments,

$$G(\vec{r}) = \frac{1}{N_{\text{env}}} \sum_i^{N_{\text{env}}} \sum_{a_i \in i} \exp\left(-\frac{\|\vec{r} - \vec{r}_{a_i}\|^2}{2\sigma^2}\right). \quad (2.11)$$

Individual atomic density maps were constructed for each element present in the set of selected environments. The Gaussian functions were not normalised, and this leads to a value of 1 at a given position if an atom of a given element is found at that position in all selected environments. Each atomic density map was evaluated on a 31x31x31 cubic grid centered at the atomic site and with 12 Å sides. This corresponds to a spatial sampling of 0.4 Å. The size of the grid was chosen to be close to the 7 Å radius sphere used to construct the descriptor to perform chemical shift predictions using ShiftML2.³⁶⁵ The atomic density maps obtained using randomly selected environment represent chemical shift-independent interaction maps (IIMs), and those obtained from environments selected around the measured chemical shifts represent chemical shift-dependent interaction maps (SIMs). All IIMs and SIMs constructed here are shown in **Figures 2.35-2.49**.

The score s_i of a local atomic environment i in a candidate crystal representing its compatibility with the measured chemical shift was evaluated as the overlap between the atomic density map of that local environment $G_i^{\text{cand}}(\vec{r})$ and the difference between the corresponding SIM ($G_i^{\text{SIM}}(\vec{r})$) and IIM ($G_i^{\text{IIM}}(\vec{r})$),

$$s_i = \int G_i^{\text{cand}}(\vec{r}) \cdot [G_i^{\text{SIM}}(\vec{r}) - G_i^{\text{IIM}}(\vec{r})] d\vec{r}. \quad (2.12)$$

This score thus represents how much the local atomic environment is promoted by the SIM compared to the IIM. In practice, we set values in the difference between SIM and IIM at a given point with a magnitude below 0.01 to zero in order to mitigate noise in the difference maps. Here, a positive value of s_i indicates that the corresponding atomic environment is more compatible with the SIM than with the IIM. A value of zero indicates that the candidate is equally promoted by the SIM and the IIM. If the atomic environment is more compatible with the IIM than with the SIM, then a negative value will be obtained. The global score for a candidate crystal was computed as the mean of all considered local atomic environment scores. Here, we discarded the maps that correspond to ambiguous assignments (e.g., aromatic rings and CH_2 groups) from the computation of global scores in order to avoid ambiguities in the scores. Ambiguity arises in such groups due to the mapping of the 2D descriptors to atomic sites in the chosen 3D conformation. It is not possible to determine a priori the assignment of, e.g., the two different protons in a CH_2 group yielding two different chemical shifts without knowledge of the crystal structure.

When comparing sets of candidate structures, we normalised the scores obtained by subtracting the mean score across all candidates from the global score obtained for each candidate. This removes any systematic tendency observed within the set of candidates, leaving only variations between candidates. The final normalised scores obtained thus indicate, within the set of candidate structures considered, which candidates are better matching the SIMs than the IIMs, corresponding to a positive score. While these scores may not be able to definitively identify the correct candidate crystal, they can allow the pre-selection of potential crystal structures by discarding structures displaying strongly negative scores.

2.4.3 Results and Discussion

The method presented here was applied to AZD8329 (forms 1 and 4), decitabine, lisinopril dihydrate, and AZD5718, using the previously reported experimental ^1H and ^{13}C chemical shifts of these compounds.^{53, 172, 177, 364}

For each atomic site considered in each compound, the database was first queried to obtain the local atomic environments matching the covalent environment queried, as well as their associated chemical shift. The IIMs and SIMs were subsequently constructed by selecting 1,000 environments either randomly or with associated shifts close to the experimental value, respectively, as described in **Section 2.4.2**. The whole process can be performed directly from the chemical structure of the molecule studied and the set of assigned chemical shifts, and can thus be performed, e.g., in parallel to the construction of CSP candidates. In general, obtaining each interaction map takes under an hour on a single CPU core and can be straightforwardly parallelised. Once the interaction maps are constructed, computing scores for candidate crystal structures typically takes up to a few seconds per structure, against hours to days of CPU time to obtain chemical shifts using DFT, and scales linearly with the number of atoms in the structure (against a cubic dependence for GIPAW DFT). The method presented here thus provides great potential to facilitate structure determination by NMR.

Figure 2.28 shows the atomic density maps obtained for the carboxylic acid proton of AZD8329 form 1. By aligning 1,000 environments randomly selected regardless of the chemical shifts (**Figure 2.28B**) or such that their predicted chemical shift is the same as the experimental value, to within the prediction error (**Figure 2.28C**), we obtain the atomic density maps shown in **Figure 2.28D-E**. Both maps were found to be similar and to predict a carboxylic acid dimer in at least 20% of the environments aligned. By displaying the difference between the maps obtained with and without the experimental chemical shift of the carboxylic acid proton (**Figure 2.28F**), the dimer was found to be promoted in the ensemble of local atomic environments that match the experimental chemical shift, by at least 5% of the total number of environments aligned. As shown in **Figure 2.28G**, the dimer is indeed present in the crystal structure of AZD8329 form 1, which is consistent with the higher atomic densities found at the positions of the atoms in the dimer in the environments selected around the experimental chemical shift compared to the environments selected regardless of the shift.

As mentioned in **Section 2.4.2**, the maps were aligned to the conformer found in the crystal structure, but can be generated around any conformation, allowing the visualisation of preferred interactions without any prior knowledge of the crystal structure of the compound studied.

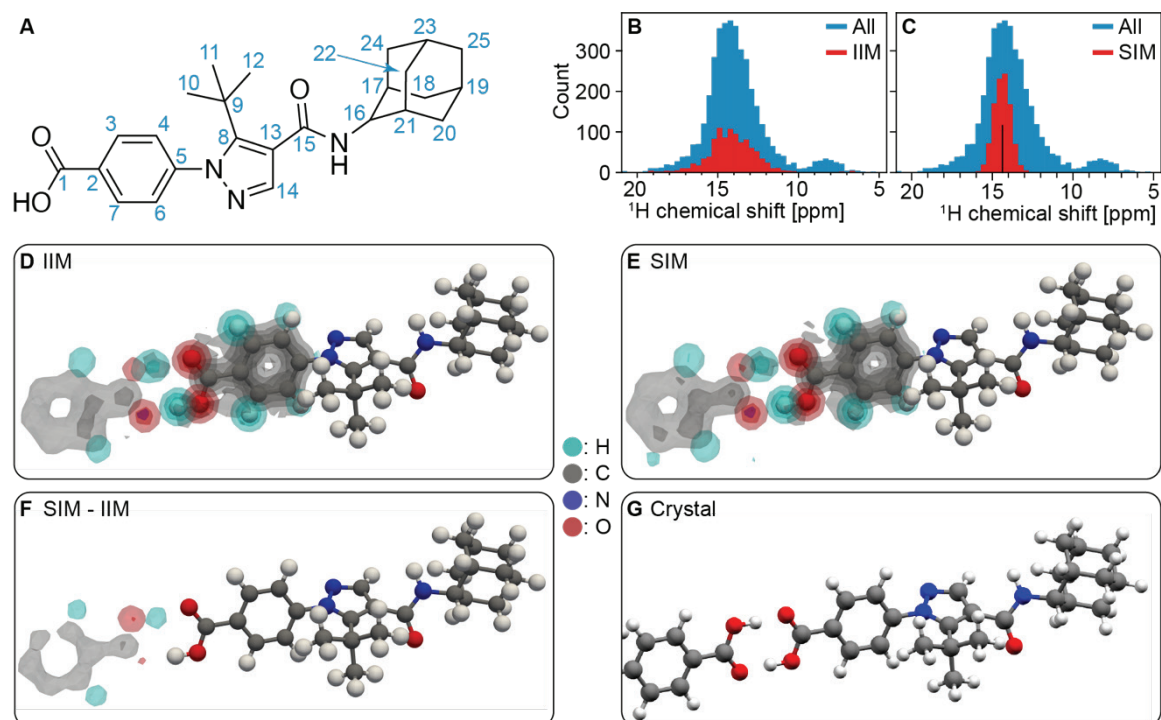


Figure 2.28. (A) Labelling scheme of AZD8329. (B), (C) Histogram of chemical shifts associated with structures from the database matching the covalent environment of proton labelled 1 (blue) and of the 1,000 structures (red) selected either randomly to construct the IIM (B) or sampled around the experimental chemical shift (vertical black line) measured in AZD8329 form 1 to construct the SIM (C). (D), (E) Three-dimensional contour levels of the IIM and SIM of proton 1 in AZD8329 obtained using Equation 2.11 from the structures selected in (B) and (C), respectively. Contour levels are drawn at values of 0.2, 0.4, 0.6 and 0.8. (F) Three-dimensional contour levels of the difference of atomic density between the SIM and IIM. Contour levels are drawn at values of 0.05, 0.1, 0.15 and 0.2. (G) Intermolecular hydrogen bonding motif of the proton labelled 1 in the crystal structure of AZD8329 form 1.

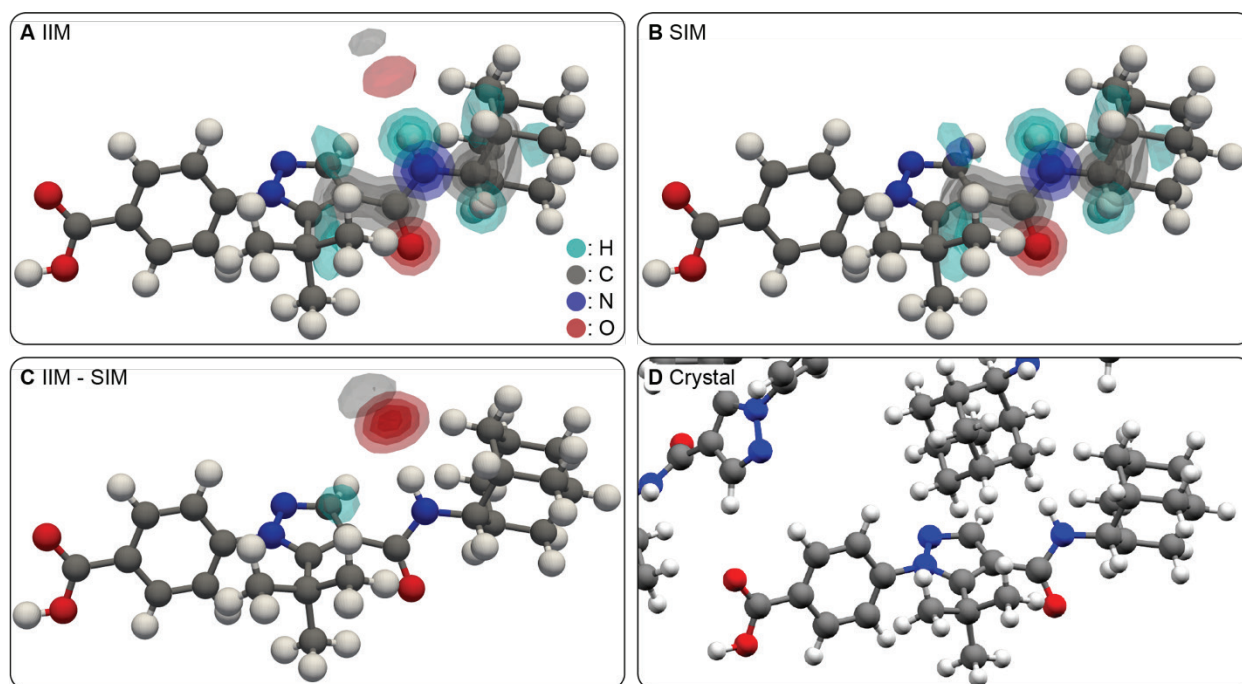


Figure 2.29. (A), (B) Three-dimensional contour levels of the IIM and SIM of the NH proton of AZD8329 form 1, respectively. Contour levels are drawn at values of 0.2, 0.4, 0.6 and 0.8. (C) Three-dimensional contour levels of the difference of atomic density between the IIM and SIM. Contour levels are drawn at values of 0.05, 0.1, 0.15 and 0.2. (D) Local atomic environment of the NH proton in the crystal structure of AZD8329 form 1.

Figure 2.29 and **Figure 2.30** show the atomic density maps obtained around the NH proton in AZD8329 forms 1 and 4, respectively. In form 1 (**Figure 2.29A-B**), the selection of local environments with associated chemical shifts around the experimental value (see **Figure 2.50**) was found to reduce the atomic density of oxygen in contact with the NH proton. The reduction in atomic density corresponds to a difference of at least 20% of the local atomic environments aligned, as seen in the difference map shown in **Figure 2.29C**. We note that **Figure 2.29C** shows the difference between atomic densities obtained from randomly selected environments and those selected around the experimental chemical shift (IIM - SIM), unlike those shown in **Figure 2.28F** and below. This allows us to identify interactions that are less likely than on average when considering the experimental chemical shift. Indeed, the NH proton is not hydrogen bonded in the crystal structure of AZD8329 form 1 (**Figure 2.29D**).

In AZD8329 form 4, the atomic density maps obtained for local atomic environments around the same NH proton with associated chemical shifts close to the experimental value (see **Figure 2.50**) were found to promote hydrogen bonding to oxygen atoms (**Figure 2.30A-C**). This is in agreement with the hydrogen bond found in the crystal structure of form 4 (**Figure 2.30D**).

For Form 4 we also note that in this case, the maps in **Figure 2.30A-B** do not capture the cis conformation of the amide group found in the crystal structure. This suggests that the overwhelming majority of amides in the database display a trans conformation, and/or that the conformation is not captured in the chemical shift of the NH proton. We note that none of the ^1H or ^{13}C shifts considered was able to capture the cis conformation

The atomic density maps obtained around the carboxylic acid proton in AZD8329 form 4, shown in **Figure 2.35**, were found to promote hydrogen bonding of the proton, which is consistent with the crystal structure of the material. However, the difference map was found to promote the carboxylic acid dimer found in the structure of form 1, and which is not present in form 4. This can be explained by bias in the database, where most hydrogen bonded carboxylic acid groups are dimers. Experimental validation of the presence of a carboxylic acid dimer can be obtained using complementary methods such as, e.g., a BABA-xy16 experiment.^{366, 367} The CH protons, as well as carbon environments obtained were not found to promote any significant interaction or conformation in the material. The superposition of interaction maps generated around all ^1H , ^{13}C and ^1H - ^{13}C sites are provided for AZD8329 form 1 and 4 in **Figures 2.35-2.40**.

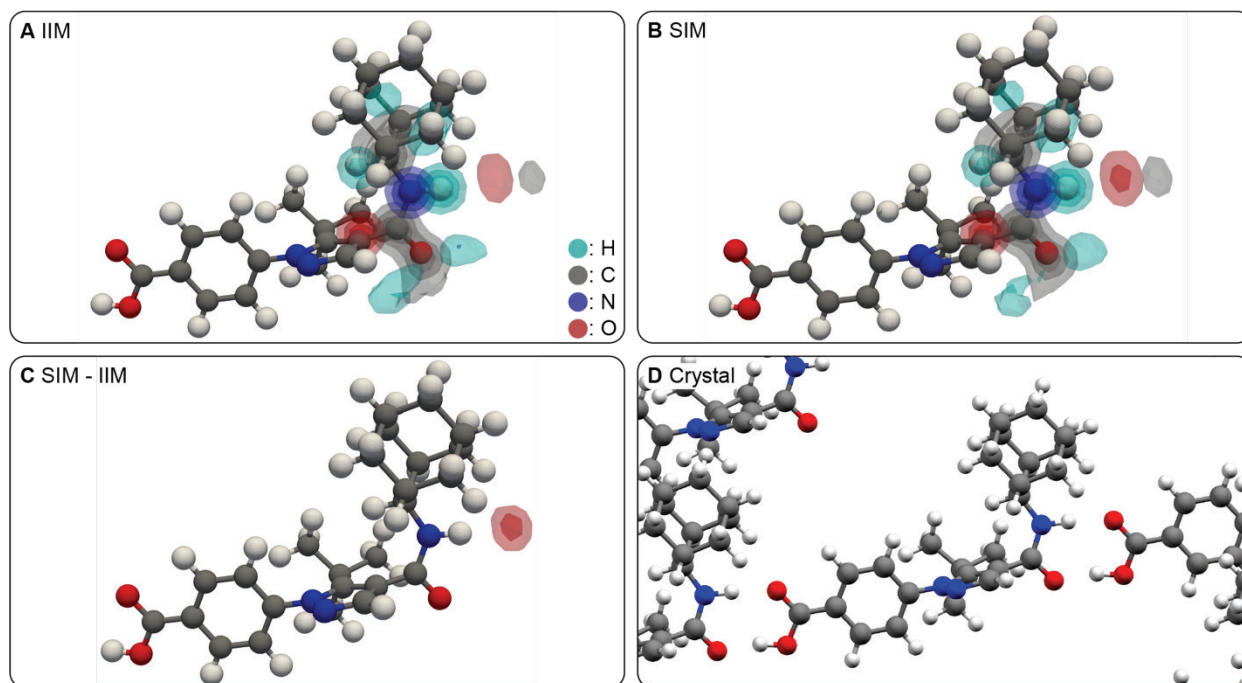


Figure 2.30. (A), (B) Three-dimensional contour levels of the IIM and SIM of the NH proton of AZD8329 form 4, respectively. Contour levels are drawn at values of 0.2, 0.4, 0.6 and 0.8. (C) Three-dimensional contour levels of the difference of atomic density between the SIM and IIM. Contour levels are drawn at values of 0.05, 0.1, 0.15 and 0.2. (D) Local atomic environment of the NH proton in the crystal structure of AZD8329 form 4.

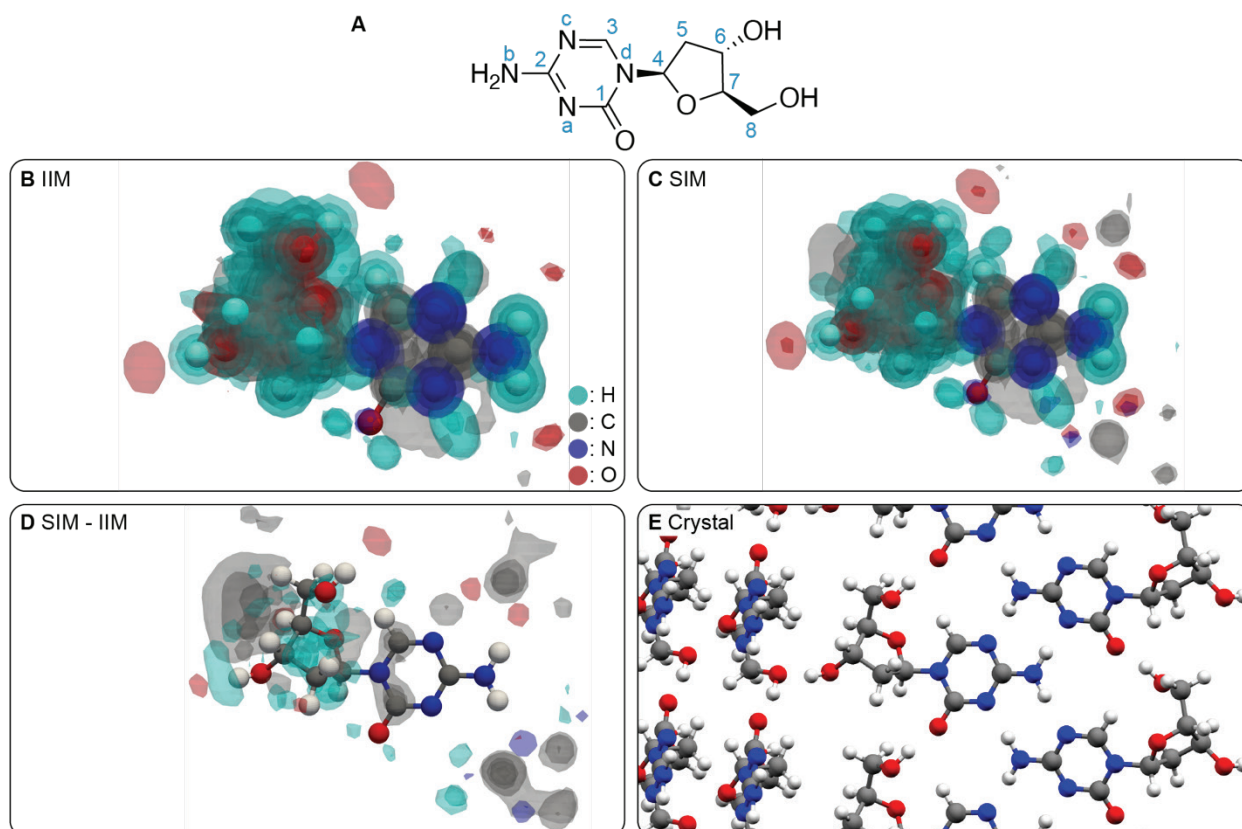


Figure 2.31. (A) labelling scheme of decitabine. (B), (C) Superposition of three-dimensional contour levels of the IIMs and SIMs of all protons in decitabine, respectively. Contour levels are drawn at values of 0.2, 0.4, 0.6 and 0.8. (D) Three-dimensional contour levels of the difference of atomic density between the SIMs and IIMs of each proton in the molecule. Contour levels are drawn at values of 0.05, 0.1, 0.15 and 0.2. (E) Local environment around a decitabine molecule in the crystal structure.

Figure 2.31 shows the atomic density maps generated around all protons in decitabine. Both density maps constructed from randomly selected environments and environments with associated shifts close to experimental values display hydrogen bonding of both protons in the amine, both OH protons, as well as of nitrogens labelled **a** and **c** in **Figure 2.31A**, and of the oxygen labelled **1** in at least 20% of the environments used to construct the atomic density maps (**Figure 2.31B-C**). The difference map shown in **Figure 2.31D** shows that the experimental ^1H chemical shifts are associated with a higher degree of all hydrogen bonding identified above than on average by at least 5% of all environments aligned. This is confirmed in the crystal structure, where all the aforementioned atomic sites are hydrogen bonded. We note that one of the NH_2 protons is expected to be H-bonded to a carboxylic acid moiety in the atomic density map, while it is H-bonded to a nitrogen in the crystal structure.

Figure 2.31D illustrates the limitations of the method presented here. First, the atomic density maps generated do not explicitly identify functional groups. For example, the hydrogen bonding partners of the OH groups in decitabine are not identified in **Figure 2.31D**, which only provides the information that the OH groups are likely to be H-bonded. Nonetheless, the shape of the atomic density maps can be used to infer the bonding partner. In addition, the method presented here is not able to disambiguate intra- or intermolecular interactions. A careful analysis of the flexibility of the molecule can however often establish the possibility of intra-molecular interactions. Another limitation of the method is the identification of the hydrogen bonding acceptors around H-bonded protons. In the case of decitabine here, one of the NH_2 protons is expected to be bonded to a carboxylic acid, although no such functional group is present in the crystal structure. This artifact is due to bias in the database used to construct the atomic density maps, where in this case most environments that match the observed chemical shift display hydrogen bonding interactions with carboxylic acid groups. However, in the absence of such a chemical group in the crystal structure, the most similar group is the aminopyrimidine-like moiety in the molecule, which is the hydrogen bonding partner observed in the crystal structure (**Figure 2.31E**). This interaction could be probed with complementary experiments such as, e.g., a ^{14}N - ^1H d-HMQC experiment.³⁶⁸

In **Figure 2.31B-D**, the proton density found around the $\text{C}=\text{O}$ group is an artifact in the maps constructed for the NH_2 protons, which predict an NH_2 group instead of the oxygen next to the carbon labelled **1**. This is due to bias in the database. The superposition of interaction maps generated around all ^{13}C and ^1H - ^{13}C sites are provided for decitabine in **Figures 2.42-2.43**.

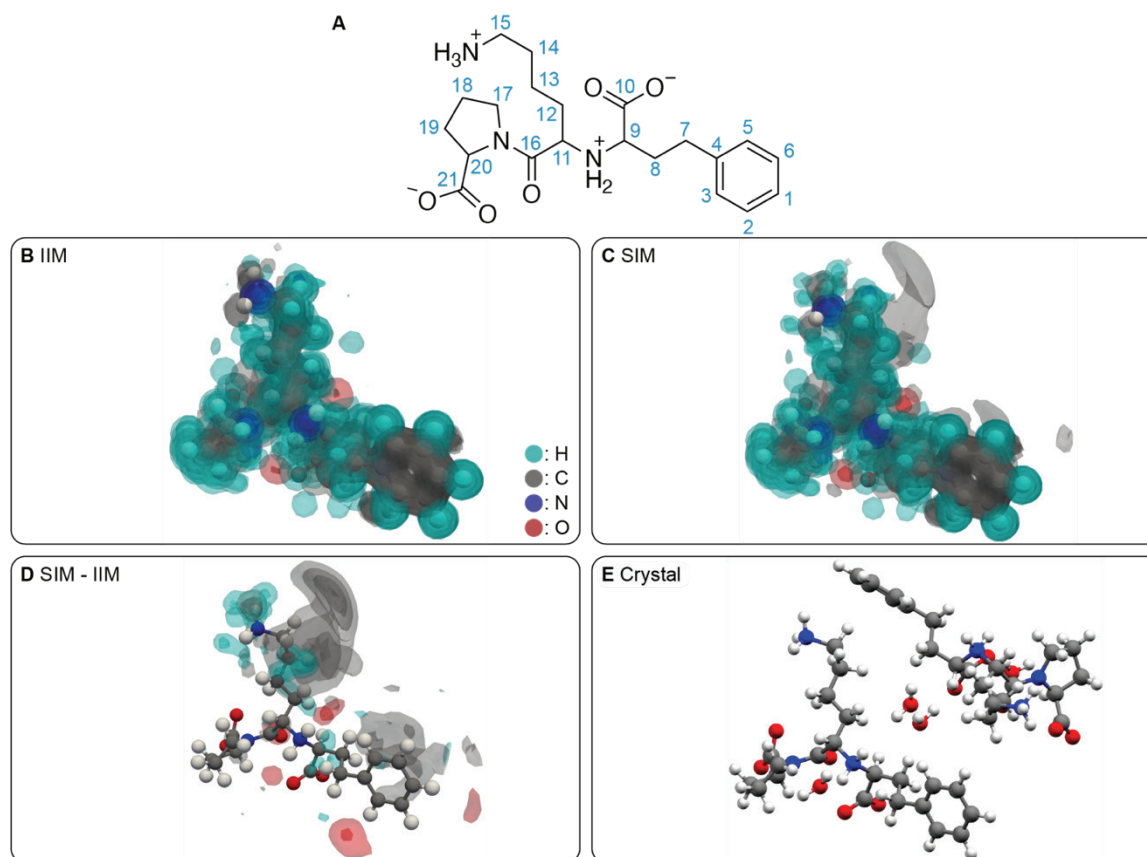


Figure 2.32. (A) labelling scheme of lisinopril dihydrate. (B), (C) Superposition of three-dimensional contour levels of the IIMs and SIMs of all protons in lisinopril, respectively. Contour levels are drawn at values of 0.2, 0.4, 0.6 and 0.8. (D) Three-dimensional contour levels of the difference of atomic density between the SIMs and IIMs of each proton in the molecule. Contour levels are drawn at values of 0.05, 0.1, 0.15 and 0.2. (E) Local environment around a lisinopril molecule in the crystal structure.

Figure 2.32 shows the atomic density maps generated around all protons in lisinopril dihydrate (excluding water protons, since their chemical shift was not reported). The map generated around the CH₂ protons labelled 15 using environments selected to have chemical shifts close to experiment (**Figure 2.32C**) displays a clear presence of carbon atomic density close to the protons, which is absent in the map generated using random local atomic environments (**Figure 2.32B**). This is confirmed in the difference map (**Figure 2.32D**), and corresponds to the presence of the phenyl ring of a neighbouring lisinopril molecule. The unusually low shift of one of the CH₂ protons (see **Table 2.15** and **Figure 2.51**) is associated with the presence of an aromatic ring in its vicinity, whose ring currents induce an increased shielding of the proton. This effect has previously been extensively studied in the context of nucleus independent chemical shift (NICS).^{151, 359-363} The superposition of interaction maps generated around all ¹³C and ¹H-¹³C sites are provided for lisinopril dihydrate in **Figures 2.45-2.46**.

The atomic density maps presented here can be used to qualitatively evaluate the likelihood of candidate structures in chemical shift-based structure determination, or can serve as the basis for the derivation of structural constraints in CSP protocols. In addition, we introduce a quantitative measure of the likelihood of candidate crystal structures based on the atomic density maps generated (see **Section 2.4.2**). **Figure 2.33A-B** shows the scores obtained for the X-ray structures of forms 1 and 4 of AZD8329 when evaluated using the maps generated from the experimental ¹H chemical shifts of all unambiguously assigned protons (see **Figure 2.52**). In addition, the evaluation of a set of ten candidate structures is shown for AZD8329 form 4. The SIMs constructed from the experimental shifts of form 1 correctly lead to a higher score for the X-ray structure of form 1 compared to form 4 (**Figure 2.33A**). In addition, using SIMs derived from the experimental shifts of AZD8329 form 4 led to the correct identification of the X-ray structure of form 4 and candidate #1 in the CSP set to have the highest scores compared to the X-ray structure of form 1 and the other CSP candidates (**Figure 2.33B**). This indicates that the method is able to identify the correct polymorphic form of AZD8329 based on experimental chemical shifts only, and highlights the ability of SIMs to identify the correct crystal structure among a set of candidates directly from the experimentally measured chemical shifts, and without the need to perform any chemical shift computation for any candidate among the set. Using ¹³C or both ¹H and ¹³C chemical shifts from AZD8329 form 1 similarly leads to a higher score for the X-ray structure form 1 compared to form 4, however using ¹³C or both ¹H and ¹³C chemical shifts from AZD8329 form 4 did not attribute the highest score to candidate #1 (see **Figure 2.52**).

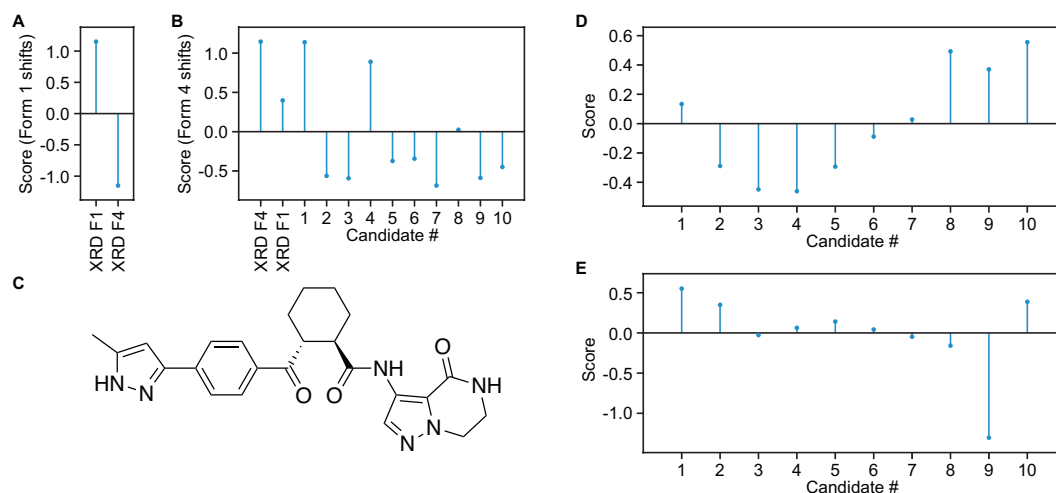


Figure 2.33. Scores obtained as described in Equation 2.12 and averaged over all atomic environments considered for the X-ray structures of AZD8329 forms 1 and 4 using SIMs constructed using the experimentally obtained chemical shifts of (A) AZD8329 form 1 and (B) form 4. In (B), the scores obtained for a CSP set of ten candidate structures of AZD8329 form 4 are also shown. Chemical structure of AZD5718 (C) and scores obtained for candidate structures of AZD5718 using SIMs constructed using the experimentally obtained (D) ^1H , and (E) ^1H and ^{13}C chemical shifts.

Figure 2.33C-E shows the scores obtained for a CSP set of candidate structures of AZD5718 using unambiguously assigned experimental ^1H (Figure 2.33D) and ^1H and ^{13}C (Figure 2.33E) chemical shifts. In this case, using protons only did not identify candidate #1 (i.e., the correct candidate) as having the highest score. However, adding ^{13}C chemical shifts led to the correct identification of candidate #1 as best matching (see Figure 2.52). The superposition of interaction maps generated around all ^1H , ^{13}C and ^1H - ^{13}C sites are provided for AZD5718 in Figures 2.47-2.49. Not unexpectedly, the scores display a weaker discriminating power as compared to DFT chemical shift computation of the candidate structures and comparison to experiments,^{53, 177} so far, and further work will focus on improving the robustness of candidate scoring.

We note that here all CSP candidate structures of AZD8329 form 4 were originally selected by Baías *et al.*⁵³ within 30 kJ/mol in total energy from the most stable predicted crystal structure with the *cis* conformation of the amide group, and ordered by increasing energy. While the lowest energy candidate corresponds to the X-ray structure of AZD8329 form 4, it lies well above the lowest energy candidate generated with a *trans* conformation of the amide group. For AZD5718, the ten candidate crystal structures were previously selected within 6 kJ/mol from the lowest energy candidate generated,¹⁷⁷ and are ordered by increasing energy. In general, there is no guarantee that the lowest energy candidate corresponds to the observed structure, and this is evident for polymorphic compounds that display several observed structures with different energies. The IIMs and the SIMs generated here do not incorporate any information or bias related to predicted energies.

2.4.4 Conclusion

In this section we have developed a method to obtain three-dimensional atomic density maps of local atomic environments based on the experimental chemical shift associated to the covalent environment queried. The maps constructed can be used to visualise preferred noncovalent interactions in molecular solids directly from any random conformation of the compound studied, without requiring any prior knowledge about the conformation of molecular packing in the solid state. This can be used to qualitatively evaluate the likelihood of candidate crystal structures in chemical shift-based structure determination, or to derive experimentally derived structural constraints in CSP protocols. It can also be used to generate structural hypotheses that can guide further experimental validations. We have also introduced a scoring system able to quantitatively evaluate candidate crystal structures based on experimental chemical shifts, which was found able to identify the correct candidate.

While we believe that the method presented here presents great potential to facilitate the structure determination of molecular solids by NMR, we expect it to become more powerful in the future, using larger and more diverse databases of structures with more accurate chemical shifts associated. Using larger and more diverse databases would also allow the use of the method for a broader range of compounds. Finally, we expect that managing bias in the database (e.g., the over-representation of particular functional groups) would allow the construction of more accurate SIMs.

The approach presented here is not limited to crystalline compounds, and can be used straightforwardly to identify preferred non-covalent interactions in disordered materials, by using experimental chemical shifts from such disordered samples and adapting the width of the shift distributions to match the observed lineshapes, potentially made more accurate by using a database comprising distorted structures.

2.4.5 Appendix III

Data availability

All data and code used are available from <https://doi.org/10.24435/materialscloud:98-sx> under the license CC-BY-4.0 (Creative Commons Attribution-ShareAlike 4.0 International)

Experimental Details

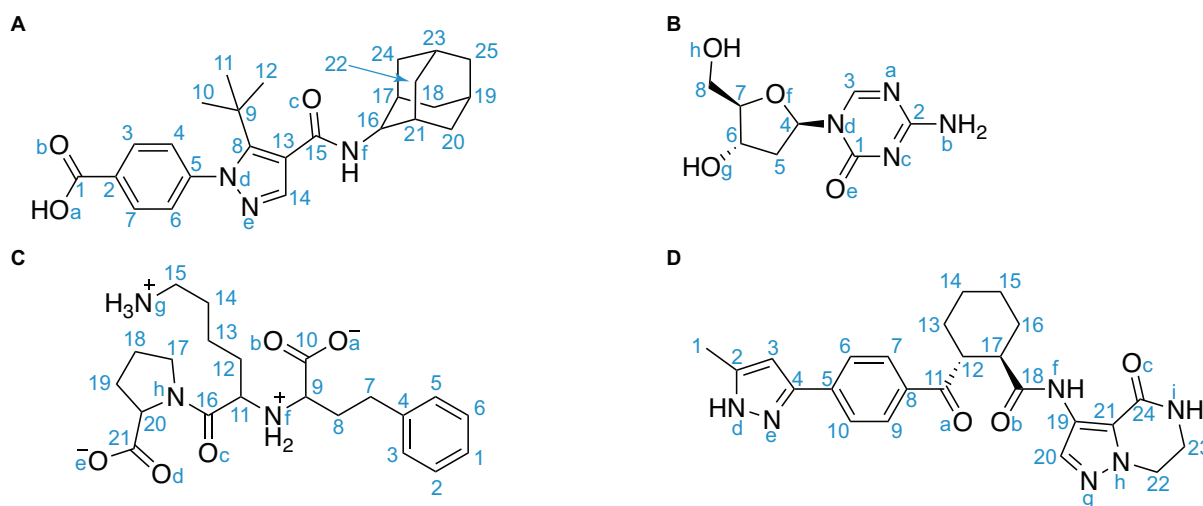


Figure 2.34. Labelling scheme of (A) AZD8329, (B) decitabine, (C) lisinopril dihydrate and (D) AZD5718.

Table 2.9. Experimental ^1H chemical shifts and atoms aligned for AZD8329.

Label	Experimental shift, Form 1 / Form 4 [ppm]	Atoms aligned
H1	14.37 / 15.37	H ^a , O ^a , O ^b , C ¹
H3	8.46 / 9.01	H ³ , C ³ , O ^a , O ^b
H4	7.08 / 8.47	H ⁴ , C ⁴ , N ^d , C ¹
H6	8.46 / 6.92	H ⁶ , C ⁶ , N ^d , C ¹
H7	8.46 / 8.69	H ⁷ , C ⁷ , O ^a , O ^b
H10	1.01 / 0.73	C ¹⁰ , C ⁹ , C ⁸
H11	1.01 / 0.73	C ¹¹ , C ⁹ , C ⁸
H12	1.01 / 0.73	C ¹² , C ⁹ , C ⁸
H14	8.28 / 7.73	H ¹⁴ , C ¹⁴ , C ¹³ , N ^e
NH	6.96 / 9.64	H ^f , N ^f , C ¹⁵ , C ¹⁶
H16	4.39 / 2.90	H ¹⁶ , C ¹⁶ , N ^f
H17	1.64 / 1.54	H ¹⁷ , C ¹⁷ , C ¹⁶
H18	1.64 / 1.60	C ¹⁸ , C ¹⁷ , C ¹⁹
H18'	0.89 / 0.44	C ¹⁸ , C ¹⁷ , C ¹⁹
H19	0.82 / 1.00	H ¹⁹ , C ¹⁹ , C ¹⁸
H20	1.64 / 0.80	C ²⁰ , C ²¹ , C ¹⁹
H20'	0.89 / 0.80	C ²⁰ , C ²¹ , C ¹⁹
H21	2.12 / 1.78	H ²¹ , C ²¹ , C ¹⁶
H22	0.82 / 1.88	C ²² , C ²¹ , C ²³
H22'	1.58 / 1.88	C ²² , C ²¹ , C ²³
H23	1.49 / 1.80	H ²³ , C ²³ , C ²²
H24	2.12 / 1.88	C ²⁴ , C ²³ , C ¹⁷
H24'	1.83 / 1.88	C ²⁴ , C ²³ , C ¹⁷
H25	0.82 / 1.74	C ²⁵ , C ²³ , C ¹⁹
H25'	-0.03 / 1.74	C ²⁵ , C ²³ , C ¹⁹

Table 2.10. Experimental ^{13}C chemical shifts and atoms aligned for AZD8329.

Label	Experimental shift, Form 1 / Form 4 [ppm]	Atoms aligned
C1	173.60 / 171.04	C ¹ , O ^a , O ^b , H ^a
C2	133.27 / 131.10	C ² , C ³ , H ³ , O ^b
C3	131.50 / 133.01	C ³ , H ³ , O ^a , O ^b
C4	127.00 / 128.05	C ⁴ , H ⁴ , N ^d , C ¹
C5	148.27 / 147.31	C ⁵ , N ^d , C ⁴ , C ⁶
C6	128.32 / 128.05	C ⁶ , H ⁶ , N ^d , C ¹
C7	131.50 / 130.48	C ⁷ , H ⁷ , O ^a , O ^b
C8	151.97 / 148.71	C ⁸ , N ^d , C ⁹ , C ¹³
C9	34.20 / 33.42	C ⁹ , C ⁸ , C ¹⁰
C10	30.13 / 29.53	C ¹⁰ , C ⁹ , C ⁸
C11	30.13 / 29.53	C ¹¹ , C ⁹ , C ⁸
C12	30.13 / 29.53	C ¹² , C ⁹ , C ⁸
C13	119.17 / 114.10	C ¹³ , C ⁸ , C ¹⁴ , C ¹⁵
C14	139.16 / 138.43	C ¹⁴ , N ^e , C ¹³ , H ¹⁴
C15	165.41 / 172.98	C ¹⁵ , N ^f , O ^c
C16	55.24 / 60.16	C ¹⁶ , N ^f , H ¹⁶
C17	32.13 / 32.45	C ¹⁷ , C ¹⁶ , C ¹⁸ , C ²⁴
C18	32.13 / 30.80	C ¹⁸ , C ¹⁷ , C ¹⁹
C19	27.26 / 27.81	C ¹⁹ , C ¹⁸ , C ²⁰ , C ²⁵
C20	32.13 / 30.80	C ²⁰ , C ²¹ , C ¹⁹
C21	32.79 / 34.14	C ²¹ , C ¹⁶ , C ²⁰ , C ²²
C22	37.32 / 37.41	C ²² , C ²¹ , C ²³
C23	26.93 / 27.81	C ²³ , C ²² , C ²⁴ , C ²⁵
C24	38.83 / 36.42	C ²⁴ , C ²³ , C ¹⁷
C25	37.13 / 37.41	C ²⁵ , C ²³ , C ¹⁹

Table 2.11. Experimental ^{13}C - ^1H chemical shifts and atoms aligned for AZD8329.

Label	Experimental shift, Form 1 / Form 4 [ppm]	Atoms aligned
C3-H3	131.50, 8.46 / 133.01, 9.01	C ³ , H ³ , O ^a , O ^b
C4-H4	127.00, 7.08 / 128.05, 8.47	C ⁴ , H ⁴ , N ^d , C ¹
C6-H6	128.32, 8.46 / 128.05, 6.92	C ⁶ , H ⁶ , N ^d , C ¹
C7-H7	131.50, 8.46 / 130.48, 8.69	C ⁷ , H ⁷ , O ^a , O ^b
C10-H10	30.13, 1.01 / 29.53, 0.73	C ¹⁰ , C ⁹ , C ⁸
C11-H11	30.13, 1.01 / 29.53, 0.73	C ¹¹ , C ⁹ , C ⁸
C12-H12	30.13, 1.01 / 29.53, 0.73	C ¹² , C ⁹ , C ⁸
C14-H14	139.16, 8.28 / 138.43, 7.73	C ¹⁴ , N ^e , C ¹³ , H ¹⁴
C16-H16	55.24, 4.39 / 60.16, 2.90	C ¹⁶ , N ^f , H ¹⁶
C17-H17	32.13, 1.64 / 32.45, 1.54	C ¹⁷ , C ¹⁶ , C ¹⁸ , C ²⁴
C18-H18	32.13, 1.64 / 30.80, 1.60	C ¹⁸ , C ¹⁷ , C ¹⁹
C18-H18'	32.13, 0.89 / 30.80, 0.44	C ¹⁸ , C ¹⁷ , C ¹⁹
C19-H19	27.26, 0.82 / 27.81, 1.00	C ¹⁹ , C ¹⁸ , C ²⁰ , C ²⁵
C20-H20	32.13, 1.64 / 30.80, 0.80	C ²⁰ , C ²¹ , C ¹⁹
C20-H20'	32.13, 0.89 / 30.80, 0.80	C ²⁰ , C ²¹ , C ¹⁹
C21-H21	32.79, 2.12 / 34.14, 1.78	C ²¹ , C ¹⁶ , C ²⁰ , C ²²
C22-H22	37.32, 0.82 / 37.41, 1.88	C ²² , C ²¹ , C ²³
C22-H22'	37.32, 1.58 / 37.41, 1.88	C ²² , C ²¹ , C ²³
C23-H23	26.93, 1.49 / 27.81, 1.80	C ²³ , C ²² , C ²⁴ , C ²⁵
C24-H24	38.83, 2.12 / 36.42, 1.88	C ²⁴ , C ²³ , C ¹⁷
C24-H24'	38.83, 1.83 / 36.42, 1.88	C ²⁴ , C ²³ , C ¹⁷
C25-H25	37.13, 0.82 / 37.41, 1.74	C ²⁵ , C ²³ , C ¹⁹
C25-H25'	37.13, -0.03 / 37.41, 1.74	C ²⁵ , C ²³ , C ¹⁹

Table 2.12. Experimental ^1H chemical shifts and atoms aligned for decitabine.

Label	Experimental shift [ppm]	Atoms aligned
N ^b H	9.38	N ^b , C ² , N ^a , N ^c
N ^b H'	10.81	N ^b , C ² , N ^a , N ^c
H3	8.30	H ³ , N ^a , N ^d
H4	5.66	H ⁴ , C ⁴ , O ^f , N ^d
H5	1.83	C ⁵ , C ⁴ , C ⁶
H5'	1.96	C ⁵ , C ⁴ , C ⁶
H6	4.08	H ⁶ , C ⁶ , O ^g
O ^h H	5.90	O ^h H, O ^g , C ⁶
H7	3.33	H ⁷ , C ⁷ , O ^f
H8	3.91	C ⁸ , O ^h , C ⁷
H8'	3.36	C ⁸ , O ^h , C ⁷
O ^h H	5.90	O ^h H, O ^h , C ⁸

Table 2.13. Experimental ^{13}C chemical shifts and atoms aligned for decitabine.

Label	Experimental shift [ppm]	Atoms aligned
C1	153.75	C1, Oe, Nc, Nd
C2	165.97	C2, Na, Nb, Nc
C3	153.75	C3, Na, Nd
C4	88.61	C4, Of, Nd, C5
C5	44.97	C5, C4, C6
C6	72.23	C6, Og, H6
C7	98.73	C7, Of, H7
C8	62.12	C8, Oh, C7

Table 2.14. Experimental ^{13}C - ^1H chemical shifts and atoms aligned for decitabine.

Label	Experimental shift [ppm]	Atoms aligned
C3-H3	153.75, 8.30	C ³ , N ^a , N ^d
C4-H4	88.61, 5.66	C ⁴ , O ^f , N ^d , C ⁵
C5-H5	44.97, 1.83	C ⁵ , C ⁴ , C ⁶
C5-H5'	44.97, 1.96	C ⁵ , C ⁴ , C ⁶
C6-H6	72.23, 4.08	C ⁶ , O ^g , H ⁶
C7-H7	98.73, 3.33	C ⁷ , O ^f , H ⁷
C8-H8	62.12, 3.91	C ⁸ , O ^h , C ⁷
C8-H8'	62.12, 3.36	C ⁸ , O ^h , C ⁷

Table 2.15. Experimental ^1H chemical shifts and atoms aligned for lisinopril dihydrate.

Label	Experimental shift [ppm]	Atoms aligned
H1	7.8	H ¹ , C ² , C ⁷
H2	6.3	H ² , C ² , C ⁷
H3	7.6	H ³ , C ³ , C ⁷
H5	7.9	H ⁵ , C ⁵ , C ⁷
H6	7.6	H ⁶ , C ⁶ , C ⁷
H7	3.8	C ⁷ , C ⁴ , C ⁸
H8	2.1	C ⁸ , C ⁷ , C ⁹
H9	4.6	H ⁹ , C ⁹ , N ^f
N ^f H	11.3	N ^f , C ⁹ , C ¹¹
H11	4.5	H ¹¹ , C ¹¹ , N ^f
H12	1.7	C ¹² , C ¹¹ , C ¹³
H13	0.7	C ¹³ , C ¹² , C ¹⁴
H14	0.2	C ¹⁴ , C ¹³ , C ¹⁵
H14'	1.5	C ¹⁴ , C ¹³ , C ¹⁵
H15	0.2	C ¹⁵ , C ¹⁴ , N ^g
H15'	2.6	C ¹⁵ , C ¹⁴ , N ^g
H17	5.2	C ¹⁷ , N ^h , C ¹⁸
H18	1.6	C ¹⁸ , C ¹⁷ , C ¹⁹
H19	1.6	C ¹⁹ , C ¹⁸ , C ²⁰
H20	4.4	H ²⁰ , C ²⁰ , N ^h

Table 2.16. Experimental ^{13}C chemical shifts and atoms aligned for lisinopril dihydrate.

Label	Experimental shift [ppm]	Atoms aligned
C1	127.4	C ¹ , C ² , C ⁷
C2	128.7	C ² , H ² , C ⁷
C3	130.1	C ³ , H ³ , C ⁷
C4	142.3	C ⁴ , C ³ , H ³
C5	128.2	C ⁵ , H ⁵ , C ⁷
C6	130.1	C ⁶ , H ⁶ , C ⁷
C7	30.9	C ⁷ , C ⁴ , C ⁸
C8	35.2	C ⁸ , C ⁷ , C ⁹
C9	56.4	C ⁹ , N ^f , H ⁹
C10	173.9	C ¹⁰ , O ^a , O ^b
C11	54.6	C ¹¹ , N ^f , H ¹¹
C12	28.3	C ¹² , C ¹¹ , C ¹³
C13	18.9	C ¹³ , C ¹² , C ¹⁴
C14	27.2	C ¹⁴ , C ¹³ , C ¹⁵
C15	35.9	C ¹⁵ , C ¹⁴ , N ^g
C16	164.4	C ¹⁶ , O ^c , N ^h
C17	47.6	C ¹⁷ , N ^h , C ¹⁸
C18	25.3	C ¹⁸ , C ¹⁷ , C ¹⁹
C19	30.9	C ¹⁹ , C ¹⁸ , C ²⁰
C20	61.2	C ²⁰ , N ^h , H ²⁰
C21	175.7	C ²¹ , O ^d , O ^e

Table 2.17. Experimental ^{13}C - ^1H chemical shifts and atoms aligned for lisinopril dihydrate.

Label	Experimental shift [ppm]	Atoms aligned
C1-H1	127.4, 7.8	C^1 , C^2 , C^7
C2-H2	128.7, 6.3	C^2 , H^2 , C^7
C3-H3	130.1, 7.6	C^3 , H^3 , C^7
C5-H5	128.2, 7.9	C^5 , H^5 , C^7
C6-H6	130.1, 7.6	C^6 , H^6 , C^7
C7-H7	30.9, 3.8	C^7 , C^4 , C^8
C8-H8	35.2, 2.1	C^8 , C^7 , C^9
C9-H9	56.4, 4.6	C^9 , N^f , H^9
C11-H11	54.6, 4.5	C^{11} , N^f , H^{11}
C12-H12	28.3, 1.7	C^{12} , C^{11} , C^{13}
C13-H13	18.9, 0.7	C^{13} , C^{12} , C^{14}
C14-H14	27.2, 0.2	C^{14} , C^{13} , C^{15}
C14-H14'	27.2, 1.5	C^{14} , C^{13} , C^{15}
C15-H15	35.9, 0.2	C^{15} , C^{14} , N^g
C15-H15'	35.9, 2.6	C^{15} , C^{14} , N^g
C17-H17	47.6, 5.2	C^{17} , N^h , C^{18}
C18-H18	25.3, 1.6	C^{18} , C^{17} , C^{19}
C19-H19	30.9, 1.6	C^{19} , C^{18} , C^{20}
C20-H20	61.2, 4.4	C^{20} , N^h , H^{20}

Table 2.18. Experimental ^1H chemical shifts and atoms aligned for AZD5718.

Label	Experimental shift [ppm]	Atoms aligned
H1	1.2	C ¹ , C ² , C ³ , N ^d
H3	5.8	H ³ , C ³ , C ² , C ⁴
N ^d H	10.6	N ^d H, N ^d , N ^e , C ²
H6	6.9	H ⁶ , C ⁶ , C ⁴
H7	6.7	H ⁷ , C ⁷ , C ¹¹
H9	7.0	H ⁹ , C ⁹ , C ¹¹
H10	7.3	H ¹⁰ , C ¹⁰ , C ⁴
H12	3.9	H ¹² , C ¹² , C ¹¹
H13	0.0	C ¹³ , C ¹² , C ¹⁴
H13'	1.7	C ¹³ , C ¹² , C ¹⁴
H14	-0.5	C ¹⁴ , C ¹³ , C ¹⁵
H14'	0.8	C ¹⁴ , C ¹³ , C ¹⁵
H15	-0.5	C ¹⁵ , C ¹⁴ , C ¹⁶
H15'	0.8	C ¹⁵ , C ¹⁴ , C ¹⁶
H16	1.6	C ¹⁶ , C ¹⁵ , C ¹⁷
H16'	1.6	C ¹⁶ , C ¹⁵ , C ¹⁷
H17	1.6	H ¹⁷ , C ¹⁷ , C ¹⁸
N ^f H	7.7	N ^f H, N ^f , C ¹⁸ , C ¹⁹
H20	7.6	H ²⁰ , C ²⁰ , N ^g , C ¹⁹
H22	1.7	C ²² , N ^h , C ²³
H22'	2.7	C ²² , N ^h , C ²³
H23	1.9	C ²³ , N ⁱ , C ²²
H23'	2.7	C ²³ , N ⁱ , C ²²
N ⁱ H	6.9	N ⁱ H, N ⁱ , C ²³ , C ²⁴

Table 2.19. Experimental ^{13}C chemical shifts and atoms aligned for AZD5718.

Label	Experimental shift [ppm]	Atoms aligned
C1	11.1	C ¹ , C ² , C ³ , N ^d
C2	141.5	C ² , C ¹ , C ³ , N ^d
C3	102.3	C ³ , C ² , C ⁴
C4	149.8	C ⁴ , C ³ , C ⁵ , N ^e
C5	139.5	C ⁵ , C ⁴ , C ⁶ , N ^e
C6	123.9	C ⁶ , H ⁶ , C ⁴
C7	130.1	C ⁷ , H ⁷ , C ¹¹
C8	133.3	C ⁸ , O ^a , C ⁷
C9	130.8	C ⁹ , H ⁹ , C ¹¹
C10	125.3	C ¹⁰ , H ¹⁰ , C ⁴
C11	201.1	C ¹¹ , O ^a , C ¹² , C ⁸
C12	46.3	C ¹² , H ¹² , C ¹¹
C13	31.2	C ¹³ , C ¹² , C ¹⁴
C14	26.6	C ¹⁴ , C ¹³ , C ¹⁵
C15	26.0	C ¹⁵ , C ¹⁴ , C ¹⁶
C16	29.2	C ¹⁶ , C ¹⁵ , C ¹⁷
C17	49.8	C ¹⁷ , H ¹⁷ , C ¹⁶
C18	174.0	C ¹⁸ , O ^b , N ^f
C19	125.8	C ¹⁹ , N ^f , C ²⁰ , C ²¹
C20	130.8	C ²⁰ , H ²⁰ , N ^g , C ¹⁹
C21	119.7	C ²¹ , N ^h , C ¹⁹ , C ²⁴
C22	43.5	C ²² , N ^h , C ²³
C23	40.1	C ²³ , N ⁱ , C ²²
C24	161.8	C ²⁴ , C ²¹ , O ^c , N ⁱ

Table 2.20. Experimental ^{13}C - ^1H chemical shifts and atoms aligned for AZD5718.

Label	Experimental shift [ppm]	Atoms aligned
C1-H1	11.1, 1.2	C^1 , C^2 , C^3 , N^d
C3-H3	102.3, 5.8	C^3 , C^2 , C^4
C6-H6	123.9, 6.9	C^6 , H^6 , C^4
C7-H7	130.1, 6.7	C^7 , H^7 , C^{11}
C9-H9	130.8, 7.0	C^9 , H^9 , C^{11}
C10-H10	125.3, 7.3	C^{10} , H^{10} , C^4
C12-H12	46.3, 3.9	C^{12} , H^{12} , C^{13}
C13-H13	31.2, 0.0	C^{13} , C^{12} , C^{14}
C13-H13'	31.2, 1.7	C^{13} , C^{12} , C^{14}
C14-H14	26.6, -0.5	C^{14} , C^{13} , C^{15}
C14-H14'	26.6, 0.8	C^{14} , C^{13} , C^{15}
C15-H15	26.0, -0.5	C^{15} , C^{14} , C^{16}
C15-H15'	26.0, 0.8	C^{15} , C^{14} , C^{16}
C16-H16	29.2, 1.6	C^{16} , C^{15} , C^{17}
C16-H16'	29.2, 1.6	C^{16} , C^{15} , C^{17}
C17-H17	49.8, 1.6	C^{17} , H^{17} , C^{16}
C20-H20	130.8, 7.6	C^{20} , H^{20} , N^e , C^{19}
C22-H22	43.5, 1.7	C^{22} , N^h , C^{23}
C22-H22'	43.5, 2.7	C^{22} , N^h , C^{23}
C23-H23	40.1, 1.9	C^{23} , N^i , C^{22}
C23-H23'	40.1, 2.7	C^{23} , N^i , C^{22}

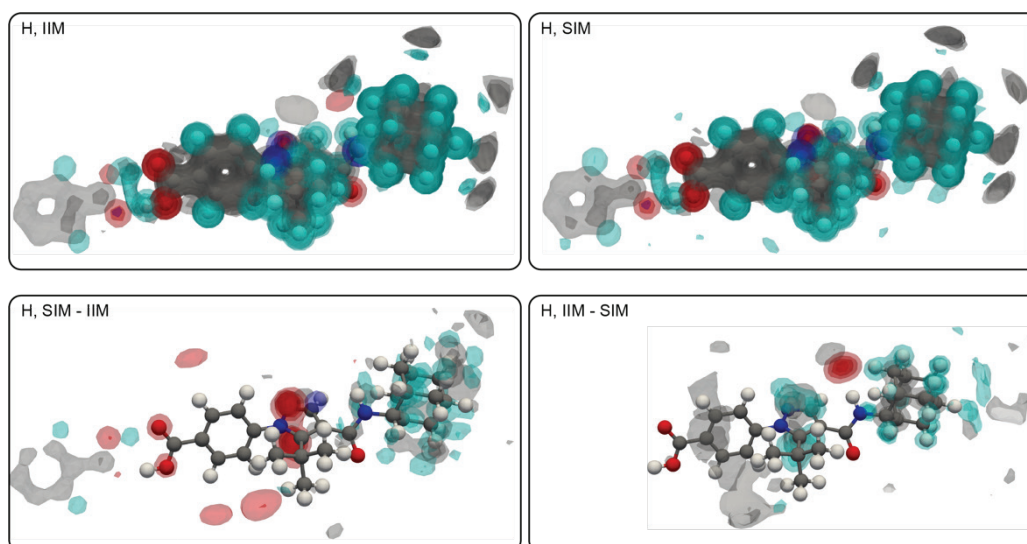


Figure 2.35. Interaction maps of AZD8329 Form 1 based on ^1H chemical shifts.

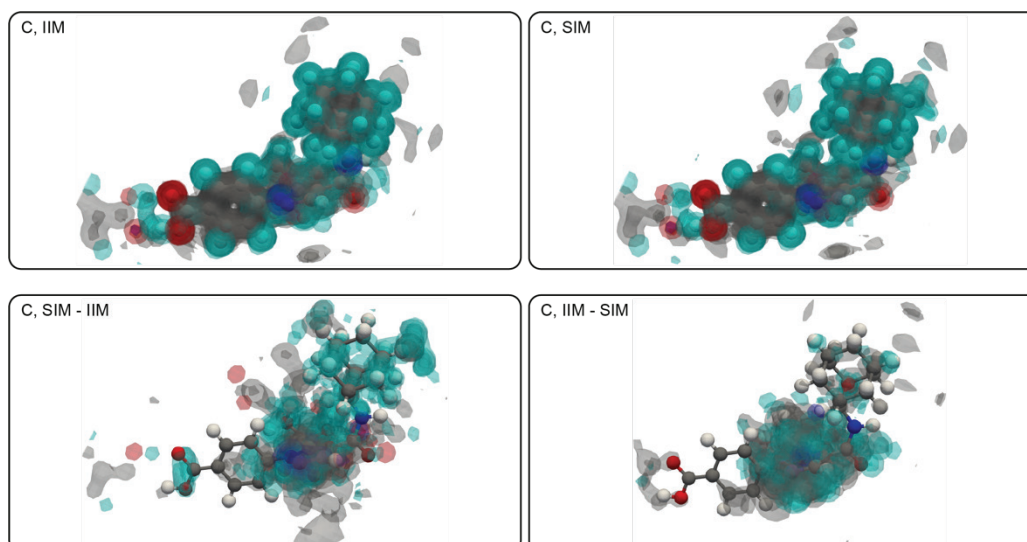


Figure 2.36. Interaction maps of AZD8329 Form 1 based on ^{13}C chemical shifts.

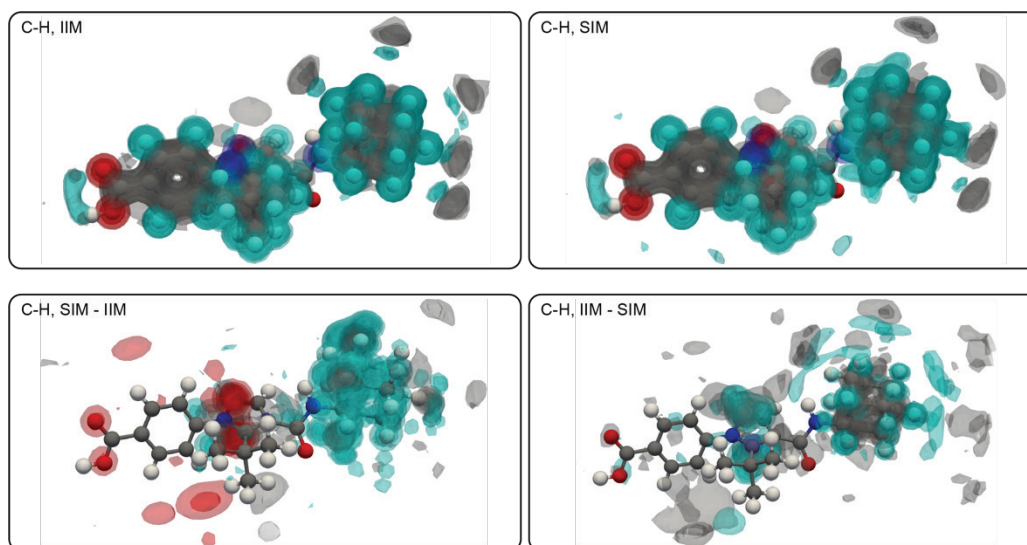


Figure 2.37. Interaction maps of AZD8329 Form 1 based on ^1H and ^{13}C chemical shifts.

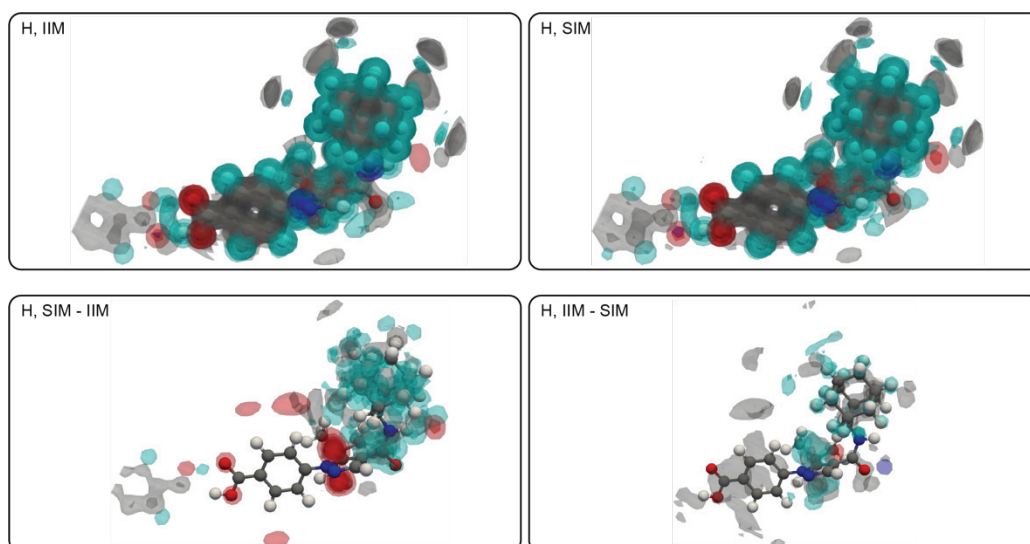


Figure 2.38. Interaction maps of AZD8329 Form 4 based on ^1H chemical shifts.

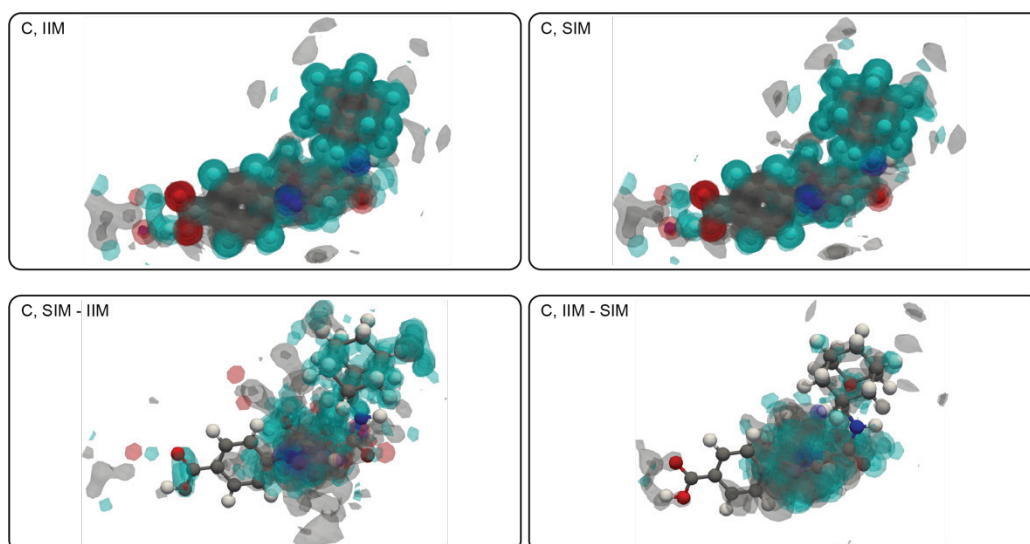


Figure 2.39. Interaction maps of AZD8329 Form 4 based on ^{13}C chemical shifts.

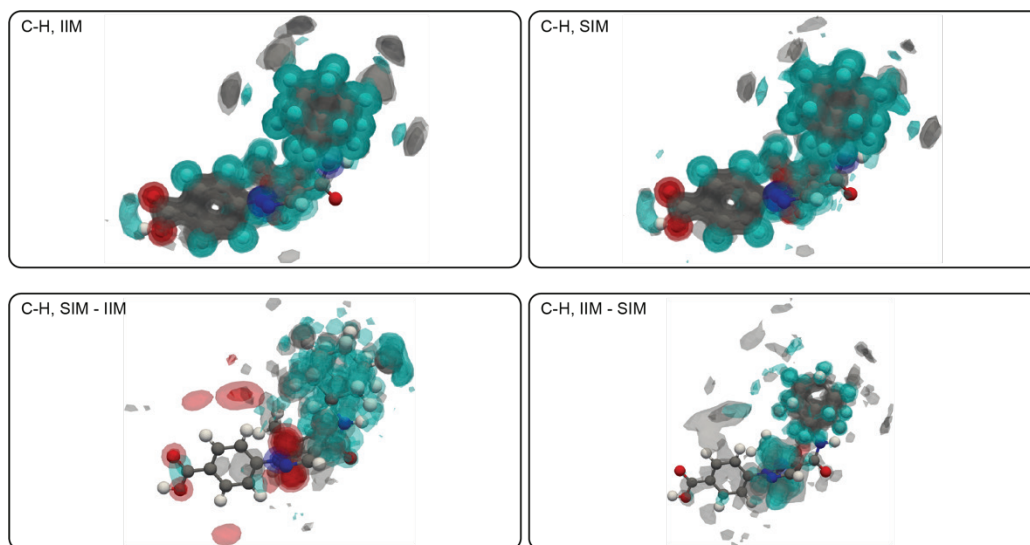


Figure 2.40. Interaction maps of AZD8329 Form 4 based on ^1H and ^{13}C chemical shifts.

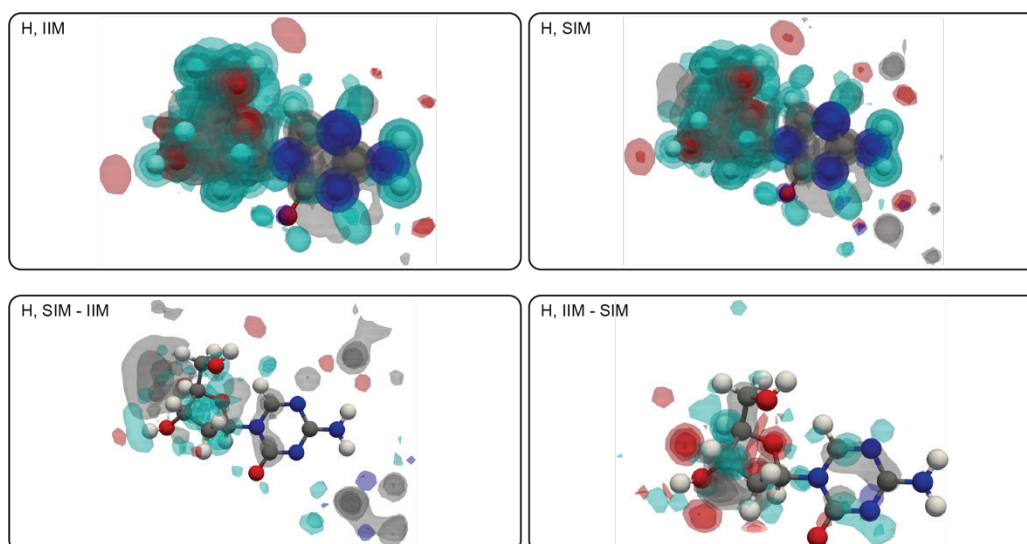


Figure 2.41. Interaction maps of decitabine based on ^1H chemical shifts.

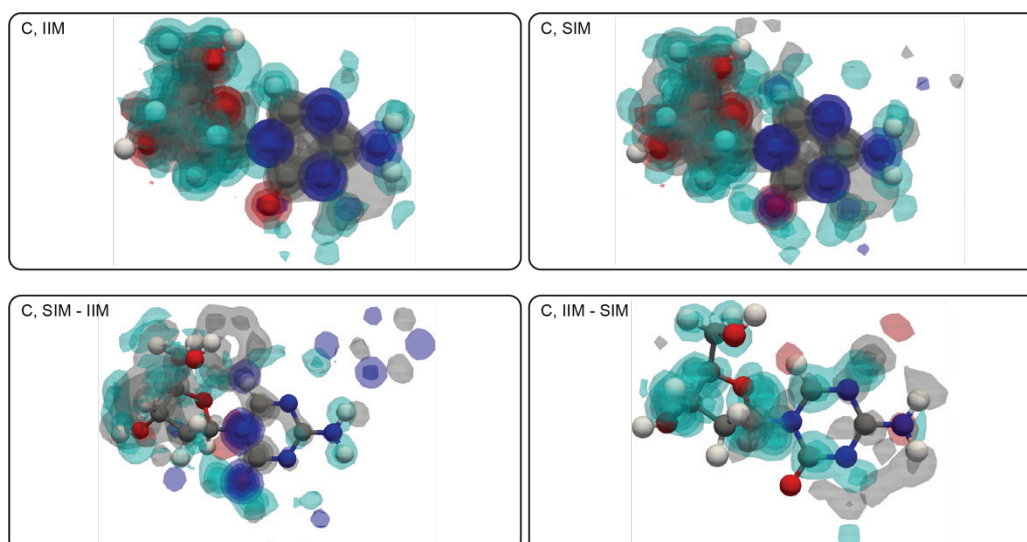


Figure 2.42. Interaction maps of decitabine based on ^{13}C chemical shifts.

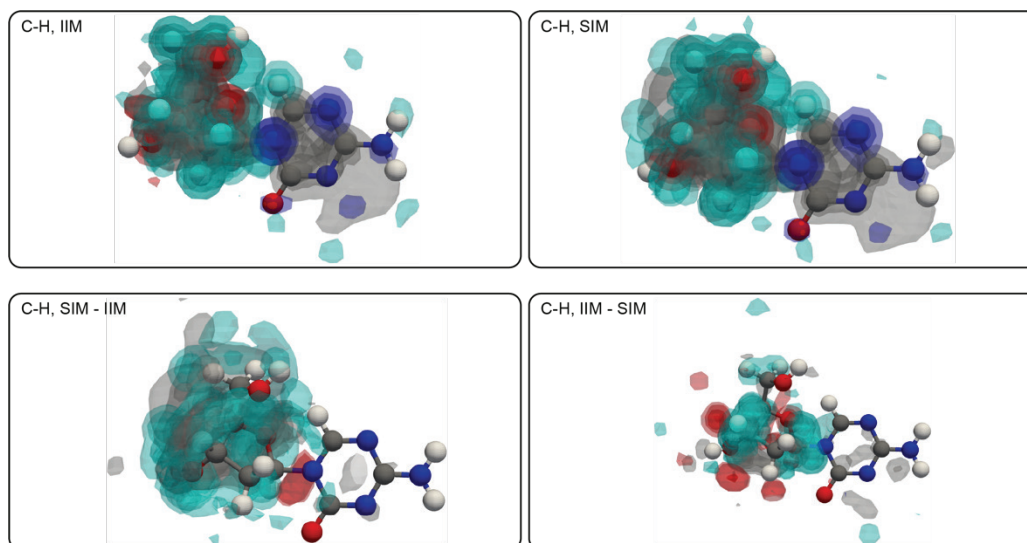


Figure 2.43. Interaction maps of decitabine based on ^1H and ^{13}C chemical shifts.

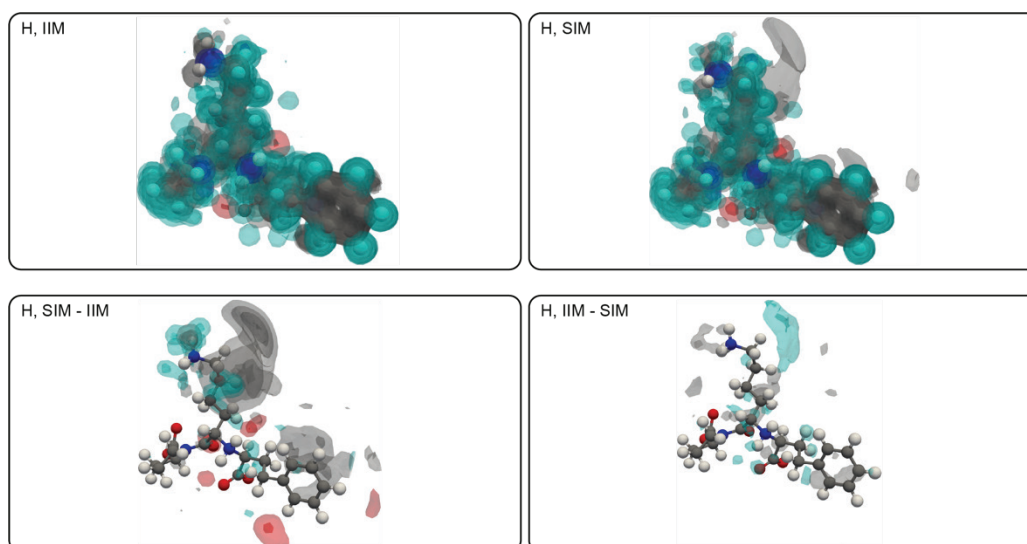


Figure 2.44. Interaction maps of lisinopril dihydrate based on ^1H chemical shifts.

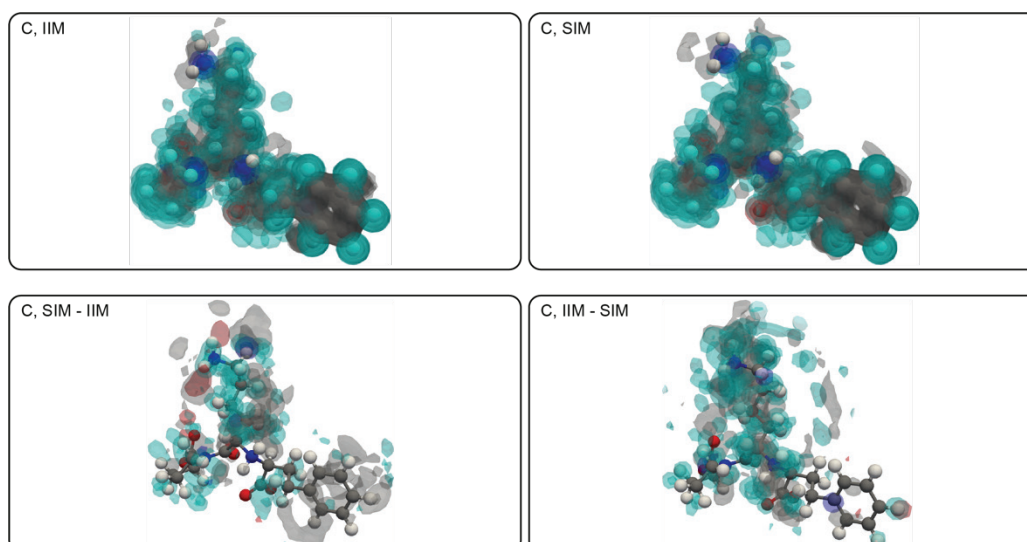


Figure 2.45. Interaction maps of lisinopril dihydrate based on ^{13}C chemical shifts.

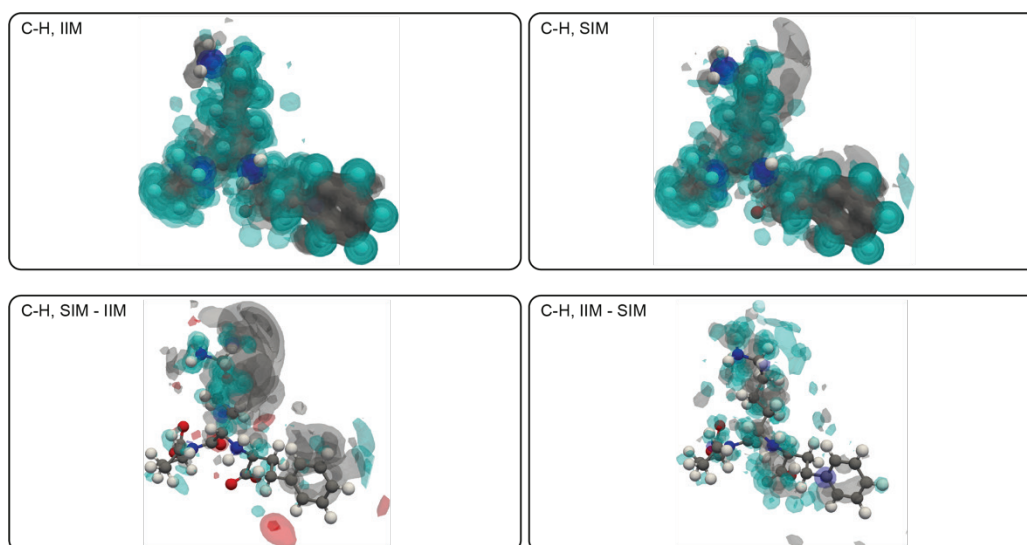


Figure 2.46. Interaction maps of lisinopril dihydrate based on ^1H and ^{13}C chemical shifts.

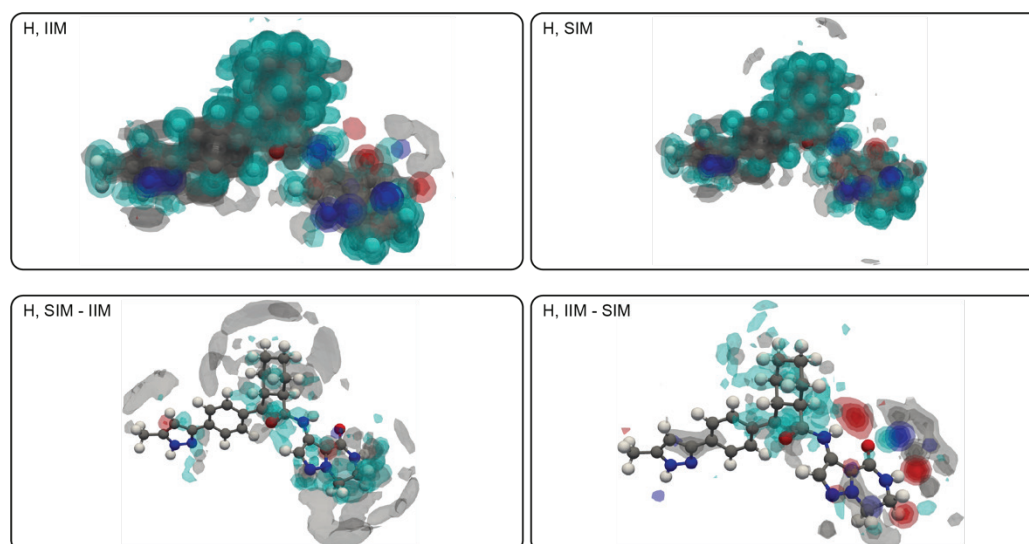


Figure 2.47. Interaction maps of AZD5718 based on ^1H chemical shifts.

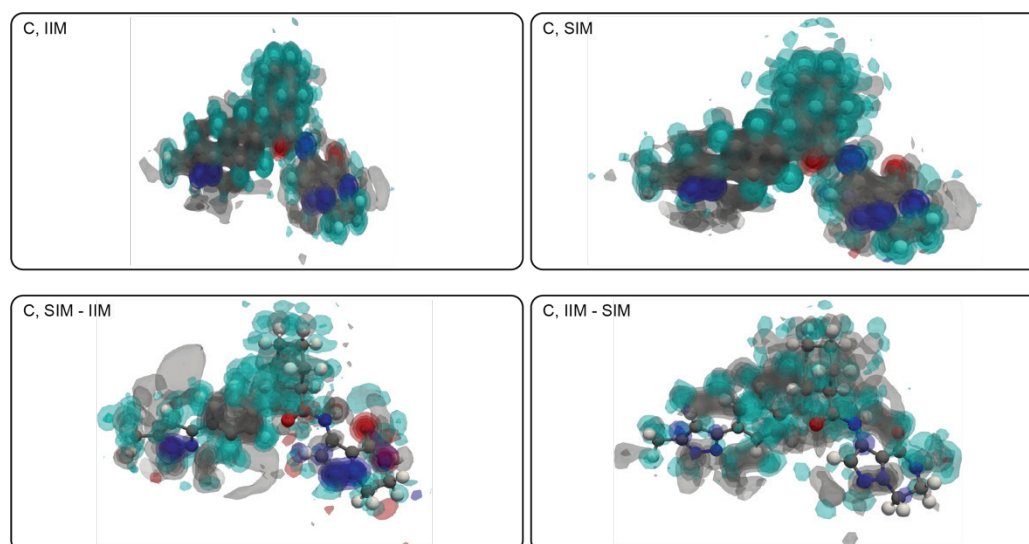


Figure 2.48. Interaction maps of AZD5718 based on ^{13}C chemical shifts.

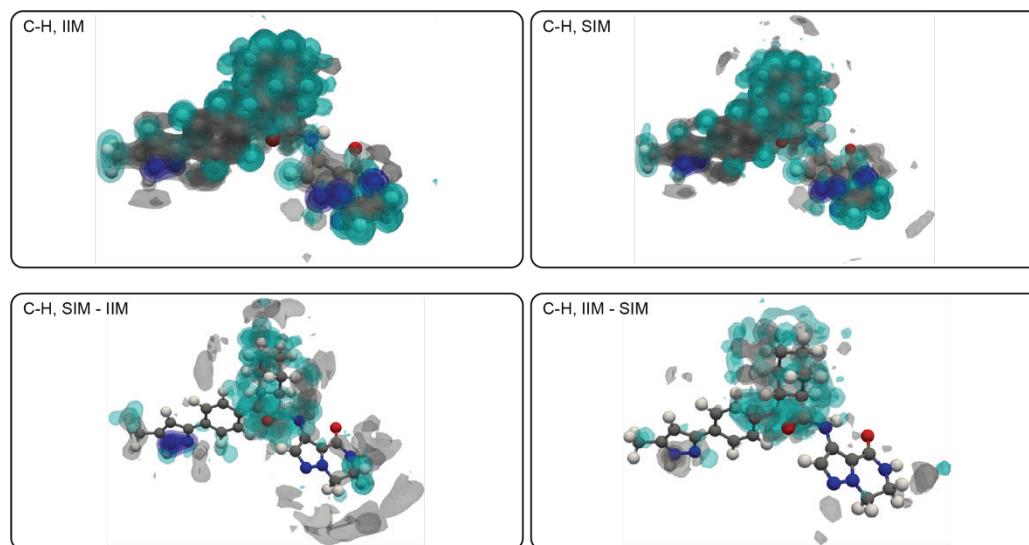


Figure 2.49. Interaction maps of AZD5718 based on ^1H and ^{13}C chemical shifts.

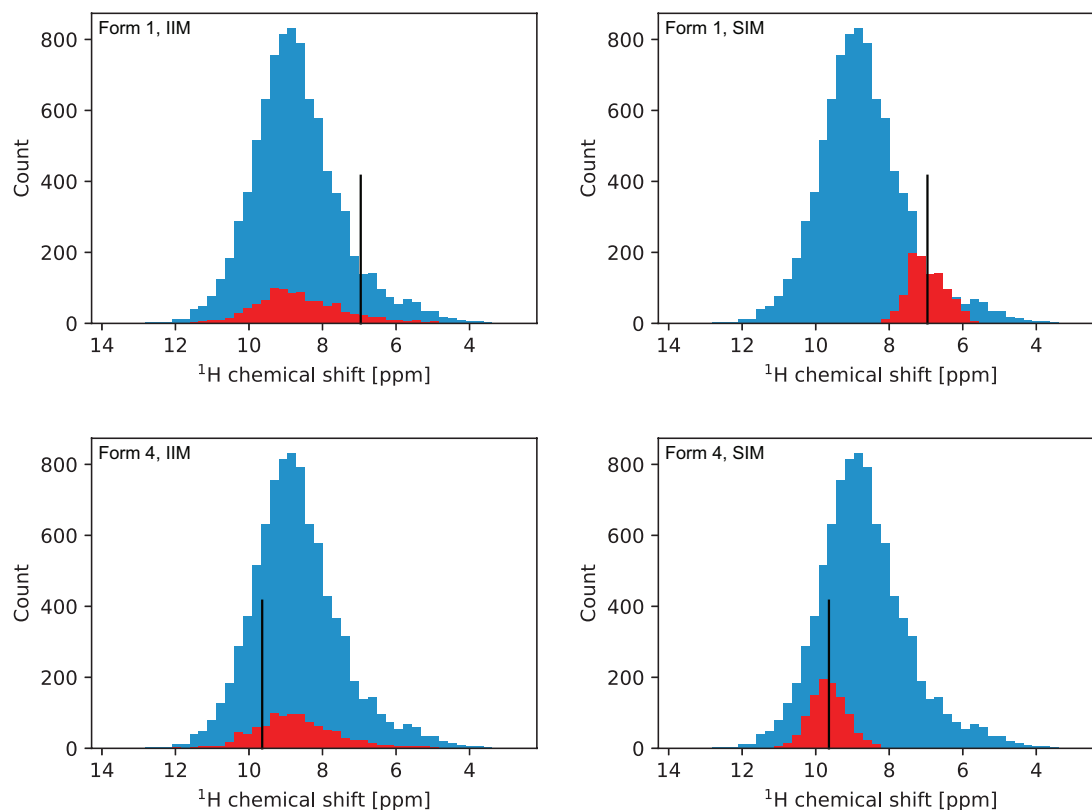


Figure 2.50. Histogram of ^1H chemical shifts from the database matching the local covalent environment of the NH proton (blue) and selected environments (red) used to construct the IIM (left) and SIM (right) for AZD8329 form 1 (top) and form 4 (bottom). The experimental shifts are indicated by the vertical black lines.

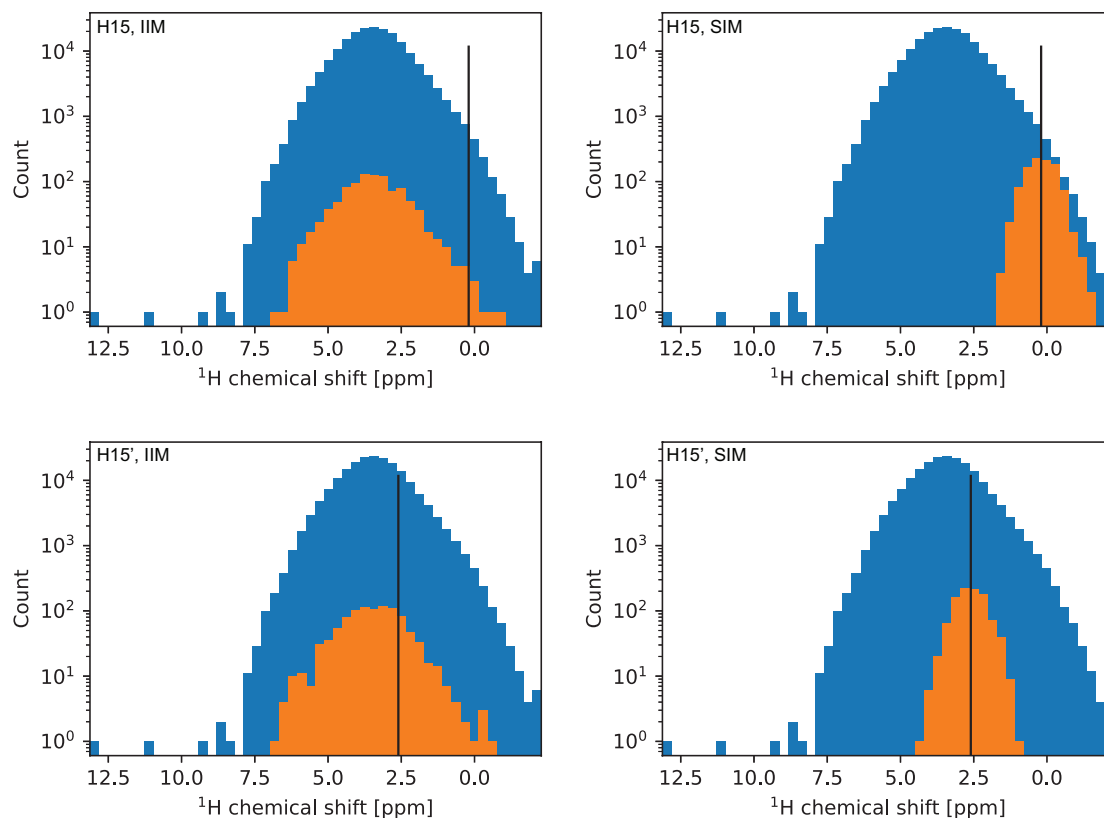


Figure 2.51. Histogram of ^1H chemical shifts from the database matching the local covalent environment of proton labelled 15 (blue) and selected environments (orange) used to construct the IIM (left) and SIM (right). The experimental shift is indicated by the vertical black line.

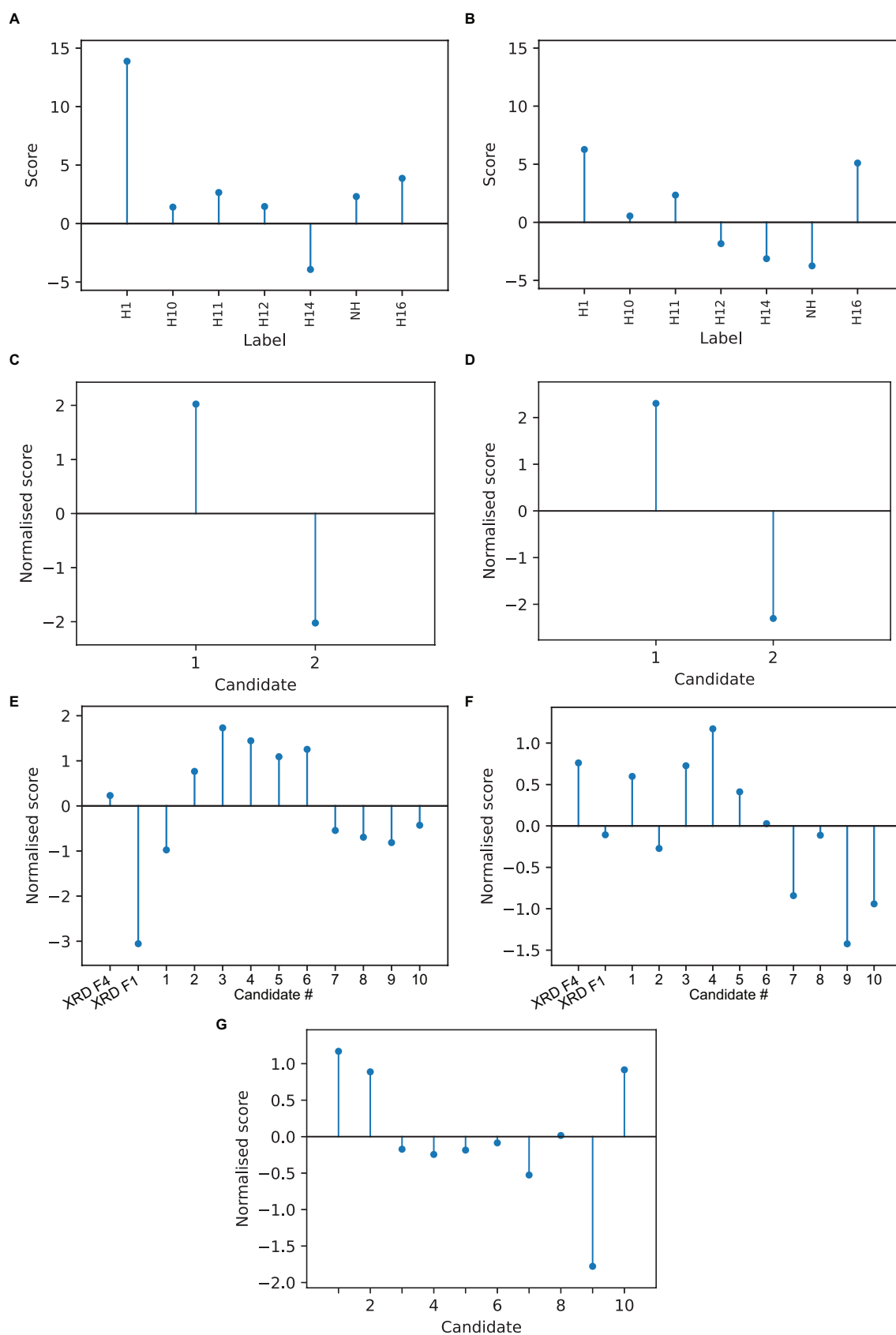


Figure 2.52. (A), (B) Scores of individual atoms of the X-ray structures of AZD8329 forms 1 and 4, respectively, using SIMs constructed using the experimentally obtained chemical shifts of AZD8329 form 1. (C), (D) Scores of the X-ray structures of AZD8329 forms 1 (candidate 1) and 4 (candidate 2) using experimental ^{13}C and ^1H - ^{13}C chemical shifts of AZD8329 form 1, respectively. (E), (F) Scores of the X-ray structures of AZD8329 forms 1 and 4 and of the CSP set for AZD8329 form 4 using experimental ^{13}C and ^1H - ^{13}C chemical shifts of AZD8329 form 4, respectively. (G) Scores of the CSP set of AZD5718 using experimental ^{13}C chemical shifts.

Chapter 3 The assignment problem

3.1 Introduction

The approaches presented in **Chapter 2** heavily rely on the use of assigned chemical shifts. Chemical shift assignment is the process of assigning each peak in a spectrum to its corresponding atomic site(s) in the molecule. In general, this procedure is the starting point of any detailed NMR study.³⁰⁸ In organic solids at natural isotopic abundance, this is still a laborious and often challenging process. In particular, ^{13}C resonance assignment typically requires the use of the through-bond ^{13}C - ^{13}C INADEQUATE experiment.^{196, 369} For materials for which the crystal structure is already known, the assignment can be determined at least partially by comparing the experimental chemical shifts with shifts computed using DFT in the gauge invariant projector augmented wave (GIPAW) method,^{117, 118, 307} or fragment-based methods.^{126, 128} However, in most applications the full structure is not known, and in particular *de novo* chemical shift-based NMR crystallography relies on chemical shift assignment in order to identify the crystal structure.^{49, 52, 176, 177}

Chemical shift assignment of biomolecules such as proteins and RNA can be obtained directly from their sequence through statistical analysis of chemical shifts.³⁷⁰⁻³⁷² In addition, simultaneous chemical shift assignment and structure determination can be obtained from matching atomic contacts to Nuclear Overhauser Effect (NOE) experiments.³⁷⁰ These approaches rely on the existence of a large database of experimental chemical shifts and molecular structures, such as the Biological Magnetic Resonance Data Bank (BMRB)²⁷⁰ and Protein Data Bank (PDB),²⁶⁸ respectively. For example, the BMRB contains over 9.4 million instances of experimental chemical shifts for 279 types of proton, carbon and nitrogen sites in the 20 amino acids that make up proteins, with, e.g., over 89,000 instances of the NH shift in alanine alone. Such large and diverse chemical shift databases however do not exist, to my knowledge, for organic crystals.

In **Section 3.2**, by combining the Cambridge Structural Database with ShiftML, we construct a statistical basis for probabilistic chemical shift assignment of organic crystals by calculating shifts for more than 200,000 compounds, enabling the probabilistic assignment of organic crystals directly from their two-dimensional chemical structure. The approach is demonstrated with the ^{13}C and ^1H assignment of 11 molecular solids with experimental shifts and benchmarked on 100 crystals using predicted shifts. The correct assignment is found among the two most probable assignments in more than 80% of cases.

The main issue preventing ^1H -based assignment and atomic-level characterisation of molecular solids is that the resolution of ^1H solid-state NMR spectra is limited by broadening due to the homonuclear dipolar interactions between the abundant ^1H spins.³⁰⁸ Magic angle spinning (MAS)^{81, 82} helps reduce dipolar broadening by spinning the sample around an axis tilted at 54.74° from the direction of the main magnetic field. This process induces coherent averaging of second-rank tensor interactions such as the homonuclear dipolar interaction while having no effect on isotropic interactions such as the chemical shift. However, the second-rank interactions cannot be completely removed even at the highest spinning rates currently available.^{85, 373-378} This results in residual broadening typically on the order of hundreds of hertz,^{84, 85, 373, 378-381} which obscures the information contained in ^1H solid-state NMR spectra.

Obtaining isotropic proton spectra in molecular solids is a key objective in order to leverage the advantage provided by ^1H NMR in solids compared to other nuclei.^{39, 151, 308, 382-385} This has fueled the advent of faster magic angle spinning (MAS), as well as the development of pulse sequences designed to remove homonuclear dipolar couplings.³⁸⁶⁻³⁹⁸ However, no such method has yet been able to completely remove dipolar interactions in proton spectra of molecular solids.

Based on the description of the dependence of residual splittings and shifts on the MAS rate ω_{MAS} ,^{85, 373, 378-381, 399} Moutzouri *et al.* introduced a two-dimensional approach to obtain the pure isotropic (infinite MAS rate) spectrum of molecular solids from a set of spectra measured at different MAS rates.⁴⁰⁰ While the method introduced provides a powerful method to obtain isotropic spectra, several assumptions and restrictions inherent to this fitting approach may limit its performance.

In **Section 3.3**, we introduce a deep learning approach to determine pure isotropic proton (PIP) spectra from a two-dimensional set of magic-angle spinning spectra acquired at different spinning rates. Applying the model to 8 organic solids yields high-resolution ^1H solid-state NMR spectra with isotropic linewidths in the 50-400 Hz range.

While high-resolution one-dimensional spectra are useful, most applications of NMR spectroscopy today require two-dimensional correlation experiments. In this respect, the possibility of measuring ultrahigh-resolution ^1H - ^1H correlations is especially attractive, as it enables both structure determination and assignment.

In **Section 3.4**, we extend the PIP approach to a second dimension, and for samples of L-tyrosine hydrochloride and ampicillin we obtain high resolution ^1H - ^1H double-quantum/single-quantum dipolar correlation and spin-diffusion spectra with significantly higher resolution than the corresponding spectra at 100 kHz MAS, allowing the identification of previously overlapped isotropic correlation peaks.

Overall, this chapter presents methods that aim at improving the assignment of NMR spectra of molecular solids, from a probabilistic approach to determining the assignment based on the chemical structure and a list of chemical shifts to machine learning models providing better resolved ^1H spectra. These methods have the potential to bypass the need for long multi-dimensional experiments typically required to obtain a confident measurement and assignment of chemical shifts, and to significantly accelerate the atomic-level characterisation of molecular solids by NMR.

3.2 Bayesian probabilistic assignment of chemical shifts in organic solids

This section has been adapted with permission from: Cordova, M.; Balodis, M.; Simões de Almeida, B.; Ceriotti, M.; Emsley, L., Bayesian probabilistic assignment of chemical shifts in organic solids. *Science Advances* **2021**, 7 (48), eabk2341. (post-print)

My contribution was to construct the database, to develop and apply the method and to analyse results. I also wrote the manuscript, with contributions of all other authors.

3.2.1 Introduction

An illustrative example of the assignment problem for ^{13}C nuclei is shown in **Figure 3.1**, with the ^{13}C cross-polarisation magic angle spinning (CPMAS) spectrum of ritonavir. The spectrum contains 32 peaks, corresponding to the 37 magnetically inequivalent carbon atoms in the molecule, and assigning the peaks to the atoms is not at all obvious. Several straightforward experimental methods can be used to simplify the assignment process in organic solids. Heteronuclear correlation (HETCOR) experiments^{401, 402} provide pairwise ^1H -X (where X = ^{13}C , ^{15}N , etc...) correlations and allow the separation of NMR signals along two dimensions, which simplifies the identification of the bonding environment associated with the observed peaks. In addition, spectral editing⁴⁰³⁻⁴⁰⁷ can be used to identify the carbon multiplicity (i.e., the number of bonded protons) associated to each observed peak, allowing the reduction of the assignment problem to subsets of peaks and corresponding atomic sites.

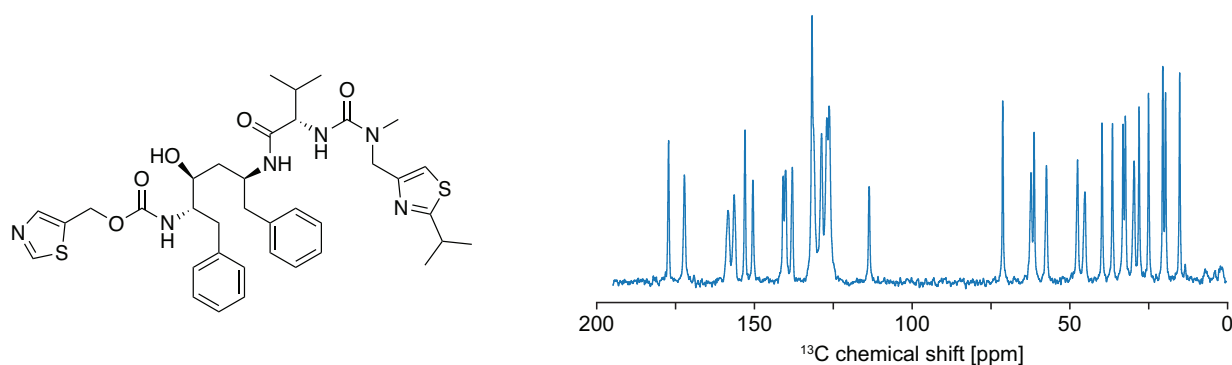


Figure 3.1. Molecular structure of ritonavir and the ^{13}C CPMAS spectrum recorded for a powder sample of ritonavir form II.

ShiftML allows chemical shifts to be obtained directly from the structure of a molecular solid, bypassing the need for an optimised wavefunction and making the shifts of large ensembles of large structures accessible with DFT accuracy.^{176, 261} Here, we show how combining this model with a database of three-dimensional structures such as the Cambridge Structural Database (CSD)³¹² enables the probabilistic assignment of organic crystals using chemical shift statistics without any knowledge of the 3D structure. We generate a large database of chemical shifts for organic crystals by predicting shifts using ShiftML on structures extracted from the CSD. By relating the shifts obtained to molecular fragment descriptors, we obtain probabilistic assignments of organic crystals directly from their molecular structure.

3.2.2 Methods

NMR spectroscopy. The sample of strychnine and ritonavir form II were purchased from Sigma-Aldrich and Tokyo Chemical Industry, respectively. Experiments were performed on Bruker Ascend 400 and Ascend 500 wide-bore Avance III, and 900 US² wide-bore Avance Neo NMR spectrometers. The spectrometers operate at ^1H Larmor frequencies of 400, 500 and 900 MHz respectively, and are equipped with H/X/Y 3.2 mm, H/X/Y 4.0 mm, H/C/N/D 1.3 mm and H/C/N 0.7 mm CPMAS probes.

1D ^1H MAS NMR spectra were recorded at a temperature of 298 K using rotor spinning rates (ν_r) up to 111 kHz. 1D ^{13}C cross-polarization (CP)⁴⁰⁸ MAS NMR spectra were acquired at 298 K with ν_r of 12.5 and 22 kHz for ritonavir and strychnine respectively. During the signal acquisition SPINAL-64 decoupling⁴⁰⁹ was applied with a ^1H rf field amplitude of 100 kHz. For ritonavir spectral editing experiments were used to distinguish carbons with different numbers of protons attached to them. To selectively remove quaternary carbons a 1D version of MAS-J-HSQC⁴⁰⁶ was used, to remove quaternary and primary carbons a double quantum filter was added to the MAS-J-HSQC⁴⁰⁶ sequence and to remove primary and secondary carbons a simple CP experiment with an inserted delay of 0.5 ms before acquisition and after the CP pulse was applied.⁴⁰³ 2D ^1H - ^{13}C HETCOR experiments were carried out at 298 K using $\nu_r = 22$ kHz. During t_1 100 kHz eDUMBO-1²² was applied to decouple the ^1H - ^1H dipolar coupling,⁴¹⁰ and during t_2 100 kHz SPINAL-64 decoupling was applied.

The natural abundance 2D ^{13}C - ^{13}C refocused INADEQUATE^{369, 411} spectra required for the direct experimental assignment for ritonavir and strychnine were acquired using a Bruker 400 MHz Ascend NMR spectrometer. The probe was configured into $^1\text{H}/^{13}\text{C}$ double resonance mode. Variable amplitude cross-polarisation⁴¹² was used to transfer polarisation from ^1H to ^{13}C . SPINAL-64⁴⁰⁹ heteronuclear ^1H decoupling with RF fields of 100 kHz was applied in all cases. The temperature of the sample for ritonavir was 250 K and a 4 mm rotor was used with a spinning frequency of 12.5 kHz. 2 x 120h experiments were acquired and combined in post processing to obtain the final spectrum (total time: 10 days). For strychnine DNP was used.⁴¹³ The sample was impregnated with 10 mM AMUPOL dissolved in 60:30:10 glycerol- d_8 : D_2O : H_2O . The spectrometer is equipped with a low temperature magic angle spinning (LTMAS) 3.2 mm probe and connected through a corrugated waveguide to a 263 GHz gyrotron capable of outputting ca. 5-10 W of continuous wave microwaves.⁴¹⁴ The sweep coil of the main magnetic field was optimised so that the microwave irradiation gave the maximum positive proton DNP enhancement with binitroxide cross effect-based polarising agents (e.g., AMUPOL, TEKPOL). The temperature of the sample for ritonavir was 92 K and a 3.2 mm rotor was used with a spinning frequency of 12.5 kHz. A DNP enhancement of 36 was determined based on the ratio of the area of the spectra acquired with and without microwave irradiation. The DNP-enhanced natural abundance 2D ^{13}C - ^{13}C refocused INADEQUATE experiment⁴¹³ was run for 45 hours.

All chemical shifts were referenced via alanine. The full set of acquisition parameters is given in **Tables 3.1-3.4**.

Selection of crystal structures. The structures used to construct the chemical shift database were obtained from the CSD.³¹² Only the organic crystal structures suitable for chemical shift predictions were selected. The corresponding selection criteria were that every structure must only contain C, H, N, O and S atoms, and that the disorder (if any) is resolvable (i.e., all atomic sites in the structure can be assigned to their major occupancy sites and the corresponding structure matches the reported chemical formula). Missing protons were added automatically using the tool built into the CSD Python API. In total, 205,069 valid structures were selected.

Relaxation and chemical shift prediction. Because proton positions in published single-crystal X-ray diffraction structures may not correspond to the actual hydrogen positions in the crystals, they have to be optimised. Due to the large number of structures selected, DFT relaxation would be prohibitively costly. The semiempirical DFTB method³²⁵ was thus chosen to relax proton positions in all structures. The structures were optimised at the DFTB3-D3H5 level of theory^{326, 351} using the 3ob-3-1 parameter set.^{350, 415} Further computational details are given in **Appendix IV**. Instances where the structure relaxation failed were discarded. 203,303 structures were successfully relaxed and considered for chemical shift prediction.

All chemical shift predictions were performed using ShiftML version 1.2 (publicly available at <https://shiftml.epfl.ch>).^{176, 261} Conversions of predicted shieldings to chemical shifts were performed by least squares fitting of the shieldings obtained for benchmark sets of DFTB-relaxed structures to their experimental chemical shifts, fixing the slope to a value of -1. The offsets obtained were found to be 30.96 ppm for ^1H , 168.64 ppm for ^{13}C , 185.99 ppm for ^{15}N and 205.08 ppm for ^{17}O . This corresponds to ^1H and ^{13}C shifts relative to TMS, ^{15}N shifts relative to NH_4Cl , and ^{17}O shifts relative to liquid H_2O . The sets of structures and isotropic chemical shifts used to determine shielding-to-shift conversions are described in **Tables 3.5-3.8**. We note that chemical shieldings are stored in the database, and converted to chemical shifts on-the-fly during the construction of chemical shift distributions. In total, the database contains 5,243,129 unique ^1H , 4,847,864 unique ^{13}C , 466,370 unique ^{15}N and 867,446 unique ^{17}O chemical shifts, respectively.

Molecular fragment descriptors. For assignment of the spectrum of a molecule of unknown structure, classification of the predicted shifts should be done such that a statistical distribution of chemical shifts can be obtained for any nucleus from the two-dimensional representation of a molecule. The molecular fragment descriptor should thus not contain any information about conformation or molecular packing in the crystal structures. Among the topological atom-centered descriptors that fit these requirements,⁴¹⁶⁻⁴¹⁸ we chose to represent topological atomic environments by graphs where vertices represent atoms and edges represent covalent connectivities. The vertices were labelled by element, and the edges were kept unlabelled. Graphs were cut to a maximum depth w of 6, defined as the maximum shortest path between the central vertex (for which the chemical shift is predicted) and any other vertex in the path.

Conversion of the three-dimensional crystal structures to their corresponding graphs was performed by identifying atom pairs as covalently bonded when the distance between the atoms in the pair is less than 1.1 times the sum of the covalent radii of the atoms involved.

Database construction and search. A given topological atomic environment can be searched by identifying which graphs in the database match the graph of the selected atomic environment. However, there is no known algorithm able to solve the graph isomorphism problem required for each database entry in polynomial time.^{419, 420} Thus, the search was simplified by using the Weisfeiler-Lehman hash⁴²¹ as a unique graph identifier. If the number of instances of a given atomic environment identified in the database was deemed too small to produce statistically significant chemical shift distributions, the atomic environment was searched again after reducing the graph depth. For this work, we chose a minimum number of instances of 10. Further details concerning the database architecture and search can be found in **Appendix IV**.

Construction of probability distributions. We use a notation and a conceptual framework extending the Bayesian selection of crystal-structure prediction candidate structures compatible with measured shifts.¹⁷⁶ From the set of chemical shifts and uncertainties $\{y_k, \sigma_k\}$ predicted by ShiftML for the CSD structures that share the same graph G_i as the atom i in the molecule of interest, we define the probability of observing a chemical shift y for that atom as proportional to the sum of Gaussian functions centered on each predicted shift y_k and with a width σ_k given by its prediction uncertainty.

$$p_i(y) \propto \sum_{k \in G_i} \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left[-\frac{(y - y_k)^2}{2\sigma_k^2}\right] \quad (3.1)$$

Similarly, we define the probability of observing a cross-peak $(y^{(1)}, y^{(2)})$ for a pair of bonded atoms (i, j) in a molecule as proportional to the sum of uncorrelated two-dimensional Gaussian functions,

$$p_{ij}(y^{(1)}, y^{(2)}) \propto \sum_{k \in G_{ij}} \frac{1}{2\pi\sigma_k^{(1)}\sigma_k^{(2)}} \exp\left[-\frac{(y^{(1)} - y_k^{(1)})^2}{2\sigma_k^{(1)2}} - \frac{(y^{(2)} - y_k^{(2)})^2}{2\sigma_k^{(2)2}}\right] \quad (3.2)$$

Where $\{y_k^{(1)}, \sigma_k^{(1)}\}$ and $\{y_k^{(2)}, \sigma_k^{(2)}\}$ are the sets of chemical shifts and predicted uncertainties computed for all the bonded atoms in the reference dataset that share the same graph G_{ij} as the pair being considered.

Probabilistic assignment. Considering the vector of observed shifts \mathbf{y} , the probability that one of its elements y_j originates from atom i is obtained by evaluating **Equation 3.1** (or **Equation 3.2**) for all elements in \mathbf{y} ,

$$p(y_j|i) = \frac{p_i(y_j)}{\sum_k p_i(y_k)}. \quad (3.3)$$

For a given assignment \mathbf{a} (defined as the vector mapping atoms in the molecule to experimental shifts such that $a_i = j$ if atom i is assigned to shift j), the probability of observing a vector of chemical shifts \mathbf{y} is given by

$$p(\mathbf{y}|\mathbf{a}) = \prod_i p(y_{a_i}|i). \quad (3.4)$$

Applying Bayes theorem on **Equation 3.4** yields the probability of an assignment \mathbf{a} given the observed vector of shifts \mathbf{y} ,

$$p(\mathbf{a}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{a})p(\mathbf{a})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\mathbf{a})p(\mathbf{a})}{\sum_{\mathbf{a}'} p(\mathbf{y}|\mathbf{a}')p(\mathbf{a}')}. \quad (3.5)$$

In **Equation 3.5**, we assume that $p(\mathbf{a})$ is a non-zero constant if the assignment is valid (i.e., if all nuclei are assigned to only one chemical shift, and if all observed shifts are assigned at least one nucleus), and zero otherwise. Whenever some of the assignments can be made according to experimental data or heuristic arguments, such prior information can be incorporated in the definition through $p(\mathbf{a})$. By combining individual assignments, the complete set of possible global assignments can be generated. Because of the combinatorial complexity of generating all possible global assignments, several procedures were implemented to reduce the global assignment generation cost while ensuring that the most probable assignments are generated, and these are described in **Appendix IV**. Note that if the probability of any shift originating from a given nucleus is lower by a set threshold (typically a factor of 100) than the maximum probability for that nucleus then it is discarded. This results in some nuclei being assigned unambiguously independently of the rest of the global assignment (e.g., shift "e" in **Figure 3.3**).

Equation 3.5 assigns a distinct probability to each possible assignment of the entries of the measured shifts vector \mathbf{y} to all the environments. It is the correct probabilistic metric to compare two assignments but is hard to interpret. A more compact indicator is given by the marginal probability that atom i is assigned to shift j , which can be extracted from the set of generated assignments by considering only the vectors \mathbf{a} containing that particular individual assignment. This is shown in **Equation 3.6** by the Kronecker delta $\delta_{a_i j}$ which selects the assignments for which $a_i = j$,

$$p(a_i = j | \mathbf{y}) = \frac{\sum_{\mathbf{a}} \delta_{a_i j} p(\mathbf{a} | \mathbf{y})}{\sum_{\mathbf{a}} p(\mathbf{a} | \mathbf{y})}. \quad (3.6)$$

For topologically equivalent nuclei, which have identical graphs and probability distributions, tuples of nuclei were assigned to tuples of experimental shifts (which can be partly or entirely identical).

Synthetic benchmark set. A set of 100 randomly selected crystal structures from the database were selected to benchmark the probabilistic assignment. The selection was restricted to crystals having between 10 and 20 unique carbon atoms. The selected structures are listed in **Appendix IV**. The ShiftML predicted shifts associated to each nucleus were used as ground-truth assignment. The structure to assign was systematically excluded from the database search performed to construct statistical distributions of chemical shifts. The synthetic benchmark set was separated into five sets containing 20 crystals each and 241, 260, 212, 259 and 242 unique carbon atoms, respectively.

3.2.3 Results and Discussion

The framework presented here was applied to a set of various organic molecules for which the carbon chemical shift assignment was already (at least partially) determined experimentally. The selected set is composed of theophylline,⁵² thymol,⁵⁰ cocaine,⁵² strychnine, AZD5718,¹⁷⁷ lisinopril,³⁶⁴ ritonavir, the K salt of penicillin G,¹³⁹ β -piroxicam,¹⁴³ decitabine¹⁷² and simvastatin.⁴²² The experimental spectra used for the assignment of strychnine and ritonavir are shown in **Figures 3.9-3.10**. Experimental shifts of lisinopril were obtained from a dihydrate form.³⁶⁴ Experimental shifts of ritonavir were obtained from the polymorphic form II.

Graph generation is the starting point of statistical assignment and can be performed directly from the two-dimensional representation of the molecule. **Figure 3.2A-B** shows the graphs generated for illustrative carbon atoms in theophylline with a depth $w = 3$. The chemical shift distributions of the carbon labelled 4 in theophylline corresponding to different graph depths are shown in **Figure 3.2C**, together with the corresponding graphs. As expected, the distribution changes significantly as w is increased, until at $w = 3$ and above where they are found to be highly similar, with a width dominated by the uncertainty in the ShiftML prediction. We thus selected a minimum number of ten instances to construct each probability distribution, and used the maximum graph depth that fulfils this requirement for each nucleus.

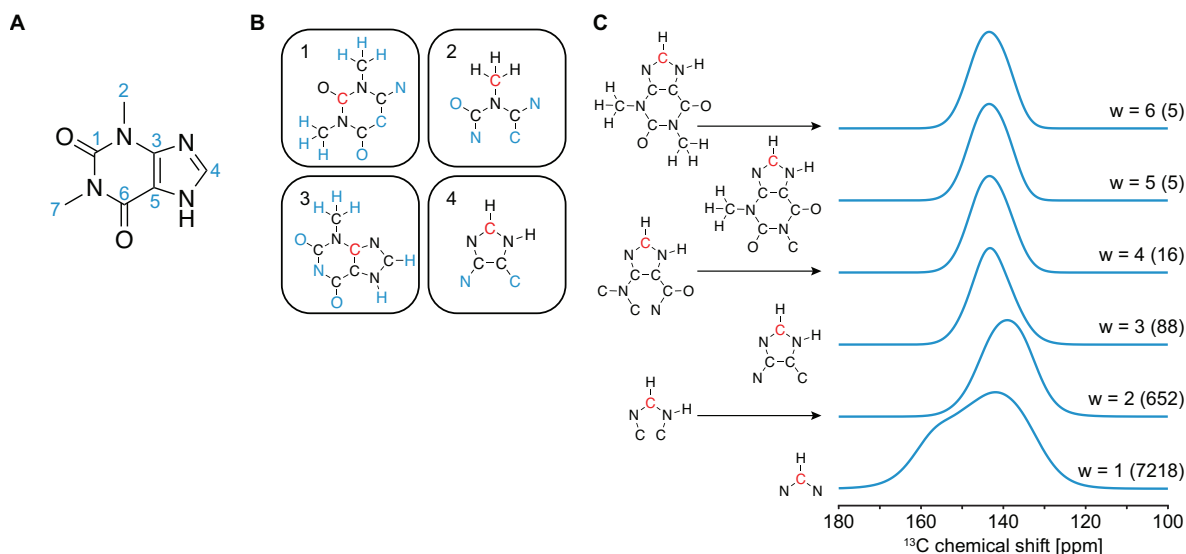


Figure 3.2. (A) Two-dimensional structure and carbon labelling scheme of theophylline. (B) Graphs of carbons 1, 2, 3 and 4 of theophylline constructed at a depth $w = 3$. In each graph, the red vertex corresponds to the central atom (for which the chemical shift distribution is extracted), and blue vertices indicate the atoms at the maximum shortest path from the central vertex. (C) Chemical shift distributions obtained corresponding to the carbon labelled 4, with different graph depths w . The number of instances from the database used to construct each distribution is indicated in parentheses.

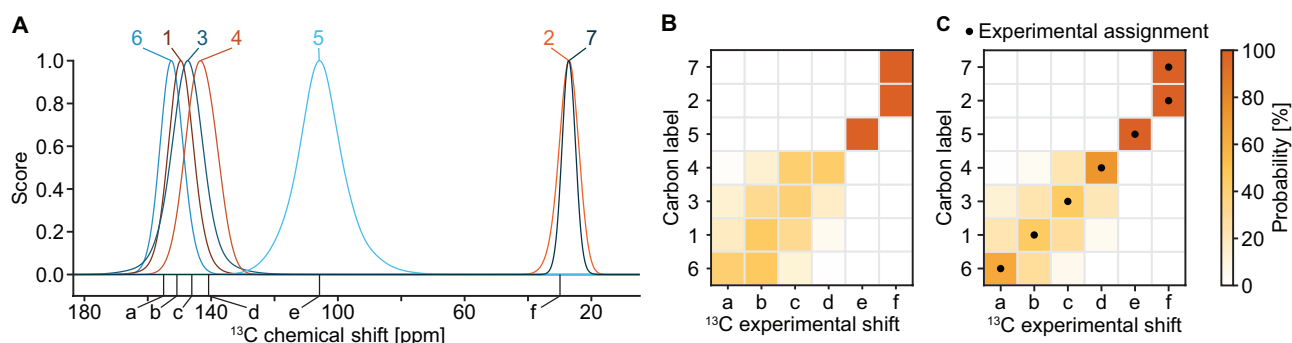


Figure 3.3. (A) Statistical ^{13}C chemical shift distributions for theophylline (coloured lines). The carbon labels follow Figure 3.2A. Experimental shifts are indicated by black vertical lines below the distributions and are labelled “a” through “f” in order of decreasing chemical shift (see Table 3.9). (B) Probabilities of observing each chemical shift of theophylline for a given carbon nucleus. (C) Marginal individual assignment probabilities of the ^{13}C chemical shifts of theophylline after Bayesian inference of the possible global assignments. The dots indicate the experimentally determined correct assignment.

The prior statistical distribution of chemical shifts for each atom in a molecule can be constructed from the shifts predicted for all atoms in the database that share the same graph. Evaluating the obtained statistical distributions at the observed shifts yields the probability of observing each shift originating from each nucleus in the molecule (Figure 3.3A-B). The possible combinations of individual assignments, based on a Bayesian construction, makes it possible to associate a probability to each global assignment of all shifts. After obtaining the probability for each global assignment in the set, marginalisation yields individual assignment probabilities (Figure 3.3C). In this case, the most probable individual assignment for each carbon, as well as the most probable global assignment, were found to correspond to the experimental assignment of theophylline (black dots in Figure 3.3C).

Overlap of the chemical shift distributions can lead to highly ambiguous assignments. A common method to separate overlapping NMR signals consists in spreading them along multiple dimensions. The HETCOR experiment yields high-sensitivity correlated ^1H and ^{13}C chemical shifts of dipolar coupled nuclei, and can be tuned to obtain a spectrum dominated by one-bond correlations.^{401, 402} The correlated statistical distributions of chemical shifts corresponding to a simulated HETCOR can be obtained by considering bonded CH pairs in the molecule. This additional dimension often helps separate overlapping one-dimensional statistical distributions and chemical shifts by incorporating the additional information given by the ^1H chemical shift. In addition, this can also be used to simultaneously assign ^{13}C and ^1H chemical shifts.

Figure 3.4 depicts the probabilistic assignment of bonded ^{13}C - ^1H chemical shifts of thymol using two-dimensional correlated statistical shift distributions. The pair of topologically equivalent bonded C-H groups (labelled 9 and 10) was assigned to a pair of experimental shifts in Figure 3.4D as the disambiguation of topologically equivalent nuclei cannot be performed from the two-dimensional representation of a molecule. As seen in Figure 3.4B, the assignment of the carbon labelled 8 would have been much more ambiguous using only ^{13}C chemical shifts. Indeed, the probability of assigning carbon 8 to chemical shift “e” is 34% using only statistical distributions of ^{13}C chemical shifts (Figure 3.4E), and 100% using correlated statistical distributions of ^1H and ^{13}C chemical shifts (Figure 3.4D). We note that the most probable assignments of carbons 6 and 7 and of the methyl groups 1, 9 and 10 do not match the experimentally determined ones. We attribute these discrepancies to substantial overlap between the corresponding statistical distributions of chemical shifts, that arise because of similar local bonding environments of carbons 6 and 7, and of methyl groups.

In addition to HETCOR, spectral editing methods are also straightforward high-sensitivity experiments that can be performed routinely to aid assignment. Such experiments are able to separate ^{13}C chemical shifts according to the number of bonded protons (multiplicity).^{403-405, 407} The method can thus be directly applied to the statistical assignment framework presented here in order to break down the statistical assignment problem into smaller sub-problems of reduced complexity. This is especially useful when considering molecules yielding substantial overlap of statistical distributions. Knowledge of the multiplicity of ^{13}C chemical shifts can also be used to select a subset of HETCOR peaks to assign.

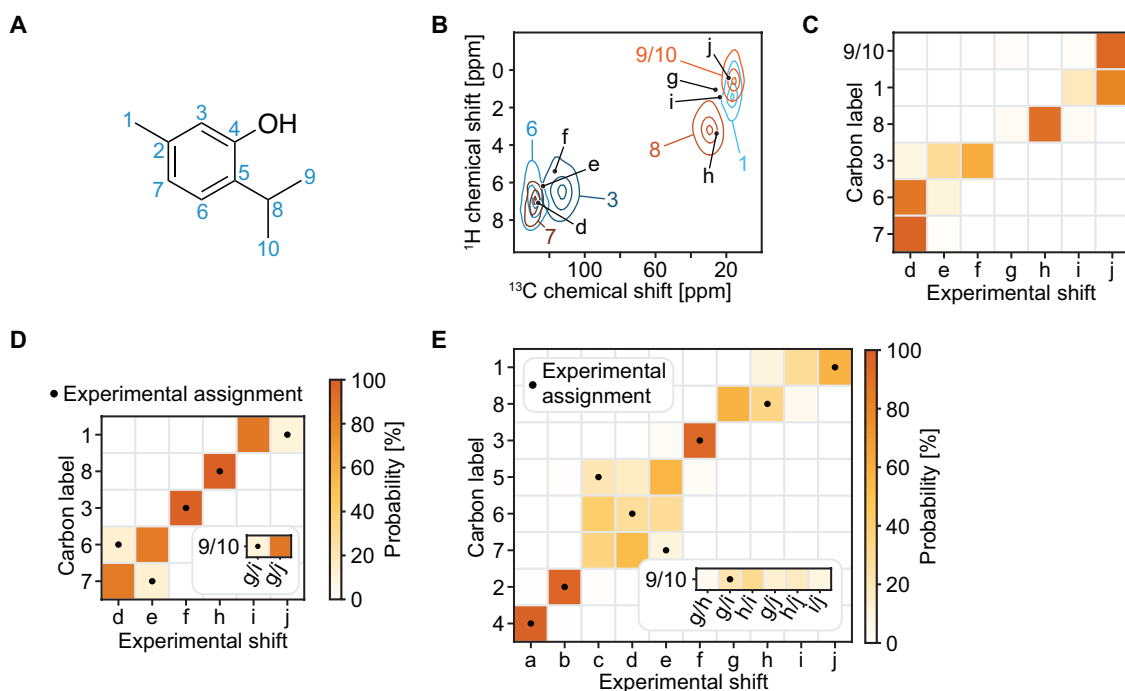


Figure 3.4. (A) Carbon labelling scheme of thymol. (B) Contour plot of the correlated statistical chemical shift distributions of bonded ^{13}C - ^1H in thymol. The carbon labels follow (A). Experimental shifts are indicated by black dots and are labelled alphabetically in order of decreasing ^{13}C chemical shift (see Table 3.10). The statistical distributions, normalised such that their maximum is one, are drawn as contour plots at levels 0.1, 0.5 and 0.9. (C) Probabilities of observing each ^{13}C - ^1H shift pair in thymol for a given carbon nucleus. (D) Marginal individual assignment probabilities of unique directly bonded CH pairs and of pairs of topologically equivalent CH pairs (inset) in thymol. (E) Marginal individual assignment probabilities of unique carbons and of pairs of topologically equivalent carbons (inset) in thymol using only ^{13}C chemical shift distributions. In (D) and (E), the dots indicate the experimentally determined correct assignment.

Figure 3.5 shows the assignment of ^{13}C and ^1H - ^{13}C chemical shifts of strychnine using the combination of spectral editing and correlated statistical distributions of chemical shifts. In **Figure 3.5D**, the chemical shifts of carbons without any proton attached were assigned using the one-dimensional ^{13}C chemical shift distributions of the associated nuclei. Carbons with a single bonded proton were assigned using the correlated ^1H - ^{13}C statistical chemical shift distributions. The carbons with two attached protons were assigned to pairs of correlated ^1H - ^{13}C chemical shifts, restricting the ^{13}C shift to be unique in each pair.

Figure 3.5E summarises the three most probable global assignments of strychnine. For each assignment, the global assignment is broken down into blocks by multiplicity, and then potentially into sub-blocks where there is no significant probability of overlap according to a threshold (here a factor 100 with respect to the highest probability for each nucleus). For each sub-assignment there is an associated probability. The most probable assignment of each block was found to match the experimentally determined one, except for the assignment of CH_2 groups, where the assignments of carbons 21 and 19 are swapped compared to the experimentally determined assignment. This is due to the large difference between the distribution of chemical shifts and experimental shift of carbon 19 (see **Figure 3.17**), which could come from an unusual intermolecular environment of that atomic site in the crystal structure.

We consider that a reliable assignment is difficult to extract from the set of global assignments and associated probabilities, especially in cases with a large number of overlapping distributions and shifts, which yield many possible global assignments. Marginalisation helps simplify the analysis of global assignments and identify ambiguities more easily. This can be seen in **Figure 3.5D**, where the assignment of carbon 7 to shift “d” is favoured compared to shifts “b” and “c”, which suggests only a pairwise uncertainty between carbon 2 and 14.

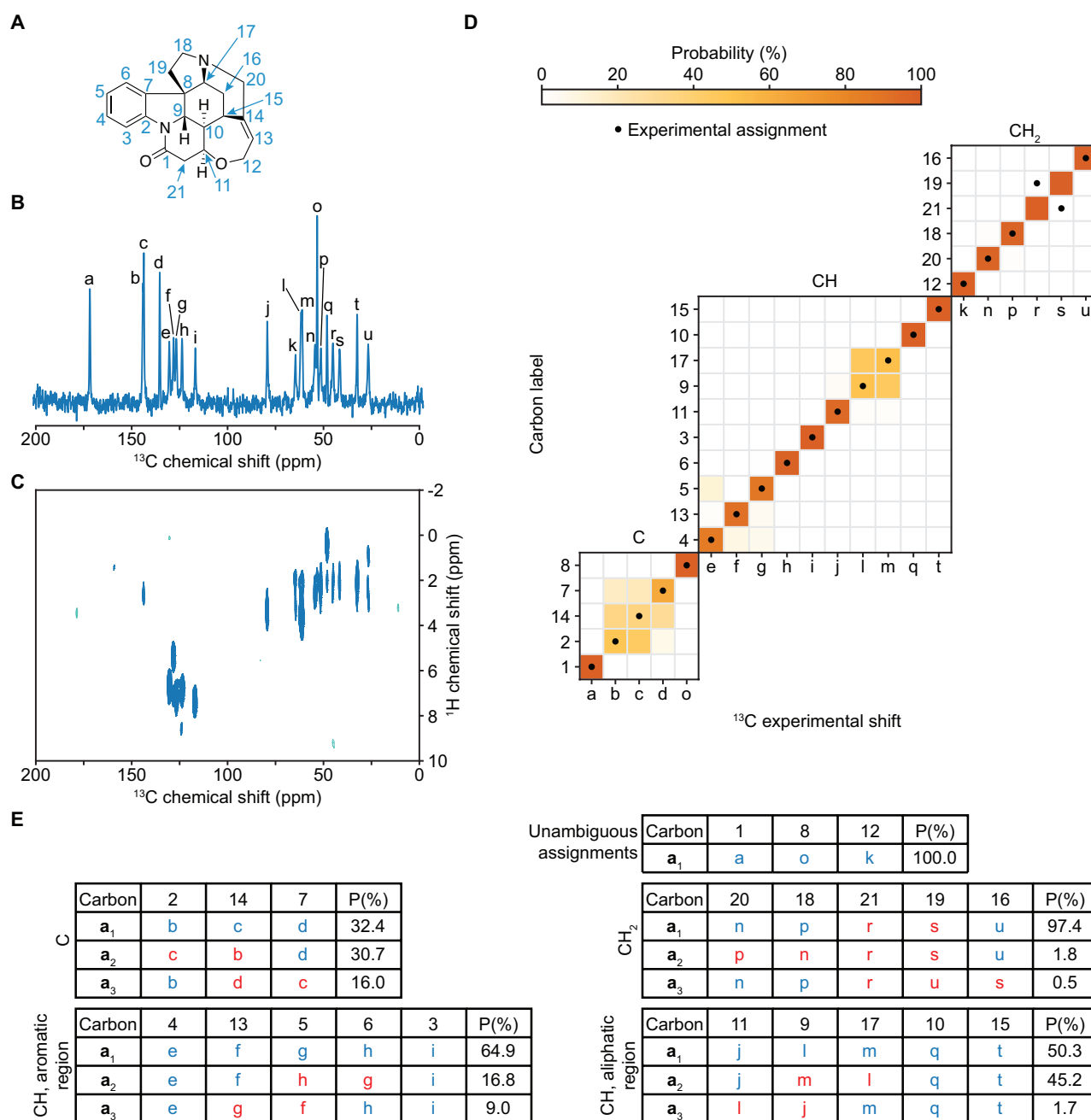


Figure 3.5. (A) Carbon labelling scheme of strychnine. (B) 125 MHz ^{13}C CPMAS NMR spectrum of strychnine. (C) ^1H - ^{13}C HETCOR spectrum of strychnine. (D) Marginal individual assignment probabilities of the carbon nuclei of strychnine. The carbon multiplicity is indicated above each probability map. The HETCOR shifts were used to assign CH and CH₂ carbons. The shifts are labelled alphabetically in order of decreasing chemical shift (see Table 3.11) (E) The three most probable global assignments for the different blocks assigned individually along with their probability. The individual assignments making up the global assignments are indicated in blue if they correspond to the experimentally determined assignment, and in red otherwise. Carbons 1, 8 and 12 were assigned without ambiguity ($P = 100\%$) directly from the evaluation of their statistical distributions of chemical shifts on the observed shifts.

In addition to strychnine, shown in Figure 3.5, the marginal individual assignment probabilities obtained for a set of 10 selected molecules with complete experimental assignments (except for the two phenyl rings of ritonavir) using spectral editing and correlated ^1H - ^{13}C statistical chemical shift distributions are shown in Figure 3.6 and Figures 3.24-3.26. The assignment of carbon nuclei without any attached proton were obtained from the one-dimensional statistical distributions of ^{13}C chemical shifts. The statistical distributions of chemical shifts for each example are shown in Figures 3.11-3.18. Notably, the assignment of lisinopril was found to be possible even when omitting the water molecules present in the crystal structure. This highlights the ability of the method to obtain probabilistic assignments without prior knowledge about the presence of solvent in the crystal lattice (we note that organic solvents in the crystal structure can be identified by the presence of additional peaks observed in, e.g., the ^{13}C spectrum of the sample).

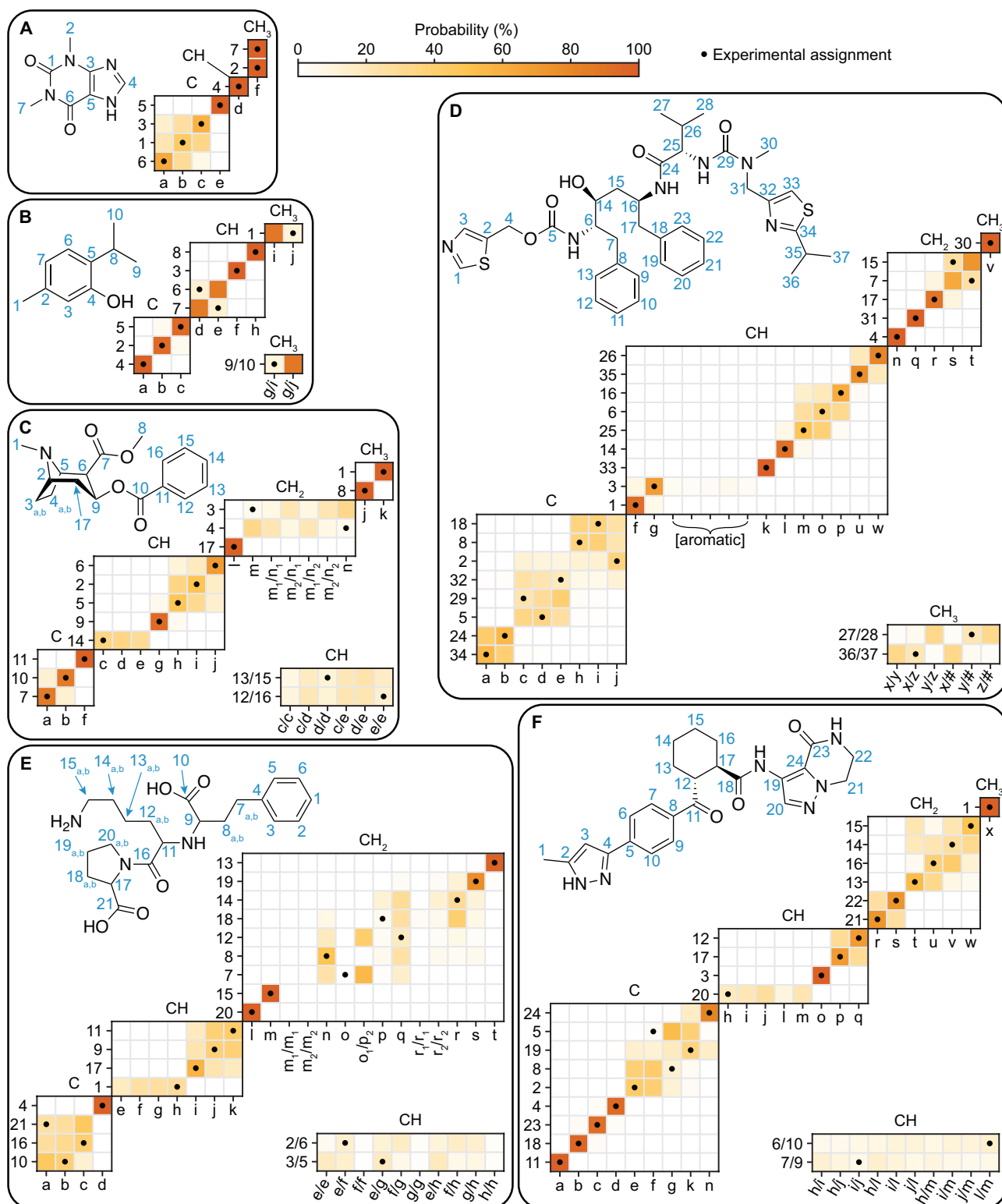


Figure 3.6. Marginal individual assignment probabilities of ^{13}C chemical shifts of (A) Theophylline, (B) thymol, (C) cocaine, (D) ritonavir, (E) lisinopril and (F) AZD5718 using correlated ^1H - ^{13}C chemical shift distributions and spectral editing. For each probability map, labels along the vertical axis indicate nuclei, and labels along the horizontal axis denote experimental shifts labelled alphabetically in order of decreasing ^{13}C shift (see Tables 3.9-3.14). The carbon multiplicity is indicated above each marginal assignment probability map. In (D), the assignment of carbons 9-13 and 19-23 is not shown as their experimental assignment is ambiguous. Nevertheless, the associated peaks were considered during the assignment process. The assignment probabilities of the aromatic CH groups of ritonavir are shown in Figure 3.19.

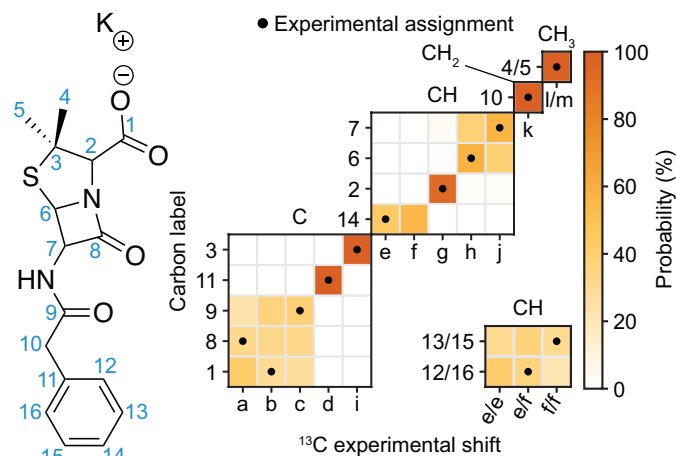


Figure 3.7. Carbon labelling scheme and marginal individual assignment probabilities of the K salt of penicillin G. The shifts are labelled “a” through “m” in order of decreasing chemical shift (see Table 3.16).

Figure 3.7 shows the assignment of the K salt of penicillin G. Only the organic ion was considered to construct the graph descriptors used to extract statistical distributions of chemical shifts. As for the presence of the water molecule in the case of lisinopril above, here the presence of the potassium ion, which is absent from the database, did not lead to a significant decrease in the ability of the model to predict the assignment, highlighting its generality beyond molecules for which chemical shifts can be computed by ShiftML. While ShiftML would not be able to compute shifts for crystals where even only one atom is different from C, H, N, O and S, this model only requires the molecule to be assigned to only contain these elements in order to obtain the probabilistic assignment. Of course, if the additional component in a salt or a co-crystal were to lead to a very different crystalline environment from those included in the database, this might lead to poor performance of the probabilistic assignment.

The marginal individual assignment probabilities obtained directly from the two-dimensional representation of the molecules were found to match the experimentally determined assignment in most cases. We observe that assignment ambiguities generally involve pairs or triplets of nuclei and shifts, leaving only a few possibilities for the NMR spectroscopist to further investigate in order to obtain the complete chemical shift assignment. Out of the 178 experimental individual assignments considered in **Figures 3.5-3.7** and **Figures 3.24-3.26**, only eight were associated with a probability below 10%, and two below 1%. These low probabilities were generally associated with crowded regions in experimental spectra, or with statistical outlier shifts compared to the distributions, which could have originated from unusual intermolecular environments.

In order to validate these results in a statistically significant manner, we evaluated the performance of the framework presented here on a benchmark set of a hundred crystal structures having between 10 and 20 different carbon atoms, randomly selected from the CSD database. In total, this corresponds to 1214 inequivalent carbon atoms. We used the ShiftML predicted shifts for each atom as the correct assignment, and excluded those shifts from the statistical distributions used to assign the molecules. The benchmark set was separated into five subsets containing 20 structures each that were evaluated independently in order to obtain standard deviations. Although using shifts predicted by ShiftML may introduce a bias, as the same method was used to construct the database of shifts, we assumed that the Gaussian width used to construct the statistical distributions of chemical shifts as well as the exclusion of the shifts assigned from the sets of shifts used to construct those distributions mitigate this issue.

Figure 3.8 summarises the performance of the probabilistic assignment model on the experimental (**Figure 3.8A**) and synthetic (**Figure 3.8B**) sets of molecules selected. The use of spectral editing and correlated ^1H - ^{13}C chemical shift distributions was found to improve the ability of the model to correctly assign carbon chemical shifts. Using either two-dimensional statistical distributions of chemical shifts, spectral editing, or combining both led to the experimental assignment being among the two most probable marginal assignments in over 80% of cases. Overall, the performances on the experimental benchmark set were consistent with the synthetic benchmark set, except when using spectral editing where a slight improvement in the experimental set compared to the synthetic set was observed.

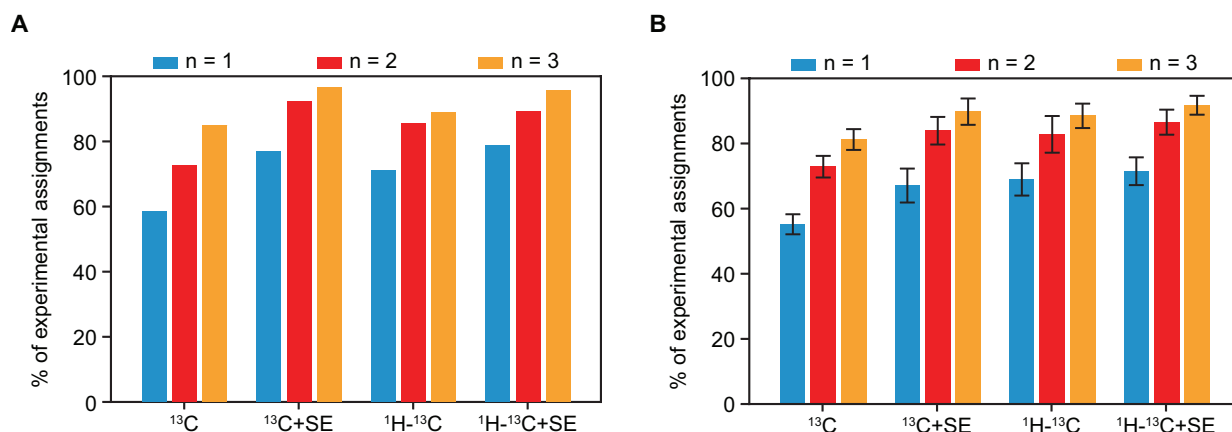


Figure 3.8. Comparison of probabilistic assignment performances using one-dimensional (^{13}C) or two-dimensional ($^1\text{H}-^{13}\text{C}$) statistical distributions, and including spectral editing (SE). Proportion of the experimental assignments being within the n ($n = 1, 2, 3$) most probable marginal individual assignments in the (A) experimental and (B) synthetic benchmark sets of molecules. Error bars indicate the standard deviation over the five subsets making up the synthetic benchmark set.

3.2.4 Conclusion

The framework presented in this section allows chemical shift assignment of organic crystals directly from their two-dimensional structure. This was achieved through the chemical shift prediction for over 200,000 organic crystal structures, which yields statistical distributions of chemical shifts for given covalent environments. A Bayesian framework was then used to obtain probabilistic marginal assignments of individual nuclei from the probabilities of the set of global assignments generated. Overall, using correlated $^1\text{H}-^{13}\text{C}$ chemical shift distributions in tandem with spectral editing, the method was found to include the experimental assignment among the two most probable marginal assignments in more than 80% of cases.

Furthermore, in most cases any ambiguity is found in small subgroups of shifts. This is highlighted in lisinopril in, for example, the CH_2 carbons because of significant overlap between the corresponding statistical distributions of chemical shifts and due to similar experimental shifts (see **Figure 3.15**).

In summary, the approach presented here can provide marginal assignments based only on the two-dimensional molecular structure, where typically most of the resonances will be assigned with high probabilities, and only a few resonances will show ambiguities among doubles or triples that can then be the subject of targeted experiments for disambiguation if needed, or left unresolved and assigned such that the error is minimised when compared with computed shifts for model structures (e.g., when performing NMR-driven crystal structure determination). This can greatly accelerate the assignment process. In particular the method is shown to provide assignments for molecules such as strychnine, lisinopril, AZD5718 and ritonavir, which have crowded ^{13}C spectra with between 20 and 40 distinct carbons, and which would have been previously completely unaddressable without resorting to natural abundance $^{13}\text{C}-^{13}\text{C}$ correlations. For example, in strychnine, of the 21 carbons, 14 are correctly assigned with more than 75% confidence. The model was also successfully applied to the assignment of a hydrate and an organic salt, with no significant performance loss compared to the benchmark set. We expect that a more accurate model of chemical shifts could lead to improved probabilistic assignment through the framework presented here.

The method shown here is not restricted to ^1H and ^{13}C , and can be used to assign the isotropic shifts of any NMR-active isotope of hydrogen, carbon, nitrogen and oxygen in principle. To illustrate that, **Figure 3.27** and **Appendix IV** describe the probabilistic assignment of the ^{15}N shifts of AZD5718.

The code is publicly available at <https://github.com/manucordova/ProbAsn> and a user guide is available in **Appendix IV** as well as on the Github webpage. A suggested workflow to assign an organic solid is also described in **Appendix IV**.

Further improvements to the method presented here can be achieved from more accurate chemical shift predictions and larger databases of structures and associated chemical shifts. Since the original publication of this project, we have recomputed the shifts of the structures in the database using ShiftML2 to improve the prediction accuracy and expanded it to all structures accessible by the updated model, yielding 338,341 structures in total, and allowing the application of the method to compounds containing C, H, N, O, S, F, P, Cl, Na, Ca, Mg and K atoms.

3.2.5 Appendix IV

DFTB relaxation of proton positions. The hydrogen positions of the crystal structures selected were optimised using the semiempirical DFTB3-D3H5^{325, 326, 351} method and the 3ob-3-1^{350, 415} parameter set. The maximum angular momenta were set to s for hydrogen, p for carbon, oxygen and nitrogen, and d for sulphur. We used a Monkhorst-Pack grid of k-points³³⁸ corresponding to a maximum spacing of 0.05 Å⁻¹ in reciprocal space for all computations. Hubbard derivatives were set following Ref. 326.

Conversion of isotropic shielding to chemical shift. The conversion from isotropic shielding to chemical shift was performed by linear regression between ShiftML predicted shieldings on DFTB3-D3H5 proton relaxed structures from the database and their corresponding experimental shifts, keeping the slope fixed to -1. The lists of crystal structures and chemical shifts considered for each NMR nucleus were taken from Ref. 127 and 128, and are shown in **Tables 3.5-3.8**. Ambiguous assignments of chemical shifts were solved by selecting the assignment yielding the lowest shift RMSE.

Database construction and architecture. Because graphs of small molecules centered on different atoms may be identified as isomorphic, the central vertex was assigned the label “Y” prior to hashing. Similarly, the graphs used to construct statistical distributions of correlated chemical shifts of neighbouring atoms were constructed by assigning the labels “Y” to the central vertex and “Z” to the neighbouring vertex where the correlated chemical shift is considered.

The database is made of several directories named according to the type of nucleus considered for chemical shift prediction (H, C, N and O). For correlated shifts, the directories are named “X-Y”, where X and Y are the two nuclei containing chemical shift predictions, X being the central nucleus and Y being its neighbour. Inside each directory, comma-separated files contain the predicted shifts of nuclei of a given graph of depth $w = 1$, indicated by all first neighbours of the central nucleus, sorted alphabetically and separated by a dash in the filename. Each entry in 1D database files contains comma-separated fields corresponding to the CSD REFCODE of the crystal, the index of the nucleus, the associated predicted chemical shift and error, and hashes corresponding to graphs of depth $w = 2$ through $w = 6$ for the nucleus considered. Each entry in 2D database files contains comma-separated fields corresponding to the CSD REFCODE of the crystal, the index of the central nucleus and associated predicted chemical shift and error, the index of the neighbouring nucleus and associated predicted chemical shift and error, and hashes corresponding to graphs of depth $w = 2$ through $w = 6$ for the nucleus considered.

Database search was performed by generating graphs of depth $w = 1$ through $w = 6$ corresponding to the nucleus of interest, identifying the file to search using the graph of depth $w = 1$, then searching for the pattern given by comma-separated Weisfeiler-Lehman hashes of the graphs of depth $w = 2$ through $w = 6$ in the file. If the number of database entries was found to be less than 10, the search was reiterated after removing the last hash in the pattern to search, until this condition was met.

Generation of global assignments. In total, there are $N^{M-N} \cdot N!$ possible ways to assign M nuclei to N chemical shifts. For instance, this represents a total of over $1.7 \cdot 10^{14}$ possible global assignments of the 17 carbon nuclei of cocaine to the 13 ¹³C NMR shifts present in its spectrum. Several procedures were implemented to reduce the complexity of generating possible global assignments while ensuring that the most probable assignments are generated.

Reducing the complexity of generating plausible global assignments can be done by considering only the most probable individual assignments as possible. This was done by setting a threshold p_{thresh} such that only the shifts for which the probability of being observed originating from a given nucleus is higher than the maximum probability for this nucleus divided by p_{thresh} are considered. In most cases shown in this work, the threshold was set to a value of 100.

Reducing the number of possible individual assignments in this way may generate independent sets of nuclei and experimental shifts to generate possible global assignments for. Each such set can be considered individually, which breaks down the global assignment generation problem into smaller sub-problems, reducing the overall complexity of the process.

Global assignments were generated recursively by considering all possible assignments for a given set of nuclei and experimental shifts. Where needed, after generating a partial global assignment of the first r nuclei considered to r experimental shifts, only the N_r possible assignments yielding the highest partial scores were selected for the subsequent individual assignments. This was done to reduce the complexity of generating global assignments.

List of structures in the synthetic benchmark set. The CSD Refcodes of the structures selected in the synthetic benchmark set are:

Subset 1: HECFIE01, BXPLOL01, FIQDOA, QATKAY, WARYIX, WURJEA, BINHIS, CAJBOG, FAGFUP, LERGUM, PEYSIV01, TBMYSO, VUJWEC, AKUGUK, COYDOJ, RUFZEY, PCYPOL19, TEVYEV, FELDIM, FUHNUR.

Subset 2: AXAZUW, SLFNMB08, YIDHEA, QAJSUP, WEXWOL01, PALDUB, YOXDYI, TIGPEF, HEMCEK, VOXNIF, CABWEH10, HATQAX, SABNOA, CIKDAD, KEFCIH, DINLIV10, HIVBOE, DEMZON, PESQOW, OJIGET.

Subset 3: YAVFIL, SACRET, XIXQED, KEDJAE, HURBUR, QUPNIZ, OPUBAE01, NUHNUB, TCMUR, HYFURA, WOCFUP, ASPLOL, XANCAQ, MAZBEV, YIXQAY01, OVIJOU, LIKNAW, SADRAS, ACNTBP, YEYTEC.

Subset 4: UTOHUH, YAKGEW, IKONON, YURKIH, ZEJVES, FINWAD, YETPAR, PONSES, NAJBAB, XAVGAF, MIQKOM01, MUGLIJ, NUT-SOK, ZINLUF, LUTHUE, KAYXUD, CEVKUM, ZIKKEJ, MEYVAP, GAMNEO.

Subset 5: XEDWUY, EDARIN, ROTSOK, CAYPEY, QIQPEO, OPOJEJ, NIWNNUC, RIMPUA, JOWMAL01, MIHROK, EBUYIK, HABXEP, WIGYOC, WIZZEL, VERSOL10, HAVMIA, XATWUK, BINAPH18, WOBGOK, OPUSOI.

Assignment of other nuclei. In principle, assignment of isotropic shifts of any NMR-active isotope of hydrogen, carbon, nitrogen or oxygen is implemented. As an example, **Figure 3.27** shows the probabilistic assignment of ^{15}N shifts of AZD5718. We note that the number of nitrogen environments used to train ShiftML is lower than the number of ^1H or ^{13}C environments which makes ShiftML slightly less reliable for the prediction of nitrogen shifts, which translates into slightly less reliable chemical shift distributions for our probabilistic model. In order to circumvent these issues, we would advise users to require at least 100 instances from the database to construct smooth statistical distributions of ^{15}N chemical shifts. The direct assignment of all ^{15}N shifts yields pairwise ambiguities (see **Figure 3.27B**), which can be partially resolved by separating the assignment of the NH groups from the non-protonated nitrogens (which can be determined from simple CP experiments) (**Figure 3.27D-E**).

User guide for the code. The code is freely available at <https://github.com/manucordova/ProbAsn>. Installation instructions can be found on the Github webpage. The software is written in Python and uses a database in CSV format, which is available at <https://doi.org/10.24435/materialscloud:vp-ft>. The assignment can be performed by running the “run.py” script (python run.py input_file.in). Input file construction and examples are found in the Github repository. The minimal input to provide is the chemical structure of the molecule to assign (we strongly recommend the SMILES format, which explicitly contains connectivities between atoms and can easily be extracted from molecular drawing softwares such as Chemdraw and MarvinSketch), as well as the list of shifts or cross peaks to which the atoms of the molecule should be assigned.

The software can output the distribution of shifts of each nucleus in the molecule, as well as prior and marginal individual assignment probabilities, and the list of all global assignments generated and associated probabilities.

The first step is to generate the graph corresponding to each nucleus to assign in the molecule. This is done using the three functions “make_mol()”, “get_bonds()” and “generate_graphs()” from the “graph” module. This will also save the structure to a file for visualisation of the labels.

Once the graphs are constructed, the database is searched for corresponding matches. This is done using the “fetch_entries()” function from the “db” module. This function extracts the predicted shifts and errors to use for the construction of the statistical distributions, and generates the labels for the assignment. The labels start from “1” and increase following the order of the atoms in the input structure. Only the element to assign is considered for the labelling. This corresponds to the labels displayed by the software VESTA when opening the structure output in the previous step.

The distributions are then cleaned up by removing duplicate protons in, e.g., methyl/ NH_3 groups and gathering topologically equivalent nuclei. This is done by the functions “cleanup_methyl_protons()”, “cleanup_methyls()” and “cleanup_equivalent()” in the “sim” module.

By evaluating the statistical distributions at the observed shifts, the software extracts prior individual assignment probabilities. This is done using the “compute_scores_1D()” or “compute_scores_2D()” function in the “sim” module for 1D and 2D shifts, respectively.

Then, the possible individual assignments for each nucleus are extracted using the “get_possible_assignments()” function in the “assign” module. This uses the threshold p_{thresh} described above to evaluate which assignments are possible.

From the possible individual assignment for each nucleus, the function “get_probabilistic_assignment()” in the “assign” module generates the possible global assignments and evaluates their probabilities. This is done in several pools, which correspond to sets of distributions and shifts without any significant overlap, that can thus be assigned independently. This is typically the limiting step of the assignment process, as generating all possible global assignments has a factorial complexity in principle.

From the set of global assignments generated, the marginal individual assignment probabilities are extracted using the “update_split_scores()” function in the “assign” module.

Suggested assignment workflow. We propose a workflow aimed at optimising the time required to obtain the ^{13}C assignment of an organic solid as follows:

- (Optional) Generate the statistical distributions of shifts for all nuclei in the molecule using the software in order to identify the regions where each experimental shift is expected. This can also be used to evaluate whether spectral editing experiments are required.
- Record the ^{13}C CPMAS and ^1H - ^{13}C HETCOR spectra. If the molecule contains more than 15 distinct carbon atoms or statistical distributions of shifts overlap substantially, record ^{13}C spectral editing experiments by default.
- Use the software to assign the molecule with the extracted shifts.
- If too many ambiguities remain, record ^{13}C spectral editing experiments if not already done.
- Use the software to assign the molecule including the spectral editing data.
- If needed, perform targeted experiments to resolve any remaining ambiguity.

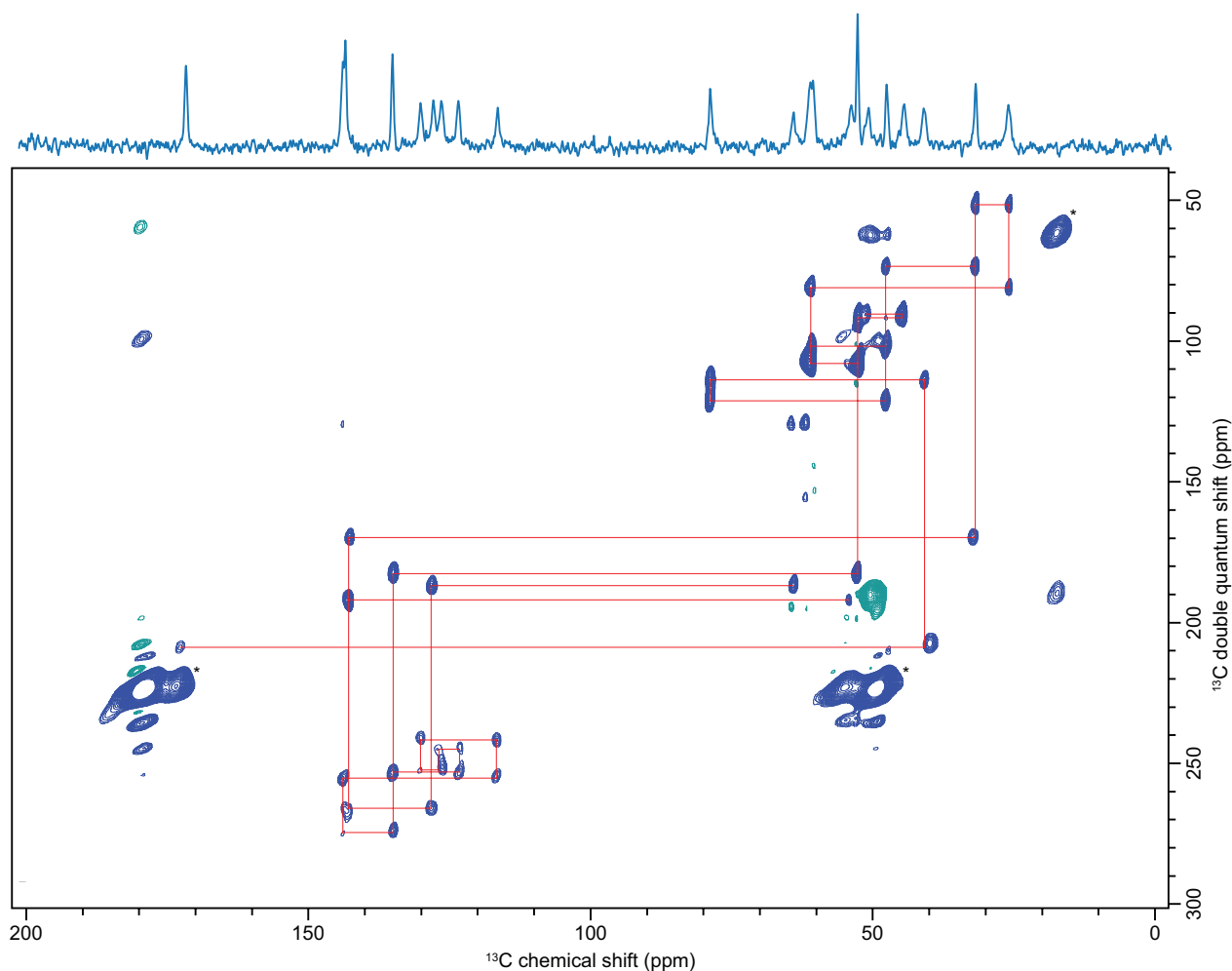


Figure 3.9. DNP-enhanced ^{13}C - ^{13}C refocused INADEQUATE spectrum of strychnine. The peaks indicated with a “*” are assigned to impurities introduced during sample preparation for DNP NMR.

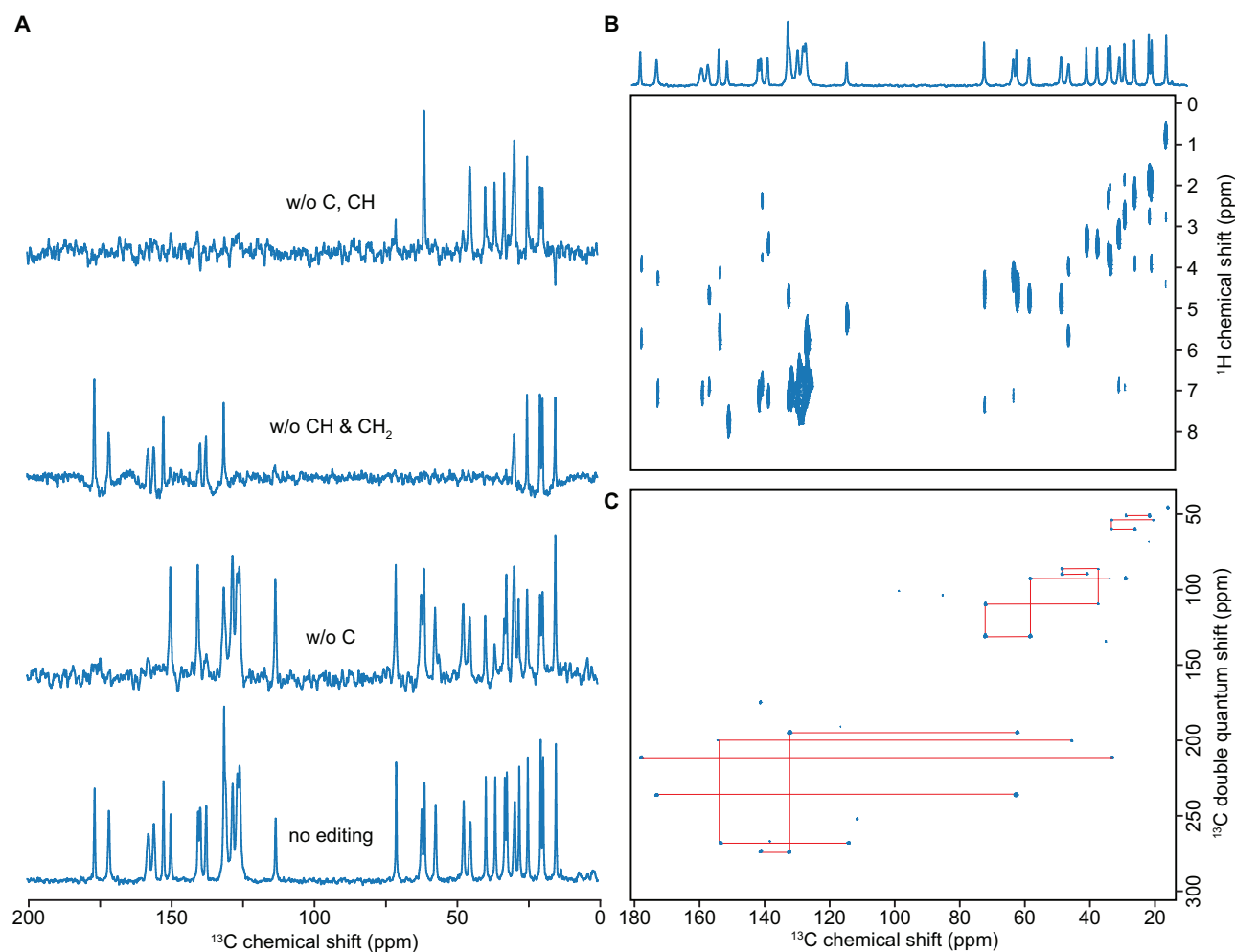


Figure 3.10. (A) Spectral editing of the ^{13}C CPMAS spectrum of ritonavir. (B) ^1H - ^{13}C HETCOR spectrum of ritonavir. (C) ^{13}C - ^{13}C refocused INADEQUATE spectrum of ritonavir.

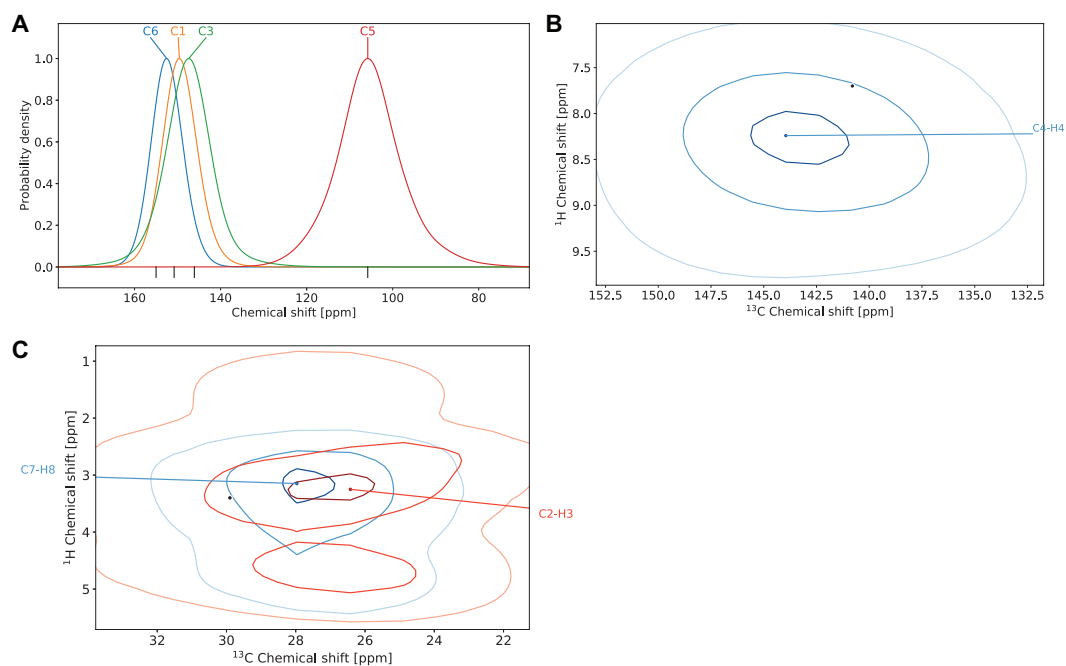


Figure 3.11. Chemical shift distributions of (A) C, (B) CH and (C) CH₃ carbons of theophylline. Experimental shifts are indicated by black vertical lines under ^{13}C chemical shift distributions, and as black dots in correlated ^1H - ^{13}C chemical shift distributions.

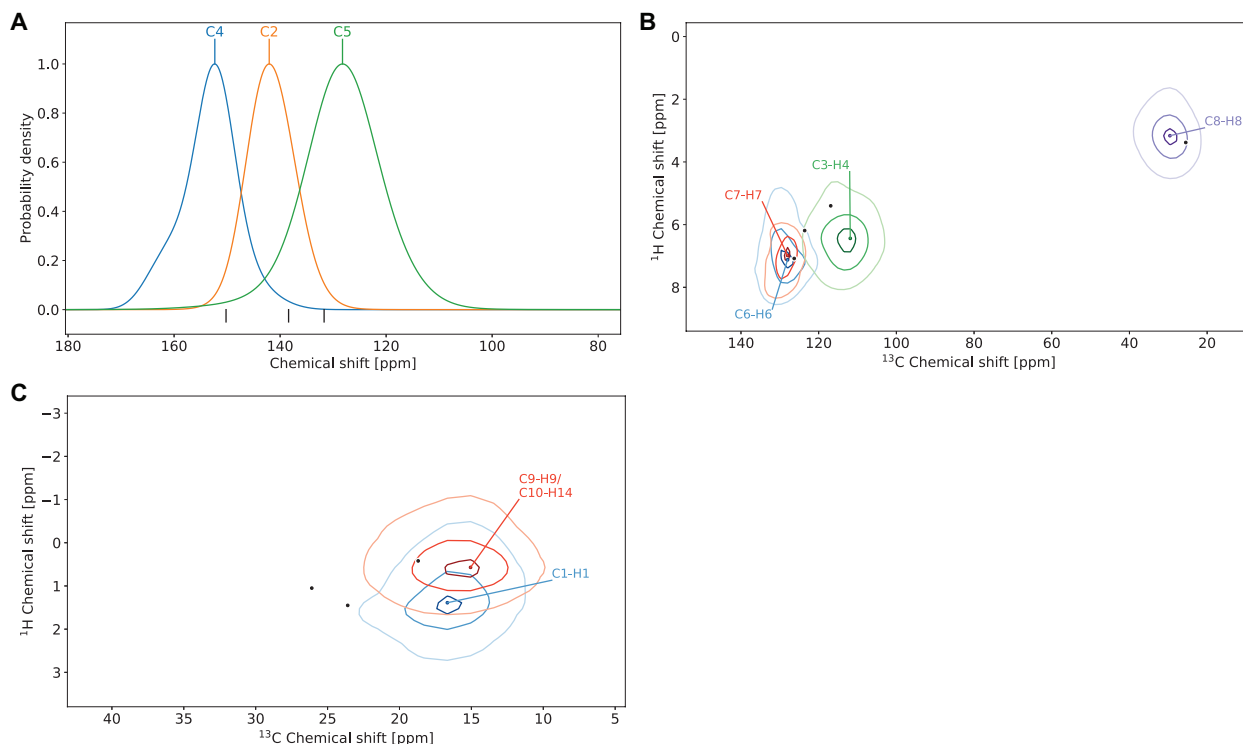


Figure 3.12. Chemical shift distributions of (A) C, (B) CH and (C) CH_3 carbons of thymol. Experimental shifts are indicated by black vertical lines under ^{13}C chemical shift distributions, and as black dots in correlated ^1H - ^{13}C chemical shift distributions.

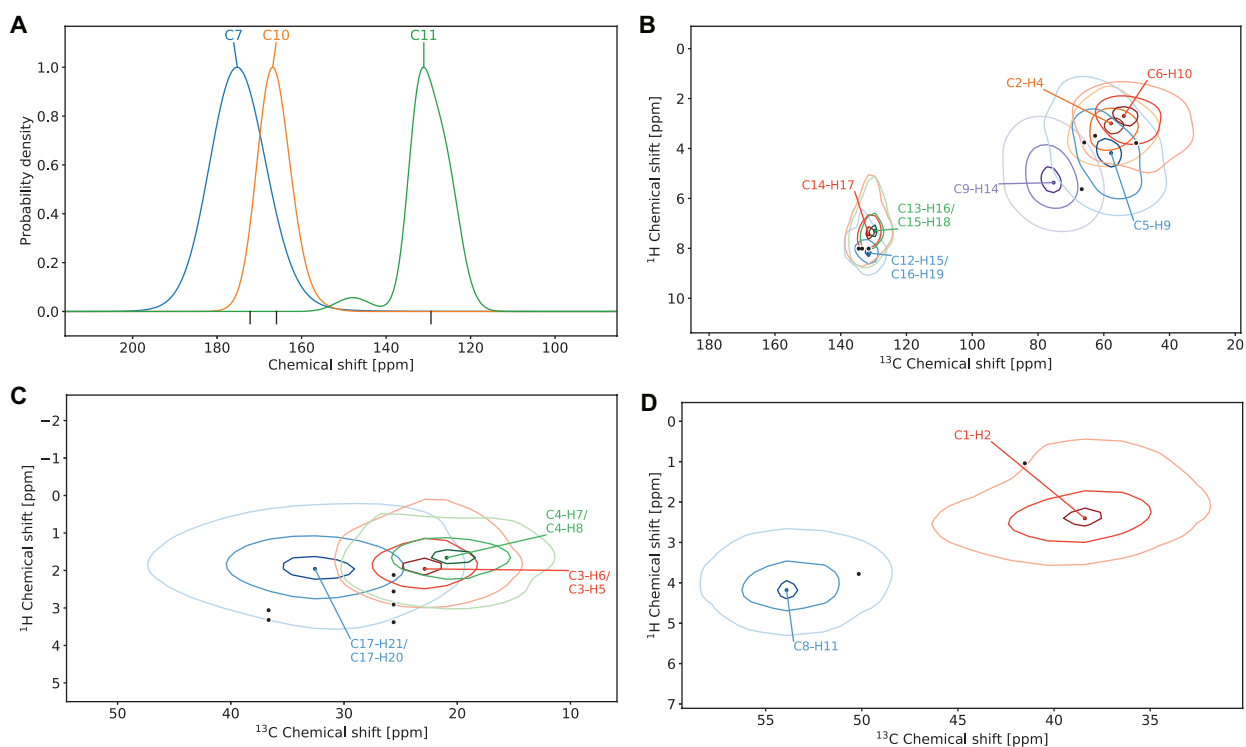


Figure 3.13. Chemical shift distributions of (A) C, (B) CH, (C) CH_2 and (D) CH_3 carbons of cocaine. Experimental shifts are indicated by black vertical lines under ^{13}C chemical shift distributions, and as black dots in correlated ^1H - ^{13}C chemical shift distributions.

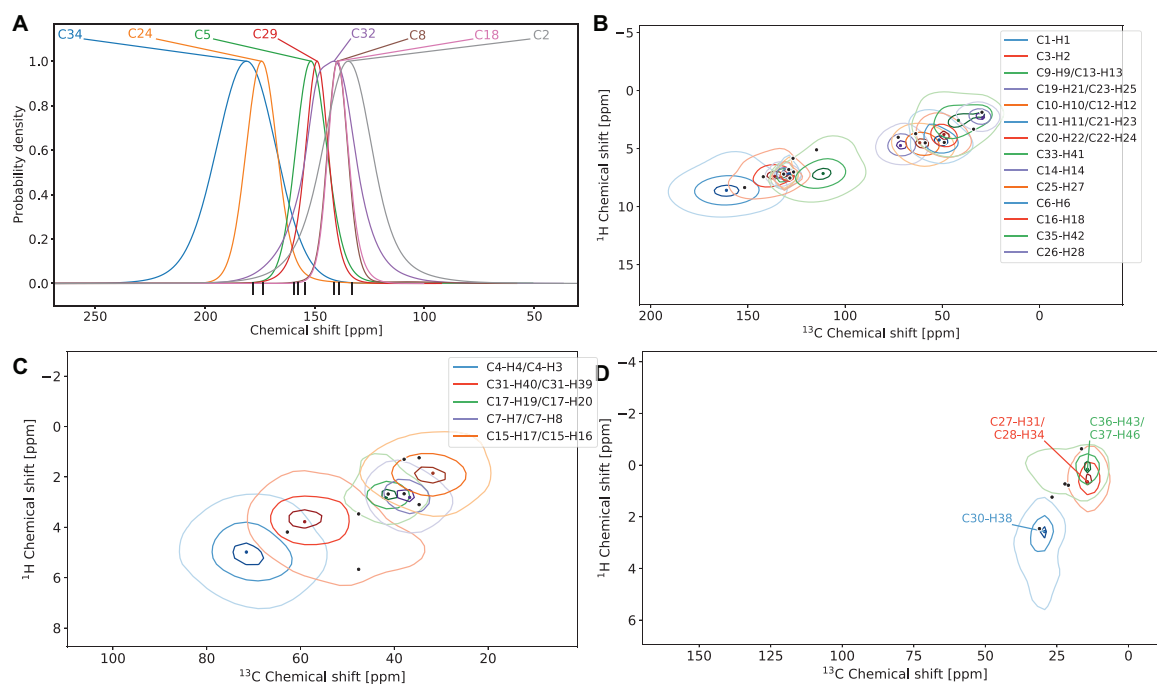


Figure 3.14. Chemical shift distributions of (A) C, (B) CH, (C) CH₂ and (D) CH₃ carbons of ritonavir. Experimental shifts are indicated by black vertical lines under ¹³C chemical shift distributions, and as black dots in correlated ¹H-¹³C chemical shift distributions.

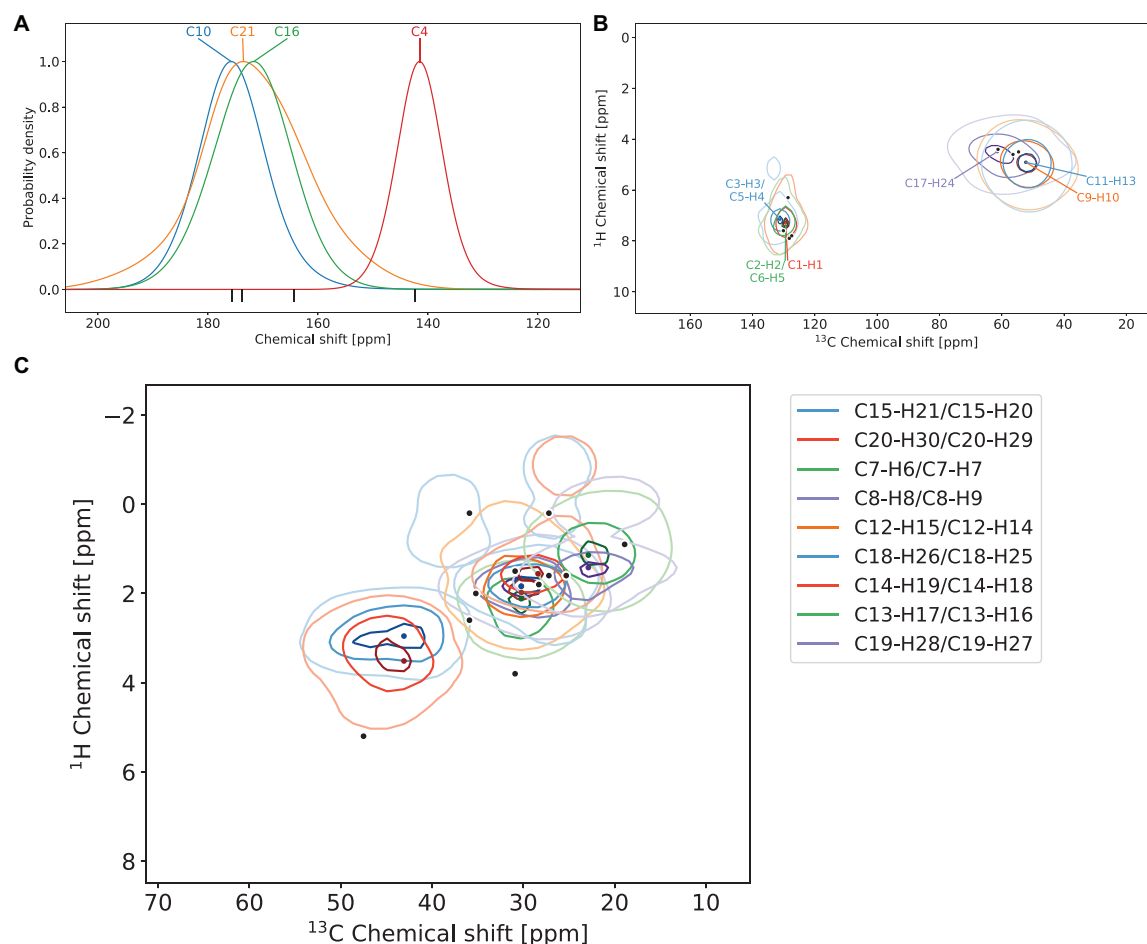


Figure 3.15. Chemical shift distributions of (A) C, (B) CH and (C) CH₂ carbons of lisinopril. Experimental shifts are indicated by black vertical lines under ¹³C chemical shift distributions, and as black dots in correlated ¹H-¹³C chemical shift distributions.

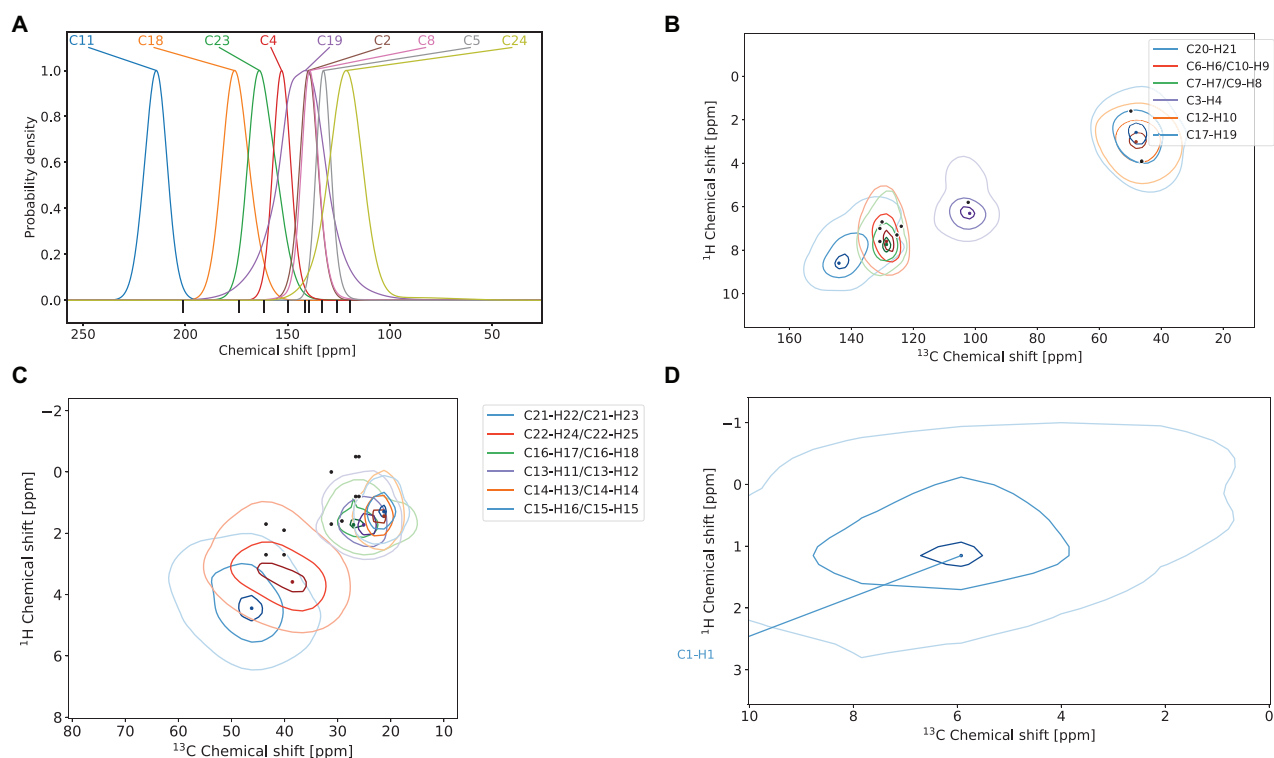


Figure 3.16. Chemical shift distributions of (A) C, (B) CH, (C) CH₂ and (D) CH₃ carbons of AZD5718. Experimental shifts are indicated by black vertical lines under ¹³C chemical shift distributions, and as black dots in correlated ¹H-¹³C chemical shift distributions.

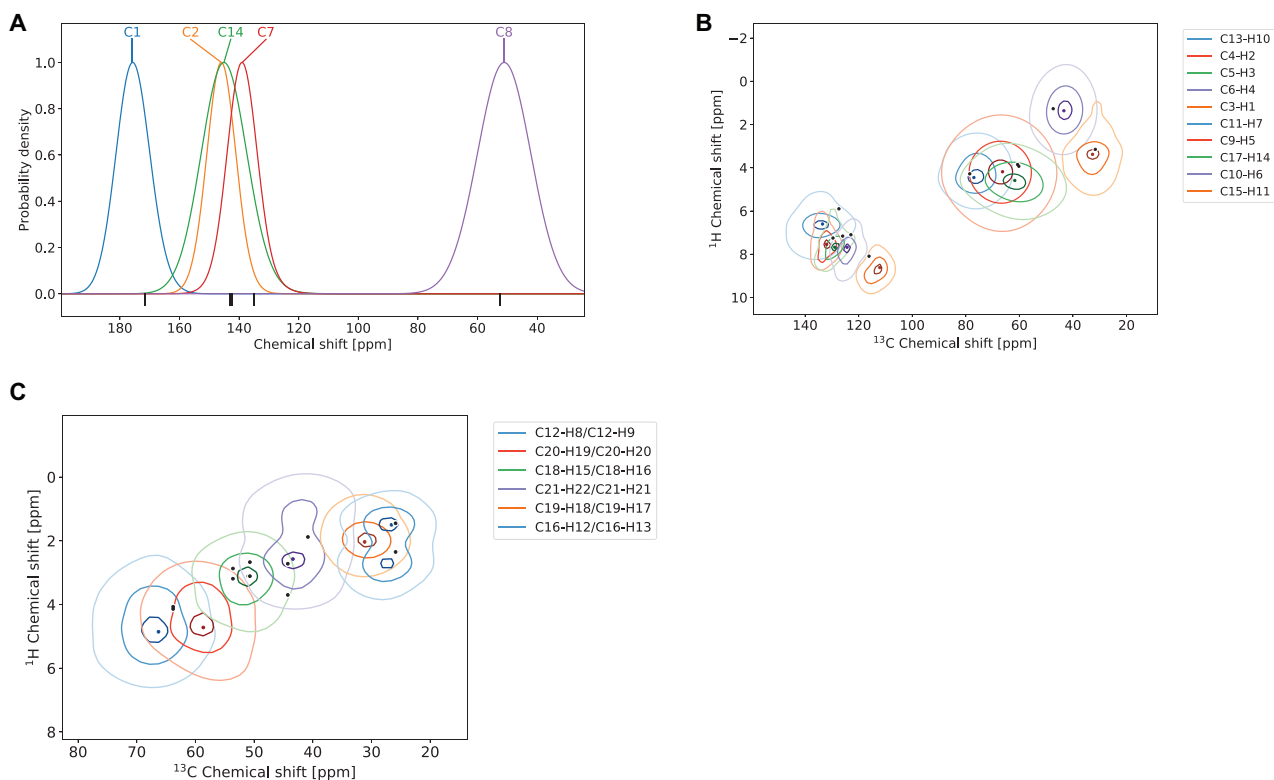


Figure 3.17. Chemical shift distributions of (A) C, (B) CH and (C) CH₂ carbons of strychnine. Experimental shifts are indicated by black vertical lines under ¹³C chemical shift distributions, and as black dots in correlated ¹H-¹³C chemical shift distributions.

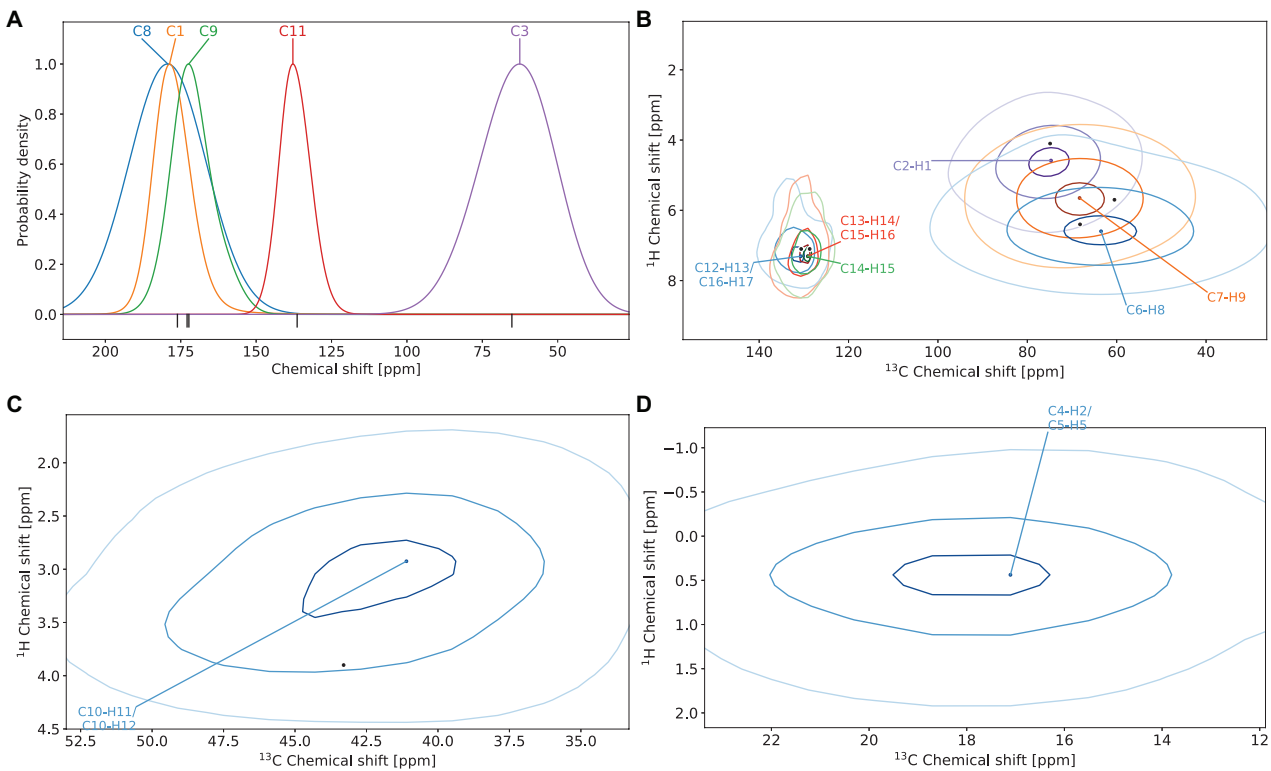


Figure 3.18. Chemical shift distributions of (A) C, (B) CH, (C) CH₂ and (D) CH₃ carbons of the K salt of penicillin G. Experimental shifts are indicated by black vertical lines under ¹³C chemical shift distributions, and as black dots in correlated ¹H-¹³C chemical shift distributions.

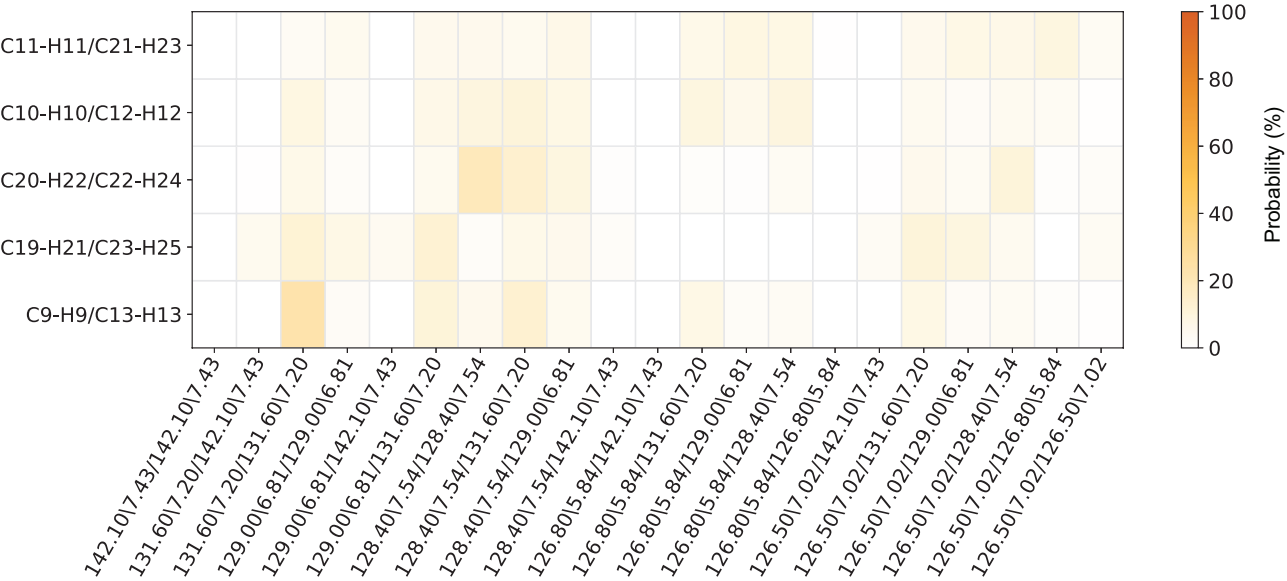


Figure 3.19. Marginal individual assignment probabilities of pairs of aromatic CH groups in ritonavir to their corresponding shifts.

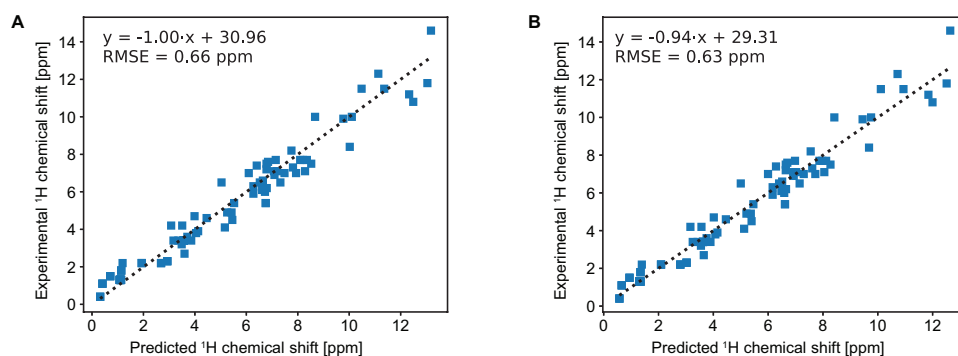


Figure 3.20. Linear regression between predicted ^1H shielding and experimental shifts **(A)** with a slope fixed to -1, and **(B)** allowing the slope to vary.

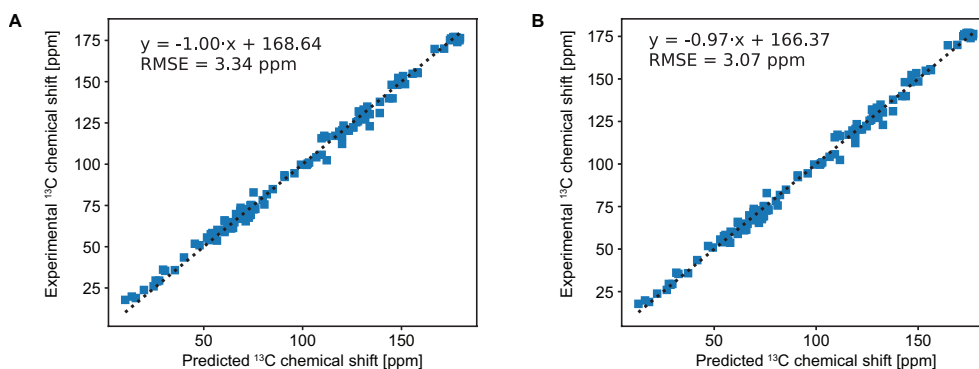


Figure 3.21. Linear regression between predicted ^{13}C shielding and experimental shifts **(A)** with a slope fixed to -1, and **(B)** allowing the slope to vary.

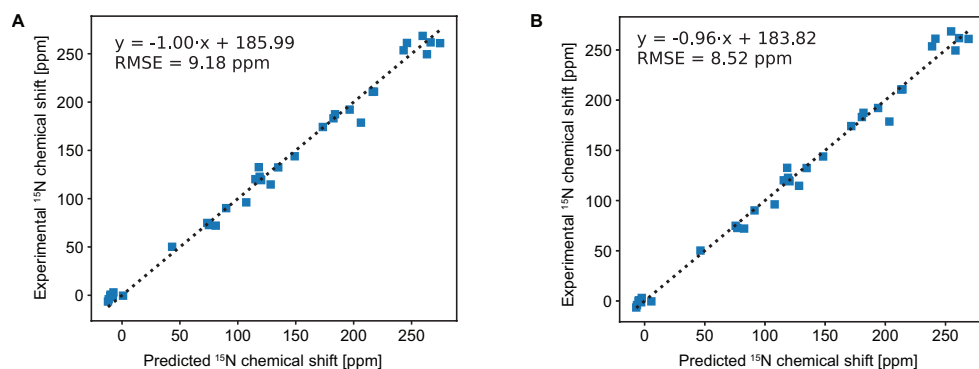


Figure 3.22. Linear regression between predicted ^{15}N shielding and experimental shifts **(A)** with a slope fixed to -1, and **(B)** allowing the slope to vary.

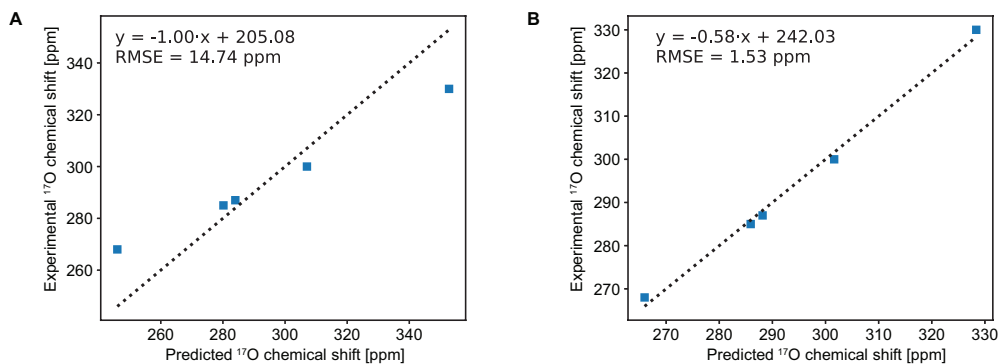


Figure 3.23. Linear regression between predicted ^{17}O shielding and experimental shifts **(A)** with a slope fixed to -1, and **(B)** allowing the slope to vary.

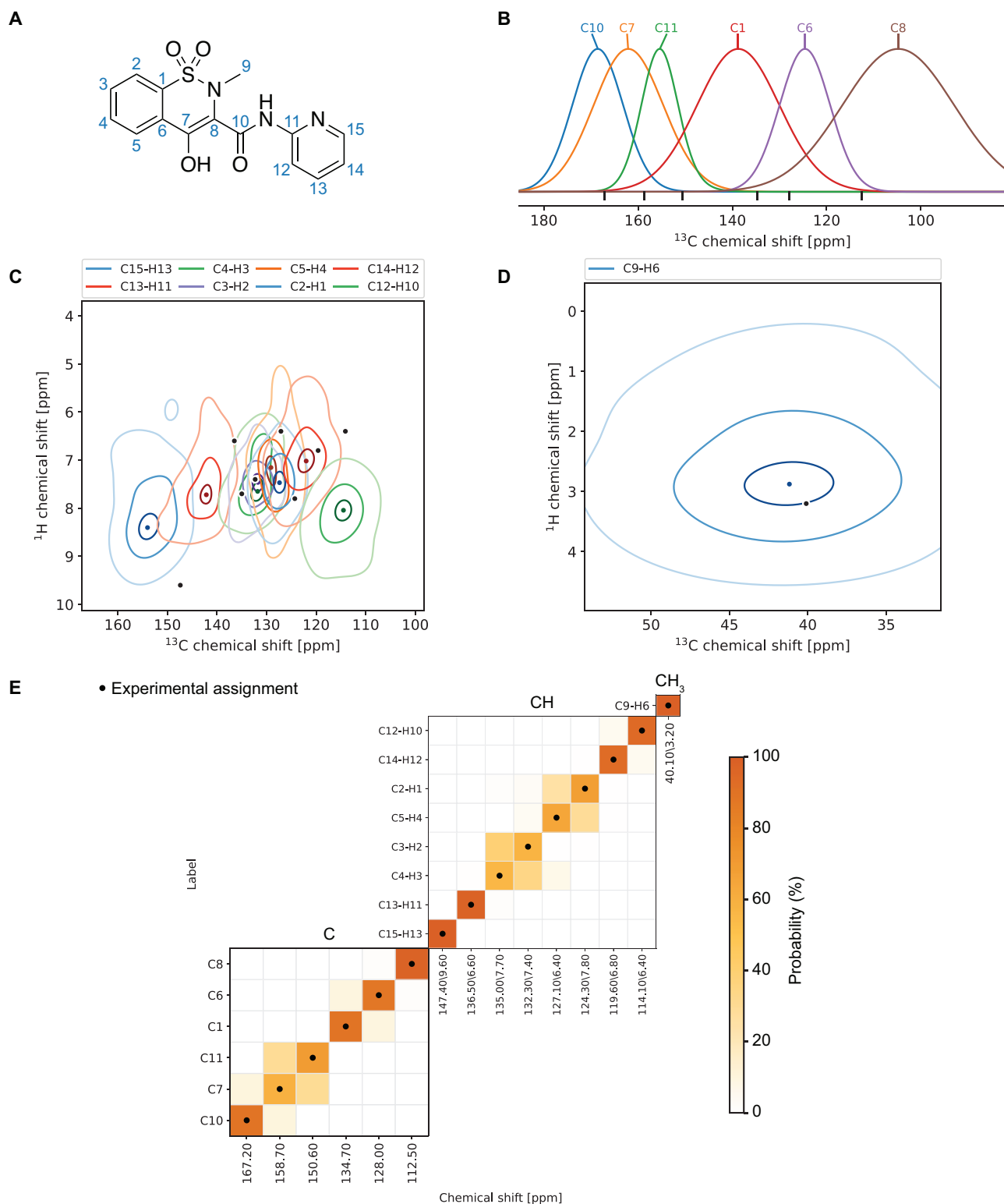


Figure 3.24. (A) Carbon labelling scheme and chemical shift distributions of the (B) C, (C) CH and (D) CH₃ carbons of β -piroxicam. (E) Marginal individual assignment probabilities of ^{13}C chemical shifts using correlated ^1H - ^{13}C chemical shift distributions and spectral editing.

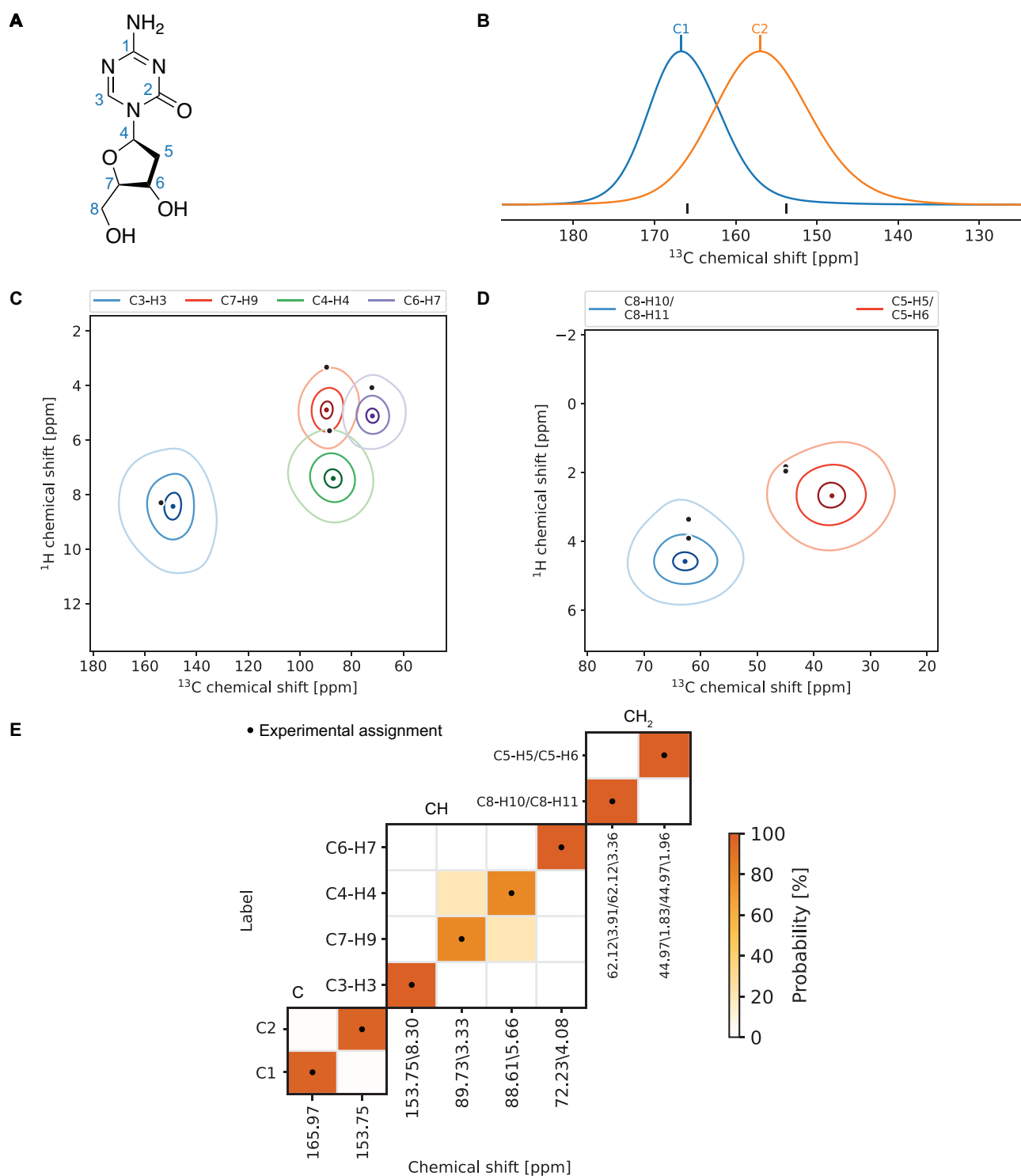


Figure 3.25. (A) Carbon labelling scheme and chemical shift distributions of the (B) C, (C) CH and (D) CH₂ carbons of decitabine. (E) Marginal individual assignment probabilities of ¹³C chemical shifts using correlated ¹H-¹³C chemical shift distributions and spectral editing.

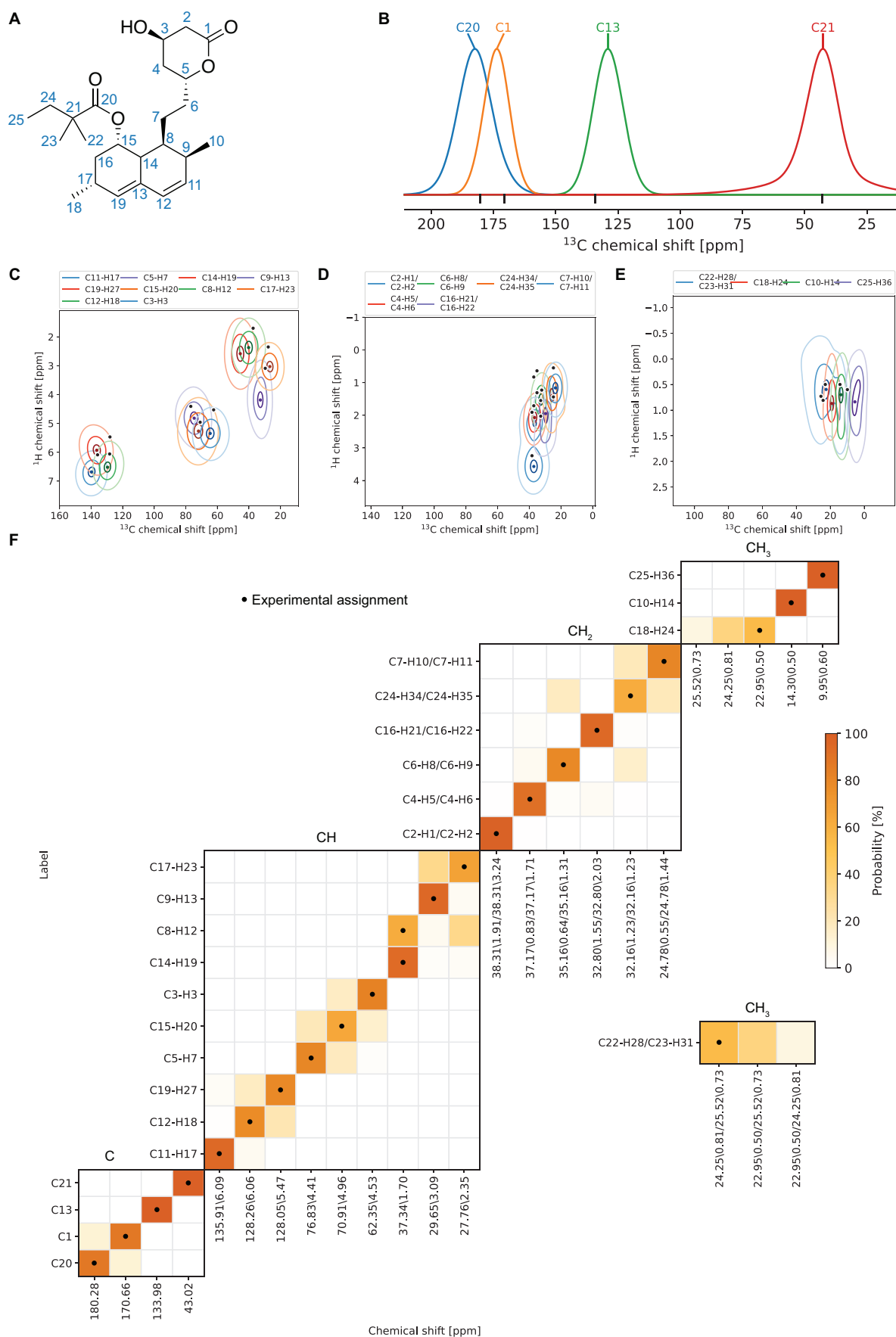


Figure 3.26. (A) Carbon labelling scheme and chemical shift distributions of the (B) C, (C) CH, (D) CH₂ and (E) CH₃ carbons of simvastatin. (F) Marginal individual assignment probabilities of ¹³C chemical shifts using correlated ¹H-¹³C chemical shift distributions and spectral editing.

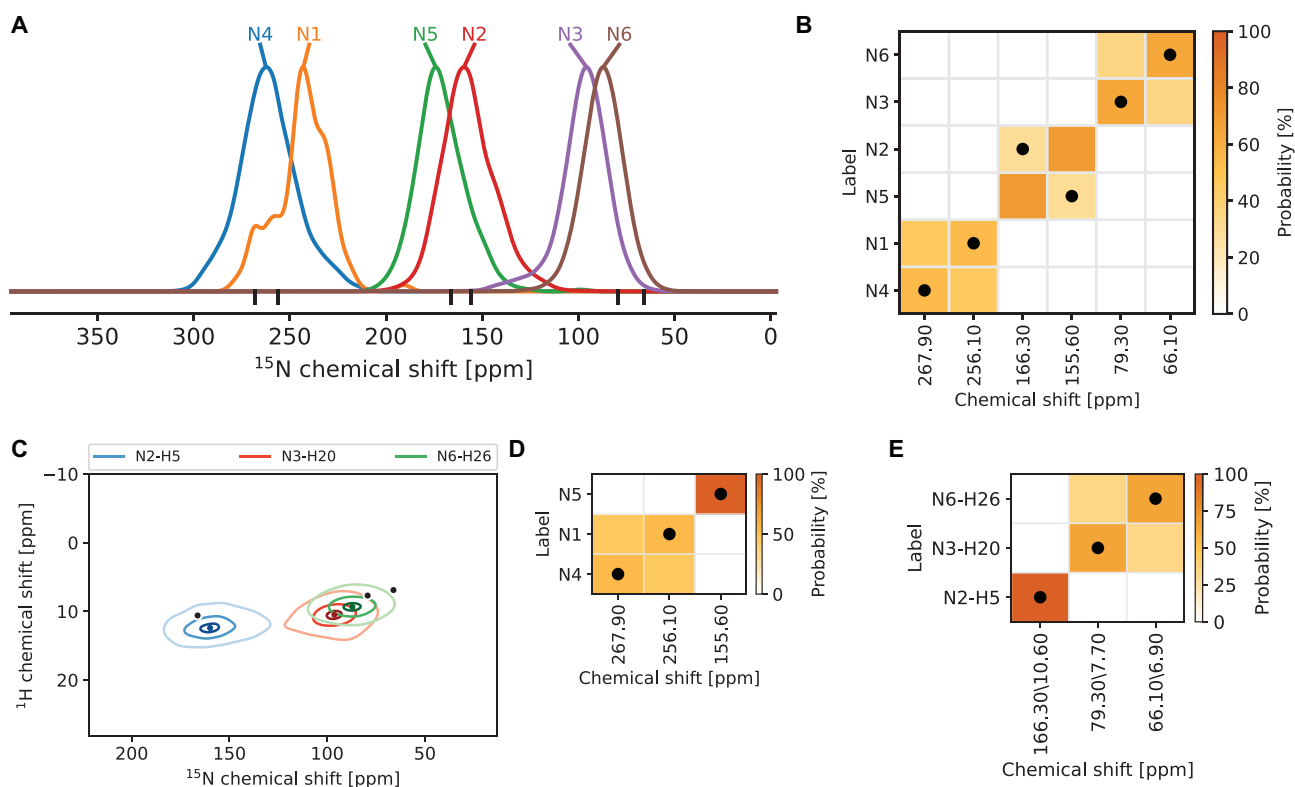


Figure 3.27. (A) Statistical chemical shift distributions of nitrogen atoms of AZD5718. The black vertical lines indicate the experimental shifts. (B) Marginal individual assignment probabilities of the ^{15}N shifts of AZD5718. The black dots indicate the experimentally determined assignment.

Table 3.1. Experimental parameters for 1D experiments on ritonavir.

	^1H	^{13}C	^{13}C w/o primary and secondary	^{13}C w/o quaternary	^{13}C w/o quaternary and primary
MAS rate	22 kHz	12.5 kHz	12.5 kHz	12.5 kHz	12.5 kHz
Recycle delay (d_1)	5 s	2 s	2 s	3s	3s
^1H to X CP					
Spin lock duration	-	3 ms	3 ms	3 ms	3 ms
Delay after acquisition for primary and secondary carbon filtering	-	-	0.5 ms	-	-
Total acquisition time	4 ms	30 ms	30 ms	30 ms	30 ms
Dwell time	1 μs	5 μs	5 μs	5 μs	5 μs
Number of points	4096	6144	6144	6144	6144
Number of scans	4	512	64	512	12288
Acquisition mode	qsim	qsim	qsim	qsim	qsim

Table 3.2. Experimental parameters for 2D experiments on ritonavir.

	^1H - ^{13}C HETCOR	^{13}C - ^{13}C INADEQUATE
MAS rate	22 kHz	12.5 kHz
Recycle delay (d_1)	2.7 s	2.15 s
^1H to X CP		
Spin lock duration	0.1 and 1.0 ms	3.5 ms
Acquisition in the indirect dimension (t_1)		
Total acquisition time	4.6 ms	2.6 ms
Dwell time	96 μs	20 μs
Number of points	96	256
Acquisition in the direct dimension (t_2)		
Total acquisition time	40.5 ms	25 ms
Dwell time	13.2 μs	5 μs
Number of points	256	2494
Number of scans per increment	16	1'536
Acquisition mode	States-TPPI	States-TPPI
Delay t	-	3.6 ms

Table 3.3. Experimental parameters for 1D experiments on strychnine.

	^1H	^{13}C
MAS rate	111 kHz	22 kHz
Recycle delay (d_1)	26 s	120 s
^1H to X CP		
Spin lock duration	-	1 ms
Total acquisition time	8.4 ms	30 ms
Dwell time	2.8 μs	13.2 μs
Number of points	2998	2268
Number of scans	4	32
Acquisition mode	DQD	qsim

Table 3.4. Experimental parameters for 2D experiments on strychnine.

	^1H - ^{13}C HETCOR	^{13}C - ^{13}C INADEQUATE
MAS rate	22 kHz	12.5 kHz
Recycle delay (d_1)	120 s	5 s
^1H to X CP		
Spin lock duration	0.1 ms	3 ms
Acquisition in the indirect dimension (t_1)		
Total acquisition time	2.8 ms	2.6 ms
Dwell time	88 μs	20 μs
Number of points	64	256
Acquisition in the direct dimension (t_2)		
Total acquisition time	30 ms	15 ms
Dwell time	13.2 μs	5 μs
Number of points	2268	2988
Number of scans per increment	34	128
Acquisition mode	States-TPPI	States-TPPI
Delay t	-	4 ms

Table 3.5. List of experimental and predicted ^1H chemical shifts of reference compounds used for shielding to shift conversion. All shifts are reported relative to TMS.

CSD REFCODE	Experimental shift (ppm)	ShiftML predicted shielding (ppm)
CIMETD ⁴²³	11.8	17.92
	7.6	24.12
	8.4	20.94
	9.9	21.19
	2.2	29.03
	4.2	27.46
	4.7	26.98
	2.2	29.78
	4.2	27.89
	2.7	27.37
	3.6	27.26
	2.2	28.28
URACIL ⁴²⁴	7.5	22.44
	10.8	18.47
	11.2	18.63
	6.0	24.24
AMBACO05 ⁴²⁵	6.5	24.44
	7.7	22.61
	6.5	25.93
	6.5	23.64
	5.4	24.21
	12.3	19.83
IPMEPL ⁴⁹	5.4	25.44
	6.2	24.17
	7.1	23.89
	3.4	27.61
	1.1	30.56
	1.5	30.26
	0.4	30.65
	10.0	22.29

COYRUD11 ⁴²⁶	7.0	23.03
	6.1	24.35
	3.8	26.92
	4.5	25.51
	4.1	25.80
	5.9	24.69
	3.2	27.48
	1.8	29.83
	2.3	28.04
	11.5	19.60
BAPLOT01 ⁵²	14.6	17.78
	7.7	23.82
	3.4	27.40
ZIVKAQ ¹³⁶	6.9	23.87
	6.6	24.32
	7.0	23.48
	7.5	24.18
	10.0	20.86
	4.9	25.71
	8.2	23.21
	3.4	27.80
	3.4	27.11
	7.7	22.87
	7.7	22.76
	1.3	29.89
	7.3	23.15
	7.2	24.18
	6.3	24.69
	7.4	24.56
	11.5	20.48
	4.9	25.54
	7.0	24.87
	3.9	26.84
	4.6	26.51
	7.1	22.68
	7.1	23.77
	1.3	29.84

Table 3.6. List of experimental and predicted ¹³C chemical shifts of reference compounds used for shielding to shift conversion. All shifts are reported relative to TMS.

CSD REFCODE	Experimental shift (ppm)	ShiftML predicted shielding (ppm)
MBDGAL02 ⁴²⁷	105.7	59.01
	71.2	99.75
	72.1	95.81
	69.3	94.88
	75.6	87.97
	62.8	106.64
	57.6	115.11
MEMANP11 ⁴²⁷	99.6	66.58
	71.3	95.60
	71.7	96.80
	64.8	103.15
	71.9	97.42
	58.9	108.13
	54.9	112.92

MGALPY01 ⁴²⁷	100.4	66.05
	67.6	95.9
	72.6	93.58
	70.0	98.14
	72.9	92.53
	61.4	103.82
	55.2	114.02
MGLUCP11 ⁴²⁷	101.0	65.38
	72.3	97.18
	74.6	93.78
	72.5	96.72
	75.3	94.66
	63.8	704.6
	56.5	113.92
XYLOBM01 ⁴²⁷	104.2	61.71
	72.2	95.58
	78.2	88.84
	69.5	101.54
	66.9	100.3
	57.3	114.37
SUCROS04 ⁴²⁸	93.3	77.82
	66.0	108.12
	73.7	93.05
	102.4	56.45
	72.8	97.73
	82.9	93.47
	67.9	99.72
	71.8	100.19
	73.6	100.03
	81.8	86.79
	60.0	108.17
	61.0	104.85
RHAMAH12 ⁴²⁹	94.5	72.77
	72.2	95.70
	71.0	98.68
	72.5	97.46
	69.8	102.3
	17.8	158.37
FRUCTO02 ⁴³⁰	65.4	107.97
	99.7	69.48
	67.2	101.24
	69.0	95.61
	71.4	96.39
	64.9	104.19
GLYCIN29 ⁴³¹	176.2	-6.4
	43.5	128.64
LALNIN12 ⁴³²	176.8	-9.28
	50.9	120.77
	19.8	155.03
LSERIN01 ⁴³³	175.1	-5.91
	55.6	116.98
	62.9	106.71
LSERMH10 ⁴³⁴	175.6	-9.75
	58.3	114.31
	61.8	105.28

ASPARM03 ⁴³⁵	176.4	-10.96
	51.8	123.14
	36.1	139.21
	177.1	-8.41
LTHREO01 ⁴³⁶	170.0	-2.92
	60.2	111.74
	65.4	97.42
	18.9	153.0
GLUTAM01 ⁴³³	177.0	-7.57
	54.0	116.17
	26.0	144.02
	29.0	142.14
	174.0	-9.61
LTYROS11 ⁴³³	176.0	-10.55
	123.0	34.71
	130.3	34.70
	54.7	115.46
	131.0	29.63
	155.7	10.69
	117.2	57.68
	35.8	133.26
	117.2	52.11
LCYSTN21 ⁴³³	35.4	138.16
	53.7	111.97
	175.1	-10.71
NAPHTA36 ⁴³⁷	125.4	41.47
	129.3	38.74
	134.9	35.93
	129.9	38.03
	126.0	40.65
ACENAP03 ⁴³⁸	148.1	20.73
	120.3	46.63
	129.4	39.37
	122.3	43.15
	131.9	40.00
	139.9	23.15
	29.5	141.21
	148.1	23.62
	120.3	45.32
	129.4	38.76
	122.3	42.96
	131.9	40.25
	139.9	24.95
	29.5	142.99
HXACAN09 ⁵⁴	152.3	20.27
	116.4	56.49
	120.6	48.57
	133.1	37.79
	123.4	47.95
	115.7	59.10
	169.8	1.87
	23.8	148.90
SULAMD06 ⁴³⁹	127.1	36.99
	129.5	39.69
	117.1	48.69
	153.4	17.83

	112.3	48.69
	129.5	39.69
ADENOS12 ⁴⁴⁰	154.8	13.04
	148.5	16.84
	119.7	50.14
	155.2	10.42
	137.8	29.55
	92.3	77.71
	71.2	95.02
	75.0	95.05
	84.9	83.73
	62.7	105.42

Table 3.7. List of experimental and predicted ¹⁵N chemical shifts of reference compounds used for shielding to shift conversion. All shifts are reported relative to NH₄Cl with NH₃(l) at 39.3 ppm.

CSD REFCODE	Experimental shift (ppm)	ShiftML predicted shielding (ppm)
BITZAF ⁴⁴¹	249.5	-77.29
GEHHAD ⁴⁴¹	253.6	-57.14
	261.8	-80.44
GEHHEH ⁴⁴¹	187.4	2.07
	261.0	-88.63
GEHHIL ⁴⁴¹	268.5	-73.58
	261.2	-60.02
LHISTD02 ⁴⁴²	210.8	-31.58
	132.6	67.9
LHISTD13 ⁴⁴²	210.6	-30.49
	132.4	51.11
TEJWAG ⁴⁴²	143.9	36.91
GLYCIN03 ⁴⁴³	-6.5	198.12
FUSVAQ01 ⁴⁴⁴	183.2	3.43
	174.2	12.59
	192.2	-10.56
	120.2	70.85
	50.2	142.74
THYMIN01 ⁴⁴⁴	119.5	65.88
	90.2	96.07
URACIL ⁴⁴⁴	96.2	78.72
	120.2	66.48
BAPLOT01 ⁴⁴⁵	114.7	57.65
	72.7	110.92
	122.7	67.18
	178.7	-20.29
LSEIN01 ¹²⁷	-4.1	197.50
GLUTAM01 ¹²⁷	-1.3	193.91
ASPARM03 ¹²⁷	0.7	195.8
	74.9	112.38
LCYSTN21 ¹²⁷	-0.4	196.16
ALUCAL04 ¹²⁷	3.0	193.4
LGLUAC11 ¹²⁷	-0.4	185.11

Table 3.8. List of experimental and predicted ^{17}O chemical shifts of reference compounds used for shielding to shift conversion. All shifts are reported relative to liquid H_2O .

CSD REFCODE	Experimental shift (ppm)	ShiftML predicted shielding (ppm)
LALNIN12 ⁴⁴³	285.0	-75.08
	268.0	-40.91
ACANIL03 ⁴⁴⁶	330.0	-147.73
BZAMID07 ⁴⁴⁷	300.0	-101.97
MBNZAM10 ⁴⁴⁶	287.0	-78.91

Table 3.9. Experimental chemical shift assignment of theophylline. The carbon labels follow **Figure 3.2**.

Carbon label	^{13}C shift (ppm)	^1H shift (ppm)	Multiplicity	Shift label
1	150.8	-	0	b
2	29.9	3.4	3	f
3	146.1	-	0	c
4	140.8	7.7	1	d
5	105.8	-	0	e
6	155.0	-	0	a
7	29.9	3.4	3	f

Table 3.10. Experimental chemical shift assignment of thymol. The carbon labels follow **Figure 3.4**. Superscript “a” indicates topologically equivalent carbon nuclei, for which the assignment cannot be resolved by the probabilistic assignment model.

Carbon label	^{13}C shift (ppm)	^1H shift (ppm)	Multiplicity	Shift label
1	18.7	0.42	3	j
2	138.4	-	0	b
3	116.9	5.40	1	f
4	150.2	-	0	a
5	131.7	-	0	c
6	126.3	7.08	1	d
7	123.6	6.19	1	e
8	25.5	3.38	1	h
9 ^a	26.1	1.05	3	g
10 ^a	23.6	1.45	3	i

Table 3.11. Experimental chemical shift assignment of strychnine. The carbon labels follow **Figure 3.5**.

Carbon label	^{13}C shift (ppm)	^1H shift (ppm)	Multiplicity	Shift label
1	171.70	-	0	a
2	142.95	-	0	b
3	116.14	8.09	1	i
4	129.60	7.25	1	e
5	125.97	7.16	1	g
6	122.91	7.10	1	h
7	134.90	-	0	d
8	52.63	-	0	o
9	60.67	3.85	1	l
10	47.39	1.27	1	q
11	78.53	4.28	1	j
12	63.80	4.14, 4.07	2	k
13	127.38	5.90	1	f
14	142.4	-	0	c
15	31.67	3.15	1	t
16	25.89	2.35, 1.45	2	u
17	60.20	3.93	1	m
18	50.72	3.11, 2.67	2	p
19	44.26	3.70, 2.72	2	r
20	53.61	3.19, 2.87	2	n
21	40.83	1.88	2	s

Table 3.12. Experimental chemical shift assignment of cocaine. The carbon labels follow **Figure 3.6**. Superscript “a” and “b” indicate pairs of topologically equivalent carbon nuclei, for which the assignment cannot be resolved by the probabilistic assignment model.

Carbon label	^{13}C shift (ppm)	^1H shift (ppm)	Multiplicity	Shift label
1	41.52	1.04	3	k
2	62.63	3.49	1	i
3	25.62	3.38, 2.91	2	m (m_1 , m_2)
4	25.62	2.56, 2.12	2	n (n_1 , n_2)
5	65.95	3.76	1	h
6	50.16	3.78	1	j
7	172.18	-	0	a
8	50.16	3.78	3	j
9	66.70	5.63	1	g
10	165.94	-	0	b
11	129.37	-	0	f
12 ^a	131.50	8.01	1	e
13 ^b	133.50	8.01	1	d
14	134.53	8.01	1	c
15 ^b	133.50	8.01	1	d
16 ^a	131.50	8.01	1	e
17	36.66	3.32, 3.06	2	l

Table 3.13. Experimental chemical shift assignment of lisinopril dihydrate. The carbon labels follow **Figure 3.6**. Superscript “a” and “b” indicate pairs of topologically equivalent carbon nuclei, for which the assignment cannot be resolved by the probabilistic assignment model.

Carbon label	¹³ C shift (ppm)	¹ H shift (ppm)	Multiplicity	Shift label
1	127.4	7.8	1	h
2 ^a	128.6	6.3	1	f
3 ^b	130.1	7.6	1	e
4	142.3	-	0	d
5 ^b	128.2	7.9	1	g
6 ^a	130.1	7.6	1	e
7	30.9	3.8	2	o
8	35.2	2.0	2	n
9	56.4	4.6	1	j
10	173.9	-	0	b
11	54.6	4.5	1	k
12	28.3	1.8	2	q
13	18.9	0.9	2	t
14	27.2	1.6, 0.2	2	r
15	35.9	2.6, 0.2	2	m
16	164.4	-	0	c
17	61.2	4.4	1	i
18	30.9	1.5	2	p
19	25.3	1.6	2	s
20	47.5	5.2	2	l
21	175.7	-	0	a

Table 3.14. Experimental chemical shift assignment of AZD5718. The carbon labels follow **Figure 3.6**. Superscript “a” and “b” indicate pairs of topologically equivalent carbon nuclei, for which the assignment cannot be resolved by the probabilistic assignment model.

Carbon label	¹³ C shift (ppm)	¹ H shift (ppm)	Multiplicity	Shift label
1	11.1	1.2	3	x
2	141.5	-	0	e
3	102.3	5.8	1	o
4	149.8	-	0	d
5	139.5	-	0	f
6 ^a	123.9	6.9	1	m
7 ^b	130.8	7.0	1	i
8	133.3	-	0	g
9 ^b	130.1	6.7	1	j
10 ^a	125.3	7.3	1	l
11	201.1	-	0	a
12	46.3	3.9	1	q
13	31.2	1.7, 0.0	2	t
14	26.6	0.8, -0.5	2	v
15	26.0	0.8, -0.5	2	w
16	29.2	1.6	2	u
17	49.8	1.6	1	p
18	174.0	-	0	b
19	125.8	-	0	k

20	130.8	7.6	1	h
21	43.5	2.7, 1.7	2	r
22	40.1	2.7, 1.9	2	s
23	161.8	-	0	c
24	119.7	-	0	n

Table 3.15. Experimental chemical shift assignment of ritonavir. The carbon labels follow **Figure 3.6**. Superscript “a” and “b” indicate pairs of topologically equivalent carbon nuclei, for which the assignment cannot be resolved by the probabilistic assignment model.

Carbon label	¹³ C shift (ppm)	¹ H shift (ppm)	Multiplicity	Shift label
1	151.7	8.36	1	f
2	132.9	-	0	j
3	142.1	7.43	1	g
4	62.8	4.19	2	n
5	157.5	-	0	d
6	58.8	4.51	1	o
7	34.7	3.10, 1.24	2	t
8	141.3	-	0	h
9-13	126.5, 126.8, 128.4, 129.0, 131.6	7.02, 5.84, 7.54, 6.81, 7.20	1	-
14	72.7	4.03	1	l
15	37.9	2.67, 1.30	2	s
16	49.0	4.47	1	p
17	41.3	2.68	2	r
18	139.1	-	0	i
19-23	126.5, 126.8, 128.4, 129.0, 131.6	7.02, 5.84, 7.54, 6.81, 7.20	1	-
24	173.4	-	0	b
25	63.7	3.72	1	m
26	29.6	1.87	1	w
27 ^a	16.4	-0.63	3	#
28 ^a	22.2	0.73	3	y
29	159.5	-	0	c
30	31.1	2.46	3	v
31	47.6	5.67, 3.47	2	q
32	154.2	-	0	e
33	114.7	5.11	1	k
34	178.3	-	0	a
35	33.9	3.33	1	u
36 ^b	26.7	1.24	3	x
37 ^b	21.0	0.78	3	z

Table 3.16. Experimental chemical shift assignment of the K salt of penicillin G. The carbon labels follow **Figure 3.7**. Superscript “a”, “b” and “c” indicate pairs of topologically equivalent carbon nuclei, for which the assignment cannot be resolved by the probabilistic assignment model.

Carbon label	¹³ C shift (ppm)	¹ H shift (ppm)	Multiplicity	Shift label
1	172.9	-	0	b
2	74.9	4.1	1	g
3	65.2	-	0	i
4	37.6	0.9	3	l
5	27.1	1.7	3	m
6	68.2	6.4	1	h
7	60.5	5.7	1	j
8	176.1	-	0	a
9	172.3	-	0	c
10	43.3	4.7, 3.9	2	k
11	136.4	-	0	d
12	128.7	7.1	1	f
13	128.7	7.1	1	f
14	130.6	7.1	1	e
15	128.7	7.1	1	f
16	130.6	7.1	1	e

3.3 Pure isotropic proton NMR spectra in solids using deep learning

This section has been adapted with permission from: Cordova, M.; Moutzouri, P.; Simões de Almeida, B.; Torodii, D.; Emsley, L., Pure Isotropic Proton NMR Spectra in Solids using Deep Learning. *Angewandte Chemie-International Edition* **2023**, 62 (8), e202216607. (post-print)

My contribution was to develop and apply the method and to analyse results. I also wrote the manuscript, with contributions of all other authors.

3.3.1 Introduction

In cases where the resolution in the proton spectrum is sufficient, the advantage provided by ^1H NMR in solids compared to other nuclei is clearly established.^{39, 151, 308, 382-385} The advent of faster magic angle spinning (MAS), which usually leads to better resolved ^1H spectra, has been a key factor in enabling ^1H detection in a broader range of systems. Nevertheless, poor ^1H resolution is still the main bottleneck for widespread application of ^1H based schemes in rigid organic materials at natural isotopic abundance.

In the CRAMPS approach,³⁸⁶ pulse sequences designed to remove homonuclear dipolar couplings can be combined with low or fast MAS to produce extra narrowing.³⁸⁷⁻³⁹⁴ At 100 kHz the linewidths obtained with MAS alone are about the same as the best results from state-of-the-art CRAMPS at slower MAS rates, and so far CRAMPS schemes have yielded no significant improvement for MAS rates above 65 kHz.

The anti-z-COSY experiment³⁹⁵⁻³⁹⁸ is an alternative approach to homonuclear decoupling which exploits a simple 2D scheme that yields correlations between remote transitions of the coupling partners, and which removes the non-refocusable part of the residual dipolar broadening.^{395, 396} This approach does not rely on complex multiple-pulse averaging sequences and typically provides a factor two reduction in linewidth compared to MAS alone, but contributions due to refocusable interactions will remain.

The dependence of residual splittings and shifts on the MAS rate ω_{MAS} has previously been described as polynomial with respect to the inverse of the MAS rate, typically dominated by first- and, to a lesser extent, second-order terms.^{85, 373, 378-381, 399} In this light, Moutzouri *et al.* recently introduced a two-dimensional approach to obtain the pure isotropic (infinite MAS rate) spectrum of a solid from a set of spectra measured at different MAS rates.⁴⁰⁰

To process these two-dimensional datasets, they used a method of fitting an amplitude vector (i.e., the isotropic spectrum), together with parametric broadenings and shifts, to reproduce the set of ^1H spectra acquired at varying MAS rates by convoluting the MAS-independent amplitude vector with the MAS-dependent shift and broadening function. While this provides a powerful method to obtain isotropic spectra, several assumptions and restrictions inherent to this fitting approach may limit its performance.

Deep learning (DL) has tremendously improved many areas of science and technology over the last decade, thanks to the ability of deep neural networks (NNs) to learn complex functions in an automated manner.^{275, 295} In particular, convolutional neural networks (CNNs) are popular models to extract information from images or spectral data^{272, 448, 449} and have been used in the context of NMR to denoise or deconvolute spectra, to reconstruct under-sampled spectra, to virtually decouple spectra, and to perform automated peak picking.^{281-286, 450}

Recurrent neural networks (RNNs) are a class of neural network developed to process time series data. The “long short-term memory” (LSTM) architecture has been shown to outperform other types of RNNs in many applications, including language modeling.^{290, 293, 294, 451} In NMR, models based on the LSTM architecture have been used to reconstruct under-sampled free induction decays (FIDs).²⁹¹

In this section, by encoding two-dimensional dataset of MAS spectra recorded at different spinning rates as a series, we infer the isotropic ^1H NMR spectrum (i.e., the spectrum that would be obtained at infinite rate) using a modified convolutional LSTM neural network trained on millions of synthetic datasets. The model, dubbed PIPNet, yields isotropic spectra that display linewidths in the 50-400 Hz range, in line with expectations, from experimental sets of MAS spectra for eight molecular solids, β -aspartylalanine (β -AspAla), flutamide, thymol, L-tyrosine hydrochloride, ampicillin, L-histidine hydrochloride monohydrate, \pm -N, α -Dimethyl-3,4-methylenedioxyphephenethylamine hydrochloride (MDMA) hydrochloride and molnupiravir. The model bypasses assumptions about the MAS-dependent broadening and shift parameters of neighbouring peaks, suppresses artifacts arising from inconsistencies between spectra acquired at different MAS rates, and inferences of full spectra can be performed in seconds.

3.3.2 Methods

Previous approach. Moutzouri *et al.*⁴⁰⁰ proposed a method to obtain the isotropic spectrum from a two-dimensional dataset of MAS spectra measured at different rates. To transform these data they assumed, based on predictions from average Hamiltonian theory,^{83-85, 373, 399} that the lineshape of a peak in a ^1H MAS spectrum can be described as a convolution of the intrinsic (isotropic) lineshape of the peak, subject to a MAS-dependent frequency shift, with a MAS-dependent broadening function. Assuming an inverse linear MAS dependence of the shift and broadening, they optimised a vector of amplitudes (the isotropic spectrum), as well as the MAS-dependent shift and widths of Lorentzian and Gaussian broadening functions, such that the difference between the resulting simulated MAS spectra and experiment was minimised.

This approach to transforming the data involves assumptions and restrictions that limit its performance. Notably, there is an assumption that the MAS dependent part of the lineshape is the same across the whole spectrum. Because of this, the spectrum is broken down into a series of resolved regions, in order to minimise its impact. Further, the number of variables to fit (the amplitude vector and the MAS-dependent broadening and shift), usually ranges between 50 and 300 for each separate region, resulting in intensive computations to obtain the isotropic spectrum, typically taking several CPU hours. Separating the spectra into regions can introduce artifacts in the isotropic spectra arising from inconsistent integrals between the spectra recorded at different MAS speeds due to truncation of the spectra within the selected regions at low MAS rates. Finally, with the large number of points used, convergence can be an issue, which thus requires to perform several fits with different starting guesses, which increases the robustness of the method but at the cost of performing the fitting several times.

Data generation. Due to the substantial amount of data required to train deep neural networks and the impossibility to record isotropic spectra of solids experimentally to use as targets for the predictions, synthetic isotropic and variable MAS rate spectra were generated according to a theoretical description of the dependence of the spectra on the MAS rate. Here, to maintain the highest level of generality, a MAS spectrum is composed of a sum of peaks (intensity I against frequency ν), where each peak $I_{\omega_r}(\nu)$ is described as a convolution between the corresponding (Gaussian) peak in the isotropic spectrum $I_{\infty}(\nu)$ and a Gaussian-Lorentzian sum (GLS) function⁴⁵²

$$GLS(\nu; w, p, m) = (1 - m) \exp \left[-\frac{4 \ln(2) (\nu - p)^2}{w^2} \right] + \frac{m}{1 + 4 \frac{(\nu - p)^2}{w^2}}, \quad (3.7)$$

where w is the width of the GLS function, p is the peak position of the GLS, here always set to the middle of the spectrum, such that convoluting the GLS with the isotropic spectrum does not affect the position of the peak, and m is the mixing factor describing the lineshape of the function. A mixing of 0 corresponds to a pure Gaussian function, while a mixing of 1 corresponds to a pure Lorentzian function. The ranges of possible mixing and width of the GLS were set based on observed MAS dependence of the lineshape. This function is particularly well suited to describe ^1H MAS broadening, where the expectation is that the lineshape is a mixture of Gaussian and Lorentzian components, and where the mixture for each spin is a function of both the MAS rate and the local dipolar coupling network.^{453, 454}

In addition, a MAS-dependent shift of the frequency of the peak s_{ω_r} was added to capture the residual shift observed in MAS spectra.^{85, 373, 380, 381, 395} The generation of a peak in an MAS spectrum is thus described as:

$$I_{\omega_r}(\nu) = \text{FT}[I_{\infty}(t) \cdot e^{i2\pi s_{\omega_r} t}] * GLS(\nu; w_{\omega_r}, p, m_{\omega_r}), \quad (3.8)$$

where $\text{FT}[\cdot]$ is the Fourier transform and $*$ denotes the convolution operation. The multiplication of the isotropic FID $I_{\infty}(t)$ with a complex exponential shifts the frequency of the corresponding peak after the Fourier transform, which is then convoluted with the MAS-dependent GLS function. Spectra made up of 512 points with a time-domain sampling frequency of 12.8 kHz were generated, corresponding to a frequency domain resolution of 25 Hz.

Isotropic spectra were generated as the sum of between 1 and 15 peaks made up of one Gaussian function each, with a linewidth sampled from a uniform distribution between 50 and 200 Hz (70% probability), between 100 and 500 Hz (20% probability), or between 100 and 1000 Hz (10% probability). This was done to ensure the representation of both sharp and broad isotropic peaks in the training spectra. The sharpest linewidth (50 Hz) was selected to be twice the frequency domain resolution (25 Hz), as we observed that sharper linewidths led to artifacts seen as negative points in the isotropic spectra. To ensure all isotropic spectra are positive, we set any points of negative intensity to zero. In order to allow different intrinsic intensities to be represented, we re-scaled each isotropic peak by a random factor sampled from a uniform distribution in the range [0.5, 1]. The intensity of the obtained isotropic spectra was then divided by 256 in order to obtain peak intensities roughly between 0.1 and 1. Generating spectra in this way by summing a series of Gaussian peaks with different isotropic widths and frequencies, we can include both well resolved spectra and more complex isotropic lineshapes that result from superpositions.

Twenty-four MAS rates were then selected randomly between 30 and 100 kHz, and the corresponding MAS spectra were constructed by shifting and convoluting each peak with a GLS function with parameters s_{ω_r} , w_{ω_r} and m_{ω_r} following a second-order inverse MAS dependence subject to noise,

$$w_{\omega_r} = \frac{w_1}{\omega_r} + \frac{w_2}{\omega_r^2} + \Delta w, \quad (3.9)$$

$$s_{\omega_r} = \frac{s_1}{\omega_r} + \frac{s_2}{\omega_r^2} + \Delta s, \quad (3.10)$$

$$m_{\omega_r} = \frac{m_1}{\omega_r} + \frac{m_2}{\omega_r^2}. \quad (3.11)$$

For each peak, the value of w_1 in **Equation 3.9** was drawn from a uniform distribution in the range $[10^7, 5 \cdot 10^7]$ Hz² with 80% probability, and in the range $[5 \cdot 10^7, 10^8]$ Hz² otherwise (corresponding to a contribution of 100 to 1,000 Hz to the width of the GLS at an MAS rate of 100 kHz), w_2 was set to be either 0 with a 50% probability, or a random value between 10^{11} and $5 \cdot 10^{11}$ Hz³ (contributing to the width of the GLS by 10 to 50 Hz at an MAS rate of 100 kHz). In addition, for each MAS rate the GLS width w_{ω_r} was randomly perturbed by a value drawn from a normal distribution $\Delta w \sim \mathcal{N}(0, \sigma_w)$ Hz where σ_w was set to be 5% of the range of widths generated between the lowest and highest MAS rates with the selected values of w_1 and w_2 . The value of s_1 in **Equation 3.10** was randomly sampled between -10^7 and 10^7 Hz² (introducing a shift contribution between -100 and 100 Hz at 100 kHz MAS), and s_2 was set to zero with an 80% probability, or to a random value between $-2 \cdot 10^{10}$ and $2 \cdot 10^{10}$ Hz³ (corresponding to a shift between -2 and 2 Hz at 100 kHz MAS). Δs was randomly drawn from a normal distribution $\Delta s \sim \mathcal{N}(0, 25)$ Hz for each peak and each MAS rate. The GLS width and frequency shift applied to a peak in an MAS rate is described in **Equations 3.9-3.10**. The mixing m_{ω_r} was set to follow the inverse MAS dependence described in **Equation 3.11** with a probability of 50% and with the value of m_1 set to zero with a 10% probability or randomly sampled from a uniform distribution in the range $[0, 10^4]$ Hz with 10% probability, or in the range $[10^4, 5 \cdot 10^4]$ Hz, and with m_2 set to zero with an 80% probability or randomly sampled between 10^8 and $5 \cdot 10^8$ Hz². Resulting values of m_{ω_r} above one were capped to one. Otherwise, the mixing m_{ω_r} was set to be either constant, monotonously increasing with random values between 0 and 1, or monotonously decreasing with random values between 0 and 1 with increasing MAS rate. These three dependences were considered with equal probabilities.

For each peak and each MAS rate, a random phase is drawn from a normal distribution $\mathcal{N}(0, 0.05)$ and used to distort that peak in the corresponding MAS spectrum with a probability of 50%. Additionally, each isotropic peak was assigned a 10% chance to have a decreasing intensity in MAS spectra with increasing rate by multiplying the intensity of the corresponding peak in each MAS spectrum by a value linearly decreasing between 1 and a final value sampled uniformly in the range $[0.3, 0.7]$.

After generating the MAS spectra as described in **Equation 3.8**, the intensity of the spectra was divided by 64 in order to obtain peak intensities roughly between 0.1 and 1. The ten leftmost and rightmost points of the obtained spectra were linearly smoothed to a value of zero.

Deep convolutional LSTM model. The model used to encode isotropic spectra from the set of variable rate MAS spectra is a recurrent neural network with six convolutional LSTM layers adapted from a model used for precipitation nowcasting.⁴⁵⁵ In particular, the output gate was found to be unnecessary and was removed from the LSTM cells (see **Figure 3.36**). The vectors C_0 and H_0 describing the initial state of the cell are initialised as zero-filled vectors of the length of the spectra and with 64 channels. In each layer and for each step in the recurrent process, the previous hidden state H_{t-1} of the LSTM cell is combined with the input X_t , which is the next MAS spectrum in the series, and fed into two one-dimensional CNN layers with a sigmoidal (σ) activation function to yield the forget and input gates f_t and i_t , respectively, as well as into a CNN layer with a hyperbolic tangent (\tanh) activation function to yield the vector of new candidate values G_t to add to the state. The previous state vector C_{t-1} is first weighted element-wise by f_t before the vector of candidate values G_t , weighted element-wise by i_t , is added to form the updated cell state C_t . Taking the hyperbolic tangent of C_t then yields H_t , which is used as the input for the next LSTM layer. The process is summarised in **Equation 3.12** where $*$ and \circ indicate convolutions and element-wise multiplication, respectively.

$$\begin{aligned}
 i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + b_i), \\
 f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + b_f), \\
 G_t &= \tanh(W_{xg} * X_t + W_{hg} * H_{t-1} + b_g), \\
 C_t &= f_t \circ C_{t-1} + i_t \circ G_t, \\
 H_t &= \tanh(C_t).
 \end{aligned} \tag{3.12}$$

Each CNN layer corresponds to convoluting a matrix of weights W with the input and the previous hidden state, adding a bias b , and applying the activation function. We used a kernel size of 5 for all convolutional layers. After the final layer, the resulting hidden cell state H_t after each step is fed into a CNN layer with a kernel size of 5 and a sigmoidal activation function to yield the prediction of the isotropic spectrum. Detailed model and training parameters are given in **Table 3.25**.

MAS spectra are encoded as vectors with two channels, the first being the real part of the spectrum and the second being a constant pseudo-spectrum containing the MAS rate divided by 100 kHz in each element. At each step, a new spectrum X_t is fed into the network, in order of increasing MAS rate.

Model training. We trained a committee of 16 models with the same architecture on different generated data in order to evaluate the confidence of predictions. Each model was trained for a total of 250,000 batches of 16 different samples, each comprising one isotropic spectrum and 24 corresponding MAS spectra at different simulated MAS rates. After every 1,000 batches (16,000 training samples), the model was evaluated on 200 batches of 16 isotropic spectra. The model was trained by minimising the mean absolute error (MAE) between the output of the model $\hat{I}_\infty(v)$ and the ground-truth isotropic spectrum $I_\infty(v)$.

Due to the sparsity of isotropic spectra, we initially convoluted the whole of each target isotropic spectrum with a Gaussian function (G_L) in order to increase the proportion of spectra containing signal and prevent the network from initially predicting spectra containing no signal. In addition, we introduced a weight to the loss function for each frequency v_i in the isotropic spectrum to be the maximum between 1 and k times the value at frequency v_i in the isotropic spectrum (after convolution with the Gaussian function) in order to bias the training towards correctly predicting regions that contain signal. The resulting loss function is:

$$\mathcal{L} = \frac{1}{N} \sum_i |I_\infty(v_i) - (I_\infty * G_L)(v_i)| \cdot \max(1, k \cdot (I_\infty * G_L)(v_i)) \tag{3.13}$$

We set the width of G_L to 75 Hz and k to 100 during the first 320,000 training samples, then reduced the width of G_L to 25 Hz and k to 10 for 480,000 additional training samples, before removing the convolution with G_L completely and setting k to zero for the rest of the training.

Random noise was introduced in the generated MAS spectra such as to match the typical signal to noise ratio observed in experimental ^1H MAS spectra (between 100 and 1,000 for the most intense peak at 100 kHz). The predictions were compared to the generated isotropic spectra after each step. The final predictions were obtained as the mean over the 16 models, and the uncertainties were estimated as the standard deviation of the prediction of each model.

Model evaluation. The model was evaluated by computing the MAE between synthetic ground-truth and predicted isotropic spectra for batches of 1,024 samples generated with different parameters. We investigated the effect of the number of peaks in the isotropic spectra, different MAS dependences (first-order only, first- and second-order, second-order only, or independent) of the linewidth and MAS-dependent shift, the range of MAS rates generated, the number of MAS spectra used, as well as the amount of noise introduced in the spectra themselves and in the linewidth and shift dependences.

The propensity of PIPNet to produce false positive signals was evaluated for the spectra generated with different noise levels by computing the false positive rate, defined here as the percentage of points containing signal in the inferred isotropic spectra but not containing signal in the corresponding ground-truth spectra, among all the points in the predicted isotropic spectra identified as containing signal (**Figure 3.37**). A point is considered to contain signal if its value is at least 1% of the maximum intensity in the whole spectrum. False positives are thus the points containing signal in the predicted isotropic spectrum but not in the ground-truth. In practice, to prevent considering lineshapes that are predicted to be broader or at a slightly different frequency with respect to the ground-truth as false positives, we only consider points containing signal in the predicted isotropic spectra as false positives if they are further than 250 Hz away from any point containing signal in the ground-truth.

Application to experimental spectra. The digital resolution of all experimental spectra (acquired as described below) was set to about 25 Hz by zero-padding the end of the FID to match the spectral resolution of the training spectra. After phasing and correcting the baseline of the recorded spectra, the spectral range between 20 and -5 ppm was extracted. All spectra were normalised by integral and scaled to a maximum amplitude of the most intense spectrum set to 0.5 in order to reproduce intensities present in the generated training spectra. Spectra with MAS rates of 30 kHz and above were used. Predictions were subsequently performed in the same manner as for synthetic spectra.

NMR experiments. The method was applied to eight different microcrystalline organic solids: β -AspAla, flutamide, thymol, L-tyrosine hydrochloride, ampicillin, L-histidine hydrochloride monohydrate, MDMA hydrochloride and molnupiravir. The assignment and MAS datasets of β -AspAla, flutamide, thymol, L-tyrosine hydrochloride and ampicillin have already been reported previously.^{49, 55, 400, 406, 410, 456} The samples of MDMA hydrochloride and molnupiravir were purchased from Lipomed AG and MedChemExpress, respectively.

Assignment. Here we report the ^1H and ^{13}C assignment of MDMA hydrochloride and molnupiravir based on the acquisition of 100 kHz 0.7 mm 1D ^{13}C CPMAS and ^1H MAS spectra, 2D ^1H - ^{13}C hCH³⁸³ and ^1H - ^{13}C INADEQUATE spectra,^{196, 369} DFT chemical shift calculations for MDMA and the probabilistic assignment approach of Cordova *et al.*³⁵⁸ (see **Figure 3.34**). The assignments of β -AspAla, flutamide, thymol, L-tyrosine hydrochloride, ampicillin and L-histidine hydrochloride monohydrate have been previously reported.^{49, 55, 406, 410, 456} A DNP-enhanced INADEQUATE spectrum^{196, 369} was recoded for MDMA (**Figure 3.34**) at 9.4 T in a 3.2 mm DNP probe at 100 K. The spectra of MDMA hydrochloride were referenced by setting the ^1H chemical shift of proton 17 of ampicillin to 0.6 ppm and the ^{13}C chemical shifts of MDMA hydrochloride according to Ref. 457. The spectra of molnupiravir were referenced by setting the ^1H chemical shift of proton 1 of L-tyrosine hydrochloride to 12.4 ppm and the ^{13}C chemical shift of carbon 6 of ampicillin to 175 ppm.

For MDMA, quaternary carbons were first identified by their absence in the short range hCH spectrum, and then further assigned by comparison to predictions of DFT shifts (**Table 3.23**). This led to no significant ambiguity, except for carbons labelled 8 and 9 in MDMA hydrochloride. The INADEQUATE spectrum shown in **Figure 3.34E** was used to fully assign the aromatic carbons of MDMA hydrochloride without ambiguity. The aliphatic carbon-proton pairs in MDMA hydrochloride were then assigned by comparison of the observed joint ^1H and ^{13}C shifts to DFT shifts (**Table 3.23**). The remaining protons were assigned using the ^1H - ^{13}C hCH spectrum (**Figure 3.34C**).

For molnupiravir, quaternary carbons were first identified by their absence in the short range hCH spectrum, and then further assigned by comparison to probabilistic distributions (**Figure 3.34F-G**)³⁵⁸ without ambiguity. Carbon 13 was assigned based on its correlation with two distinct protons (**Figure 3.34D**). The remaining carbon-proton pairs were assigned based on the most probable assignment given by **Figure 3.34G**. Protons 9 and 10 were assigned based on long-range correlations to carbons 8 and 11, respectively, observed in the hCH spectrum (**Figure 3.34D**, inset). The assignment of protons 1 and 2 was not resolved.

Variable rate MAS spectra. For each sample, spectra were acquired with a Bruker 0.7 mm room temperature HCN CP-MAS probe on an 18.81 T Bruker Avance Neo spectrometer corresponding to a ^1H frequency of 800 MHz, except for L-histidine hydrochloride monohydrate and molnupiravir which were acquired on a 21.14 T Bruker Avance Neo spectrometer (900 MHz ^1H frequency). The spectra of ampicillin, flutamide, L-histidine hydrochloride monohydrate, β -AspAla, L-tyrosine hydrochloride and thymol were recorded at MAS rates between 30 and 100 kHz. The spectra of MDMA hydrochloride and molnupiravir were measured from 40 to 100 kHz MAS rates. All spectra were measured in steps of 2 kHz. The magic angle was set for each sample by maximising the T_2' of the proton signal at 100 kHz, the 90° pulse width was optimised, and the data was recorded with active temperature regulation to keep the sample temperature at about 295 K across the range of MAS rates. The thymol and molnupiravir data were acquired at a constant VT temperature of 275 K. The pulse sequence used was a rotor synchronised spin echo for background suppression. The echo delay was equal to one rotor period for all samples except for molnupiravir, for which two rotor periods were used. For L-histidine hydrochloride monohydrate, an additional dataset of ^1H MAS spectra was recorded using a Bruker 1.3 mm room temperature HDCN CP-MAS probe on a 21.14 T Bruker Avance Neo spectrometer (900 MHz ^1H frequency). All the acquisition parameters and raw NMR data are available in **Tables 3.17-3.20**. The spectra of ampicillin, flutamide, β -AspAla, L-tyrosine hydrochloride and thymol used are those already reported in Ref. 400. The two-dimensional datasets of MAS spectra recorded at different MAS rates for all compounds are shown in **Figure 3.35**.

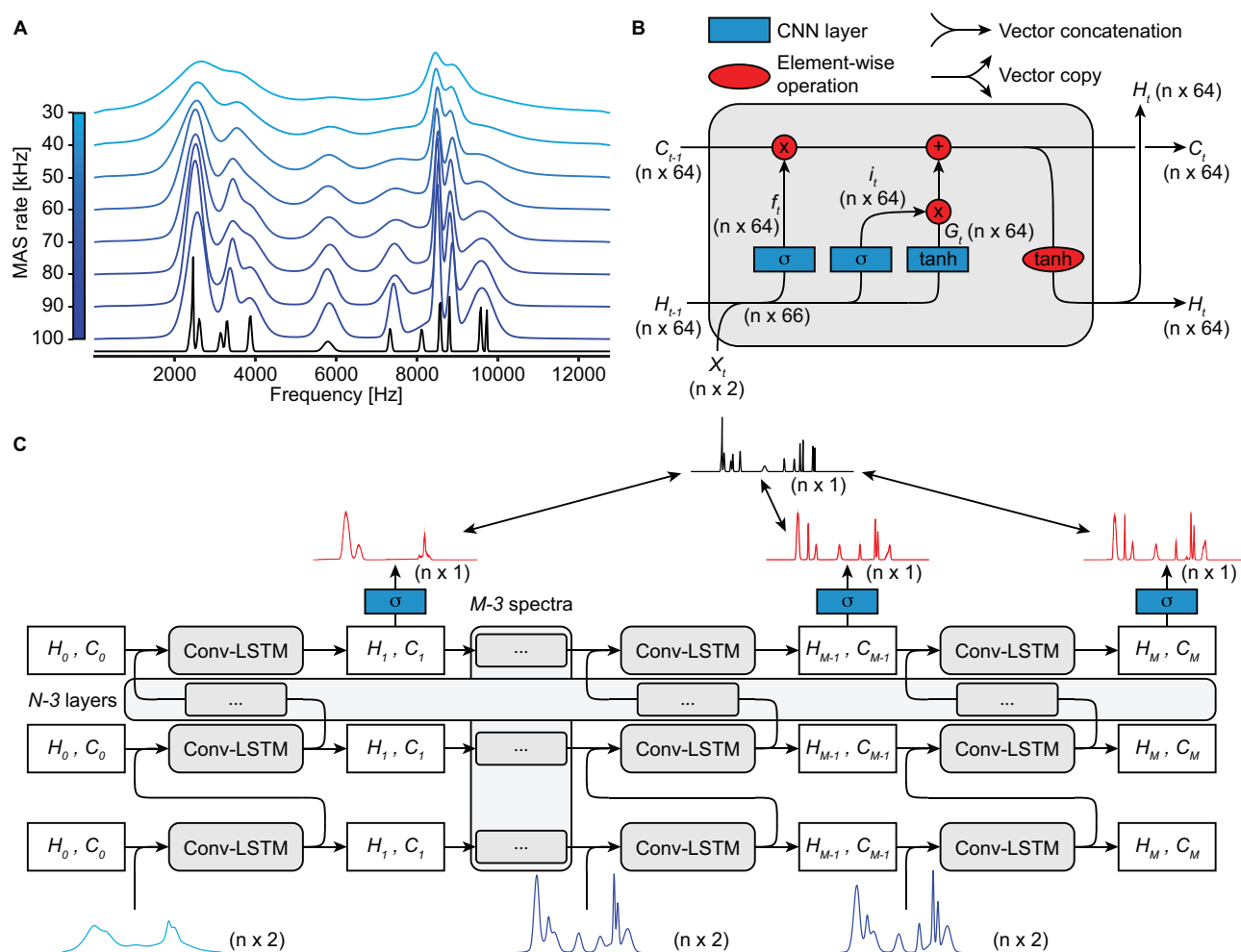


Figure 3.28. Data generation and signal processing. **(A)** Example of generated isotropic (black) and variable-rate MAS (blue) spectra. Here the top and bottom-most MAS spectra correspond to MAS rates of 30 and 100 kHz, respectively. **(B)** Convolutional LSTM cell used to process a spectrum, according to the overall scheme shown in **(C)** which describes the complete processing of the $M = 8$ MAS spectra each containing n points in **(A)** with N LSTM layers to obtain predicted isotropic spectra. The initial cell state vectors H_0 and C_0 are described in the text. MAS spectra are encoded as vectors with two channels, the first one being the real part of the spectrum and the second one containing the MAS rate at each point. After the last layer, the hidden state H_t is processed with a CNN with sigmoidal activation function (blue square) to produce the prediction at each step (red spectra). When training, the predictions at each step are compared with the target isotropic spectrum (black spectrum), as indicated by the double arrow lines.

GIPAW DFT computation of chemical shifts of MDMA hydrochloride. All DFT computations were performed using the plane-wave DFT software Quantum ESPRESSO version 6.5.^{328, 329} Atomic positions of the crystal structure of MDMA (Refcode: NEDMIS)⁴⁵⁸ were first optimised at the PBE⁹⁷ level of theory using Grimme D2 dispersion correction³³⁰ and ultrasoft pseudopotentials obtained from the PSLibrary version 1.0.0.³³² Wavefunction and charge density energy cutoffs were set to 160 and 1,280 Rydberg, respectively. A 2x2x1 Monkhorst-pack grid of k-points was used.³³⁸ Shielding computation was subsequently performed using the GIPAW method.^{117, 118} The computed shieldings were converted to chemical shift by linear regression against experimentally measured shifts.

3.3.3 Results and Discussion

Training deep neural networks requires substantial amounts of data. Given the relatively low number of available experimental datasets of ^1H MAS spectra recorded at different spinning rates, and the lack of any method to independently acquire the target isotropic spectra, synthetic data were used to train the model. **Figure 3.28A** shows an example set of eight synthetic MAS spectra and the associated isotropic spectrum. Such sets of spectra are generated in a few tens of milliseconds, allowing the training of the model on millions of sets of synthetic variable MAS spectra that include all the possible parameter variations in peaks positions, peak shapes, MAS dependences, phase and intensity errors, and noise described in **Section 3.3.2**.

The architecture of the LSTM cell used here is described in **Figure 3.28B** (and is described in detail in **Section 3.3.2**). The two main differences compared to the original description of LSTM²⁹⁰ are the use of CNN layers to process the inputs and the removal of the output gate. The former allows the processing of spectral data without the need for fixed input size and independently of the particular frequencies of peaks observed, while presence of the latter was found to be unnecessary since the isotropic spectrum is directly encoded into the memory of the LSTM network C_t and H_t , and does not require decoding that depends on the last MAS spectrum fed to the network (see **Figure 3.36**).

Figure 3.28C shows the complete processing pipeline performed by PIPNet in order to obtain the predicted isotropic spectrum from the set of MAS spectra shown in **Figure 3.28A**. At each step, another MAS spectrum from the set with a different rate is used as input to the network to update the state vectors of the LSTM cells. The spectra are fed into the process in order of increasing MAS rate, until all the spectra in the set (8 in **Figure 3.28A**, but 24 in the actual model training process) have been input. After each step, the state vector H_t of the final layer is processed by a final CNN layer to yield the predicted isotropic spectrum.

In order to obtain the uncertainty of the predicted isotropic spectra, PIPNet is a committee model made up of 16 neural networks with identical architectures, but each trained on completely independent synthetic data. At inference, the mean over the 16 predictions yields the predicted isotropic spectrum, and the standard deviation gives an indication of the uncertainty associated to the prediction at each point in the spectrum. Notably, uncertainty on the order of the predicted intensity highlights regions where the predicted spectrum is unreliable.

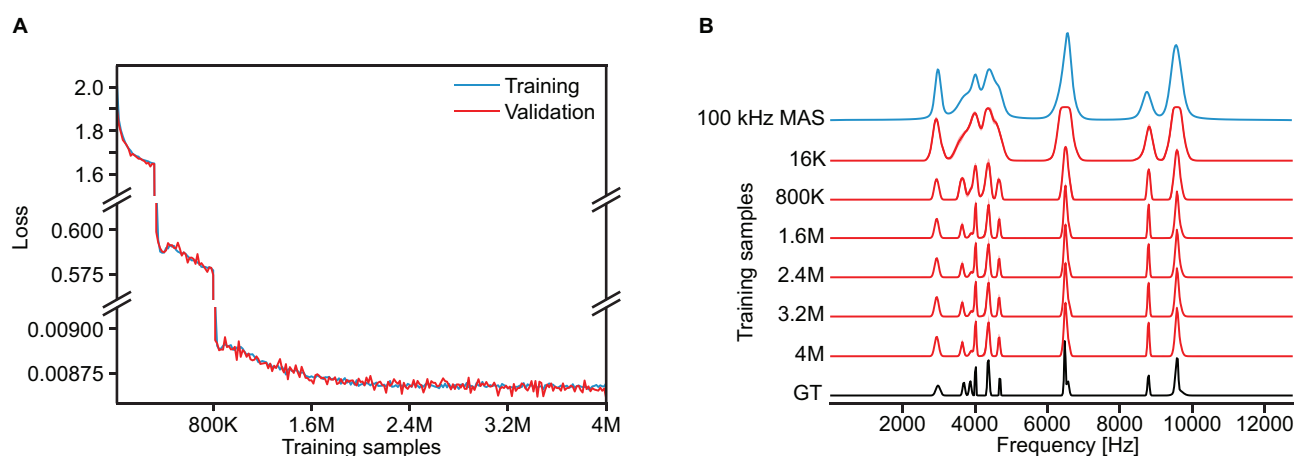


Figure 3.29. Model training. (A) Evolution of the loss function during model training. The large changes after 320,000 and 800,000 training samples per model correspond to the changes in loss function applied as described in **Section 3.3.2**. (B) Synthetic 100 kHz MAS spectrum (blue) and its associated ground-truth (GT) isotropic spectrum (black) compared to predictions of the model trained on 16,000, 800,000, 1,600,000, 2,400,000, 3,200,000 and 4,000,000 samples (red). The shaded areas in the predicted spectra indicates the standard deviation between the 16 neural networks making up the committee model.

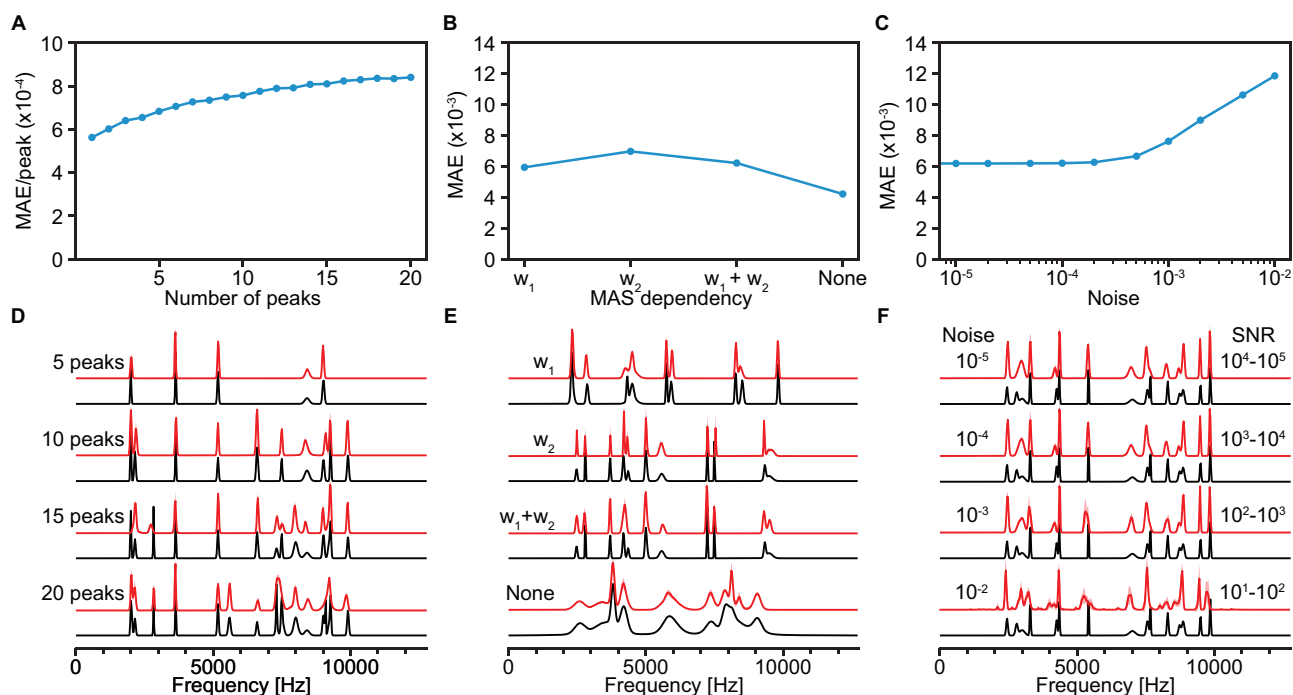


Figure 3.30. Model evaluation. (A), (B), (C) MAE between predictions and ground-truth isotropic spectra for 1,024 samples and (D), (E), (F), illustrative comparisons of predicted (red) and ground-truth (black) isotropic spectra with various (A), (D) number of peaks, (B), (E) MAS dependencies (w_1 : first-order, w_2 : second-order), and (C), (F) noise levels. The MAE in (A) is normalised by the number of peaks. The typical range of signal-to-noise ratio (SNR) in 100 kHz MAS spectra corresponding to each level of noise is indicated in (F).

Figure 3.29A shows the evolution of the loss function, corresponding to the mean absolute error (MAE) between the predicted and ground-truth isotropic spectra, during the model training. The significant changes of scale in the loss after 320,000 and 800,000 training samples per model reflects the change in the weighting of the loss function (see **Section 3.3.2**). This was performed in order to promote the detection of peaks at the beginning of the training by decreasing the importance of the prediction in empty regions of the isotropic spectra. **Figure 3.29B** shows the comparison of the predictions obtained after training each model on 16,000, 800,000, 1,600,000, 2,400,000, 3,200,000 and 4,000,000 sets of MAS spectra. Significant improvement of the predictions can be seen until 1,600,000 training samples, after which the model was considered to have converged. This is reflected both by the plateau in the loss function in **Figure 3.29A** and by the obtained predictions for the example shown in **Figure 3.29B**, where the five peaks between 3,500 and 5,000 Hz were found to be captured by the model only after 1,600,000 training samples (although with different intensities). After that point, the isotropic spectra obtained did not display any significant change with increased amounts of training data. Nonetheless, we selected the final model at the end of the full training process, i.e., after 4,000,000 training samples per model.

Figure 3.30 displays the behaviour of the model with different numbers of isotropic peaks, MAS dependences and levels of noise. The model was found to be robust to the number of peaks, with each peak resulting in a similar increase of the MAE between predicted and ground-truth isotropic spectra (see **Figure 3.30A**). As seen in **Figure 3.30D**, the number of peaks is generally correctly captured both for sparse and more crowded spectra. In instances where different peaks are not captured, they are typically found to coalesce into a single, broader peak.

Figures 3.30B and **3.30E** highlight the ability of the model to capture both first- and second-order MAS dependence of linewidths and shifts in MAS spectra, as well as combined first- and second-order dependence. A purely second-order MAS dependence was found to slightly raise the error between inferred and ground-truth isotropic spectra. In addition, using a set of identical spectra with different MAS rates (no MAS dependence) was found to result in only very marginal amounts of unexpected sharpening of peaks arising from overfitting of the model, seen in the region around 8,000 Hz in **Figure 3.30E**. This indicates that PIPNet is robust to different MAS dependences.

Figures 3.30C and 3.30F show the robustness with respect to the level of noise in the MAS spectra. The MAE between inferred and ground-truth isotropic spectra was found to increase with noise levels of 10^{-3} and above, corresponding to a signal-to-noise ratio (SNR) of 1,000 and below in 100 kHz MAS spectra. We find that the predicted spectra visually display no significant perturbations down to a SNR of 100, below which some noise and uncertainties start to appear in the inferred spectra. Importantly, the model is still able to correctly identify the regions containing signals from pure noise down to a SNR of 10. Experimental fast MAS spectra of pure organic solids typically have SNR ~ 1000 . Artifacts are thus expected to appear only with low signal-to-noise ratio spectra, and would typically be associated with a high uncertainty (see **Figure 3.30F**). We found that the false positive rate in predicted spectra was under 1% up to a noise level of 10^{-3} (SNR down to 100), and was found to be 22% with a noise level of 10^{-2} (SNR of 10) (see **Figure 3.37**).

The model was also found to be robust to perturbations in the MAS-dependent linewidth, shift, number of MAS spectra and range of MAS rates selected (see **Figures 3.37-3.38**).

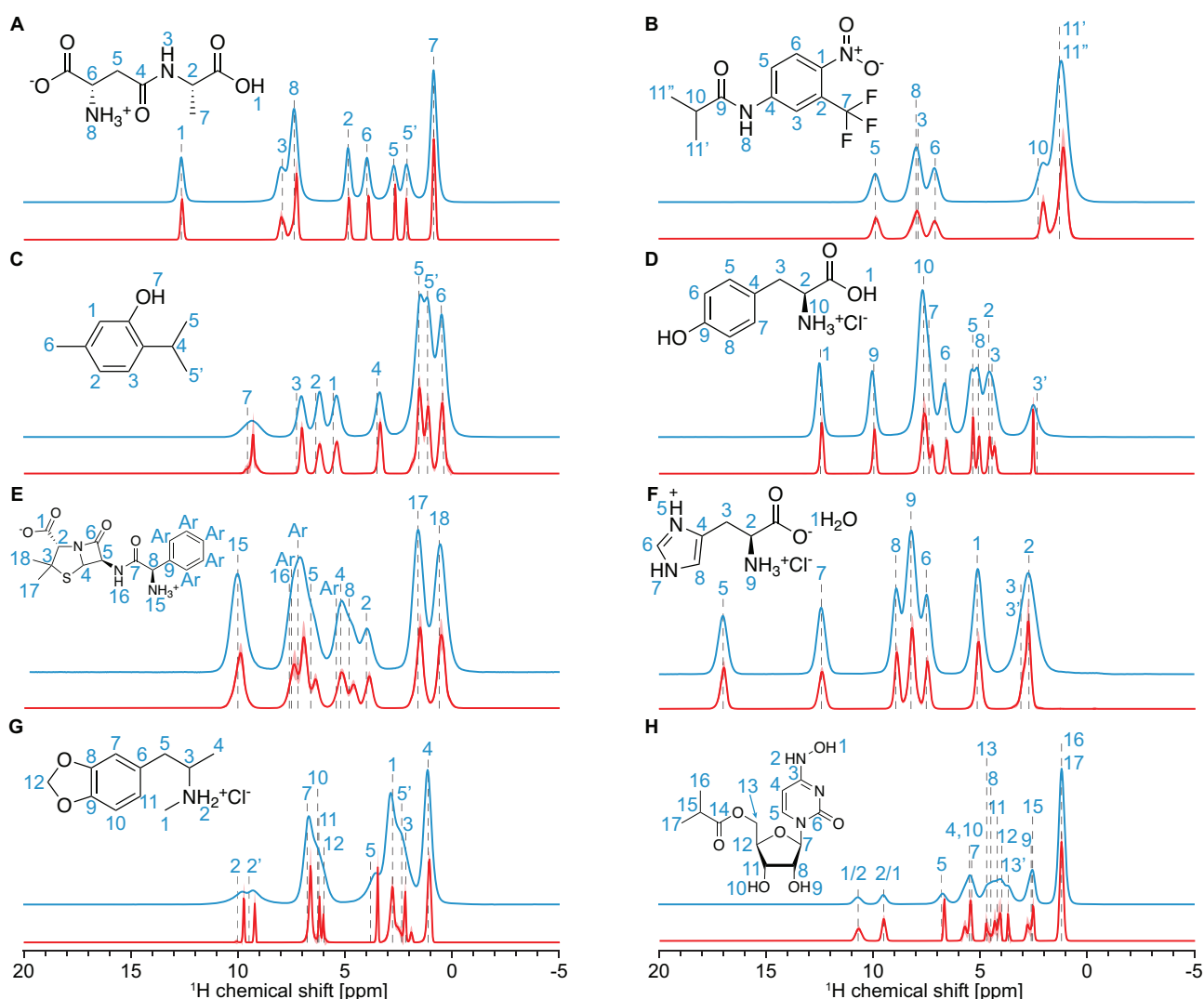


Figure 3.31. Predictions on experimental data. Experimental 100 kHz MAS spectra (blue) and isotropic spectra (red) of (A) β -AspAla, (B) flutamide, (C) thymol, (D) L-tyrosine hydrochloride, (E) ampicillin, (F) L-histidine hydrochloride monohydrate, (G) MDMA hydrochloride and (H) molnupiravir. The isotropic spectra were obtained for each compound from a set of between 31 and 41 MAS spectra recorded at rates between 20 and 100 kHz (as detailed in **Section 3.3.2**) using the model presented here. The assignment of the 100 kHz spectra, based on two-dimensional experiments (see **Section 3.3.2**), is indicated with vertical dashed black lines and blue numbers. In (H), the assignment of protons 1 and 2 is ambiguous.

Figure 3.31 shows the isotropic spectra obtained from eight experimental sets of variable rate MAS spectra recorded on different compounds (detailed in **Section 3.3.2**) in comparison to the corresponding 100 kHz MAS spectra. We note that the spectra obtained using PIPNet from these two-dimensional datasets were found to be similar to the parametrically fitted isotropic spectra, while notably displaying fewer artifacts (see **Figure 3.39**).

The samples are microcrystalline forms of β -AspAla, flutamide, thymol, L-tyrosine hydrochloride, ampicillin, L-histidine hydrochloride monohydrate, MDMA hydrochloride and molnupiravir. The assignment and MAS datasets of β -AspAla, flutamide, thymol, L-tyrosine hydrochloride and ampicillin have already been reported previously.^{49, 55, 400, 406, 410, 456} The ^1H and ^{13}C assignment of MDMA hydrochloride and molnupiravir are done here as described in **Section 3.3.2**, based on the acquisition of 100 kHz 0.7 mm 1D ^{13}C CPMAS and ^1H MAS spectra, 2D ^1H - ^{13}C hCH³⁸³ and ^1H - ^{13}C INADEQUATE spectra,^{196, 369} DFT chemical shift calculations for MDMA and the probabilistic assignment approach of Cordova *et al.*³⁵⁸ (see **Figure 3.34**). A DNP-enhanced INADEQUATE spectrum^{196, 369} was recorded for MDMA (**Figure 3.34**) at 9.4 T in a 3.2 mm DNP probe at 100 K.

If we look at the case of L-tyrosine hydrochloride (**Figure 3.31D**) as a representative example, the expected number of peaks is retrieved and the peak positions match expectations from assignments carried out using 2D methods (see **Appendix V**). The observed linewidths in the isotropic spectrum are very significantly narrower than the 100 kHz MAS spectrum, with full widths at half maximum (FWHM) between 62 and 250 Hz (0.08 and 0.32 ppm).

We note that while the obtained isotropic spectra generally display sharper lineshapes than those measured at the highest MAS rates, finite linewidths are still anticipated, as we expect distributions of the isotropic shifts due to the presence of more or less structural disorder in the sample. Isotropic shift distributions will also be caused by anisotropic bulk magnetic susceptibility (ABMS) effects,^{459, 460} or as a result of imperfect B_0 homogeneity. Importantly, no significant artifacts are seen in the isotropic spectra obtained. This behaviour is seen in general across all eight samples in **Figure 3.31**. The linewidths observed in the isotropic spectrum are in line with expectations from a previous analysis of the MAS dependence of the measured linewidths of the resolved peaks (1 and 3') given in Figure 14 of Ref. 85. It is especially notable that the isotropic spectrum predicted from the variable MAS dataset correctly identifies the 7 peaks present in the crowded spectral between 3 and 9 ppm, that are not resolved in the 100 kHz MAS spectrum.

It is important to note that PIPNet is not just identifying potential peaks and replacing them with uniformly narrow lines. We notably see in the cases of flutamide (**Figure 3.31B**) that there is only minor narrowing in the isotropic spectrum as compared to the 100 kHz MAS spectrum (see **Table 3.21** for a list of resolved linewidths measured in both the 100 kHz MAS spectra and the isotropic spectra, for all eight compounds). Similar behaviour was also observed using the parametric fitting approach (**Figure 3.39**), and we thus conclude that these peaks are dominated by chemical shift broadening and not MAS dependent dipolar broadening. We note that for ampicillin we also see only limited narrowing in the isotropic spectrum from PIPNet. In this case this is in contrast with the parametric fitting approach, which produced narrower lines (**Figure 3.39**, **Table 3.21**). Here we suspect that the parametric fitting approach might be overfitting, and this is a good example of the more general and robust nature of the PIPNet model.

In addition to the six compounds previously reported,⁴⁰⁰ here we have obtained isotropic spectra for two additional molecular solids, MDMA hydrochloride and molnupiravir (**Figures 3.31F** and **3.31G**, respectively). For MDMA hydrochloride, the expected peaks were clearly identified, except potentially for one in the aromatic (protons 7-12) regions of the spectrum. However, as seen in the HETCOR spectrum of the compound (**Figure 3.34** and dashed vertical lines in **Figure 3.31G**), protons 10 and 11 display very similar chemical shifts, suggesting that the peaks might overlap in the isotropic spectrum. The isotropic spectrum of molnupiravir displayed all expected peaks. In particular, the two peaks in the region between 2 and 3 ppm are clearly identified from the PIPNet spectrum, while the fitted isotropic spectrum predicts only one broader peak (**Figure 3.39**). This highlights one limitation of the fitted model, as the presence of two isotropic peaks can be seen from the increasing asymmetry of the corresponding peak in the set of MAS spectra with increasing rate (**Figure 3.35H**), which corresponds to the two underlying isotropic peaks having different MAS-dependent shifts and/or broadenings. Considering multiple different MAS-dependent shifts in a single spectral region is thus critical here in order to correctly describe the experimental spectra as a function of MAS rate. While this is not captured by the fitted model due to the assumption that only one MAS-dependent shift is assigned to each fitted region, PIPNet makes no such assumption and thus is able to identify the two distinct resonances.

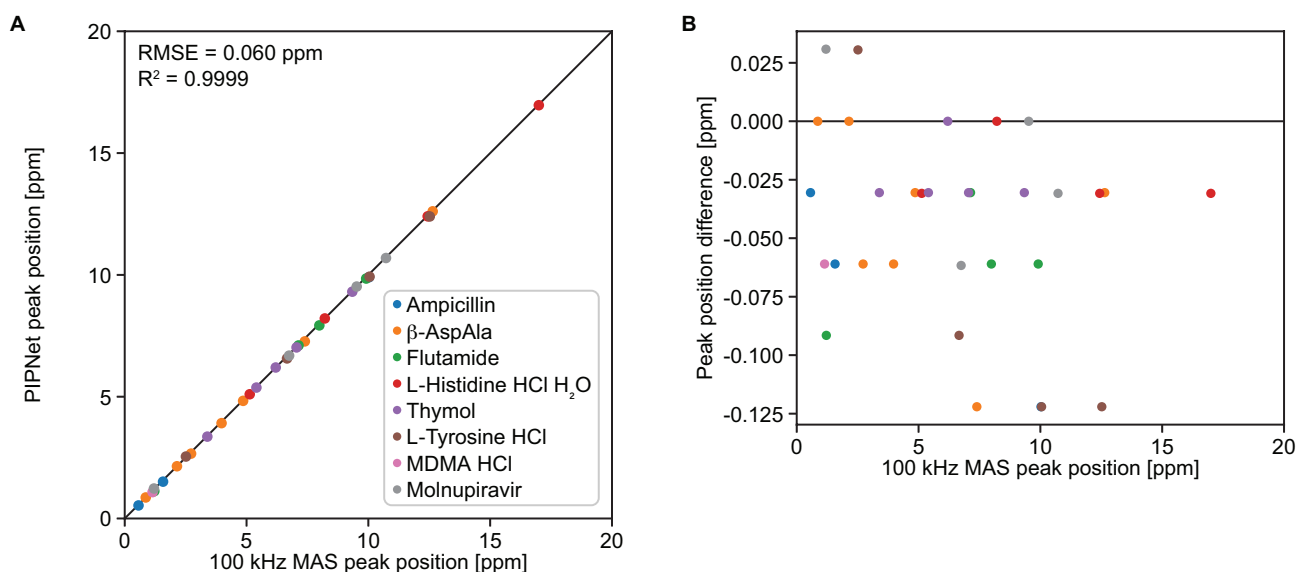


Figure 3.32. Predicted peak positions. (A) Comparison of the position of selected isolated peaks in the 100 kHz MAS and those in the isotropic spectra. (B) Difference in position of the peaks shown in (A). The black lines indicate perfect correlation.

In summary, the isotropic spectra inferred by PIPNet from the experimental variable rate MAS datasets were found to display the expected number of peaks, without any prior information given about the sample measured. No significant artifacts were identified in the inferred spectra, highlighting the increased generality and robustness of the deep learning model compared to the previous fitted approach.

Figure 3.32 shows the comparison between the chemical shifts of 32 selected peaks from the compounds studied here, and **Figure 3.40** compares the linewidths obtained, using the deep learning approach presented here with the positions observed at 100 kHz MAS, and with the fitted approach. We find strong correlations between the observed peak positions with the three methods, with a root-mean-square error of 0.060 ppm (~ 50 Hz) and a R^2 coefficient above 0.999. We note that this is not a direct metric of the accuracy of our model, as the 100 kHz MAS peak positions are not at exactly the isotropic shifts,^{85, 373, 380, 381, 395} and the fitting approach also suffers from shortcomings as mentioned above. The consistency between the three results does however strongly suggest that the method allows measurement of resolved isotropic shifts to within an error of 0.060 ppm. In **Figure 3.32B**, differences appear at discrete values in steps of ~ 0.03 ppm, corresponding to the spectral resolution. If needed, more accurate peak maxima could be obtained by increasing the spectral resolution through zero-padding before running the prediction, or by applying more advanced peak picking approaches to the isotropic spectrum.²⁸¹

Figure 3.40 and **Table 3.21** suggests that linewidths are generally predicted to be broader using PIPNet than with the fitting approach. Because the model was found able to predict linewidths as low as 57 Hz (0.07 ppm, see **Table 3.22**), we expect these differences to arise from approximations present in the fitted approach leading to a degree of overfitting, and not from incomplete removal of MAS-dependent broadening by the machine learning model.

As seen in **Figure 3.41**, the relative integrals of different spectral regions are retained in the predicted isotropic spectra compared to the experimental 100 kHz MAS spectra, with a deviation of less than $\sim 5\%$ of the total integral for all compounds.

Figure 3.44 suggests that using a lower number of experimental spectra while retaining the range of MAS rates used (40 to 100 kHz in **Figure 3.44**) leads to only marginal changes in the isotropic spectra obtained, up to increments of 10 kHz between measured MAS spectra. In light of this result and those shown in **Figure 3.33** and **Figure 3.42**, we consider that converged isotropic spectra can typically be obtained using spectra from 30 to at least 80 kHz MAS rate, in increments of up to 10 kHz. We note that, in general, a minimum of 5 spectra are typically required to obtain meaningful isotropic spectra (see **Figure 3.36**). The quality of the isotropic spectra obtained is nonetheless expected to be higher using a larger range of MAS rates and a higher number of MAS spectra.

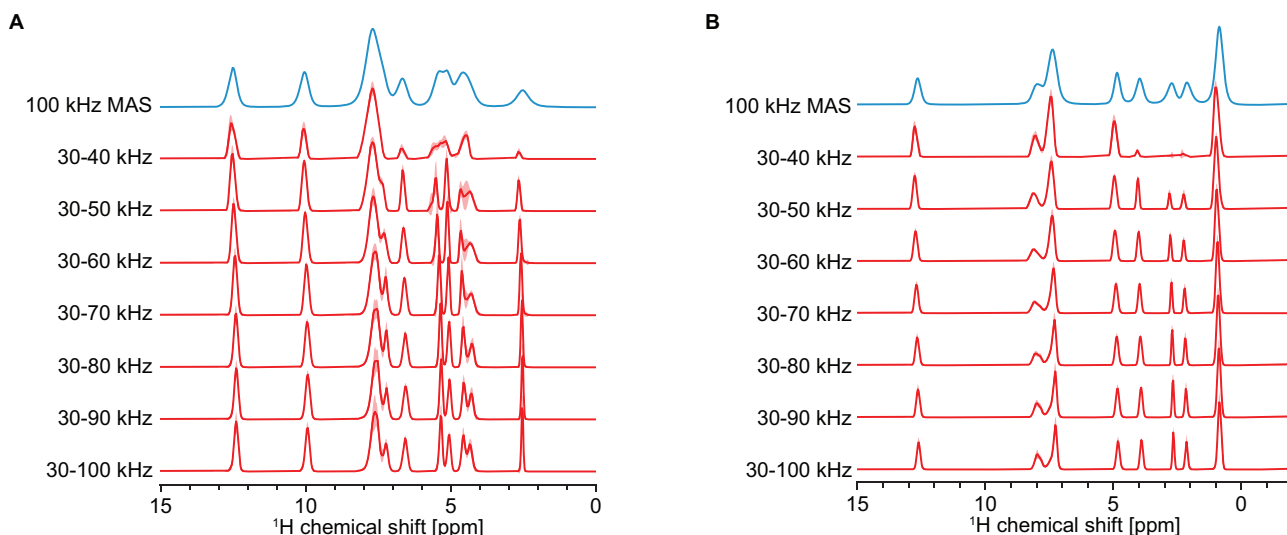


Figure 3.33. Isotropic spectra from different ranges of experimental MAS rates collected in 2 kHz increments. 100 kHz MAS spectra (blue) and isotropic spectra (red) obtained from different ranges of MAS rates for (A) L-tyrosine hydrochloride and (B) β -AspAla. The range of MAS rates used is indicated next to each isotropic spectrum.

3.3.4 Conclusion

In this section we have developed PIPNet, a deep learning approach to predict isotropic ^1H NMR spectra of solids from a two-dimensional dataset of variable rate MAS spectra, using a convolutional recurrent deep neural network with modified convolutional LSTM layers. The model is able to reliably predict linewidths on par with a previously introduced fitting approach, while bypassing several limitations of the latter approach and at a fraction of the computational cost, leading to faster and improved predictions.

The model was applied to sets of MAS spectra for eight molecular solids and showed marked resolution improvements compared to the highest MAS rates available, with linewidths down to 57 Hz. In addition, even using only relatively low MAS rates (30-50 kHz) as input for the model led to predictions with linewidths comparable to fast MAS (>100 kHz), which opens up the possibility to obtain well-resolved ^1H spectra for molecular solids from experimental data recorded at spinning rates accessible on less specialised hardware, such as, e.g., 1.3 mm MAS probes.

The spectra obtained using the PIPNet model from complete sets of variable rate MAS data up to 100 kHz MAS yield the best resolved ^1H spectra of molecular solids recorded to date, and we anticipate that access to even faster spinning rates in the future will further improve the robustness of this method. The model is freely available at <https://github.com/manucordova/PIPNet>.

3.3.5 Appendix V

Raw data statement. The NMR raw data are available from <https://doi.org/10.24435/materialscloud:a7-59> in JCAMP-DX version 6.0 standard format and original TopSpin format, as well as an archived version of the code and the pre-trained model used to obtain the results presented in this work. The code and pre-trained model are also available in the GitHub repository <https://github.com/manucordova/PIPNet>. All data and code are available under the license CC-BY-4.0 (Creative Commons Attribution-ShareAlike 4.0 International).

Table 3.17. Experimental details of all spectra datasets used in this study. Raw data is available at the link given above. The data for β -AspAla, flutamide, thymol, L-tyrosine hydrochloride and ampicillin are the same as previously reported.⁴⁰⁰

Sample	MAS range (kHz)	Step Size (kHz)	VT (K)	90° RF amplitude (kHz)	d1(s)	Number of FID points	SW (kHz)
β -AspAla	100 -30	2	275-295	277	6	2048	100
Flutamide	100 -30	2	275-295	286	18	2048	100
Thymol	100 -30	2	275	277	6	1024	100
L-tyrosine hydrochloride	100 -30	2	275-295	294	5	2048	100
Ampicillin	100 -30	2	275-295	286	3	4096	100
L-histidine hydrochloride monohydrate	100 -30	2	285-295	305	16	4096	227
MDMA hydrochloride	100 -40	2	275-295	277	7	2048	100
Molnupiravir	100 -40	2	275	333	30	4096	227
L-histidine hydrochloride Monohydrate (1.3mm probe)	60 -30	2	265-290	114	13.5	2048	100

Table 3.18. Experimental details of ^{13}C CP spectra of MDMA hydrochloride and molnupiravir. Exponential line broadening of 100 Hz was applied prior to Fourier transform of MDMA hydrochloride ^{13}C CP spectrum.

Sample	B0 field (MHz)	MAS rate (kHz)	VT (K)	d1(s)	Number of FID points	^1H - ^{13}C CP contact time (ms)	Acquisition time (ms)	^1H decoupling during acquisition	Size of real spectrum
MDMA hydrochloride	900	100	295	8	2776	3	10	Waltz16 (10 kHz)	16384
Molnupiravir	900	100	275	15	2776	3	10	Waltz16 (10 kHz)	16384

Table 3.19. Experimental details of hCH spectra of MDMA hydrochloride and molnupiravir. Exponential line broadening of 50 Hz was applied in the direct dimension prior to Fourier transform of MDMA hydrochloride hCH. Exponential line broadening of 50 and 500 Hz was applied in the direct and indirect dimension, respectively, prior to Fourier transform of molnupiravir hCH.

Sample	B0 field (MHz)	MAS rate (kHz)	VT (K)	d1(s)	Number of FID points (F2; F1)	^1H - ^{13}C CP contact time (ms)	^1H - ^{13}C back CP contact time (ms)	^1H decoupling during t_1	^{13}C decoupling during t_2	Size of real spectrum	Acquisition time (ms)
MDMA hydrochloride	800	100	275	7	1638 in F2 128 in F1	2.5	0.5	Waltz16 (10 kHz)	-	8192 in F2 512 in F1	1 in F2 10 in F1
Molnupiravir	900	100	275	15	4096 in F2 128 in F1	3	1.5	Waltz16 (10 kHz)	Waltz16 (6 kHz)	8192 in F2 256 in F1	1 in F2 22.5 in F1

Table 3.20. Experimental details of ^{13}C INADEQUATE spectrum of MDMA hydrochloride impregnated with AMUPOL in DNP juice. Exponential line broadening of 300 Hz was applied both in the direct and indirect dimension prior to Fourier transform.

Sample	B0 field (MHz)	MAS rate (kHz)	VT (K)	d1(s)	Number of FID points	^1H - ^{13}C CP contact time (ms)	^1H decoupling during t_1 and t_2	Acquisition time (ms)	Size of real spectrum
MDMA hydrochloride	400	10	100	2.2	1024 in F2 50 in F1	2	Spinal64 (100 kHz)	5 in F2 0.75 in F1	8192 in F2 1024 in F1

Table 3.21. Measured frequencies and linewidths of selected isolated peaks. Comparison of the position and linewidth (FWHM) of selected isolated peaks in experimentally measured 100 kHz MAS spectra, PIPNet predicted isotropic spectra, and spectra predicted using the previous fitting approach (PIP). The 100 kHz spectra for β -AspAla, flutamide, thymol, L-tyrosine hydrochloride and ampicillin are the same as previously reported.⁴⁰⁰

Compound	Label	Peak position [ppm]			Linewidth (FWHM) [ppm / Hz]		
		100 kHz MAS	PIPNet	PIP	100 kHz MAS	PIPNet	PIP
Ampicillin	18	0.57	0.54	0.46	0.66 / 529	0.38 / 307	0.16 / 128
	17	1.57	1.51	1.47	0.60 / 480	0.34 / 275	0.11 / 92
	15	10.03	9.90	9.98	0.72 / 579	0.44 / 355	0.11 / 90
β -AspAla	7	0.86	0.86	0.79	0.34 / 268	0.16 / 129	0.07 / 52
	5'	2.14	2.14	2.01	0.39 / 308	0.11 / 90	0.06 / 48
	5	2.72	2.66	2.68	0.39 / 309	0.09 / 73	0.03 / 27
	6	3.97	3.91	3.87	0.34 / 270	0.14 / 110	0.08 / 66
	2	4.86	4.83	4.75	0.28 / 224	0.15 / 122	0.04 / 31
	8	7.39	7.27	7.26	0.44 / 355	0.17 / 135	0.07 / 59
	1	12.64	12.61	12.54	0.29 / 229	0.15 / 123	0.11 / 89
Flutamide	11'/11"	1.21	1.12	1.11	0.70 / 559	0.38 / 301	0.11 / 90
	6	7.13	7.10	7.17	0.51 / 407	0.41 / 326	0.08 / 67
	3/8	7.99	7.93	7.84	0.58 / 461	0.46 / 368	0.08 / 62
	5	9.91	9.85	9.91	0.49 / 393	0.37 / 295	0.40 / 319
L-histidine hydrochloride monohydrate	1	5.13	5.10	5.10	0.52 / 472	0.29 / 256	0.07 / 65
	9	8.21	8.21	8.21	0.65 / 588	0.30 / 270	0.23 / 210
	7	12.44	12.40	12.40	0.46 / 412	0.34 / 308	0.35 / 316
	5	17.00	16.97	16.97	0.49 / 444	0.31 / 278	0.32 / 287
Thymol	4	3.39	3.36	3.31	0.42 / 333	0.20 / 161	0.15 / 122
	1	5.41	5.37	5.23	0.43 / 340	0.27 / 216	0.27 / 215
	2	6.20	6.20	6.11	0.39 / 315	0.27 / 212	0.23 / 183
	3	7.05	7.02	6.91	0.41 / 330	0.20 / 159	0.26 / 211
	7	9.34	9.31	9.29	0.94 / 749	0.14 / 112	0.08 / 62
L-tyrosine hydrochloride	3'	2.51	2.54	2.47	0.50 / 396	0.08 / 62	0.08 / 68
	6	6.66	6.57	6.49	0.44 / 353	0.18 / 147	0.03 / 25
	9	10.05	9.93	9.88	0.35 / 282	0.17 / 139	0.07 / 58
	1	12.52	12.40	12.32	0.33 / 264	0.17 / 132	0.07 / 53
MDMA hydrochloride	4	1.14	1.08	1.04	0.43 / 341	0.19 / 151	0.07 / 59
Molnupiravir	13/14	1.20	1.23	1.18	0.31 / 277	0.20 / 180	0.04 / 36
	4	6.75	6.69	6.64	0.43 / 387	0.10 / 89	0.07 / 63
	1/2	9.52	9.52	9.51	0.35 / 318	0.19 / 173	0.20 / 179
	2/1	10.72	10.69	10.65	0.42 / 380	0.33 / 295	0.46 / 414

Table 3.22. Additional frequencies and linewidths of selected isolated peaks in isotropic spectra. Position and linewidth (FWHM) of selected isolated peaks in PIPNet predicted isotropic spectra, excluding peaks already present in **Table 3.21**.

Compound	Label	Peak position [ppm]	Linewidth (FWHM) [ppm / Hz]
Ampicillin	2	3.89	0.36 / 290
β -AspAla	3	7.97	0.30 / 238
Flutamide	10	2.04	0.28 / 226
L-histidine hydrochloride monohydrate	6	7.50	0.25 / 203
	8	8.92	0.25 / 201
Thymol	6	0.46	0.28 / 225
L-tyrosine hydrochloride	8	5.05	0.09 / 71
	5	5.35	0.12 / 98
	10	7.61	0.32 / 252
MDMA hydrochloride	3	1.94	0.12 / 98
	5'	2.21	0.10 / 76
	1	2.82	0.22 / 173
	5	3.49	0.07 / 58
	12	6.02	0.07 / 57
	10/11	6.21	0.08 / 61
	7	6.64	0.14 / 112
	2'	9.23	0.08 / 65
	2	9.75	0.08 / 62
Molnupiravir	13'	3.73	0.09 / 71
	7	5.48	0.14 / 109

Table 3.23. Chemical shift assignment of MDMA hydrochloride. Assignment of ^1H and ^{13}C chemical shifts of MDMA hydrochloride.

Label	^1H exp. [ppm]	^{13}C exp. [ppm]	^1H DFT [ppm]	^{13}C DFT [ppm]
1	2.8	34	3.1	28
2	9.5, 10.0	-	9.9, 10.6	-
3	2.2	60	2.2	59
4	1.1	21	1.2	14
5	2.4, 3.8	38	2.5, 3.8	32
6	-	132	-	129
7	6.8	112	7.2	109
8	-	149	-	151
9	-	148	-	149
10	6.3	110	7.4	106
11	6.2	125	6.5	122
12	6.0	105	7.3, 7.9	110

Table 3.24. Chemical shift assignment of MDMA hydrochloride. Assignment of ^1H and ^{13}C chemical shifts of molnupiravir. A slash “/” indicates ambiguous assignments.

Label	^1H chemical shift [ppm]	^{13}C chemical shift [ppm]
1	9.5 / 10.7	-
2	10.7 / 9.5	-
3	-	140
4	5.5	104
5	6.8	128
6	-	154
7	5.4	86
8	4.5	70
9	2.7	-
10	5.4	-
11	4.2	68
12	10.0	80
13	3.7, 4.7	63
14	-	176
15	2.6	34
16	1.2	19 / 20
17	1.2	20 / 19

Table 3.25. Model and training parameters for PIPNet.

Parameter	Value
Number of models trained	16
Number of Conv-LSTM layers	6
Number of CNN filters (channels) per Conv-LSTM layer	64
Kernel size of Conv-LSTM layers	5
Number of filters (channels) of the output CNN	1
Kernel size of output CNN layer	5
Batch size	16
Number of training batches per epoch	1,000
Number of evaluation batches per epoch	200
Number of epochs	250
Optimiser	Adam
Initial learning rate	10^{-3}
Learning rate scheduler	Reduction on plateau of the evaluation loss by a factor 0.5, with patience of 10 epochs

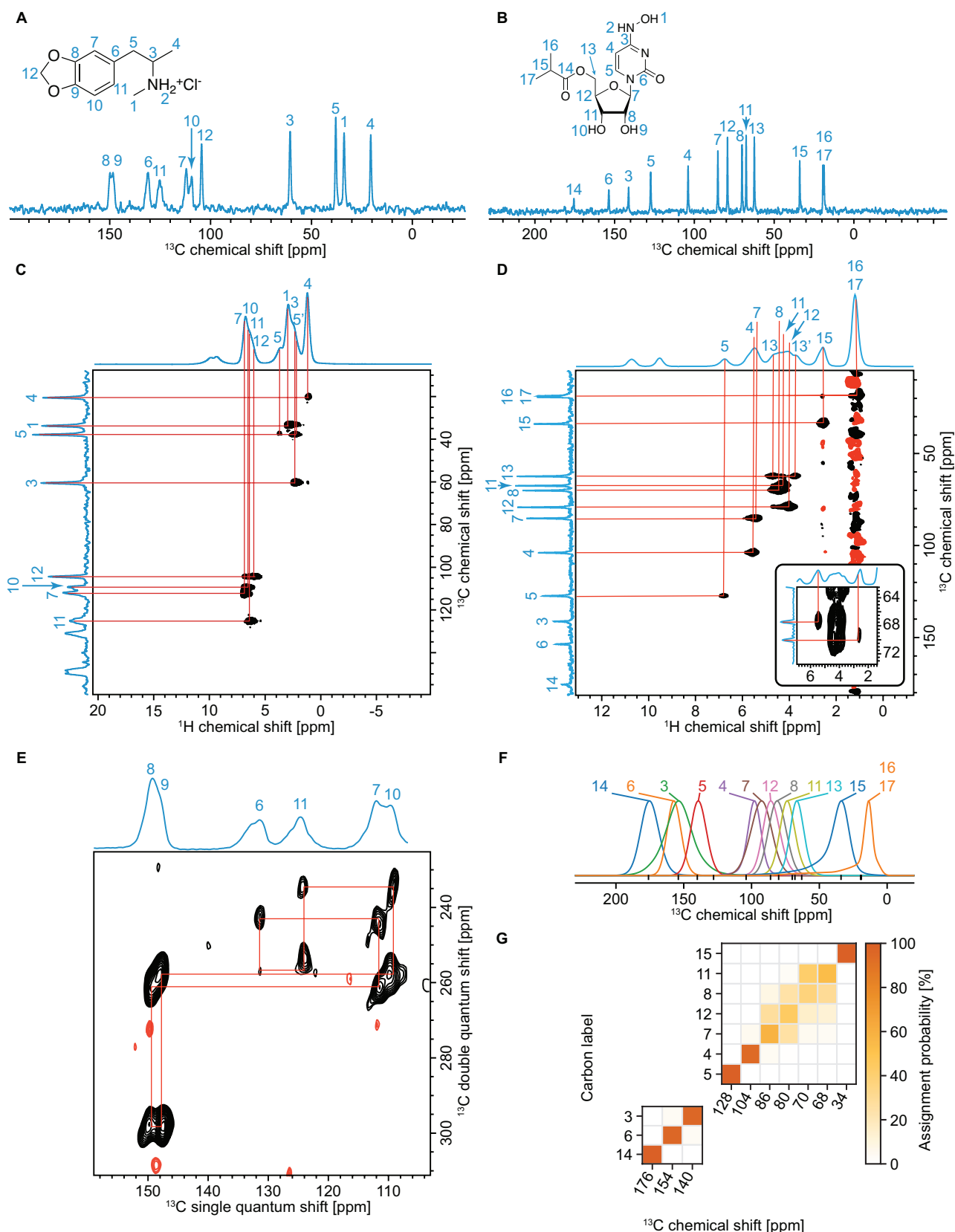


Figure 3.34. NMR experiments used for assignment. 100 kHz MAS (A), (B) ^{13}C CPMAS and (C), (D) ^1H - ^{13}C hCH spectra of (A), (C) MDMA hydrochloride (18.8 T) and (B), (D) molnupiravir (21.1 T). (E) ^{13}C - ^{13}C INADEQUATE spectrum of the aromatic region of MDMA hydrochloride. (F) Statistical distributions of ^{13}C chemical shifts obtained from the bonding environment of each carbon in molnupiravir and compared to the experimentally measured shifts (black vertical lines). (G) Probabilistic assignment of ^{13}C chemical shifts of quaternary and CH carbons of molnupiravir. Complete parameter sets and pulse sequences used to obtain the spectra are available with the raw data at the link given above. Blue numbers and red lines indicate the assignments obtained.

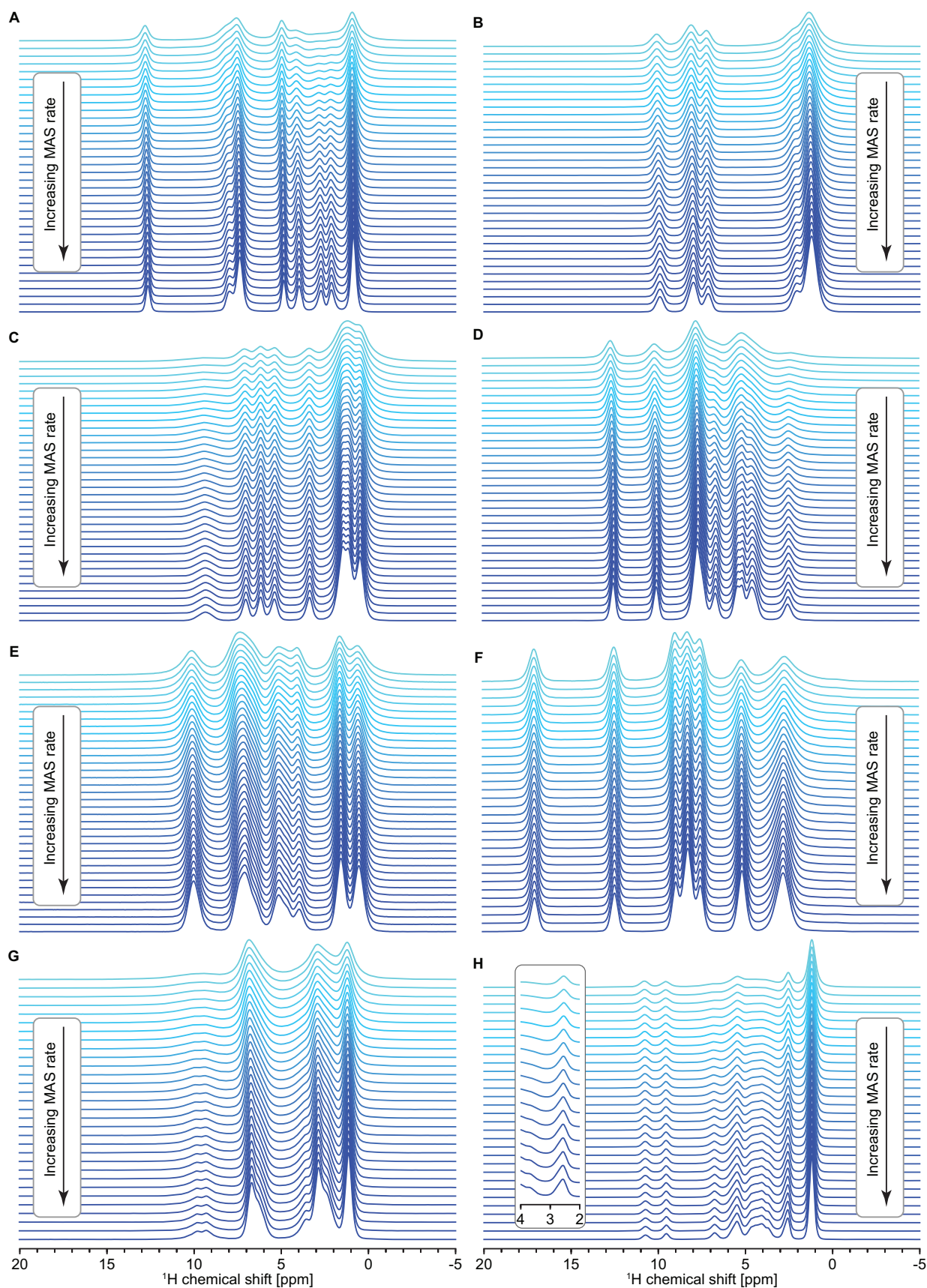


Figure 3.35. Experimental sets of variable-rate MAS spectra recorded for (A) β -AspAla, (B) flutamide, (C) thymol, (D) L-tyrosine hydrochloride, (E) ampicillin, (F) L-histidine hydrochloride monohydrate, (G) MDMA hydrochloride and (H) molnupiravir. The inset in (H) shows the 2-4 ppm region of every other spectrum recorded for molnupiravir, displaying the increasing asymmetry of the peak observed with increasing MAS rate. The data for (A-E) are the same as previously reported.⁴⁰⁰

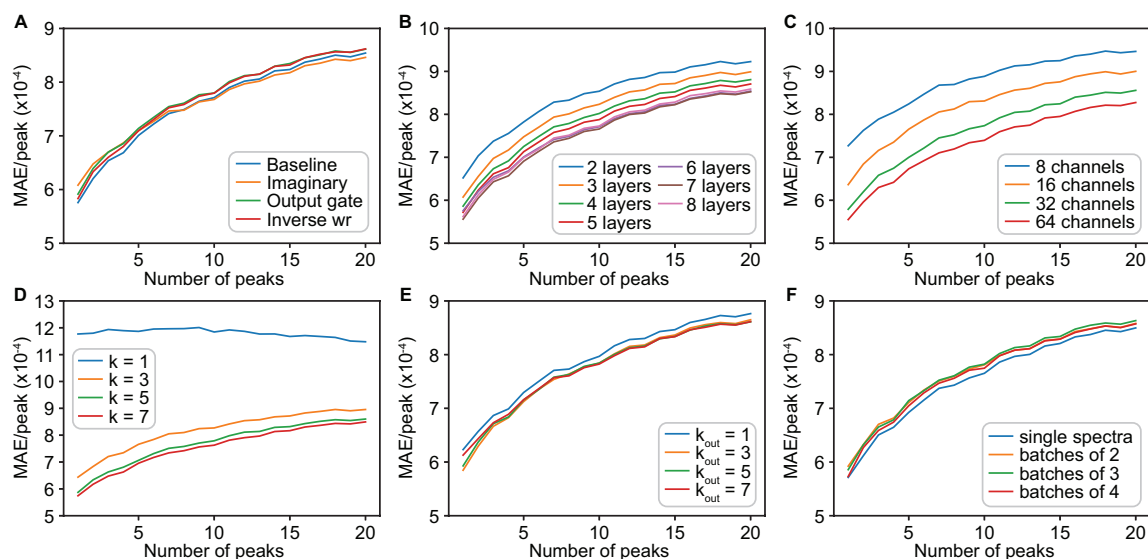


Figure 3.36. Model optimisation. MAE between predictions and ground-truth isotropic spectra for 1,024 samples with various numbers of peaks obtained with models trained with different (A) encoding of MAS spectra and with LSTM cells containing the output gate (green line), (B) number of layers, (C) number of hidden channels, (D) kernel sizes in the LSTM cells, (E) kernel sizes to convert the state vector of the last layer to the inferred isotropic spectra, and (F) numbers of spectra fed to the network at each step. In (A), “baseline” (blue line) corresponds to the encoding described in Section 3.3.2. “Imaginary” (orange line) corresponds to the addition of the imaginary part of the MAS spectra as an additional channel to the input vector. “Output gate” (green line) corresponds to the original convolutional LSTM cell with an output gate. “Inverse wr” (red line) corresponds to encoding the inverse of the MAS rate instead of the (normalised) MAS rate as described in Section 3.3.2.

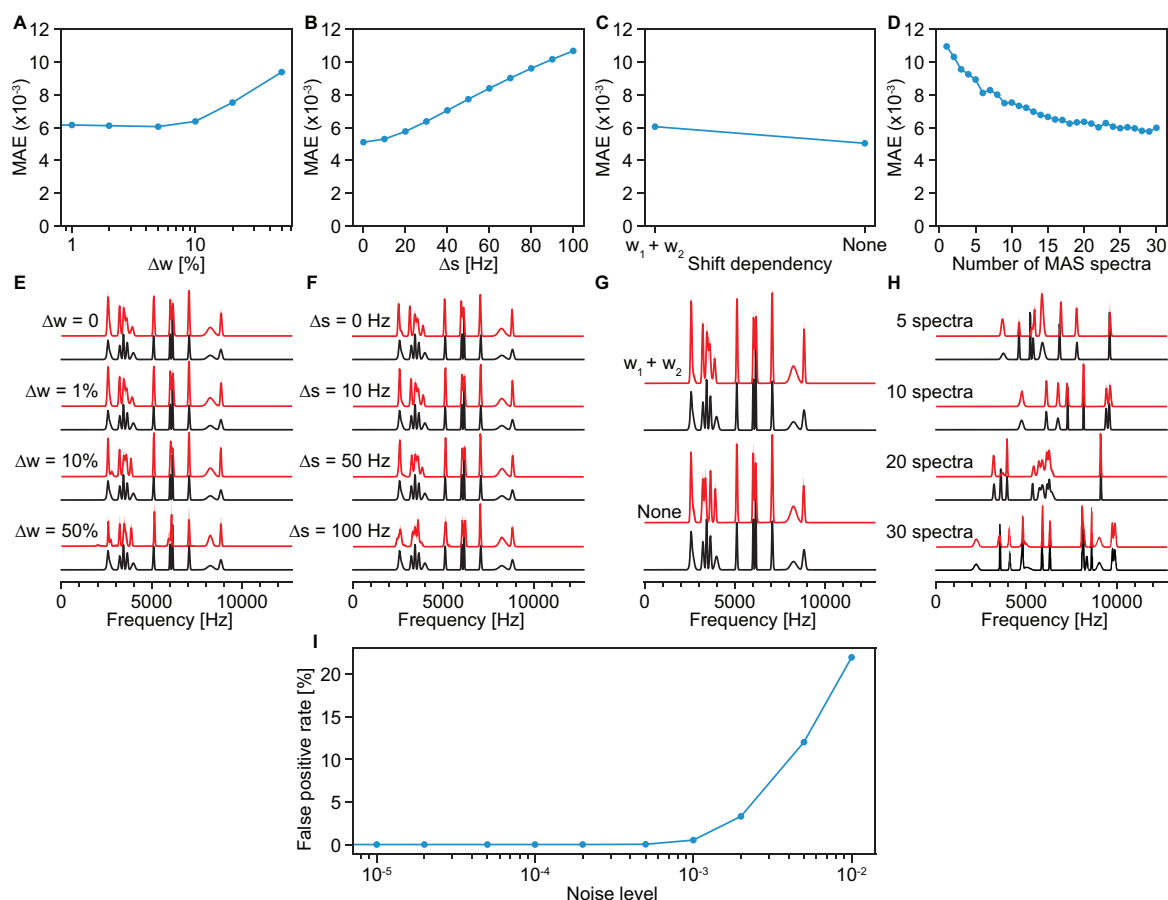


Figure 3.37. Model evaluation: MAS dependencies, number of spectra and false positive rate. (A), (B), (C), (D) MAE between predictions and ground-truth isotropic spectra for 1,024 samples and (E), (F), (G), (H) illustrative comparisons of predicted (red) and ground-truth (black) isotropic spectra with various (A), (E) levels of noise in the MAS-dependent width (see Equation 3.9), (B), (F) levels of noise in the MAS-dependent shift parameter (see Equation 3.10), (C), (G) shift dependencies and (D), (H) numbers of MAS spectra fed to the network. (I) False positive signal rate (as defined in Section 3.3.2) observed in the isotropic spectra predicted from synthetic MAS spectra as a function of the noise level.

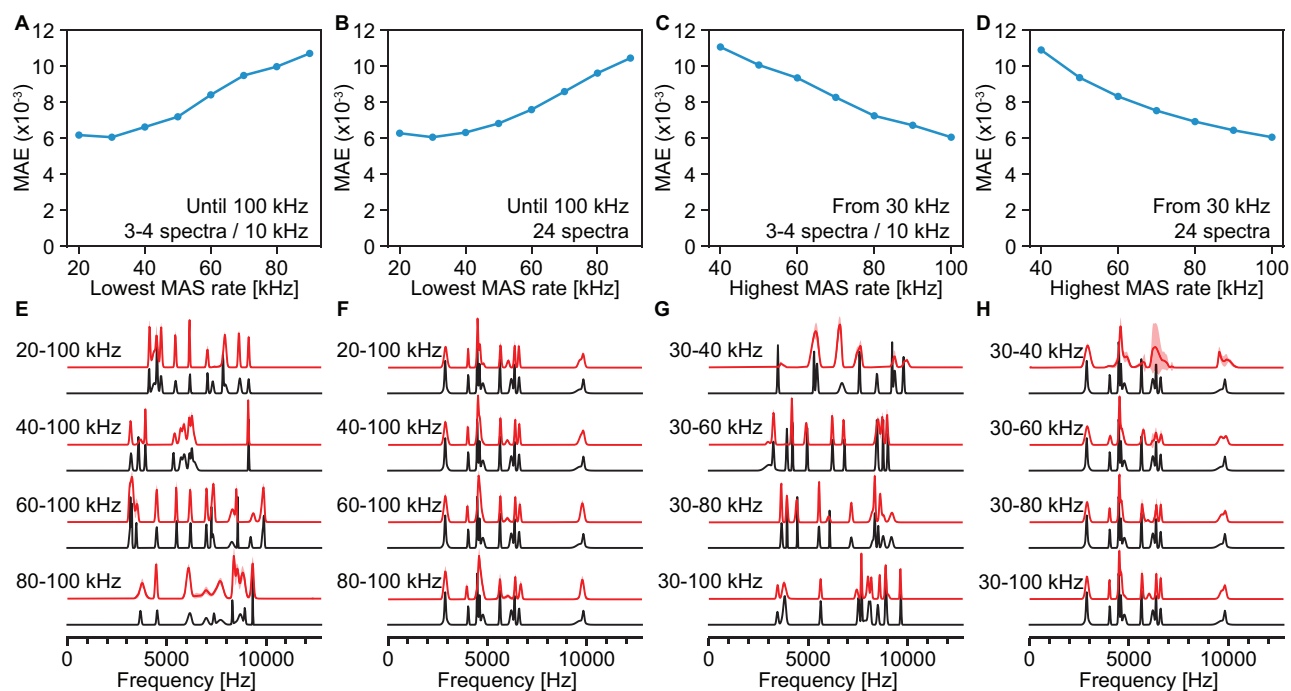


Figure 3.38. Model evaluation: MAS rate ranges. (A), (B), (C), (D) MAE between predictions and ground-truth isotropic spectra for 1,024 samples and (E), (F), (G), (H) illustrative comparisons of predicted (red) and ground-truth (black) isotropic spectra obtained from MAS datasets containing various (A), (B), (E), (F) lower bounds for the MAS rate while keeping the higher bound to 100 kHz and (C), (D), (G), (H) higher bounds for the MAS rate while keeping the lower bound to 30 kHz. In (A), (C), (E) and (G) the number of MAS spectra generated was set such that there are on average between 3 and 4 spectra per 10 kHz MAS rate range. In (B), (D), (F) and (H), the number of MAS spectra generated was set to 24 regardless of the range of MAS rates considered.

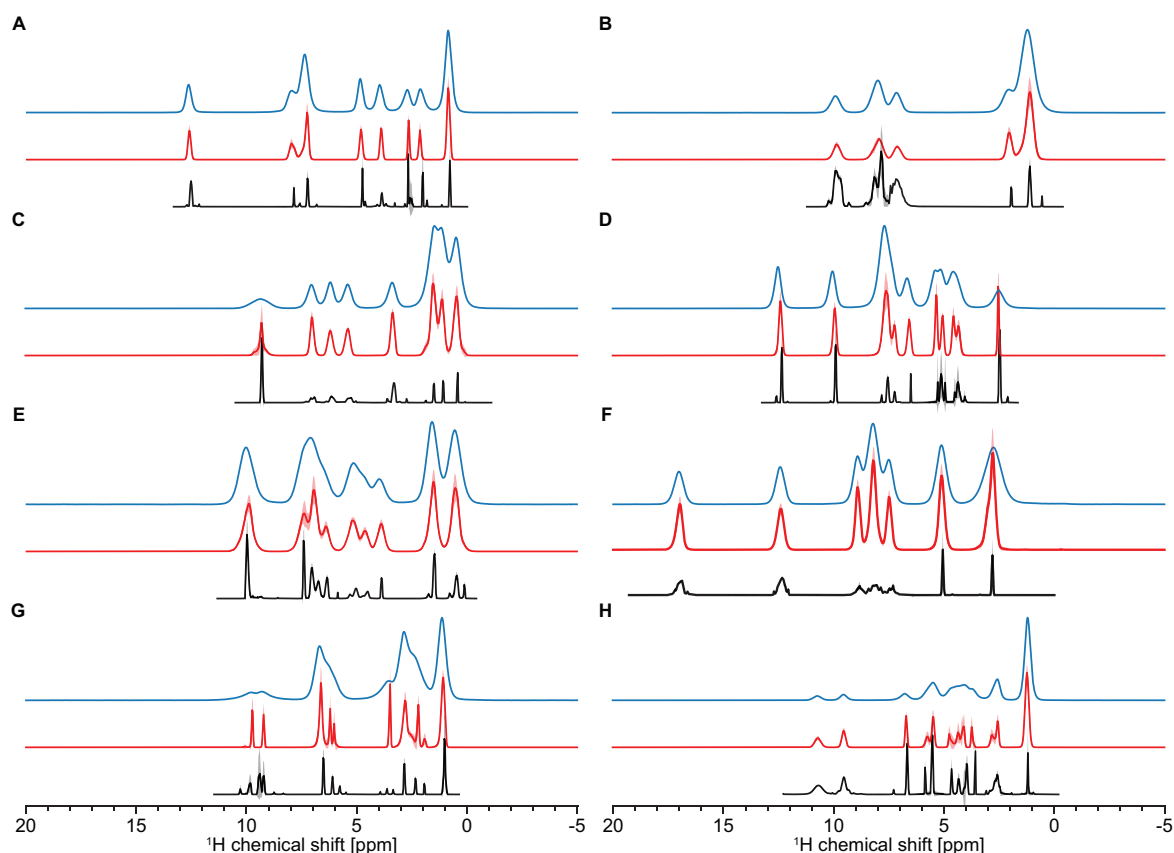


Figure 3.39. Comparison with the previous approach. 100 kHz MAS spectra (blue) and isotropic spectra of (A) β -AspAla, (B) flutamide, (C) thymol, (D) L-tyrosine hydrochloride, (E) ampicillin, (F) L-histidine hydrochloride monohydrate, (G) MDMA hydrochloride and (H) molnupiravir obtained using PIPNet (red) and the previous approach (black).

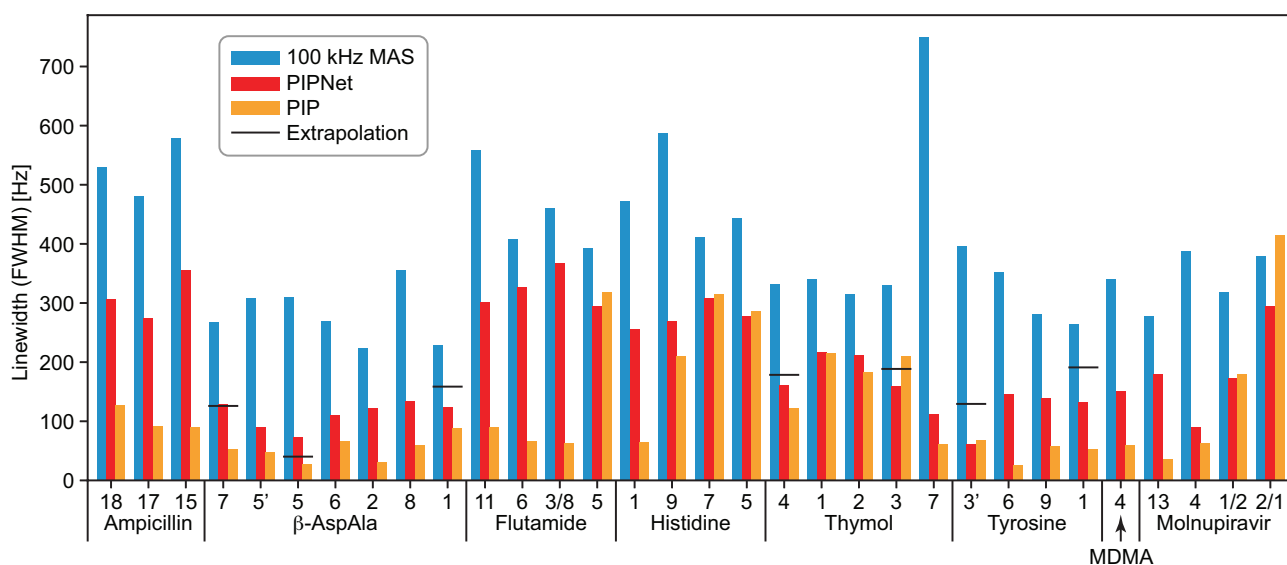


Figure 3.40. Linewidth comparison. Comparison of the linewidth of selected isolated peaks in 100 kHz MAS spectra (blue) and isotropic spectra obtained using PIPNet (red) and the previous fitting approach (orange). Black lines indicate the linewidths obtained from extrapolation of the linewidths in MAS spectra, assuming a first-order inverse MAS dependence (Figure 14 of Ref. 85).

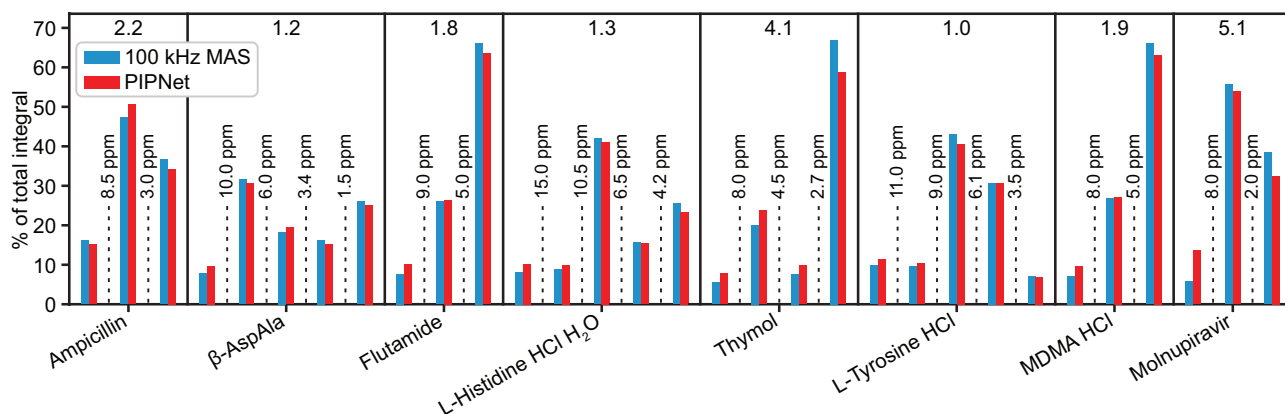


Figure 3.41. Relative integrals. Comparison of the integral of separate regions from the 100 kHz MAS spectra (blue) and predicted isotropic spectra (red). The mean absolute error on the percentage of total integral of regions for each compound is indicated at the top of each panel.

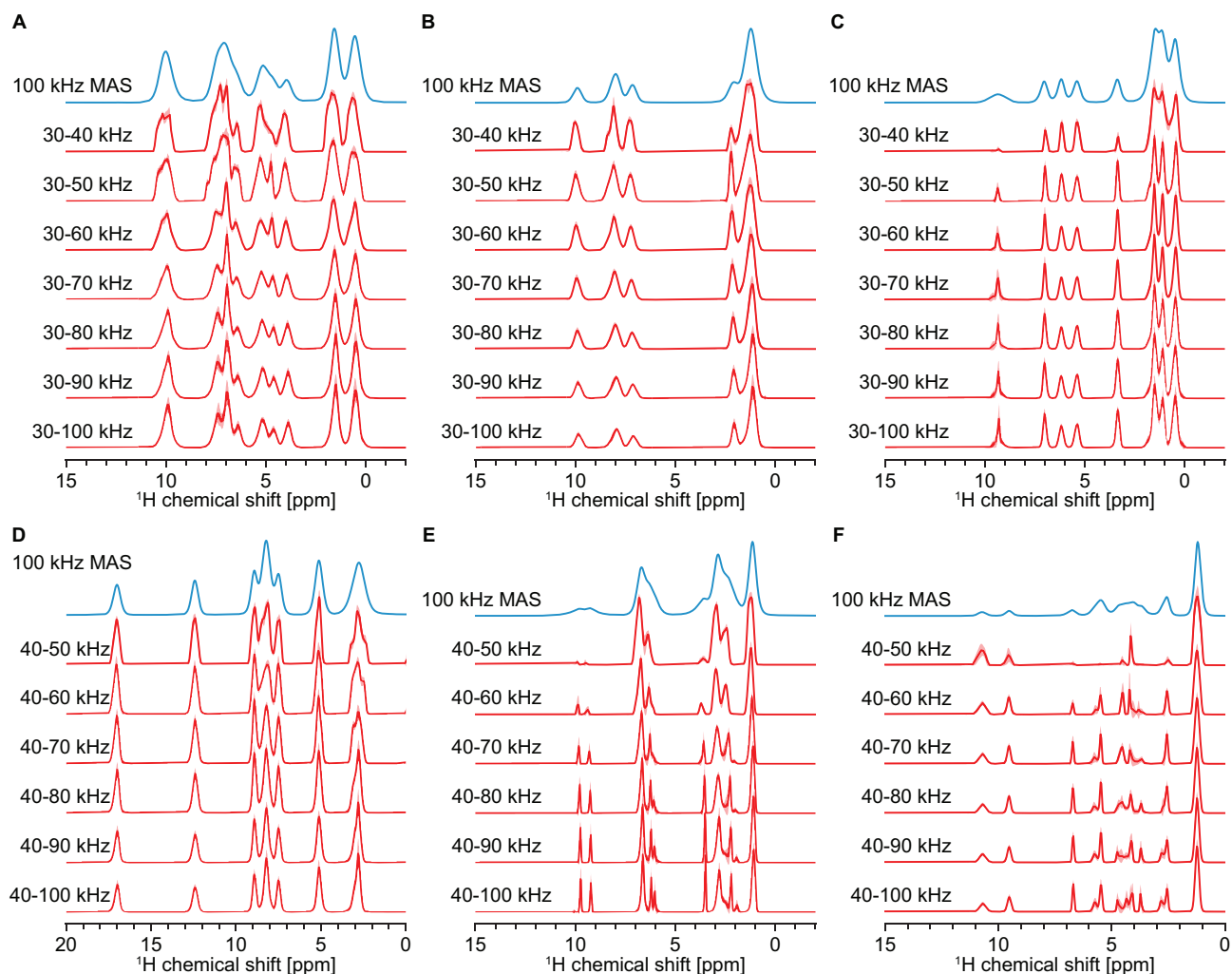


Figure 3.42. Isotropic spectra from different ranges of experimental MAS rates collected in 2 kHz increments. 100 kHz MAS spectra (blue) and isotropic spectra (red) obtained from different ranges of MAS rates for (A) ampicillin, (B) flutamide, (C) thymol, (D) L-histidine hydrochloride monohydrate, (E) MDMA hydrochloride and (F) molnupiravir. The range of MAS rates used is indicated next to each isotropic spectrum.

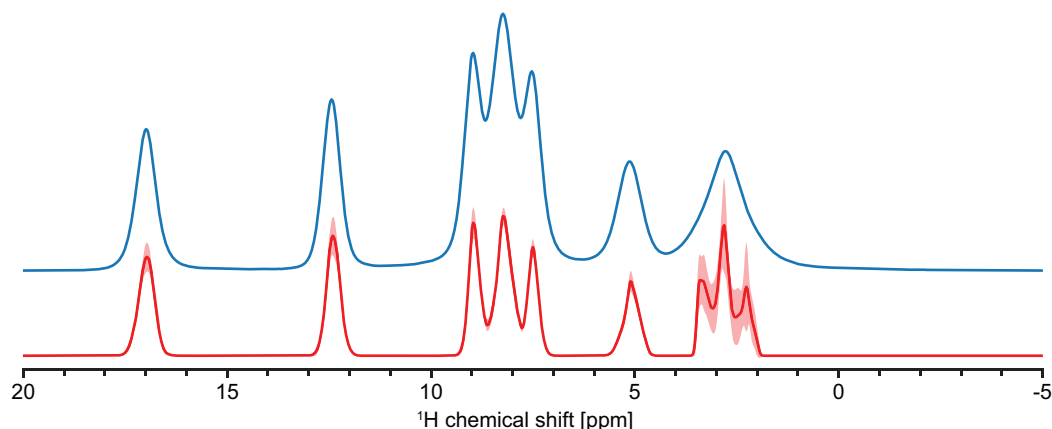


Figure 3.43. 60 kHz MAS spectrum (blue) and isotropic spectrum (red) obtained from a dataset of 16 MAS spectra of L-histidine hydrochloride obtained using a 1.3 mm rotor.

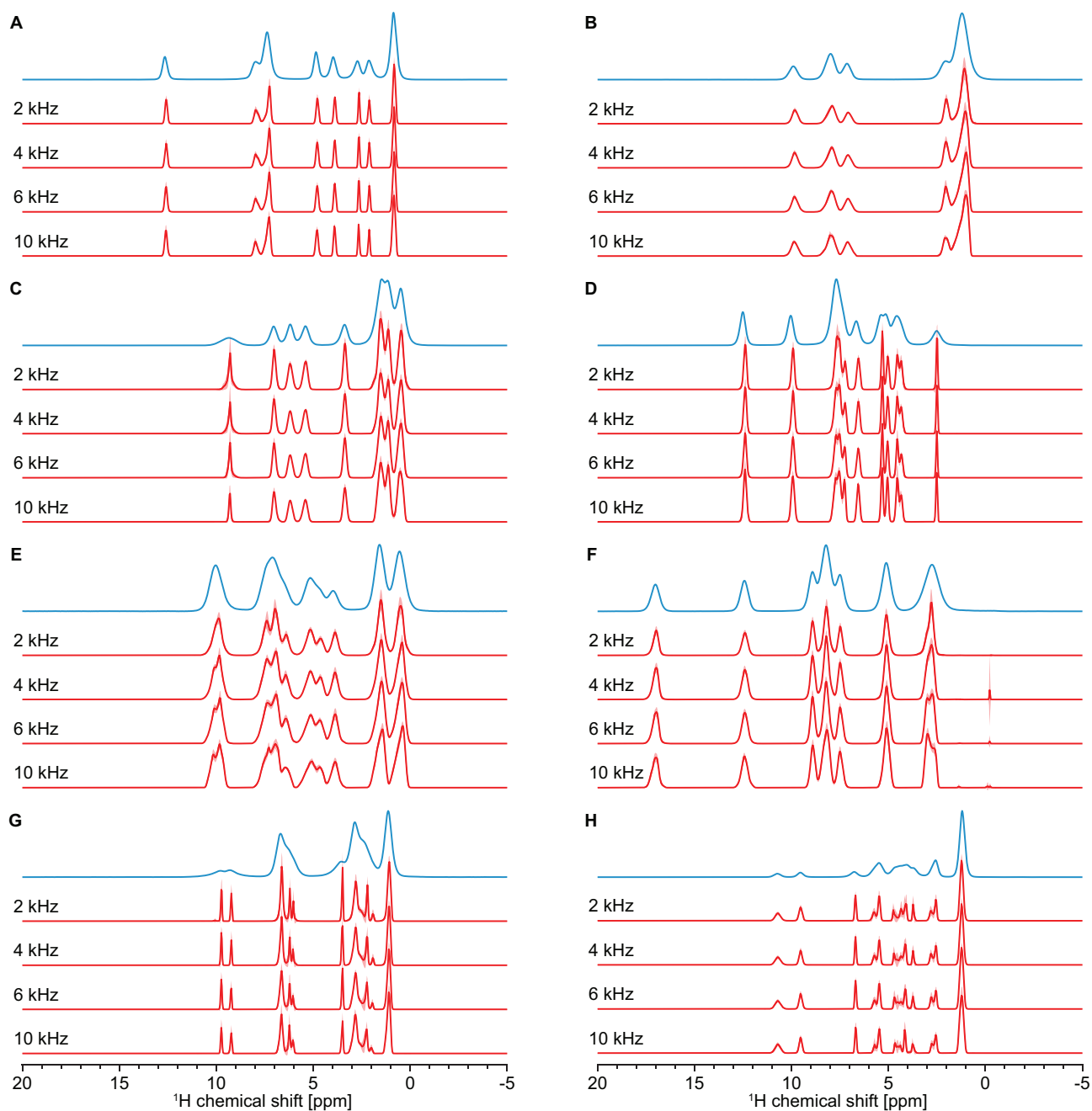


Figure 3.44. Isotropic spectra from different increments of experimental MAS rates measured between 40 and 100 kHz. 100 kHz MAS spectra (blue) and isotropic spectra (red) obtained from different MAS rate increments for (A) β -AspAla, (B) flutamide, (C) thymol, (D) L-tyrosine hydrochloride, (E) ampicillin, (F) L-histidine hydrochloride monohydrate, (G) MDMA hydrochloride and (H) molnupiravir. The MAS rate increment used is indicated next to each isotropic spectrum.

3.4 Two-dimensional pure isotropic proton solid state NMR

This section has been adapted with permission from: Moutzouri, P.; Cordova, M.; Simões de Almeida, B.; Torodii, D.; Emsley, L., Two-dimensional Pure Isotropic Proton Solid State NMR. *Angewandte Chemie International Edition* **2023**, 62 (21), e202301963. (post-print)

My contribution was to develop and apply the method and to analyse results. I also contributed to the writing of the manuscript.

3.4.1 Introduction

In this section we extend the PIP approach to a second dimension in order to obtain ultra-high resolution ^1H - ^1H double-quantum / single-quantum³⁶⁶ (DQ/SQ) dipolar correlation spectra and ^1H - ^1H spin-diffusion^{461, 462} (PSD) spectra. We illustrate the method on L-tyrosine hydrochloride and ampicillin, where we obtain two-dimensional spectra with significantly higher resolution as compared to corresponding spectra acquired at 100 kHz MAS. The spectral resolution is very significantly increased in both dimensions, allowing the identification of resolved isotropic correlation peaks that were overlapped in the 100 kHz MAS spectra.

The PIP approach works by obtaining a one-dimensional pure isotropic spectrum from a two-dimensional set of MAS spectra recorded at variable spinning rates (VMAS).^{400, 463} In this 2D dataset, the isotropic part of the interactions remains constant as a function of spinning rate, while the anisotropic parts that lead to broadening and shifts are scaled by the spinning. The isotropic part can be separated out by a suitable transform, so far shown either by parametric fitting,⁴⁰⁰ or more recently by a deep learning method.⁴⁶³ In the latter approach a modified convolutional LSTM neural network, dubbed PIPNet, was trained on millions of synthetic VMAS datasets to infer isotropic ^1H NMR spectra. Both approaches, yield isotropic spectra that display linewidths in the 50-400 Hz range for crystalline molecular solids.

Here, we use three-dimensional datasets made up of two-dimensional DQ/SQ or spin-diffusion spectra acquired at different MAS rates to obtain two-dimensional ^1H - ^1H DQ/SQ or ^1H - ^1H PSD correlation spectra with pure isotropic lineshapes in both dimensions by transforming the data with a suitable deep learning prediction network, dubbed PIPNet2D.

3.4.2 Methods

NMR experiments. The pure isotropic 2D approach is applied here to experimental datasets from two microcrystalline organic solids: L-tyrosine hydrochloride and ampicillin. For each sample a set of 2D BABA³⁶⁶ (Figure 3.52 and 3.53) or spin-diffusion^{461, 462} spectra (Figure 3.54) was acquired at MAS rates ranging from 50 to 100 kHz, using a Bruker 0.7 mm room temperature HCN CP-MAS probe at a magnetic field of 21.1 T corresponding to a ^1H frequency of 900 MHz. For each sample, prior to acquisition the magic angle was set by maximising the T_2' of the proton signals at the fastest MAS rate, and the 90° pulse width was optimised. All the data were acquired with active temperature regulation to maintain the sample temperature at about 295 K across the range of spinning rates. For each MAS rate, the BABA 2D experiments were rotor-synchronised in the indirect dimension and the number of increments was adjusted to achieve a t_1^{max} of 2.5 and 2.4 ms for L-tyrosine hydrochloride and ampicillin, respectively. For L-tyrosine hydrochloride the spin-diffusion 2D experiments were acquired with a t_1^{max} of 14.6 ms and with a mixing time chosen to produce cross peaks with similar intensities throughout the series. More specifically, as the spinning rate was increased, the proton spin diffusion (PSD) mixing time was varied was also increased in order to compensate for slower spin diffusion at faster MAS rates. A States-TPPI acquisition scheme was used to obtain phase-sensitive two-dimensional spectra. 16-step phase cycling was used for BABA and EXSY (PSD) experiments. All experimental parameters are given in Tables 3.26-3.28.

All spectra were phased, baseline corrected, their full integrals were normalised (either with respect to the total integral, for the DQ/SQ spectra, or with respect to a selected cross peak intensity, for the PSD spectra), and the spectra were scaled to the maximum amplitude of the 100 kHz spectrum of 1 in order to match the typical amplitudes of the generated training spectra. The experimental DQ/SQ MAS spectra were sheared to an SQ/SQ representation prior to processing. The SQ/SQ representation is exactly equivalent to the DQ/SQ but gives an easier to visualise rendition of the two-dimensional lineshapes. No weighting function was applied prior to Fourier transformation.

The samples of ampicillin, and L-tyrosine hydrochloride were purchased from Sigma Aldrich. Both samples were used without further recrystallisation, after mild crushing with a mortar and a pestle.

Data generation. As described previously,⁴⁶³ we generate synthetic isotropic and variable MAS rate 1D spectra according to a theoretical description of the dependence of the spectra on the MAS rate. A MAS spectrum is composed of a sum of peaks $I_{\omega_r}(\nu)$, each resulting from the convolution between the corresponding Gaussian peak in the isotropic spectrum $I_{\infty}(\nu)$ and a Gaussian-Lorentzian sum (GLS) function:

$$GLS(\nu; w, p, m) = (1 - m) \exp \left[-\frac{4 \ln(2) (\nu - p)^2}{w^2} \right] + \frac{m}{1 + \frac{4(\nu - p)^2}{w^2}}, \quad (3.14)$$

where w and p are the width and position of the GLS function, respectively, and m is the mixing factor describing the lineshape of the function (purely Gaussian with $m = 0$, and purely Lorentzian with $m = 1$). We set p to be in the middle of the generated spectrum, such that convoluting the GLS with the isotropic peak does not affect the position after convolution. After the convolution, a MAS-dependent shift of the frequency of the peak s_{ω_r} was added to capture the residual shift observed in MAS spectra. The generation of a peak in an MAS spectrum is thus described as:

$$I_{\omega_r}(\nu) = \text{FT}[I_{\infty}(t) \cdot e^{i2\pi s_{\omega_r} t}] * GLS(\nu; w_{\omega_r}, p, m_{\omega_r}), \quad (3.15)$$

where $\text{FT}[\cdot]$ is the Fourier transform and $*$ denotes the convolution operation.

Here, we generated spectra made up of 128 points with a time-domain sampling frequency of 3.2 kHz, corresponding to a frequency domain resolution of 25 Hz.

1D isotropic spectra were generated as the sum of between 1 and 5 Gaussian peaks, with a linewidth sampled from a uniform distribution between 50 and 200 Hz (60% probability), between 100 and 500 Hz (20% probability), or between 100 and 1000 Hz (20% probability). Potential negative points in the isotropic spectra generated were set to zero. In order to randomise peak intensities, we rescaled the height of each isotropic peak to a random value between 0.1 and 1.

For each pair of 1D isotropic spectra, 12 MAS rates were selected randomly between 50 and 100 kHz, and the corresponding MAS spectra were constructed as described in **Equation 3.15**, using a GLS function with parameters s_{ω_r} , w_{ω_r} and m_{ω_r} following a second-order MAS dependence subject to noise,

$$w_{\omega_r} = \frac{w_1}{\omega_r} + \frac{w_2}{\omega_r^2} + \Delta w, \quad (3.16)$$

$$s_{\omega_r} = \frac{s_1}{\omega_r} + \frac{s_2}{\omega_r^2} + \Delta s, \quad (3.17)$$

$$m_{\omega_r} = \frac{m_1}{\omega_r} + \frac{m_2}{\omega_r^2}. \quad (3.18)$$

For each peak, the value of w_1 in **Equation 3.16** was drawn from a uniform distribution in the range $[10^7, 2 \cdot 10^7]$ Hz² with 60% probability, in the range $[10^7, 5 \cdot 10^7]$ Hz² with 20% probability, and in the range $[5 \cdot 10^7, 10^8]$ Hz² otherwise. w_2 was set to be 0 with 50% probability, or a random value between 10^{11} and $5 \cdot 10^{11}$ Hz³. In addition, for each MAS rate the GLS width w_{ω_r} was randomly perturbed by a value drawn from a normal distribution $\Delta w \sim \mathcal{N}(0, \sigma_w)$ Hz where σ_w was set to be 5% of the range of width generated between the lowest and highest MAS rates with the selected values of w_1 and w_2 . The value of s_1 in **Equation 3.17** was randomly sampled between -10^7 and 10^7 Hz², and s_2 was set to zero with a 50% probability, or randomly sampled between $-2 \cdot 10^{10}$ and $2 \cdot 10^{10}$ Hz³. Δs was randomly drawn from a normal distribution $\Delta s \sim \mathcal{N}(0, 25)$ Hz for each peak and each MAS rate. The GLS mixing was set to follow the inverse MAS dependence described in **Equation 3.18** with a probability of 50% and with the value of m_1 set to zero with a 20% probability, drawn from a uniform distribution in the range $[0, 10^4]$ Hz with 20% probability, or drawn from a uniform distribution in the range $[10^4, 5 \cdot 10^4]$ Hz, and with m_2 set to zero with a 50% probability, or drawn from a uniform distribution in the range $[10^8, 5 \cdot 10^8]$ Hz². Resulting values of m_{ω_r} above one were capped to one. Otherwise, the mixing m_{ω_r} was set to be either constant, monotonously increasing with random values between 0 and 1, or monotonously decreasing with random values between 0 and 1 with increasing MAS rate. These three dependences were considered with equal probability.

In addition, each peak in each 1D MAS spectrum was subject to phase distortion drawn from a normal distribution $N(0, 0.05)$ with a 50% probability, and each isotropic peak was assigned a 30% chance to have a decreasing intensity in MAS spectra with increasing rate by multiplying the intensity of the corresponding peak in each MAS spectrum by a value linearly decreasing between 1 and a final value sampled uniformly in the range $[0.3, 0.7]$.

After generating the pairs of isotropic and MAS spectra for each rate, convolution and rotation yields the final 2D isotropic and sets of MAS spectra.

3.4.3 Results and Discussion

In the absence of any extensive experimental databases of NMR spectra, training machine learning models on synthetic datasets (of shifts or spectra) has been shown to be an efficient way forward.^{251, 252, 255, 259, 261, 262, 276, 279, 281-286, 291, 311, 365, 464} Here, the generation of synthetic three-dimensional datasets used to train a LSTM neural network was based on a protocol analogous to that used previously for two-dimensional VMAS datasets.^{400, 463} The overall approach is illustrated schematically in **Figure 3.45**. Specifically, synthetic two-dimensional pure isotropic spectra (ground truth) were generated as the outer product of two randomly generated one-dimensional isotropic spectra. The component one-dimensional isotropic spectra and associated VMAS spectra were generated assuming that the dipolar couplings lead to a MAS rate-dependent broadening, with a shape that is a sum of Gaussian and Lorentzian components, and that they also lead to a MAS rate-dependent shift in the peak positions.^{85, 400} We also include random parameter variations in peak positions, peak shapes, MAS dependences, phase and intensity errors, and noise, as described previously in Ref. 463 but with an increased probability to generate broad isotropic peaks in order to promote diversity in the two-dimensional isotropic lineshapes.

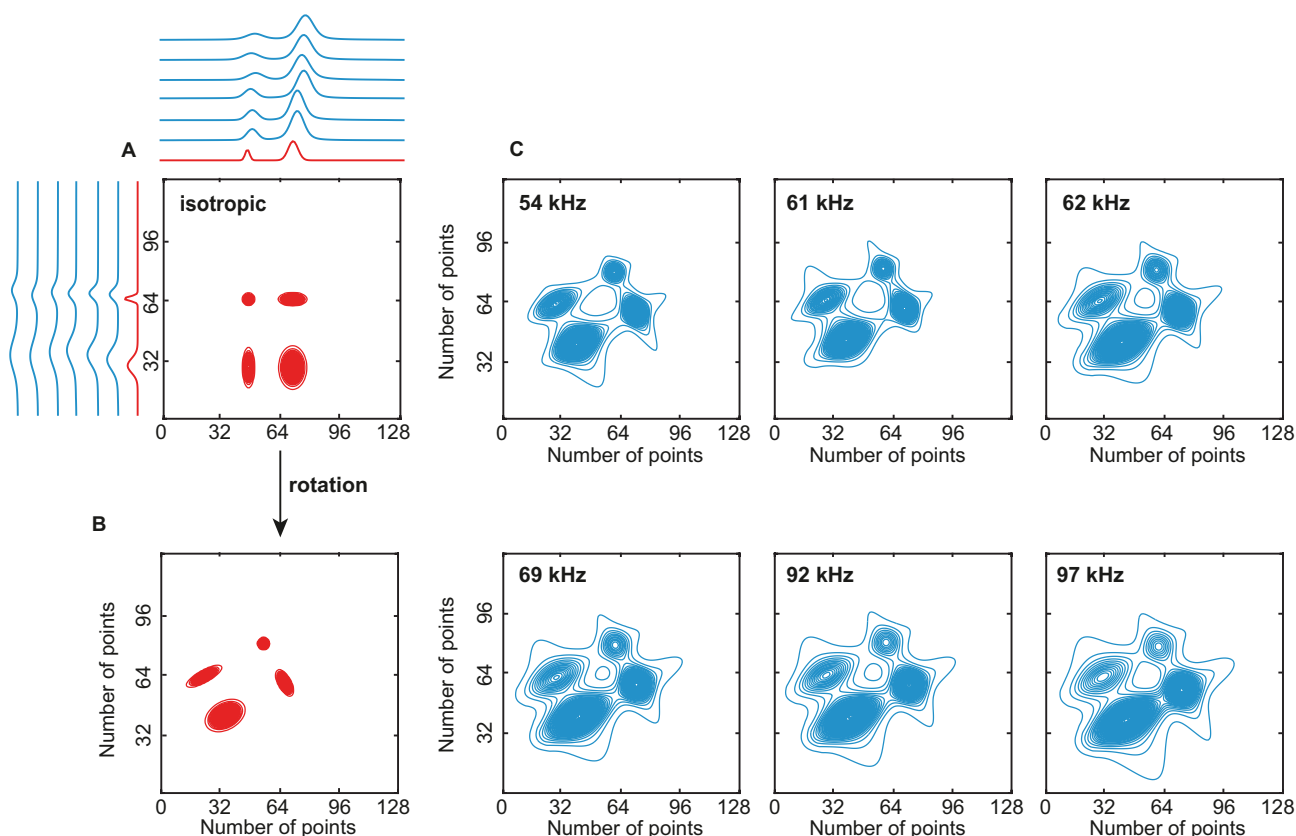


Figure 3.45. Representative example of a synthetic isotropic (red) two-dimensional spectrum (A) before and (B) after rotation and (C) a three-dimensional dataset (blue) that consists of two-dimensional spectra at different MAS rates. In (A), the one-dimensional isotropic spectra whose outer product leads to this two-dimensional isotropic spectrum and their corresponding variable MAS rate spectra are also shown. Here the full dataset contains 12 spectra at different MAS rates but only six selected spectra are shown. The rotation angle applied here during the data generation process was 61.5° .

The corresponding synthetic three-dimensional datasets of two-dimensional spectra at variable MAS rates were generated by the outer product of the 2D VMAS datasets. To mimic varying degrees of correlation in the 2D lineshapes, the sets of isotropic and associated MAS spectra were then rotated with a probability of 50% by a random angle uniformly sampled between 0 and 90°, in order to produce lineshapes with elongated shapes along different orientations in the 2D spectra. (Examples are shown in **Figure 3.50**). Each three-dimensional dataset generated contained 12 MAS spectra, each of which generated with a random MAS rate sampled from a uniform distribution between 50 and 100 kHz. Complete details about the data generation are given in **Section 3.4.2**. An example of a synthetic MAS dataset and its isotropic counterpart typically used for the training of the model is shown in **Figure 3.45**.

Note that the synthetic spectra generated here do not actually make any assumptions or follow any particular rules associated with a type of experimental 2D correlation spectrum. That is, they do not correspond to, e.g., COSY, or DQ/SQ spectra, with diagonal and/or cross peaks in well-defined positions. The synthetic spectra only consist of a set of two-dimensional peaks in randomised positions in the spectra, and with lineshapes that obey the rules described above. As such, the model could be applied to any 2D correlation spectrum.

In the problem at hand, the LSTM type of network appears suitable since it has been shown to outperform other recurrent neural networks in processing time series,^{290, 293, 294, 451} and since it was shown to work well to predict 1D isotropic spectra.⁴⁶³ The only changes used here with respect to the model to predict 1D spectra is the use of two-dimensional convolutional layers, using 4 layers instead of 6, and the use of only one model instead of a committee of 16 models. The latter being done in order to reduce the computational requirement at inference. (A link to the code used is given in **Appendix VI**). The model was trained on a total of 1,000,000 datasets, corresponding to 12,000,000 spectra. To process each three-dimensional dataset of MAS spectra in order to obtain the isotropic 2D spectrum, the network is incrementally given the next MAS spectrum in the series in order of increasing MAS rate, and produces an output after each step, as described previously.⁴⁶³ The model was trained by minimising the mean absolute error (MAE) between the prediction after each step and the ground-truth two-dimensional isotropic spectrum. Detailed model and training parameters are given in **Table 3.29**.

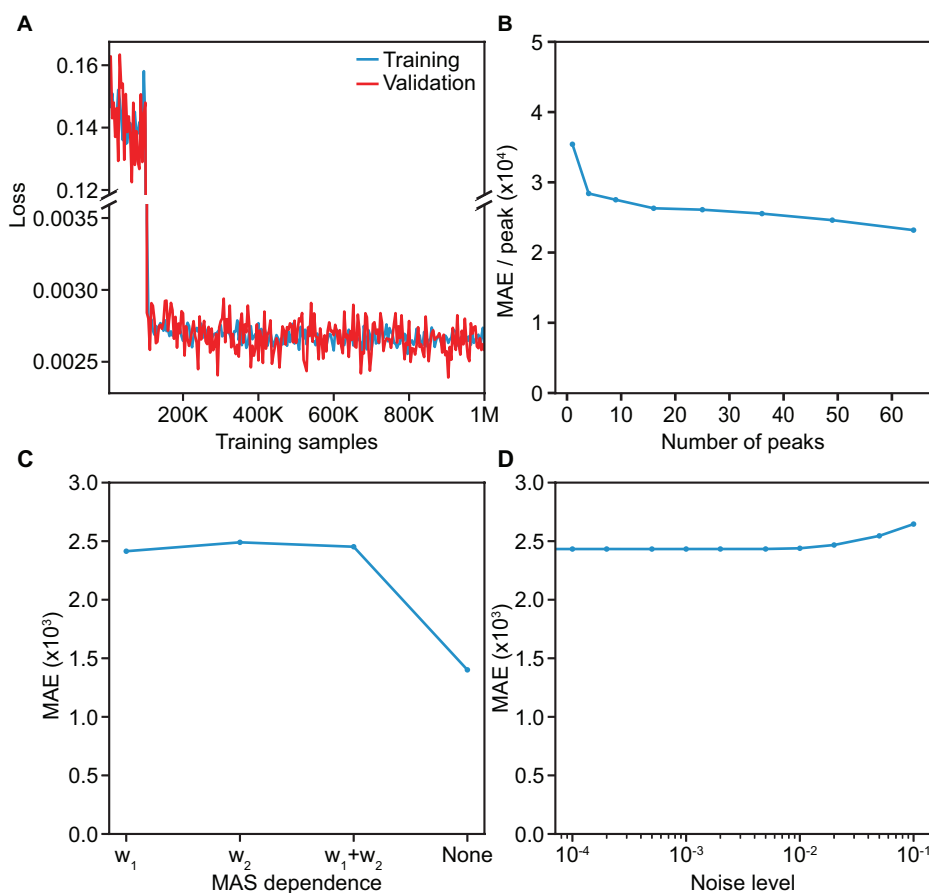


Figure 3.46. (A) Evolution of the loss function during model training. (B)-(D) MAE between predictions and ground-truth isotropic spectra for 1,024 samples of isotropic spectra with various (B) numbers of peaks, (C) MAS dependence (w_1 : first-order, w_2 : second-order), and (D) noise levels.

As before, due to the sparsity of signal in the two-dimensional isotropic spectra, we initially convoluted the entire target isotropic spectrum with a 2D Gaussian function with a width of 25 Hz and weighted the loss function by the maximum between 1 and 10 times the value of the target isotropic spectrum (after convolution with the Gaussian) in order to promote the identification of signal in the spectra. After 200,000 sets of spectra, this convolution and weighting were removed for the rest of the training.

Random noise was also introduced into the generated MAS two-dimensional spectra following the typical signal-to-noise ratios observed in experimental ^1H - ^1H correlation spectra (between 10 and 100 for the most intense peak at 100 kHz). **Figure 3.46A** shows the evolution of the loss function during the model training.

The model was evaluated by computing the MAE between the synthetic ground truth and the predicted isotropic spectra for samples generated with different parameters. We investigated the effect of (i) the number of peaks in the two-dimensional isotropic spectra, (ii) different MAS dependencies of the linewidths and MAS-dependent shift (first-order, second-order, mixed first- and second-order, or MAS independent), the range of MAS rates generated, the number of MAS spectra used, as well as the amount of noise introduced into the spectra themselves and into the linewidth and shift dependences. Mean absolute errors between predictions and ground-truth isotropic spectra for 1,024 samples of isotropic spectra with various numbers of peaks, MAS dependences (first-order, second-order, combined, or constant), and noise levels are shown in **Figure 3.46B-D**, and some selected examples are shown in **Figure 3.47** (with more examples given in **Figure 3.51**) in order to provide a more visual appreciation of the expected changes in the spectra corresponding to the changes in MAE shown in **Figure 3.46**.

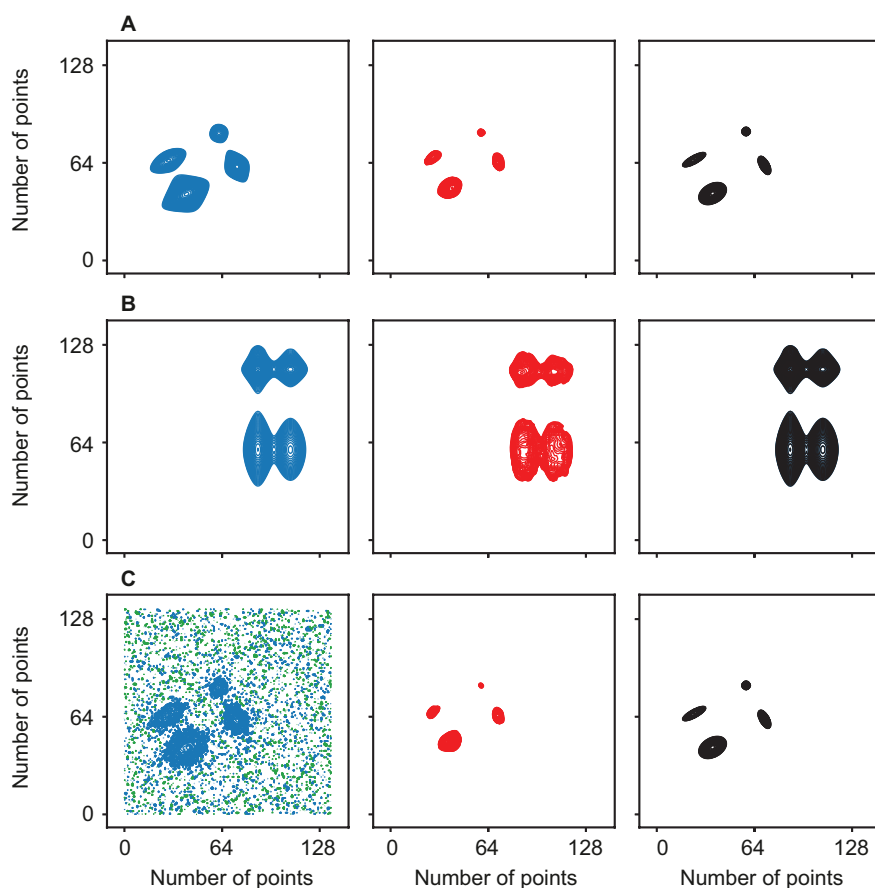


Figure 3.47. (A)-(C) Illustrative comparisons of synthetic highest MAS rate spectra (blue), predicted isotropic (red), and ground-truth isotropic (black) spectra with (A) the example of the synthetic dataset shown in **Figure 3.45**, (B) a MAS independent synthetic dataset, and (C) a synthetic dataset with a high noise level. In this example the spectra are made up of 128 x 128 points, that would correspond to a frequency range of 3 kHz with about 24 Hz/point digital resolution.

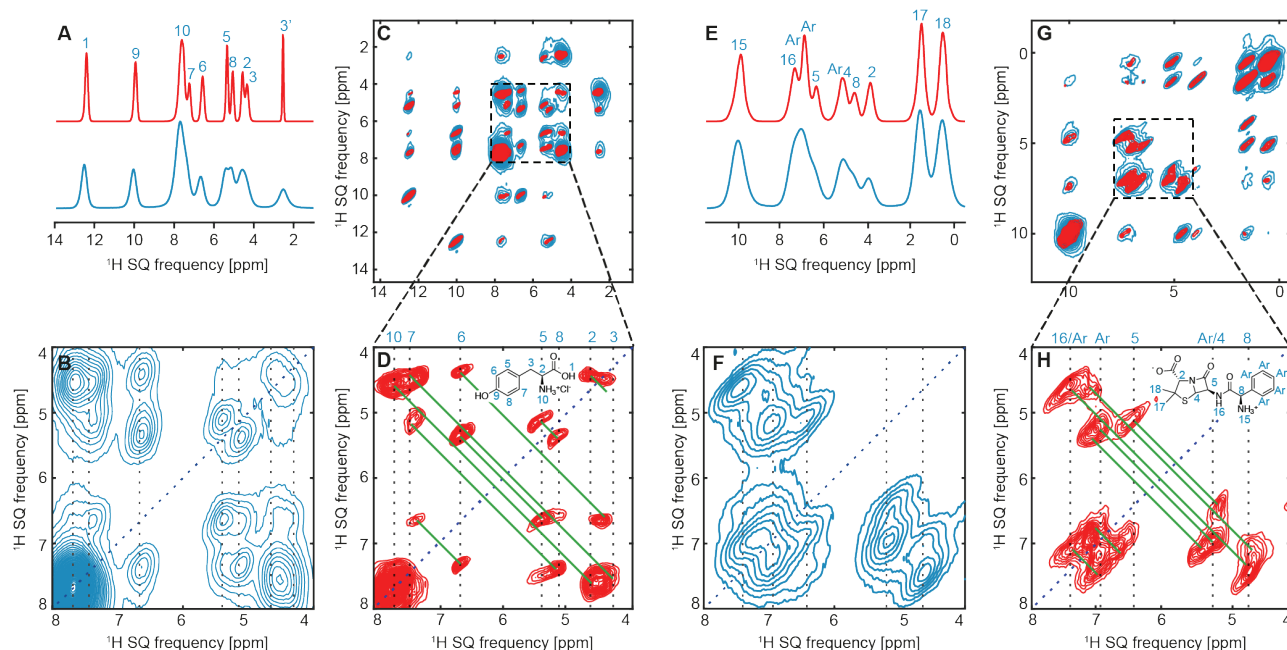


Figure 3.48. Spectra obtained from microcrystalline powdered samples of L-tyrosine hydrochloride (left) and ampicillin (right). (A) and (E) 100 kHz MAS spectra (blue) and isotropic spectra (red) inferred with the PIPNet model⁴⁶³ from a VMAS dataset of 1D spectra recorded at 36 rates between 30 and 100 kHz (reproduced from Ref. 463). (C) and (G) corresponding 100 kHz MAS 2D ^1H - ^1H DQ/SQ BABA spectra (blue) and pure isotropic 2D ^1H - ^1H DQ/SQ BABA spectra (red) inferred with the PIPNet2D model from a VMAS dataset of 11 and 9 2D spectra recorded at MAS rates between 50 and 100 kHz, both after shearing to an SQ/SQ representation, for samples of L-tyrosine hydrochloride and ampicillin, respectively, and acquired with one rotor period of DQ recoupling. (D) and (H) expansions of the pure isotropic 2D spectra, and (C) and (F) expansions of the 100 kHz 2D spectra. In (B), (D), (F) and (H), the vertical dotted lines indicate the previously assigned proton shifts at 100 kHz MAS,^{400, 463} the blue dotted line the diagonal of the spectrum, and the green solid lines the observed double quantum correlations.

Figure 3.48 shows the 1D and 2D isotropic spectra obtained from two experimental sets of variable MAS 1D and 2D spectra recorded on two small organic micro-crystalline samples of L-tyrosine hydrochloride (**Figure 3.52**) and ampicillin (**Figure 3.53**). **Figures 3.48C** and **3.48G** show the performance of the PIPNet2D model on sheared three-dimensional VMAS datasets for both compounds, consisting of two-dimensional BABA spectra recorded at 9 (ampicillin) and 11 (L-tyrosine hydrochloride) rates between 50 and 100 kHz MAS. The sheared SQ/SQ representation is exactly equivalent to the DQ/SQ but gives an easier to visualise rendition of the two-dimensional lineshapes. Full details are given in **Appendix VI**.

In **Figures 3.48C** and **3.48G**, the marked increase in resolution achieved in both dimensions of the pure isotropic 2D spectra, as compared to that obtained in the corresponding spectra at 100 kHz MAS, is clearly visible. This increase is most prominent in the crowded spectral regions between 4 and 8 ppm both for L-tyrosine hydrochloride and ampicillin (expansions of these regions in both the pure isotropic and corresponding 100 kHz MAS 2D spectra are shown in **Figures 3.48B**, **3.48D**, **3.48F**, and **3.48H**).

We note in particular that, as discussed above, the model was not specifically trained to recognise sheared DQ/SQ spectra, so that, for example, the inferred spectra are not constrained to have any particular symmetry. Furthermore, the model can be equally well applied to the unsheared DQ/SQ spectra, and very similar results are obtained as shown in **Figures 3.55** and **3.56**. Rows from the pure isotropic and 100 kHz MAS spectra are also shown in **Figure 3.55** for comparison.

The two-dimensional pure isotropic spectra were found to retain the expected number of peaks from the known assignments, without displaying any significant artifacts or any additional peaks in unexpected regions of the spectra. This is impressive, especially if we consider the reduced quality of the datasets used here as compared to typical 1D MAS spectra. Compared with the one-dimensional data used before, here the 2D spectra have lower signal-to-noise ratios and display t_1 noise, and baseline and cross-peak intensity distortions across the range of MAS rates. (Note for example that since the BABA mixing time is rotor synchronised, the mixing time systematically decreases as the MAS rate increases, which will lead to variations in cross-peak intensities.)

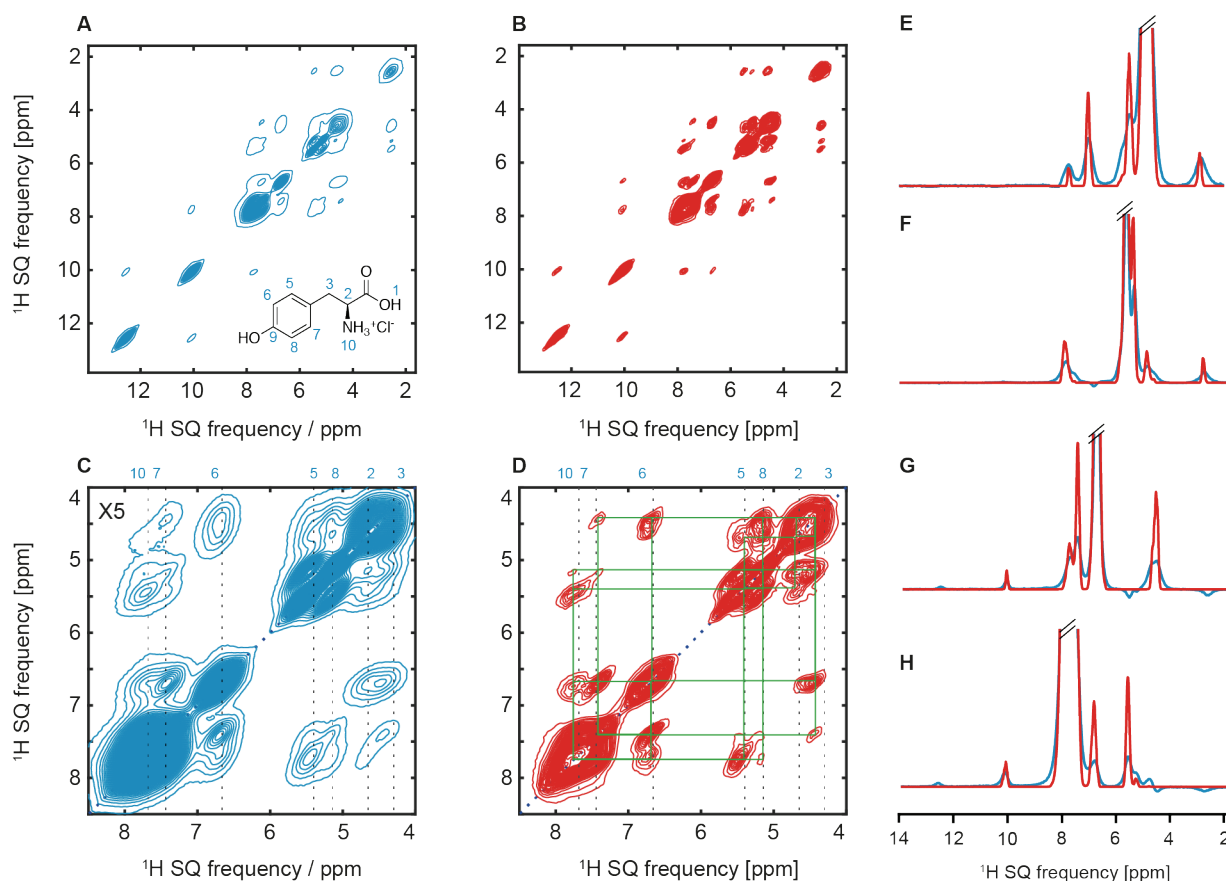


Figure 3.49. Spectra obtained from microcrystalline powdered samples of L-tyrosine hydrochloride. (A) and (B) 100 kHz MAS 2D ^1H - ^1H spin-diffusion spectra (blue) and pure isotropic 2D ^1H - ^1H spin-diffusion spectra (red) inferred with the PIPNet2D model from a VMAS dataset of 6 2D spectra recorded at the MAS rates between 50 and 100 kHz. In the VMAS experiments, the PSD mixing time was varied from 5 to 10 ms as the spinning rate was increased to maintain similar cross-peak intensities across the dataset (as described in Section 3.4.2). (C) and (D) expansions of the 100 kHz and pure isotropic 2D spectra. (E) to (H) horizontal cross sections extracted for F_1 SQ frequencies of 4.5, 5.5, 6.7, and 7.7 ppm. In (C) and (D) the vertical dotted lines indicate the previously assigned proton shifts at 100 kHz MAS,^{400, 463} the blue dotted line the diagonal of the spectrum, and the green squares the observed spin-diffusion correlations.

Another important point is that the isotropic two-dimensional peaks in the inferred spectra seem to retain the lineshape characteristics present in the 100 kHz MAS spectra, arising possibly from inhomogeneous contributions, correlated two-dimensional lineshapes,⁴⁶⁵⁻⁴⁷¹ or magnetic susceptibility effects.^{206, 459, 460, 472, 473} PIPNet2D is therefore not simply identifying potential peaks and replacing them with uniformly narrow shapes. This can be clearly seen for protons H2 and H17/H18 of ampicillin as well as the labile protons of L-tyrosine hydrochloride. Overall, the spectra obtained using the 1D PIPNet model and PIPNet2D were found to be coherent, with good agreement between the 1D isotropic spectra obtained from a set of 1D MAS spectra using PIPNet and the projection of the 2D isotropic spectra obtained here for L-tyrosine hydrochloride and ampicillin, respectively. We do note that the 2D model does not yield the same degree of line narrowing in the projections of Figures 3.57 and 3.58 as compared to the 1D model. While the increased sources of errors discussed above could contribute to the lower level of narrowing achieved here compared to the 1D model, we do not expect them to be the limiting factor since they were explicitly taken into account in the model training. This suggests that the synthetic datasets used here do not yet fully capture the whole complexity of the experimental 2D spectra, and clearly indicates that further progress can be made in the future.

Figure 3.49 shows the performance of the PIPNet2D model on a three-dimensional PSD VMAS dataset for L-tyrosine hydrochloride (Figure 3.54), consisting of two-dimensional ^1H - ^1H spin-diffusion spectra recorded at 6 MAS rates between 50 and 100 kHz MAS. In Figure 3.49D the resolution achieved in both dimensions of the pure isotropic 2D spectrum is evident and allows the clear identification of correlations between, for example, H6 and H10 or H2 and H5, that were difficult to clearly observe in the corresponding PSD spectrum at 100 kHz MAS, shown in Figure 3.49C.

The horizontal cross sections shown in **Figure 3.49E-H** are an additional direct illustration of the enhanced resolution of the pure isotropic 2D spectrum over the corresponding 100 kHz MAS 2D spectrum. PIPNet2D is expected to perform best for the cross-peaks, as compared to the diagonal peaks, since the integral normalisation of the 3D dataset was done with respect to a selected cross peak intensity.

In the VMAS experiments, the PSD mixing time was varied from 5 to 10 ms as the spinning rate was increased in order to compensate for slower spin diffusion at faster MAS rates and to maintain similar cross-peak intensities across the dataset (as described in **Section 3.4.2**). Since the spin diffusion rates between different spin pairs will have slightly different MAS rate dependencies,⁴⁷⁴ this procedure can-not be perfect and will introduce a source of error.

That the model can be successfully directly applied to DQ/SQ (whether sheared or not) or to PSD datasets illustrates that the PIPNet2D model is quite robust and can be used quite generally to obtain two-dimensional ^1H - ^1H correlation spectra with higher resolution in both dimensions from 3D VMAS datasets, and that it is not restricted to inferences for a specific type of two-dimensional spectra.

Across the three sets of 2D spectra shown here, PIPNet2D reduced observed linewidths by a factor of 3.33 ± 0.10 in both dimensions compared to 100 kHz MAS spectra (see **Table 3.30**).

3.4.4 Conclusion

In this section we have introduced PIPNet2D, a deep learning model to increase resolution in two-dimensional NMR spectroscopy by predicting pure isotropic two-dimensional correlation spectra of solids from three-dimensional datasets of 2D spectra acquired at variable MAS rates. We have illustrated the method by obtaining isotropic spectra from experimental datasets on two different microcrystalline organic solids. The resolution obtained is very significantly improved compared with the 100 kHz MAS spectra. The residual linewidths or the quantitative character of the inferred spectra (**Figure 3.56**) can in principle be limited by several factors. Some are intrinsic to the samples, such as structural disorder or magnetic susceptibility broadening, and others might be due to experimental imperfections such as systematic noise or cross-peak intensity variations, MAS instabilities or poor shimming, or limitations in the model, such as incomplete descriptions of the lineshape and MAS-dependence. All these factors will be the subject of future study.

For example, we expect that the use of more robust pulse sequences for the DQ/SQ type experiments, that might better remove some of the experimental imperfections, should potentially improve the robustness of the model.^{475, 476} Further improved results might also be obtained by training models specifically on a given type of correlation experiment.

In conclusion, the model presented here provides significant improvement in the resolution of 2D ^1H - ^1H DQ/SQ and spin-diffusion spectra, and we expect that the approach can be used to develop models for other two-dimensional correlation experiments in the future.

3.4.5 Appendix VI

The NMR raw data are available from <https://doi.org/10.24435/materialscloud:xj-5f> in JCAMP-DX version 6.0 standard format and original TopSpin format, as well as an archived version of the code and the pre-trained model, used to obtain the results presented in this work. The code and pre-trained model are also available in the GitHub repository <https://github.com/manucordova/PIPNet>. All data and code are available under the license CC-BY-4.0 (Creative Commons Attribution-ShareAlike 4.0 International). The exact pulse sequences and full parameter sets used are available with the raw data.

Table 3.26. Experimental details of the BABA VMAS 3D dataset acquired for L-tyrosine hydrochloride.

L-tyrosine hydrochloride	VT (K)	90° RF amplitude (kHz)	d1 (s)	Number of co-added transients	Number of FID points: F2/F1	SW (kHz): F2/F1	Size of real spectrum: F2/F1	DQ recoupling time (μs)	Experimental time
100 kHz	285	294	2	16	4096/100	90.9/20	4096/1024	10	57 min
96 kHz	285	294	2	16	4096/96	90.9/19.2	4096/1024	10.42	54 min
94 kHz	285	294	2	16	4096/94	90.9/18.8	4096/1024	10.64	54 min
90 kHz	285	294	2	16	4096/90	90.9/18	4096/1024	11.11	51 min
88 kHz	285	294	2	16	4096/110	90.9/22	4096/1024	11.36	1h2min
80 kHz	290	294	2	16	4096/100	90.9/20	4096/1024	12.5	57 min
78 kHz	290	294	2	16	4096/130	90.9/26	4096/1024	12.82	1h14min
72 kHz	290	294	2	16	4096/120	90.9/24	4096/1024	13.89	1h8min
66 kHz	290	294	2	16	4096/110	90.9/22	4096/1024	15.15	1h2min
60 kHz	295	294	2	16	4096/100	90.9/20	4096/1024	16.67	57 min
52 kHz	295	294	2	16	4096/130	90.9/26	4096/1024	19.23	1h14min

Table 3.27. Experimental details of the BABA VMAS 3D dataset acquired for ampicillin.

Ampicillin	VT (K)	90° RF amplitude (kHz)	d1(s)	Number of co-added transients	Number of FID points: F2/F1	SW (kHz): F2/F1	Size of real spectrum: F2/F1	DQ recoupling time (μs)	Experimental time
100 kHz	285	294	2	16	4096/160	90.9/33.3	4096/1024	10	1h31min
90 kHz	285	294	2	16	4096/144	90.9/30	4096/1024	11.1	1h22min
85 kHz	290	294	2	16	4096/136	90.9/28.3	4096/1024	11.76	1h17min
80 kHz	290	294	2	16	4096/128	90.9/26.6	4096/1024	12.5	1h13min
75 kHz	290	294	2	16	4096/120	90.9/25	4096/1024	13.3	1h2min
70 kHz	290	294	2	16	4096/168	90.9/35	4096/1024	14.28	1h35 min
60 kHz	290	294	2	16	4096/144	90.9/30	4096/1024	16.67	1h22min
55 kHz	290	294	2	16	4096/132	90.9/27.5	4096/1024	18.18	1h15min
50 kHz	290	294	2	16	4096/120	90.9/25	4096/1024	20	1h8min

Table 3.28. Experimental details of the PSD VMAS 3D dataset acquired for L-tyrosine hydrochloride.

L-tyrosine hydrochloride	VT (K)	90° RF amplitude (kHz)	d1(s)	Number of co-added transients	Number of FID points: F2/F1	SW (kHz): F2/F1	Size of real spectrum: F2/F1	DQ recoupling time (ms)	Experimental time
100 kHz	285	294	3	8	8192/366	227.3/12.5	16384/1024	10	2h31 min
90 kHz	290	294	3	8	8192/366	227.3/12.5	16384/1024	9	2h31 min
80 kHz	290	294	3	8	8192/366	227.3/12.5	16384/1024	6	2h31 min
70 kHz	290	294	3	8	8192/366	227.3/12.5	16384/1024	6	2h31 min
60 kHz	290	294	3	8	8192/366	227.3/12.5	16384/1024	6	2h31min
50 kHz	290	294	3	8	8192/366	227.3/12.5	16384/1024	5	2h31 min

Table 3.29. Model and training parameters for PIPNet2D.

Parameter	Value
Number of Conv-LSTM layers	4
Number of CNN filters (channels) per Conv-LSTM layer	64
Kernel size of Conv-LSTM layers	5
Number of filters (channels) for the output CNN	1
Kernel size of output CNN layer	5
Batch size	8
Number of training batches per epoch	500
Number of evaluation batches per epoch	100
Number of epochs	250
Optimiser	Adam
Initial learning rate	10^{-3}
Learning rate scheduler	Reduction on plateau of the evaluation loss by a factor 0.5, with patience of 10 epochs

Table 3.30. Full width at half maximum (FWHM) values of selected peaks in 100 kHz MAS and isotropic (PIP) spectra. The values were obtained by fitting Gaussian functions to selected rows or columns in the spectra .

L-tyrosine hydrochloride BABA				Ampicillin BABA				L-tyrosine hydrochloride PSD			
F1 (ppm)	F2 (ppm)	FWHM F1 100kHz / PIP (ppm)	FWHM F2 100kHz / PIP (ppm)	F1 (ppm)	F2 (ppm)	FWHM F1 100kHz / PIP (ppm)	FWHM F2 100kHz / PIP (ppm)	F1 (ppm)	F2 (ppm)	FWHM F1 100kHz / PIP (ppm)	FWHM F2 100kHz / PIP (ppm)
2.5	12.5	0.56/0.09	0.36/0.14	1.8	10.2	0.52/0.11	0.48/0.09	10.0	12.5	0.19/0.09	0.18/0.10
6.6	12.5	0.34/0.08	0.27/0.10	10.1	10.1	0.79/0.43	0.58/0.30	12.4	9.9	0.17/0.09	0.18/0.07
10.0	12.4	0.35/0.19	0.28/0.18	4.8	10.1	0.61/0.27	0.52/0.24	7.6	9.9	0.35/0.17	0.28/0.14
5.1	12.4	0.47/0.20	0.28/0.24	7.5	10.0	0.57/0.29	0.44/0.21	7.3	9.9	0.33/0.10	0.26/0.08
4.4	12.3	0.43/0.16	0.29/0.09	10.1	7.5	0.67/0.16	0.48/0.14	6.5	9.9	0.25/0.06	0.23/0.07
2.4	12.3	0.36/0.08	0.31/0.12	4.6	7.4	0.54/0.24	0.50/0.31	10.0	7.7	0.26/0.12	0.37/0.14
6.6	9.9	0.49/0.11	0.29/0.19	6.9	7.1	0.76/0.23	1.22/0.35	5.4	7.6	0.31/0.12	0.37/0.14
5.1	9.9	0.39/0.11	0.30/0.19	0.6	7.1	0.86/0.17	0.72/0.17	6.6	7.6	0.38/0.14	0.53/0.17
4.4	9.9	0.44/0.13	0.31/0.17	5.2	7.0	0.74/0.31	0.64/0.26	5.0	7.6	0.50/0.11	0.48/0.15
7.5	9.9	0.80/0.10	0.32/0.29	4.0	6.5	0.47/0.14	0.44/0.13	4.6	7.6	0.39/0.06	0.47/0.12
2.4	9.9	0.76/0.30	0.18/0.06	5.2	6.5	0.50/0.23	0.61/0.25	5.1	7.3	0.31/0.11	0.35/0.10
12.3	9.9	0.40/0.07	0.28/0.17	1.7	6.5	0.42/0.12	0.39/0.13	6.6	7.3	0.29/0.12	0.34/0.12
9.9	9.9	0.36/0.15	0.26/0.05	7.0	5.2	0.75/0.29	0.51/0.26	4.3	7.3	0.35/0.08	0.33/0.07
12.3	7.6	0.25/0.06	0.45/0.10	0.6	5.2	0.49/0.17	0.41/0.21	9.9	7.3	0.23/0.07	0.40/0.07
7.6	7.6	0.51/0.17	0.46/0.33	1.5	5.0	0.40/0.15	0.38/0.12	7.6	6.6	0.56/0.12	0.39/0.14
4.4	7.6	0.73/0.41	0.51/0.36	10.1	4.7	0.54/0.19	0.43/0.16	4.4	6.6	0.58/0.17	0.34/0.12
2.4	7.6	0.59/0.11	0.47/0.17	7.4	4.6	0.61/0.38	0.45/0.17	7.3	6.6	0.29/0.12	0.28/0.13
10.0	7.5	0.52/0.12	0.60/0.24	1.8	4.2	0.34/0.19	0.30/0.14	9.9	6.5	0.20/0.05	0.23/0.05
6.5	7.2	0.53/0.11	0.46/0.17	10.1	4.1	0.45/0.18	0.54/0.12	2.6	5.5	0.34/0.08	0.32/0.05
4.9	7.2	1.23/0.22	0.55/0.19	6.4	4.1	0.38/0.14	0.42/0.14	7.6	5.4	0.35/0.15	0.29/0.10
4.4	7.2	0.70/0.19	0.64/0.34	0.6	4.0	0.32/0.11	0.25/0.10	4.6	5.4	0.48/0.12	0.45/0.12
10.0	6.7	0.45/0.12	0.35/0.10	1.4	3.8	0.33/0.22	0.29/0.21	2.5	5.1	0.42/0.10	0.77/0.11
12.3	6.6	0.39/0.09	0.29/0.07	7.1	1.7	0.49/0.18	0.51/0.08	7.3	5.1	0.29/0.09	0.36/0.09
7.2	6.6	0.48/0.12	0.38/0.11	10.0	1.7	0.53/0.15	0.38/0.11	4.5	5.1	0.83/0.35	0.61/0.20
4.3	6.6	0.54/0.09	0.39/0.15	5.2	1.7	0.41/0.22	0.34/0.17	7.6	4.6	0.39/0.06	0.29/0.05
2.4	6.6	0.59/0.07	0.34/0.08	4.0	1.7	0.34/0.22	0.31/0.14	5.3	4.6	0.41/0.18	0.43/0.16
5.2	6.6	0.59/0.16	0.35/0.18	1.6	1.7	0.46/0.28	0.40/0.25	2.5	4.5	0.45/0.13	0.53/0.14
5.1	5.4	0.50/0.13	0.35/0.12	0.6	1.7	0.52/0.25	0.39/0.22	5.1	4.4	0.40/0.16	0.58/0.22
2.5	5.4	0.50/0.15	0.36/0.14	3.8	1.4	0.33/0.19	0.31/0.18	6.6	4.3	0.30/0.12	0.44/0.17
10.0	5.3	0.57/0.11	0.52/0.13	1.6	0.7	0.60/0.25	0.49/0.27	7.3	4.3	0.28/0.06	0.30/0.07
4.3	5.3	0.60/0.07	0.49/0.16	0.6	0.7	0.61/0.35	0.48/0.29	5.4	2.5	0.29/0.07	0.32/0.09
6.5	5.2	0.53/0.13	0.47/0.23	2.7	0.6	0.77/0.16	0.45/0.14	5.1	2.5	0.38/0.08	0.35/0.10
12.4	5.2	0.50/0.13	0.76/0.14	7.0	0.6	0.66/0.14	0.44/0.17	4.5	2.5	0.52/0.12	0.42/0.14
5.3	5.1	0.70/0.15	0.62/0.15	5.1	0.6	0.52/0.27	0.41/0.21				
7.3	5.1	0.53/0.13	0.55/0.19								
4.3	4.5	0.67/0.11	0.56/0.29								
7.5	4.5	0.90/0.30	0.56/0.31								
12.2	4.4	0.53/0.08	0.84/0.08								
2.4	4.4	0.68/0.22	0.60/0.33								
6.5	4.4	0.71/0.13	0.55/0.22								
12.4	2.5	0.45/0.10	0.51/0.07								
4.4	2.5	0.67/0.22	0.45/0.23								
7.5	2.4	0.72/0.12	0.40/0.17								
5.3	2.4	0.57/0.14	0.38/0.17								

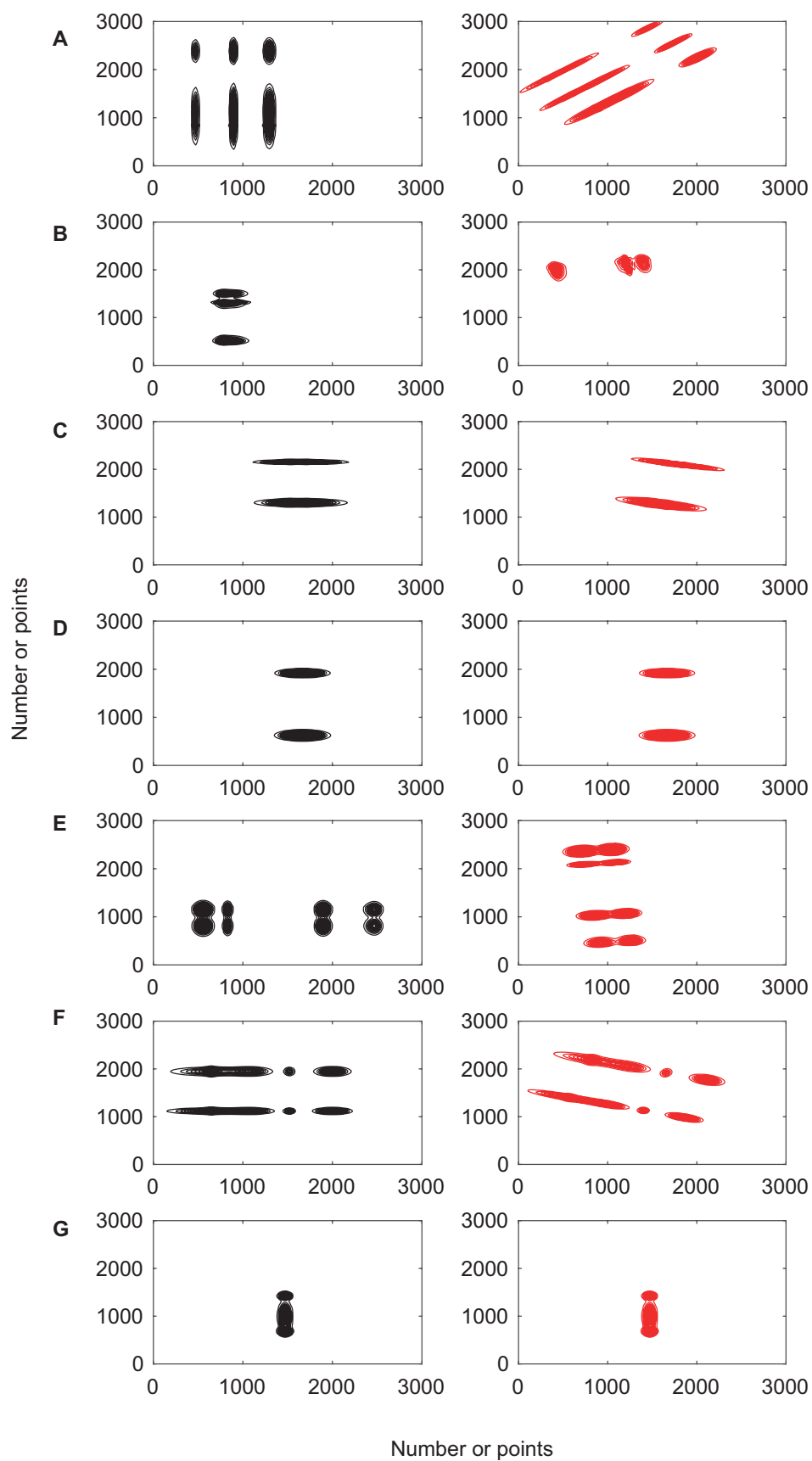


Figure 3.50. Representative examples of synthetic isotropic two-dimensional spectra (A) before (black) and (B) after (red) rotation. The rotation angles applied here during the data generation process were 47°, 80.7°, 12.8°, 0°, 21.3°, 17.9°, and 0°.

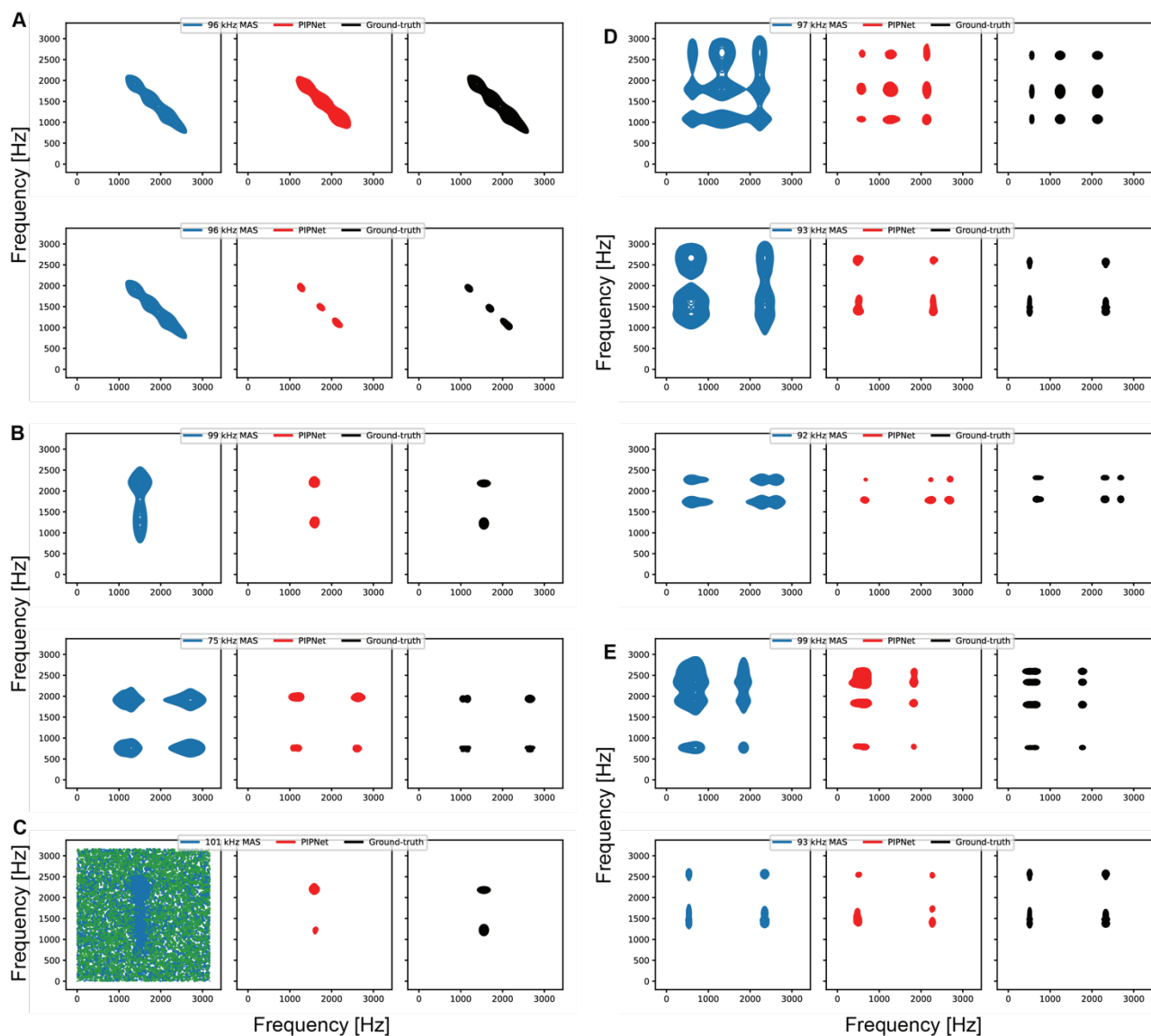


Figure 3.51. Representative sets of examples of synthetic isotropic two-dimensional spectra where the fastest 2D spectrum used in the series is indicated in blue, the PIPNet2D inferred spectrum is shown in red, and the ground truth is shown in black. **(A)** Example of synthetic data without (top) and with (bottom) MAS-dependent broadening. **(B)** Example of synthetic data in which the highest MAS 2D spectrum used in the series is 99 kHz (top) and 75 kHz (bottom). **(C)** Example of synthetic data with a high level of noise. **(D)** Example of synthetic data in which 16 (top), 12 (middle), and 6 (bottom) 2D spectra are used in the series. **(E)** Example of synthetic data with a first order MAS-dependence (top) and a second order MAS-dependence (bottom).

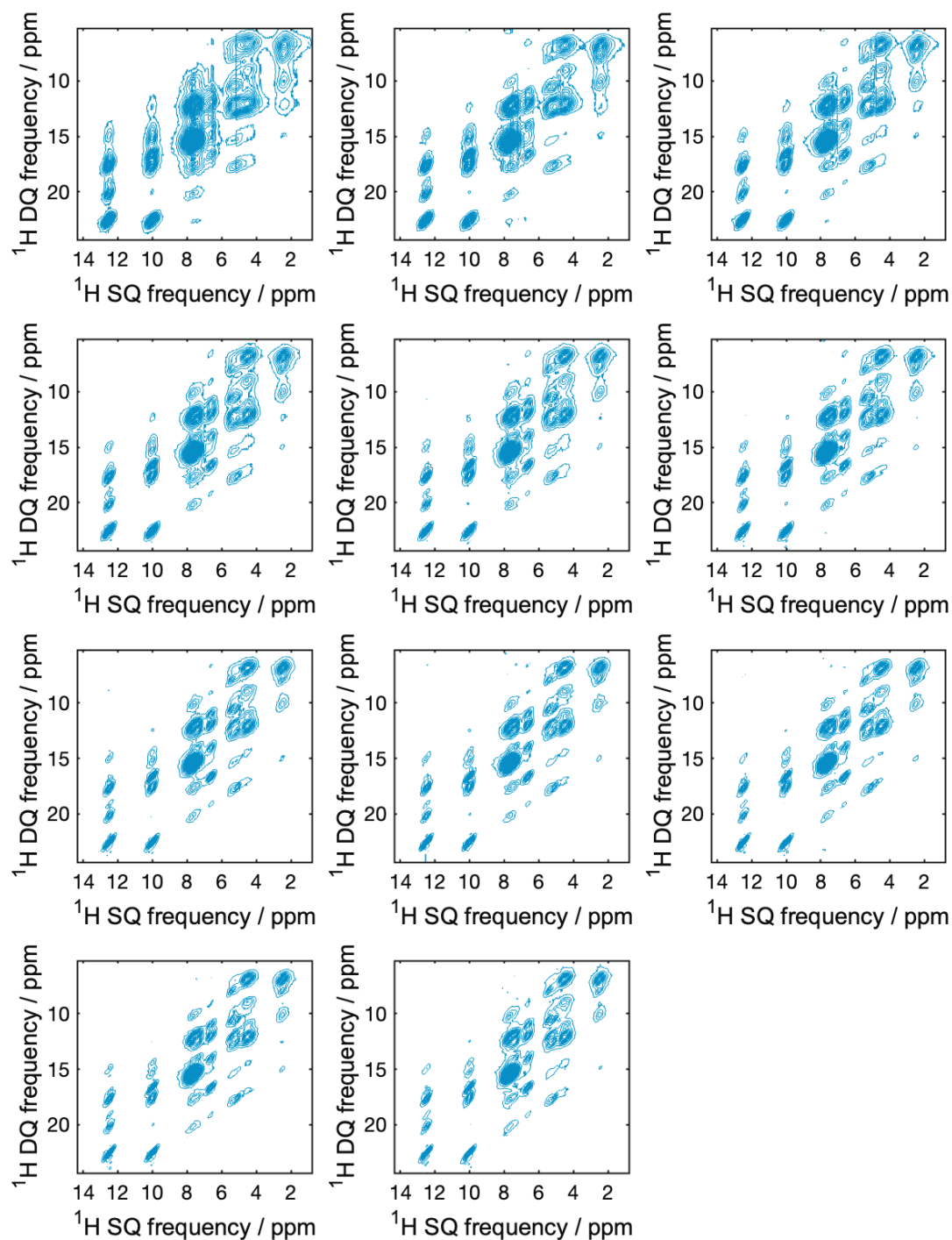


Figure 3. 52. 3D dataset of L-tyrosine hydrochloride consisting of 11 unsheared 2D ^1H - ^1H DQ/SQ BABA spectra acquired at MAS rates of 50, 62, 66, 72, 78, 80, 88, 90, 94, 96, and 100 kHz (top left to bottom right).

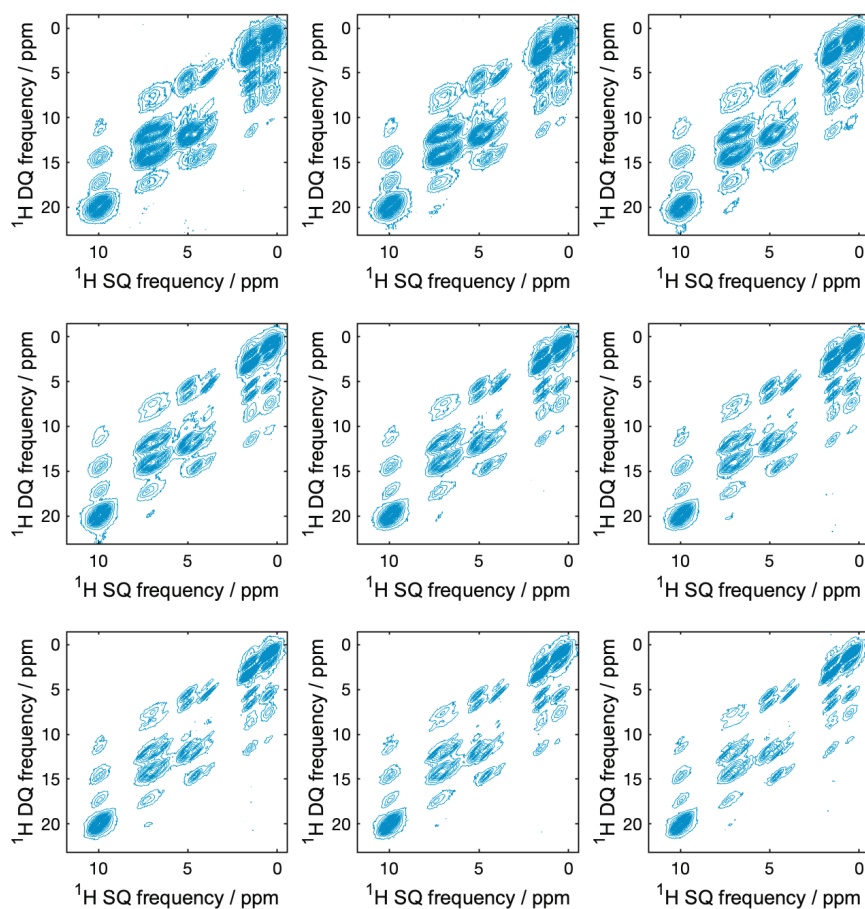


Figure 3.53. 3D dataset of ampicillin consisting of 9 unsheared 2D ^1H - ^1H DQ/SQ BABA spectra acquired at MAS rates of 50, 55, 60, 70, 75, 80, 85, 90, and 100 kHz (top left to bottom right).

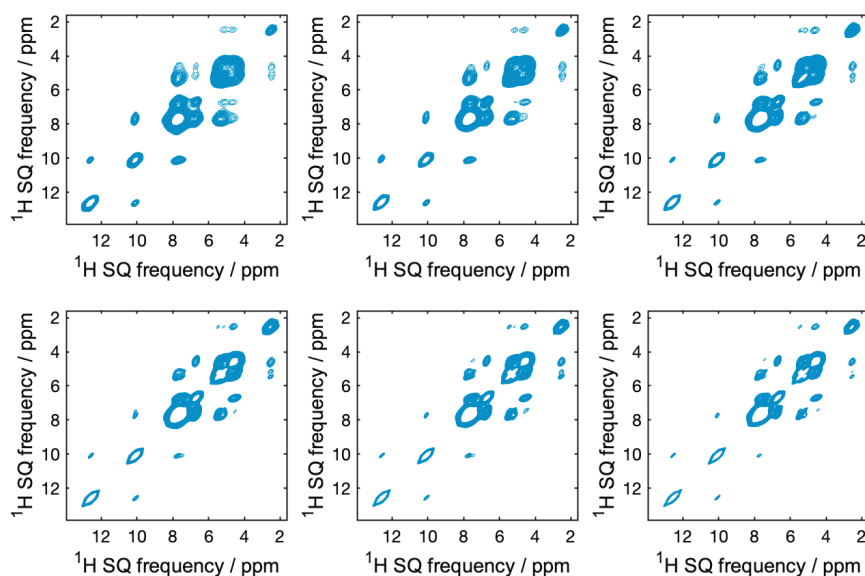


Figure 3.54. 3D dataset of L-tyrosine hydrochloride consisting of 6 unsheared 2D ^1H - ^1H spin-diffusion spectra acquired at MAS rates of 50, 60, 70, 80, 90, and 100 kHz (top left to bottom right).

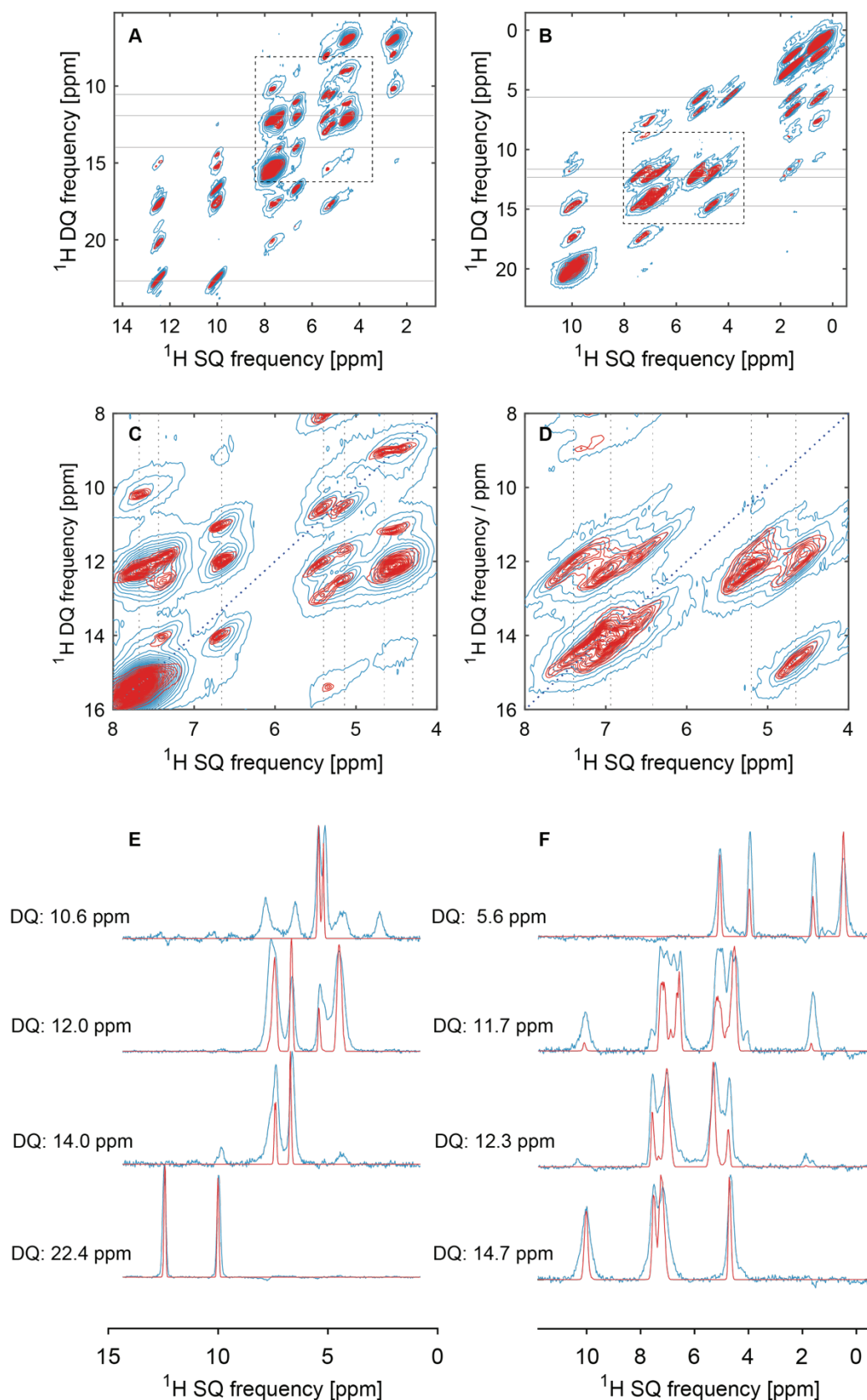


Figure 3.55. (A) and (B) 100 kHz MAS 2D ^1H - ^1H DQ/SQ BABA spectra (blue) and pure isotropic 2D ^1H - ^1H DQ/SQ BABA spectra (red) inferred with the PINet2D model from a VMAS dataset of 11 and 9 2D spectra recorded at the MAS rates between 50 and 100 kHz, both before shearing to an SQ/SQ representation, for samples of L-tyrosine hydrochloride and ampicillin, respectively. (C) and (D) expansions of the pure isotropic 2D spectra and 100 kHz 2D spectra. In (C) and (D) the vertical dotted lines indicate the previously assigned proton shifts at 100 kHz MAS and the blue dotted line the diagonal of the spectrum. In (A) and (B) the horizontal lines indicate the cross sections plotted in (E) and (F).

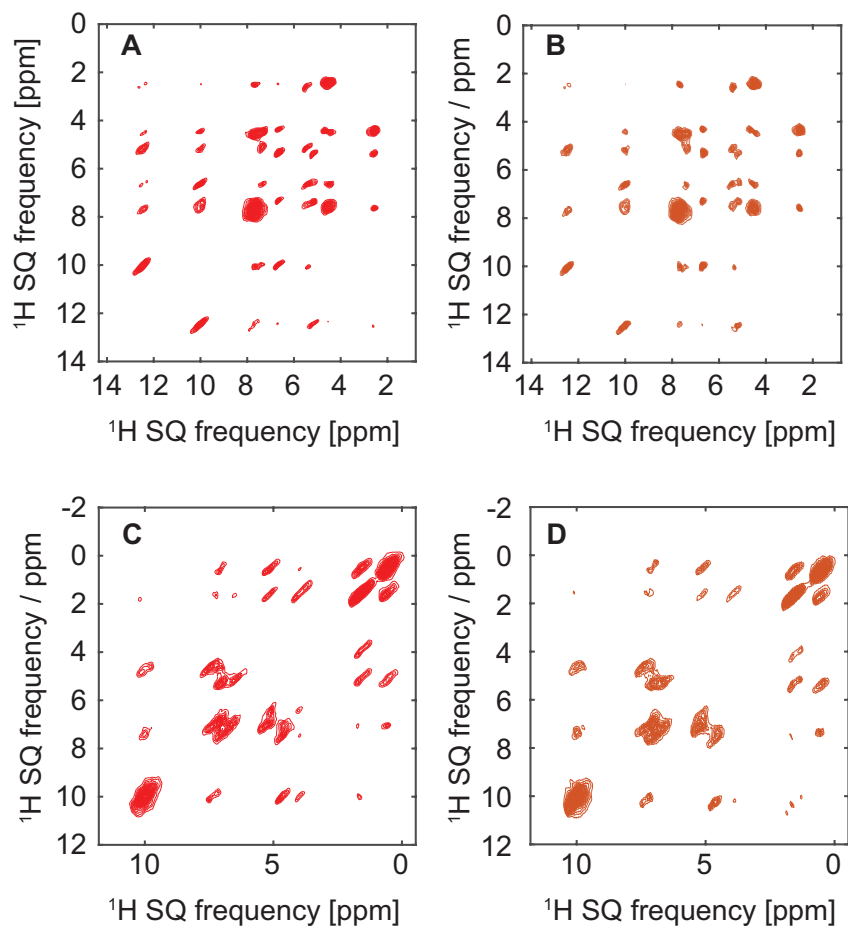


Figure 3.56. Pure isotropic 2D ^1H - ^1H DQ/SQ BABA spectra inferred with the PINet2D model from a VMAS dataset of 11 and 9 2D spectra recorded at the MAS rates between 50 and 100 kHz, (A) and (C) after shearing to an SQ/SQ representation and (B) and (D) before shearing to an SQ/SQ representation, for samples of L-tyrosine hydrochloride and ampicillin, respectively. (B) and (D) are sheared to an SQ/SQ representation after inference.

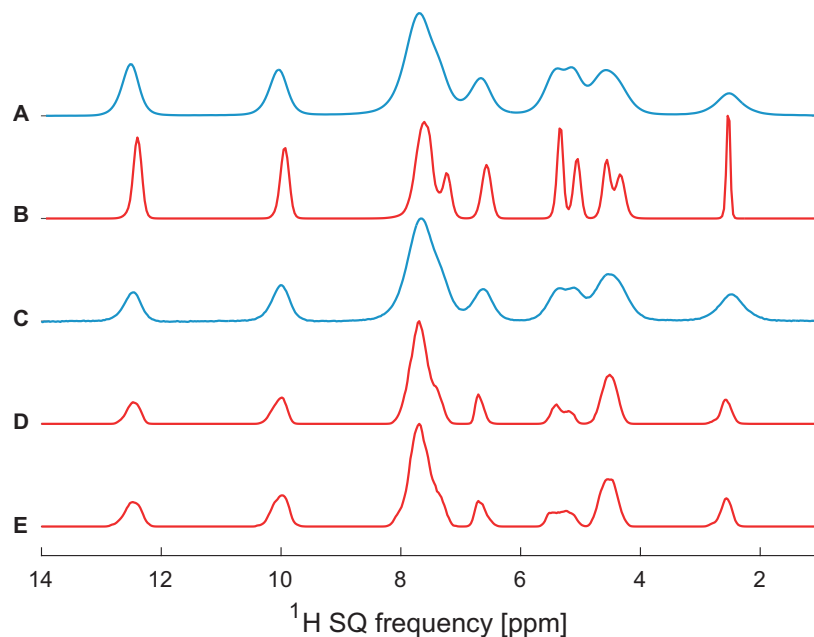


Figure 3.57. Spectra obtained from microcrystalline powdered samples of L-tyrosine hydrochloride. (A) and (B) 100 kHz MAS spectra (blue) and isotropic spectra (red) inferred with the PINet model⁴⁶³ from a VMAS dataset of 1D spectra recorded at 36 rates between 30 and 100 kHz. (C) Sum projection along F_1 of the unsheared 100 kHz MAS 2D ^1H - ^1H DQ/SQ BABA spectrum (blue), and sum projections along F_1 of the pure isotropic 2D ^1H - ^1H DQ/SQ BABA spectra inferred with the PINet2D model from a VMAS dataset of 11 2D spectra recorded at the MAS rates between 50 and 100 kHz, (D) before and (E) after shearing to an SQ/SQ representation.

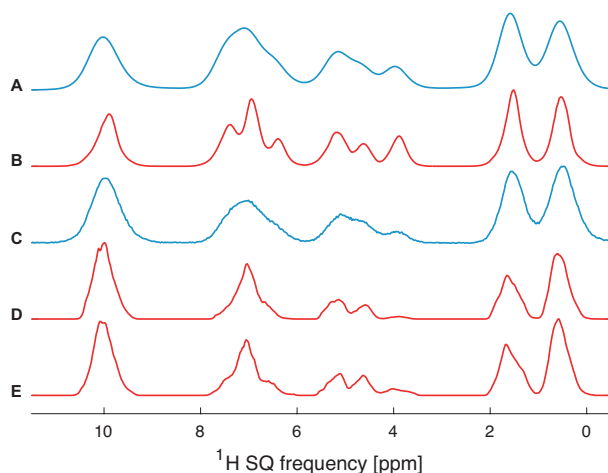


Figure 3.58. Spectra obtained from microcrystalline powdered samples of ampicillin. (A) and (B) 100 kHz MAS spectra (blue) and isotropic spectra (red) inferred with the PIPNet model⁴⁶³ from a VMAS dataset of 1D spectra recorded at 36 rates between 30 and 100 kHz. (C) Sum projection along F_1 of the unsheared 100 kHz MAS 2D ^1H - ^1H DQ/SQ BABA spectrum (blue), and sum projections along F_1 of the pure isotropic 2D ^1H - ^1H DQ/SQ BABA spectra inferred with the PIPNet2D model from a VMAS dataset of 11 2D spectra recorded at the MAS rates between 50 and 100 kHz, (D) before and (E) after shearing to an SQ/SQ representation.

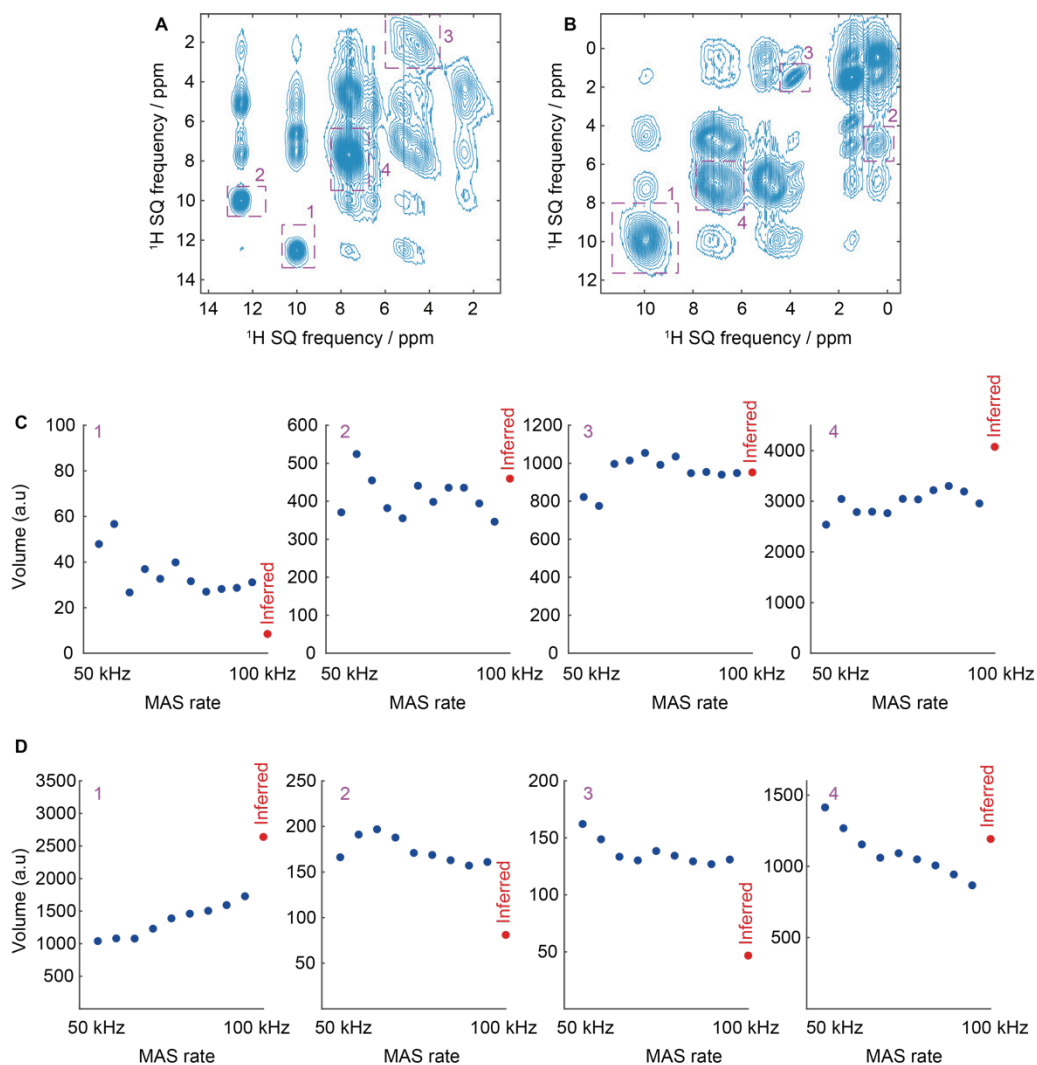


Figure 3.59. (A) and (B) 50 kHz MAS 2D ^1H - ^1H DQ/SQ BABA spectra after shearing to an SQ/SQ representation, for samples of L-tyrosine hydrochloride and ampicillin. (C) and (D) 2D peak volumes measured from the 2D ^1H - ^1H DQ/SQ BABA spectra as a function of the MAS rate (blue) and for the inferred pure isotropic 2D spectra (red). In (A) and (B) the 2D peaks chosen for volume measurements are indicated with purple dotted boxes.

Chapter 4 NMR crystallography of amorphous solids

4.1 Introduction

As mentioned in **Chapter 1**, structure-activity relations drive most areas of modern chemistry. For example, the design of efficient and safe pharmaceutical drugs can be rationalised through the understanding of their atomic-level structure. This can greatly accelerate the search for new compounds with specific properties.⁴⁷⁷⁻⁴⁷⁹ Tools to determine atomic-level structures have thus become a vital part of modern chemistry research. Whereas X-Ray diffraction (XRD) is the established gold standard when single crystals are available, atomic-level structure determination is much more challenging in powder samples, and even more so if the structures are disordered. Indeed, while there has been much progress towards complete structure determination in crystalline powders, by powder XRD^{480, 481} or particularly from solid-state nuclear magnetic resonance (NMR) approaches,^{53, 73, 308} the disorder inherent to amorphous solids makes structure determination elusive.

For example, the structure, accessible surface area, and stability of amorphous drug formulations are of high current interest,⁴⁸²⁻⁴⁸⁷ particularly because the bioavailability and/or dissolution rate of poorly soluble compounds in crystalline forms is often a severe limitation on the chemical space available for development of active pharmaceutical ingredients (APIs), and because the uptake of poorly soluble drugs can be significantly enhanced in amorphous formulations. In particular hydrated amorphous phases, wherein water molecules closely interact with the drug through hydrogen bonds, have been investigated for several systems of pharmaceutical interest.⁴⁸⁸⁻⁴⁹⁰ However, in the absence of methods for atomic-level structure determination, it is not possible to rationalise the factors that lead to the stabilisation of amorphous forms, which is a crucial step in developing stable formulations.

Solid-state NMR is among the most popular methods to study the structure of amorphous materials. While two-dimensional correlation experiments are able to identify intermolecular contacts between atom pairs,^{56, 382, 491} obtaining complete atomic-level structures is a challenge due to the disordered molecular environments present in amorphous solids. In particular, disorder leads to broadening of NMR signals, which results in significant overlap between the peaks associated to different atomic sites. Consequently, this increases the need for multidimensional experiments, which are more difficult to obtain than for crystalline materials due to the lower sensitivity associated with broader lineshapes. The assignment of chemical shifts for amorphous compounds is thus often challenging. Recent advances in dynamic nuclear polarisation (DNP)^{413, 491, 492} have resulted in remarkable gains of sensitivity in crystalline and amorphous molecular solids, leading to a significant reduction in experimental time required to obtain multidimensional NMR spectra of solids.

In addition to these experimental considerations, modelling amorphous structures of materials generally requires the use of molecular dynamics (MD) simulations of large cells typically containing hundreds of molecules. This results in a prohibitive cost for computing chemical shifts using DFT for such large systems. Several approaches have been introduced in order to circumvent this drawback, some of which consist in using small (hundreds of atoms) amorphous system sizes,^{66, 67, 74, 180, 493} isolating local environments for chemical shift computation,^{60, 165, 494, 495} or including the effect of long-range interactions by approximate methods.^{127, 128, 496-498} While these methods do enable the computation of chemical shifts at the DFT level of theory for amorphous solids, the computational cost remains significant, and prevents large-scale chemical shift computations.

Structural disorder has been investigated in proteins by a combination of solid-state NMR, structure generation algorithms and chemical shift predictions.⁴⁹⁹⁻⁵⁰¹ However, such studies have relied on models of chemical shifts in proteins based in part on the primary and/or secondary structure.^{248, 251, 252, 502} Such models are thus not directly applicable to other molecular solids. However, ShiftML provides a method to perform large-scale chemical shift computations,^{176, 261, 365} allowing the direct comparison between large ensembles of MD structures and NMR experiments measured for amorphous samples.

In **Section 4.2**, we determine the atomic-level structure of the hydrated amorphous drug AZD5718 (Atuliflapon) by combining dynamic nuclear polarisation-enhanced solid-state NMR experiments with predicted chemical shifts for MD simulations of large systems. From these amorphous structures we then identify H-bonding motifs and relate them to local intermolecular complex formation energies.

As a proof of concept, the approach presented in **Section 4.2** uses a single chemical shift to focus on the determination of the hydrogen bonding motifs in the structure. In **Section 4.3**, we extend the generality of this method, and we determine the atomic-level ensemble structure of the amorphous form of the drug AZD4625 by combining solid-state NMR experiments with molecular dynamics (MD) simulations and machine-learned chemical shifts. By considering the combined shifts of all ^1H and ^{13}C atomic sites in the molecule, we determine the structure of the amorphous form by identifying an ensemble of local molecular environments that are in agreement with the experimental observations. We then extract preferred conformations and intermolecular interactions in the amorphous sample, and examine the structure in terms of the hydrogen bonding and conformational factors that stabilise the amorphous form of the drug.

4.2 Structure determination of an amorphous drug through large-scale NMR predictions

This section has been adapted with permission from: Cordova, M.; Balodis, M.; Hofstetter, A.; Paruzzo, F.; Lill, S. O. N.; Eriksson, E. S. E.; Berruyer, P.; Simões de Almeida, B.; Quayle, M. J.; Norberg, S. T.; Svensk Ankarberg, A.; Schantz, S.; Emsley, L., Structure determination of an amorphous drug through large-scale NMR predictions. *Nature Communications* **2021**, *12* (1), 2964. (post-print)

My contribution was to develop and apply the method and to analyse the results obtained. I also wrote the manuscript, with contributions of all other authors.

4.2.1 Introduction

AZD5718 (Atuliflapon) is a 5-lipoxygenase activating protein (FLAP) inhibitor that was found promising for the treatment of diseases involving chronic inflammation, such as asthma.^{503, 504} In this section, we investigate the structure of anhydrous crystalline AZD5718^{503, 504} form A, by combining measured ^1H , ^{13}C and ^{15}N chemical shifts obtained using DNP-enhanced NMR experiments from a powder sample, CSP, and DFT chemical shift computations. The structure is validated with that obtained from single-crystal XRD. We then model the hydrated amorphous drug with different water contents using MD simulations and obtain predicted NMR spectra for large structural ensembles using machine learned chemical shifts. We then analyse the ensembles to identify the different hydrogen bonding motifs present in the amorphous structures, by comparing the experimental and predicted chemical shift distributions associated with each structural motif. From the amorphous structures we also compute the interaction energy between AZD5718 molecules and their environment, and we relate the energies to the local hydrogen bonding motifs.

4.2.2 Methods

NMR experiments. Both crystalline and amorphous forms of AZD5718 were provided by AstraZeneca. The samples were stored at equilibrium with the environment at approximately 22°C and 20% relative humidity prior to NMR analysis. The room temperature NMR experiments were performed on Bruker Ascend 500 wide-bore Avance III, Bruker 800 Ultrashield plus narrow-bore and 900 US² wide-bore Avance Neo NMR spectrometers. DNP-enhanced solid-state NMR experiments were performed on a 400 MHz Avance III HD Bruker spectrometer. The spectrometer is equipped with a low temperature magic angle spinning (LTMAS) 3.2 mm probe and is connected through a corrugated waveguide to a 263 GHz gyrotron capable of outputting ca. 5-10 W of continuous wave microwaves. All chemical shifts were referenced to alanine. For more details including experimental setup and the sample preparation see **Appendix VII**.

NMR crystallography. The candidate crystal structures were generated using a Monte-Carlo parallel tempering method⁵⁰⁵ followed by lattice energy minimisation using an internally developed force-field. The 190 most stable candidates were selected for full DFT-D optimisation at the PBE level of theory. Chemical shifts for the ten lowest energy candidates were computed at the PBE0 level of theory using the fragment- and cluster-based approach developed by Hartman *et al.*¹²⁶⁻¹²⁸ The conversion from isotropic shielding to chemical shift was performed by linear regression between the obtained shieldings and experimental isotropic chemical shifts for each candidate. The analysis of the positional uncertainty of the crystal structure was performed as described by Hofstetter *et al.*¹⁷⁵ by computing shifts of perturbed crystal structures obtained through low-temperature MD simulations of candidate #1 and relating chemical shift deviations to positional deviations. Chemical shift computations were also performed on an extended set of the 81 following lowest energy candidates, but did not lead to lower shift RMSEs than candidate #1. Further computational details are given in **Appendix VII**.

Molecular dynamics simulation of amorphous structures. The amorphous structure of AZD5718 was modelled by carrying out MD simulations on periodic amorphous cells containing 128 AZD5718 molecules and a variable number of water molecules. Five cells of each water content; 0, 0.5, 1.0 and 2.0% (w/w, 0, 16, 32 and 65 water molecules in each cell, respectively), and two cells of 4% water (w/w, 132 water molecules in each cell) were generated. After equilibration for 1 ns using the canonical NVT ensemble at 298 K followed by 10 ns using the isothermal-isobaric ensemble (NPT) at 298 K and 1 bar, production simulations were carried out for 600 ns using the NPT ensemble at 298 K and 1 bar. Models of the amorphous structure were obtained by extracting 1,001 evenly spaced snapshots from the last 100 ns of each MD simulation, corresponding to 100 ps time steps between the extracted snapshots. Further computational details are given in **Appendix VII**.

Chemical shift predictions and hydrogen bonding motifs. The predicted chemical shieldings of all snapshots extracted from the MD simulations (168,799,631 total shifts) were obtained using ShiftML version 1.2.^{176, 261} The conversion from predicted shieldings to isotropic shifts is described in **Appendix VII**.

H-bonded N-H groups were identified in 11 snapshots from each MD simulation, spaced by 10 ns each. The corresponding bonding motifs were extracted by defining hydrogen bonds as N-H...X (X = O, N) patterns with an N-H-X angle above 130° and H-X bond length shorter than 2.5 Å, typically corresponding to moderate to strong hydrogen bonds in organic solids.⁵⁰⁶ If the first H-bonded neighbour was found to be a water molecule, then secondary water-bound neighbours were searched using the same criteria to define hydrogen bonds.

In addition, the N-H groups yielding predicted ¹H chemical shifts above 11 ppm were identified within each snapshot of the 4% water MD simulations (2,002 total snapshots), and the corresponding hydrogen bonding patterns were extracted as described above.

Formation energies in the amorphous simulations. The formation energy of the intermolecular complex comprising one molecule of AZD5718 and its local environment was computed for each molecule in the same snapshots used to identify all the hydrogen bonding motifs (11 snapshots per simulation). The environment of a molecule was defined as all molecules having at least one atom within 5 Å from any atom in the probe molecule. The formation energy was computed as the difference between the energy of the total intermolecular complex and the energy of the isolated environment. The obtained formation energy thus contains both the ground-state energy of the isolated probe molecule, which includes its conformational energy, and the interaction energy between the probe molecule and its environment. The single-point energy computations were performed at the DFTB3-D3H5 level of theory using the 3ob-3-1 parameter set and the DFTB+ software version 20.1.^{325, 326, 349-351}

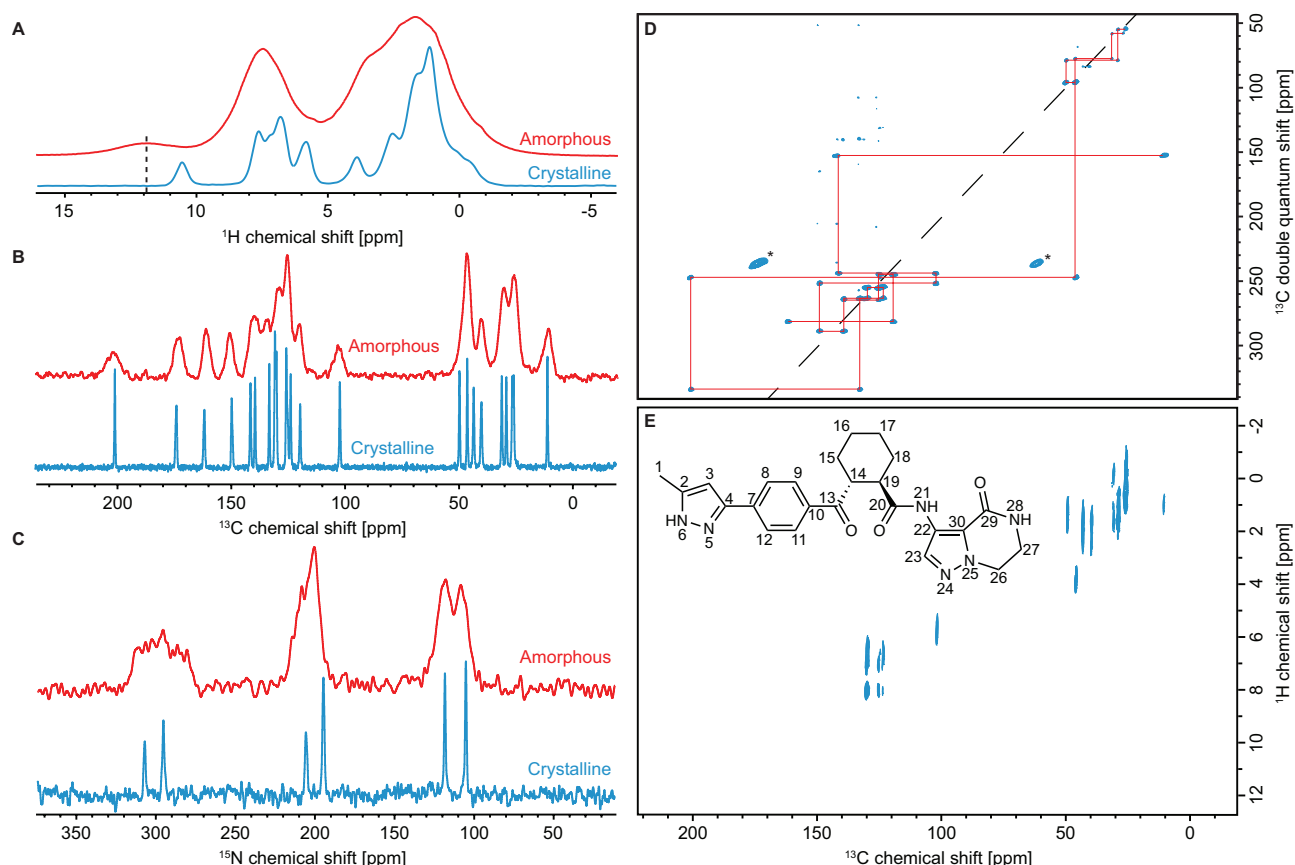


Figure 4.1. Solid-state NMR experiments. (A) ¹H, (B) ¹³C and (C) ¹⁵N MAS NMR spectra of crystalline (blue) and amorphous (red) AZD5718. (D) ¹³C-¹³C DNP-enhanced solvent suppressed INADEQUATE and (E), ¹H-¹³C HETCOR spectra of crystalline AZD5718. The dashed black line in (A) indicates the chemical shift assigned to the proton bound to N6 in the amorphous sample. In (D), the ¹³C peaks denoted by a star at 60 and 170 ppm are attributed to impurities introduced during the NMR sample preparation. The chemical structure and labelling scheme of AZD5718 is shown in (E).

4.2.3 Results and Discussion

NMR crystallography. No polymorphism was observed for anhydrous crystalline AZD5718, nor were crystalline hydrates identified. The crystal structure of anhydrous crystalline AZD5718 Form A was determined using a chemical shift-based NMR crystallographic approach. This involves the combination of the assigned experimental chemical shifts with CSP and computed chemical shifts. The ^1H , ^{13}C and ^{15}N resonances of AZD5718 (**Figure 4.1E**) were assigned using one-dimensional proton, carbon and nitrogen MAS NMR experiments (**Figure 4.1A-C**), as well as two-dimensional refocused ^{13}C - ^{13}C INADEQUATE and ^1H - ^{13}C HETCOR experiments (**Figure 4.1D-E**) as detailed in **Appendix VII**. A set of DFT-D optimised candidate structures was generated using an internally developed rapid CSP approach, then the assigned experimental chemical shifts were compared to the shifts computed using the cluster- and fragment-based DFT approach introduced by Hartman *et al.*¹²⁶⁻¹²⁸ for each structure in order to determine the experimental structure from the set of candidates. The structure determined using single-crystal X-ray diffraction was included and compared to the CSP set. The lowest energy CSP candidate (structure #1) was found to be structurally similar to the X-ray structure (as discussed in **Appendix VII**). The determination of the crystal structure of anhydrous crystalline AZD5718 Form A is briefly described in **Section 1.2.4**, and a more detailed description is presented here.

The comparisons of the experimental and computed ^1H and ^{13}C chemical shifts are shown in **Figure 1.2A**. The root-mean-square errors (RMSEs) obtained for ^1H suggest that structure #1 best matches the experiment, while ^{13}C chemical shift results identify the X-ray structure as the best match. Additionally, the DFT-D energy per molecule of structure #1 was found to be the lowest among the CSP set (x-axis in **Figure 1.2A**). This also indicates that the force field used for the CSP procedure accurately describes the crystalline system, and supports the identification of candidate #1 as being the crystal structure. In order to elucidate the ambiguity between candidate #1 and the XRD structure, and to obtain a quantitative comparison of all candidates, a Bayesian probabilistic analysis was carried out using the approach introduced by Engel *et al.*¹⁷⁶ The two main advantages of using this method to determine the structure that best matches experiment are the quantitative determination of the confidence in the identification of the experimental structure on a continuous scale from 0 to 100%, and the combined use of NMR results for several elements, which ultimately increases the accuracy of the identification.

Figure 1.2B shows the results obtained with the Bayesian approach, represented as a principal component analysis (PCA) plot. This plot is a two-dimensional representation of the similarity of the different candidate structures according to their computed chemical shifts and of the experimental chemical shifts. The computed Bayesian probability of each structure to be the experimental crystal structure is represented by the area of the blue disk around each point (here, only one disk is visible as only its probability is significant). Using both ^1H and ^{13}C chemical shifts, candidate #1 is found to be the most probable crystal structure, with 99.7% confidence. Although the structure determined by X-ray diffraction (labelled XRD) appears closer to the experimental results (red cross) in the first two chemical shift principal components in **Figure 1.2B**, the inclusion of the complete chemical shift space identifies candidate #1 as the structure that best matches experiment, as indicated by its associated confidence. **Figure 1.2B** highlights the similarity of the selected structures in terms of their chemical shifts, in the two dimensions that display the largest variance.

Comparison of the structures determined via XRD and NMR crystallography yielded a RMSD_{15} (root mean square deviation of the atomic positions in 15 molecules, ignoring hydrogen positions) of 0.42 Å. The main difference between the two structures lies in the conformation of the bicyclo ring (see **Figure 4.8**). Single-molecule heavy atom RMSD was found to be 0.22 Å, and decreased to 0.15 Å after omitting the two carbons of the bicyclo ring (labelled 26 and 27 in **Figure 4.1E**).

Unlike X-ray diffraction, NMR is highly sensitive to hydrogen nuclei, making it the method of choice for validating the tautomeric form of AZD5718. Indeed, either of the two nitrogen atoms of the pyrazole ring (labelled 5 and 6 in **Figure 4.1E**) can be protonated in the crystalline sample. After computing ^1H , ^{13}C and ^{15}N shifts for the two possible tautomers displayed in **Figure 4.2A** and comparing them with the experimental shifts (**Figure 4.2B**), the resulting chemical shift RMSE was found to be consistently lower for tautomer A, by a factor of 1.3 for ^1H , 2.6 for ^{13}C and 8.1 for ^{15}N . This unambiguously identifies tautomer A as the crystal structure. The position of the N-H proton in the pyrazole ring is crucial in setting up amorphous structures able to describe the properties of the amorphous phase of AZD5718.

The atomic displacement parameter (ADP) tensors of all atoms in the structure determined by NMR crystallography were obtained as described in Ref. 175. Simulation details are given in **Appendix VII**. **Figure 1.3B** shows the ORTEP plot of the ADP tensors corresponding to a ^1H chemical shift RMSE of 0.34 ppm. This value corresponds to the estimated error of ^1H chemical shifts computed with the fragment- and cluster-based approach.¹²⁷ The average value of the ADPs is 0.00025 Å².

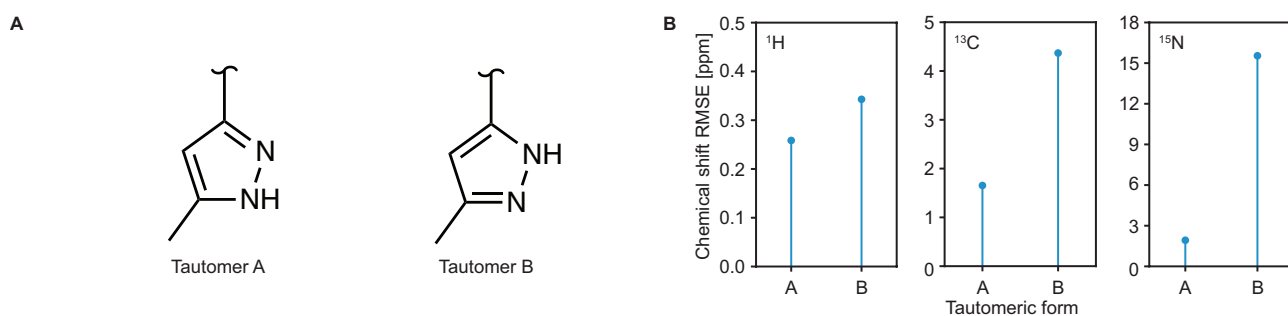


Figure 4.2. Tautomer determination of AZD5718. (A) Chemical structures of the two tautomers of AZD5718 considered, labelled as A and B. (B) Agreement between ^1H , ^{13}C and ^{15}N experimental and DFT computed chemical shifts for the two tautomers.

Hydrogen bonding motifs in the amorphous phase. Knowledge of the structure of AZD5718 in the amorphous phase is key to understanding its physicochemical properties. Investigation of the amorphous structure of AZD5718 was performed using NMR experiments combined with ShiftML-predicted chemical shifts for MD ensembles.^{176, 261}

Comparison of the proton, carbon and nitrogen NMR spectra in crystalline and amorphous AZD5718 shown in **Figure 4.1A-C** displays the overall broadening of the NMR signal typical of amorphous compounds. Apart from this observation, the chemical shifts do not display a significant change between the two phases of the compound. This suggests that AZD5718 does not undergo large amplitude structural rearrangements upon transition from the crystalline to the amorphous state. The main difference between the NMR spectra of the two phases of AZD5718 lies in the displacement of the ^1H resonance corresponding to the proton attached to the nitrogen labelled 6 (see **Figure 4.1E**) from 10.6 ppm in the crystalline sample to 11.8 ppm in the amorphous form. This suggests a change in the hydrogen bonding network in the structure.³⁹

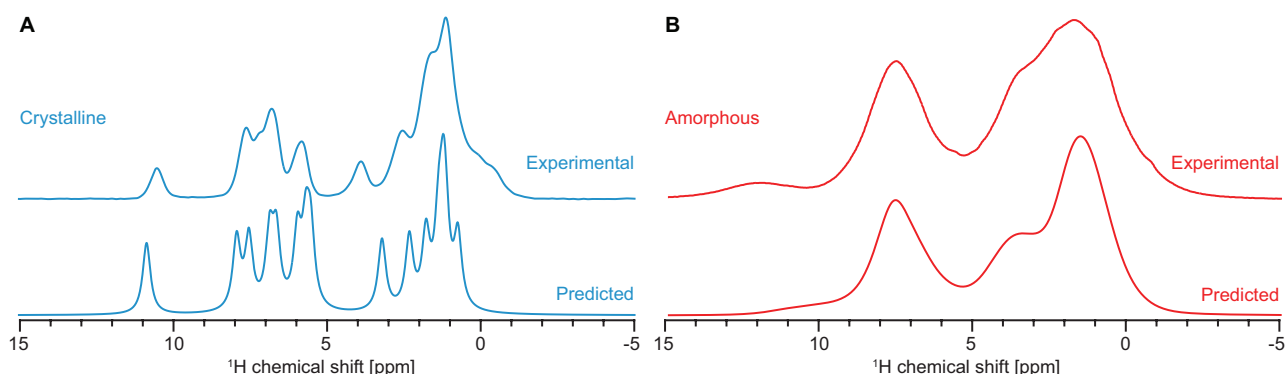


Figure 4.3. Predicted and experimental ^1H NMR spectra of (A) crystalline and (B) amorphous AZD5718. The predicted spectrum of amorphous AZD5718 was obtained by considering only the 4% w/w water MD simulations.

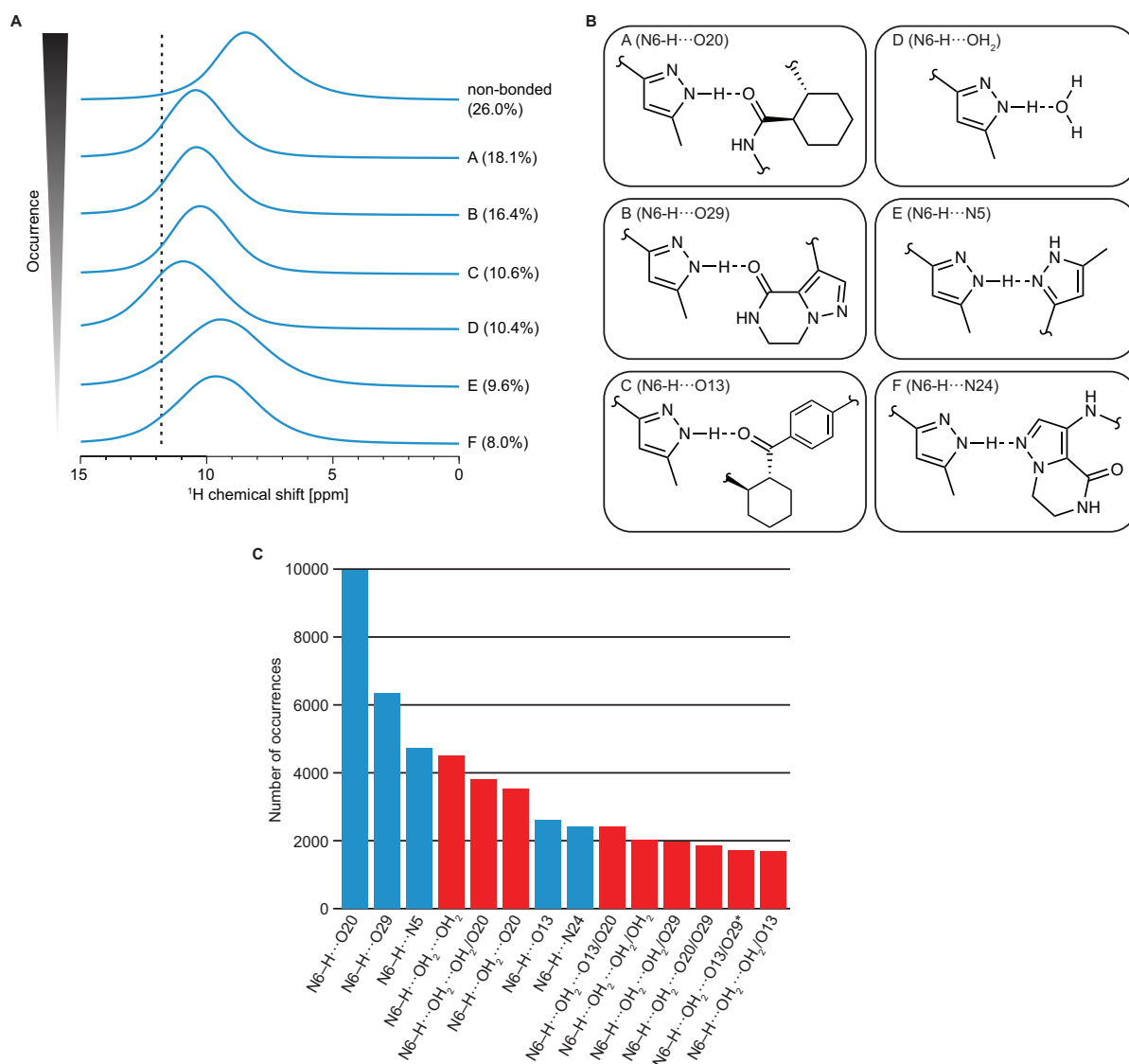


Figure 4.4. H-bonding motifs in amorphous AZD5718. **(A)** Predicted spectra obtained using the predicted ^1H chemical shifts of the most often occurring N-H bonding motifs involving N6 for 11 evenly spaced snapshots of all amorphous simulations of each water content. The percentages next to the spectra denote the fraction of bonding motifs their corresponding pattern represents, including the instances where no H-bonded neighbour was identified. The dashed vertical line indicates the experimental shift observed in amorphous AZD5718 and assigned to the proton bound to N6. **(B)** Hydrogen bonding motifs associated with the spectra in **(A)**. **(C)** Number of occurrences of extended H-bonding motifs yielding a predicted chemical shift above 11 ppm for every snapshot of the 4% water simulations. Only the patterns corresponding to the top 75% of all shifts above 11 ppm were selected. The red bars represent the bonding motifs involving water, and the blue ones correspond to the motifs that do not involve water. Two secondary neighbours from the same molecule are indicated by an asterisk. In **(B)** and **(C)**, O_n indicates the oxygen atom bonded to carbon labelled n .

To better understand the structural differences between the amorphous and crystalline phases of AZD5718, we generated MD models of the amorphous structure at different hydration levels ranging from 0% to 4% (w/w) water content. This range of water content is representative of the experimental water content under real conditions, as confirmed by dynamic vapor sorption. Considering the large size of the simulation cells (128 molecules of AZD5718 and up to 132 water molecules) and the large number of structures generated by MD, DFT computation of chemical shifts in these model systems would not be feasible. The machine learning model ShiftML was thus used to predict chemical shifts in these structures.^{176, 261} The predicted spectra obtained for the crystalline structure and obtained by summing the spectra from 202 full cell snapshots of the 4% water MD simulations (i.e., 25,856 molecules of AZD5718 and 26,664 water molecules) are displayed in **Figure 4.3**. The ^1H chemical shift RMSE obtained by comparing shifts predicted by ShiftML from the crystal structure with the experiment was found to be 0.61 ppm.

Although no clear peak is observed at 11.8 ppm for the amorphous structure, the population of predicted shifts above 11 ppm was found to increase slightly with increasing water content (see **Figure 4.9**). This behaviour suggests that interaction of AZD5718 with water molecules does promote deshielding of the proton attached to the nitrogen labelled 6.

The predicted chemical shifts obtained were related to structural motifs in the model amorphous structures by identifying the different hydrogen bonding patterns that are present in the structures (with the criteria given in **Section 4.2.2**) and the associated predicted ^1H shifts of the hydrogen bond donor groups in the hydrogen bonds. **Figure 4.4A** displays the chemical shift distributions of the most often occurring hydrogen bonding motifs. The atom most commonly bound to the N6-H group was found to be O20 (where O20 is the oxygen bound to C20), which corresponds to the hydrogen bond found in crystalline AZD5718. This is an indication that the structure of the amorphous compound is broadly similar to that of its crystalline counterpart. Over all analysed simulation snapshots (corresponding to an average water content of 1.16% (w/w), or about 3.4 times more molecules of AZD5718 than water), water was found to be the fourth most occurring hydrogen bonding partner to the N6-H group. It was also found to lead to the most pronounced deshielding of the hydrogen bond donor proton (motif D in **Figure 4.4B**).

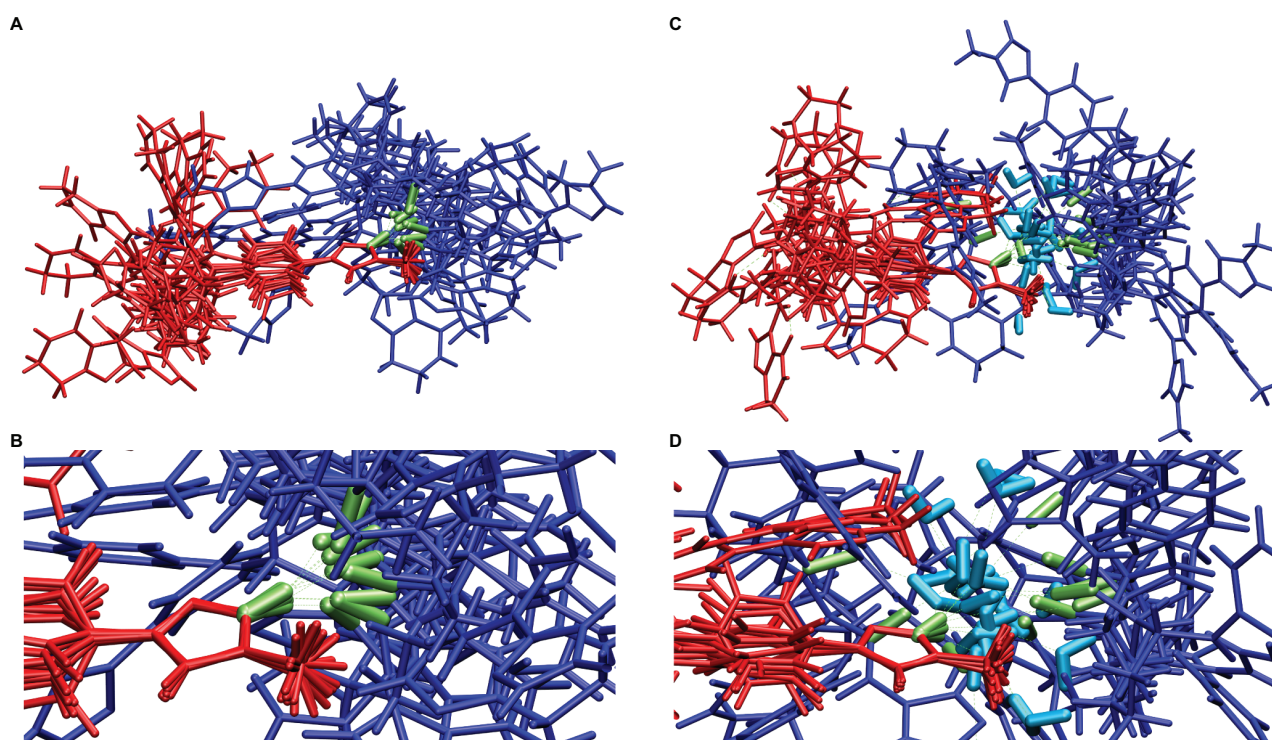


Figure 4.5. Complete structures and H-bonding motifs. (A) Superposition of 10 instances of the N6-H...O20 bonding motif. (B) close-up view of the hydrogen bonding region in (A). (C) Superposition of 10 instances of the N6-H...OH₂...OH₂/O20 bonding motif. (D) close-up view of the hydrogen bonding region in (C). The red molecule represents AZD5718 bearing the hydrogen bond donor (N6-H), the dark blue molecule represents AZD5718 bearing the hydrogen bond acceptor, water molecules are coloured in cyan and the atoms of AZD5718 involved in the hydrogen bonding motif are coloured in green.

Because a single water molecule can form two hydrogen bonds involving its hydrogen atoms and two additional ones involving its oxygen atom, more extended hydrogen bonding motifs are likely to be observed for AZD5718 molecules bound to water. In order to investigate these extended patterns, we extracted N6-H \cdots OH₂ motifs yielding a chemical shift above 11 ppm in all snapshots of the amorphous 4% water MD simulations, and obtained the secondary neighbours, bonded to the water protons. We restricted this analysis to the simulations with the highest water content, as bonding of water was found to lead to the largest deshielding of the proton attached to N6 (see **Figure 4.4A**). Moreover, a larger number of water molecules in the simulation promotes extended hydrogen bonding motifs. **Figure 4.4C** shows the occurrences of extended hydrogen bonding patterns involving water, as well as the motifs made of pairs of H-bonded molecules of AZD5718, yielding a predicted shift above 11 ppm. The most often occurring pattern is the hydrogen bond present in the crystalline phase of the compound (N6-H \cdots O20). When the H-bonded molecule is water, secondary neighbours are often found to be other water molecules, suggesting the formation of small clusters of water between AZD5718 molecules. Superpositions of ten instances of two hydrogen bonding motifs, N6-H \cdots O20 and N6-H \cdots OH₂ \cdots OH₂/O20, are shown in **Figure 4.5**. It was observed that molecules of AZD5718 being secondary H-bonded neighbours of N6-H generally lie away from the molecule bearing the hydrogen bond donor, indicating that steric clashes may constrain the possible geometries of hydrogen bonding in the amorphous form.

Formation energies of intermolecular complexes. Obtaining the formation energies of the supramolecular complexes of AZD5718 molecules and their surroundings can help determine which hydrogen bonding pairs lead to overall more favourable intra- and intermolecular interactions. After computing the formation energies (including the conformational energy of the probe molecule) using the semiempirical DFTB3-D3H5 method, the results were gathered as a function of the hydrogen bonding motifs in which the probe molecule was involved as the hydrogen bond donor. The relative formation energy was defined as the difference between the formation energy of each instance and the mean formation energy of all instances where no hydrogen bond acceptor was found for the selected hydrogen bond donor.

Figure 4.6 shows the relative formation energy for each hydrogen bond donor and acceptor identified. Bonding of any N-H group to water was found to yield the most favourable interactions. Over all the simulation snapshots analysed for hydrogen bonding motifs, 6.1% of N21-H chemical groups were found to form an intramolecular hydrogen bond with the carbonyl labelled 29. This number may however be underestimated, as the same intramolecular hydrogen bond is found in the crystal structure with a bond angle of 128.8°. This hydrogen bond would thus not be identified using the cutoff values selected here.

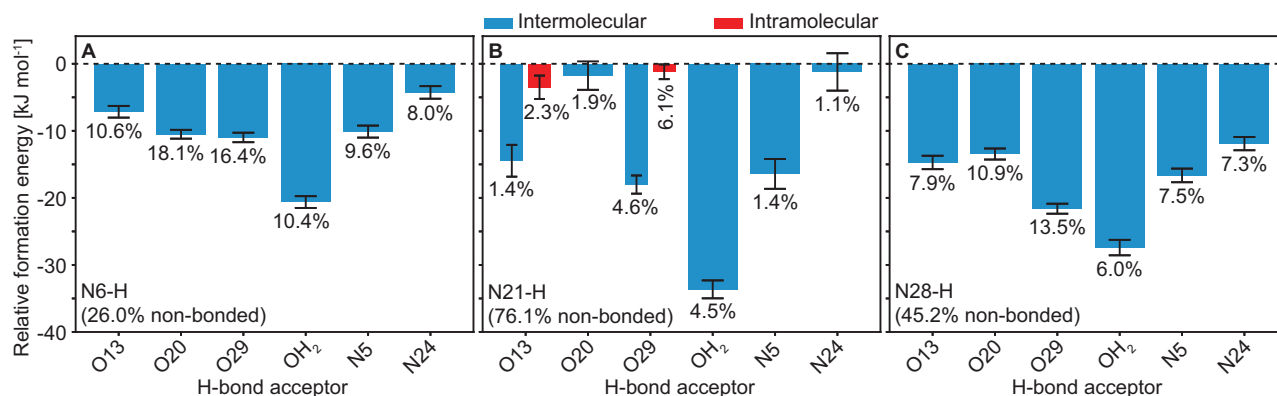


Figure 4.6. Relative formation energy of the hydrogen bonding motifs. Mean computed formation energies of intermolecular complexes for the H-bond acceptor connected to (A) N6-H, (B) N21-H or (C) N28-H of the probe molecule. The percentage under each bar indicates the fraction of the N-H group bonded to the corresponding H-bond acceptor. Only the H-bond acceptors making up at least 1% of all instances analysed are displayed. The error bars indicate the standard error of the mean of the relative formation energies.

4.2.4 Conclusion

The atomic-level structure and hydrogen bonding patterns of the hydrated amorphous phase of AZD5718 were determined through solid-state NMR chemical shifts, MD simulations with various water contents, and machine learned chemical shifts. The chemical shifts associated with possible hydrogen bonding motifs generated from MD simulations were compared to experimental NMR spectra in order to identify the most commonly occurring intermolecular interactions in the amorphous material. Bonding of N6-H to water was found to yield the largest deshielding of the proton involved in the hydrogen bond, and best described the experimental shift observed in the amorphous sample. This intermolecular bond to water was also associated with more favourable intermolecular complex formation energies as compared to direct H-bonding between two AZD5718 molecules. These favourable water-AZD5718 interactions highlight the potential ability of water in preventing physical aging of the amorphous drug.

The combination of the three techniques presented here was crucial in elucidating the structure of this amorphous material through a large-scale direct comparison of experimental chemical shifts with predicted shifts from MD structures. While solid-state NMR has already been used in tandem with MD simulations of amorphous materials, previous works have generally used molecular dynamics either to relate relative NMR peak areas to statistical ratios of different types of interactions,^{507, 508} or to generate conformational ensembles from which small supramolecular clusters are extracted for DFT shift computation.⁴⁹⁴ Overall, the method presented here can be applied to a wide range of disordered organic systems to determine their complete atomic-level structures from their NMR spectra.

The structure of the crystalline form was also determined using NMR crystallography to within a positional error of 0.1 Å and was confirmed to be almost identical to the structure obtained with single crystal X-Ray diffraction.

4.2.5 Appendix VII

Data availability. All data used in this study is freely available from <https://doi.org/10.24435/materialscloud:gg-mx>. The NMR raw data are provided in JCAMP-DX version 6.0 standard format and original TopSpin format. Data are made available under the license CC-BY-4.0 (Creative Commons Attribution-ShareAlike 4.0 International). The Python scripts used to analyse NMR crystallography and MD simulation data are available from the same link and made available under the license CC-BY-4.0 (Creative Commons Attribution-ShareAlike 4.0 International).

SUPPLEMENTARY METHODS

Sample preparation for DNP NMR. In DNP MAS experiments, the high thermal polarisation is transferred from unpaired electrons to nuclei (typically ¹H) which results in enhanced NMR signals. For organic powders, this is achieved by impregnating the powdered solid with an otherwise inert polarising solution.^{509, 510} AZD5718 dissolves both in water and in most organic solvents, so most typical polarising solutions, such as 16 mM TEKPOL⁵¹¹ in 1,1,2,2-tetrachloroethane (TCE), were found to be incompatible. Ortho-terphenyl was found to be a suitable non-solvent for AZD5718, and 16 mM TEKPOL in ortho-terphenyl 99.5%-d₁₄ (OTP-d₁₄) was used as a polarisation source. The sample was prepared according to the procedure described in Refs. 512 and 513 by mixing a solid solution of 16 mM TEKPOL in OTP-d₁₄ with powdered AZD5718, then transferring it to a sapphire rotor sealed with a PTFE insert and capped with a zirconia drive cap. The rotor was then heated at ca. 65°C in a hot water bath in order to melt the OTP and allow the liquid to impregnate the API. It was then quickly inserted into the pre-cooled LT-MAS DNP probe to rapidly freeze the sample in order for the OTP to form a glass.⁵¹³ DNP enhancements of about 5 as measured on crystalline AZD5718 signals through (¹H)¹³C DNP CPMAS were obtained, which was sufficient to allow the natural abundance INADEQUATE spectra to be recorded.

NMR spectroscopy. Experiments were performed on Bruker Ascend 400 and Ascend 500 wide-bore Avance III, and Bruker 800 Ultrashield plus narrow-bore, and 900 US² wide-bore Avance Neo NMR spectrometers. The spectrometers operate at ¹H Larmor frequencies of 400.13, 500.43, 800.13, and 900.13 MHz respectively, and are equipped with H/X/Y 3.2 mm, H/C/N/D 1.3 mm and H/C/N 0.7 mm CPMAS probes. When the 3.2 mm probe was used, the samples were restricted to the central third of a rotor with an inner diameter of 2.2 mm, in order to maximise radiofrequency (rf) homogeneity.

DNP solid-state NMR spectroscopy experiments were performed on a 400 MHz Avance III HD Bruker spectrometer. The spectrometer is equipped with a low temperature magic angle spinning (LTMAS) 3.2 mm probe and connected through a corrugated waveguide to a 263 GHz gyrotron capable of outputting ca. 5-10 W of continuous wave microwaves.⁴¹⁴ The sweep coil of the main magnetic field was optimised so that the microwave irradiation gives the maximum positive proton DNP enhancement with binitroxide cross effect-based polarising agents (e.g. AMUPOL⁵¹⁴, TEKPOL⁵¹¹). DNP enhancements were determined based on the ratio of the area of the spectra acquired with and without microwave irradiation.

1D ^1H MAS NMR spectra were recorded at a temperature of 298 K using rotor spinning rates (ν_r) up to 111 kHz. 1D ^{13}C cross-polarisation⁴⁰⁸ (CP) MAS NMR spectra were acquired at 298 K with ν_r of 22 kHz. The CP contact time was 2 ms and during the signal acquisition SPINAL-64 decoupling⁴⁰⁹ was applied with a ^1H rf field amplitude of 100 kHz. 1D ^{15}N CP NMR spectra were acquired at 100 K under DNP MAS conditions with $\nu_r = 12.5$ kHz for crystalline AZD5718, and similar measurements were made on amorphous AZD5718 using LT-MAS conditions (without DNP) in the same instrument. Variable amplitude cross-polarisation⁴¹² was used to transfer polarisation from ^1H (60% to 100% ramp) to ^{15}N (constant amplitude). For the ^{15}N CPMAS spectra of crystalline AZD5718, 360 scans were acquired with DNP spaced by a recycling delay of 20 s leading to a total acquisition time of 2 h. For amorphous AZD5718, 14,720 scans were acquired without DNP, spaced by a recycling delay of 5 s leading to a total acquisition time of 21 h.

2D ^1H - ^{13}C HETCOR experiments were carried out at 298 K using $\nu_r = 22$ kHz. 96 points were acquired in the indirect dimension with the States acquisition method,⁵¹⁵ and with indirect sampling intervals (Δt_1) of 96 μs . For the crystalline sample the recycle delay was 32 s ($T_1 \sim 22$ s) and 64 scans were collected for each t_1 point. For the amorphous sample the recycle delay was 4 s ($T_1 \sim 3$ s) and 769 scans were collected for each t_1 point. During t_1 100 kHz eDUMBO-122 was applied to decouple the ^1H - ^1H dipolar coupling,⁴¹⁰ and during t_2 100 kHz SPINAL-64 decoupling was applied.⁴⁰⁹

The 2D ^{13}C - ^{13}C refocused INADEQUATE^{369, 411} spectrum of crystalline AZD5718 was acquired using DNP MAS NMR.⁵¹⁶ For the ^{13}C - ^{13}C refocused INADEQUATE experiment, the probe was configured into $^1\text{H}/^{13}\text{C}$ double resonance mode. Variable amplitude cross-polarisation⁴¹² was used to transfer polarisation from ^1H to ^{13}C . SPINAL-64 heteronuclear ^1H decoupling⁴⁰⁹ with rf fields of 100 kHz was applied in all cases.

The DNP enhancement allowed to record a ^{13}C - ^{13}C refocused INADEQUATE spectrum at natural abundance for crystalline AZD5718 in about 2 days of signal averaging. Moreover, using a ^1H spin-lock of 30 ms between the ^1H excitation pulse and the CP, the otherwise dominant OTP solvent signal was efficiently removed,⁵¹⁷ allowing to record a 2D spectrum ^{13}C - ^{13}C refocused DNP INADEQUATE clean from the solvent signal. The spectrum was acquired in about 45 h with 128 points recorded in the indirect dimension with 256 scans each separated by recycling time of 5 s. The increment in the indirect dimension was 40 μs , allowing a total indirect acquisition time of 5.12 ms using the States-TPPI method.⁵¹⁸ The tau period for J evolution was optimised and set to 4 ms. SPINAL-64 was used for heteronuclear decoupling.⁴⁰⁹

All chemical shifts were referenced via alanine. The full set of acquisition parameters is given in **Tables 4.1-4.4**.

Table 4.1. Experimental parameters for 1D experiments on AZD5718 form A anhydrous.

	^1H	^{13}C	^{15}N
MAS rate	111 kHz	22 kHz	12 kHz
Recycle delay (d_1)	10 s	32 s	20 s
^1H to X CP			
Spin lock duration	-	2 ms	10 ms
Total acquisition time	5.5 ms	30 ms	25 ms
Dwell time	2.8 μs	13.2 μs	12.3
Number of points	1964	2268	2032
Number of scans	4	128	360
Acquisition mode	DQD	qsim	qsim

Table 4.2. Experimental parameters for 2D experiments on AZD5718 form A anhydrous.

	^1H - ^{13}C HETCOR	^{13}C - ^{13}C INADEQUATE
MAS rate	22 kHz	12.5 kHz
Recycle delay (d_1)	32 s	5 s
^1H to X CP		
Spin lock duration	0.1 ms	3 ms
Acquisition in the indirect dimension (t_1)		
Total acquisition time	4.6 ms	2.6 ms
Dwell time	96 μs	20 μs
Number of points	96	256
Acquisition in the direct dimension (t_2)		
Total acquisition time	33 ms	15 ms
Dwell time	9.9 μs	5 μs
Number of points	3328	128
Number of scans per increment	64	128
Acquisition mode	States	States-TPPI
Delay t	-	5 ms

Table 4.3. Experimental parameters for 1D experiments on AZD5718 amorphous.

	^1H	^{13}C	^{15}N
MAS rate	62.5 kHz	22 kHz	8 kHz
Recycle delay (d_1)	6.5 s	4 s	5 s
^1H to X CP			
Spin lock duration	-	2 ms	10 ms
Total acquisition time	8.2 ms	30 ms	25 ms
Dwell time	1.0 μs	9.9 μs	12.3
Number of points	8192	3024	2032
Number of scans	4	128	30720
Acquisition mode	DQD	qsim	qsim

Table 4.4. Experimental parameters for 2D experiments on AZD5718 amorphous.

¹ H- ¹³ C HETCOR	
MAS rate	22 kHz
Recycle delay (d ₁)	4 s
¹ H to X CP	
Spin lock duration	0.1 ms
Acquisition in the indirect dimension (t ₁)	
Total acquisition time	4.6 ms
Dwell time	96 μs
Number of points	96
Acquisition in the direct dimension (t ₂)	
Total acquisition time	33 ms
Dwell time	9.9 μs
Number of points	3328
Number of scans per increment	769
Acquisition mode	States

CSP protocol. To generate a predicted polymorph landscape for AZD5718, the molecular conformation determined via single-crystal XRD was optimised at the B3LYP-D3/6-31G(d,p)^{95, 99, 104, 519-522} level of theory in an implicit water environment using the Gaussian 09 Rev. D.01 program.⁵²³ The media surrounding the molecule was described using the Self Consistent Reaction Field (SCRF) PCM method⁵²⁴ with a dielectric constant ϵ set to 78.35530, as implemented in the Gaussian software. Atomic charges were obtained using the charges from electrostatic potentials using a grid-based method (CHELPG).⁵²⁵ This is a slightly modified procedure compared to the previously published in-house CSP method using an internally developed force-field (AZ-FF).⁵²⁶ The optimised geometry was then used in a single-point energy computation using the MacroModel program,⁵²⁷ where a unique force-field for AZD5718 was constructed. Conformational analysis was then performed within the GRACE program^{528, 529} in order to determine what parameters were allowed to be flexible in the molecule. For AZD5718, all single bonds were allowed to be rotated, and the two saturated rings were allowed to adopt different ring conformations. Tautomer A was assumed.

Candidate crystal structures were generated in the seven most stable chiral space groups (P2₁, P2₁2₁2₁, P1, C2, P2₁2₁2, P4₃, C222₁) employing the GRACE machinery for a flexible conformation under Z'=1 condition. The crystal structure space was searched using a Monte-Carlo (MC) parallel tempering method⁵⁰⁵ followed by lattice energy minimisation for each polymorph using the AZ-FF force field.⁵²⁶ The search was continued until the convergence criterion for statistically finding all polymorphs in the search, set to 0.7, was met.⁵²⁶ Typically, 3,000 structures are kept at this stage. A structure duplicate check allowed to reduce this number to 1,000 unique structures. From these, the top 190 candidates, named #1 through #190 by increasing force field energy, were selected for full DFT-D optimisation using the PBE functional⁹⁷ and Neumann-Perrin dispersion correction⁵²⁸ in the VASP software.⁵³⁰⁻⁵³³ The default PAW pseudopotentials and a 520 eV plane-wave energy cutoff were used. The ten most stable polymorphs (within 6 kJ/mol) were then selected for NMR computation. An extended set of the following 81 most stable structures (within 23.3 kJ/mol) was also selected for NMR computation, but did not lead to a better match of the experimental chemical shifts than structure #1. These 81 structures were thus not included in the set of structures used for the Bayesian analysis displayed in **Figure 1.2B**.

Chemical shift computation of candidate crystal structures. The proton positions of the candidates selected for NMR computations were optimised using the plane-wave DFT software Quantum ESPRESSO version 6.5.^{328, 329} The constrained optimisations were performed at the PBE level of theory⁹⁷ using Grimme D2 dispersion correction³³⁰ and projector augmented wave scalar relativistic pseudopotentials with GIPAW reconstruction, H.pbe-tm-new-gipaw-dc.UPF and C.pbe-tm-new-gipaw-dc.UPF,³⁷ and N.pbe-n-kjpaw_psl.1.0.0.UPF and O.pbe-n-kjpaw_psl.1.0.0.UPF.³³² The wavefunction and charge density energy cutoffs were set to 60 and 240 Ry, respectively, and the relaxations were carried out without k-point.

Chemical shifts were computed for the candidate crystal structures obtained through the CSP procedure at the PBE0 level of theory¹⁰⁶ using the cluster- and fragment-based approach introduced by Hartman *et al.*¹²⁶⁻¹²⁸ (computational details are provided in **Table 4.5**). Direct linear regression between the chemical shieldings computed for each candidate and the experimental chemical shifts was performed in order to obtain computed chemical shifts. The computations were run using the hybrid-many-body-interaction (HMBI) code^{534, 535} with Gaussian 16 Revision A.03 as the DFT engine.⁵³⁶ The computed chemical shieldings σ_{calc} were converted to isotropic chemical shifts δ_{calc} through the relationship

$$\delta_{\text{calc}} = \sigma_{\text{ref}} - b\sigma_{\text{calc}} \quad (4.1)$$

For each candidate crystal structure, the value of σ_{ref} and b were determined by linear regression between computed and experimental shifts, permuting the ambiguously assigned shifts to obtain the lowest root-mean-square error (RMSE).

Table 4.5. Cutoffs and basis sets used in the cluster/fragment DFT computations.

Cutoff description	Cutoff [Å]	Basis set
Cluster cutoff	0	
Pair-wise interaction cutoff	6	
Electrostatic embedding cutoff	30	
Basis set 1	2	6-311+G(2d, p)
Basis set 2	4	6-311G**
Basis set 3	12	6-31G

Positional uncertainty of the crystal structure. Perturbed crystal structures were obtained by performing molecular dynamics simulations of the crystal structure at 1, 5, 10, 15, 20 and 25 K. 300 ps simulations were carried out with a time step of 0.5 fs and using the canonical (NVT) ensemble, and 21 snapshots were extracted from the last 150 ps of each simulation. The force-field and parameters used are the same as the ones used to model the amorphous structure, except for the electrostatic and Van der Waals interaction cutoffs, which were set to 2.8 Å to avoid self-interaction. No constraint on the bond lengths to hydrogen was set. The correlation between chemical shift RMSD $\langle\delta\rangle$ and the average positional RMSD of atom i along the l^{th} principal axis of its ensemble of positional deviations $\langle r_{i,l} \rangle$ was obtained by maximising the log-likelihood between the computed correlation points and the Gaussian distribution described by **Equation 4.2** as a function of the Gaussian parameters $\mu_{i,l}$ and $\Sigma_{i,l}$,

$$G(\langle r_{i,l} \rangle, \langle \delta \rangle) = \frac{1}{\sqrt{2\pi\Sigma_{i,l}^2\langle \delta \rangle^2}} \exp\left(-\frac{(\langle r_{i,l} \rangle - \mu_{i,l}\langle \delta \rangle)^2}{2\Sigma_{i,l}^2\langle \delta \rangle^2}\right) \quad (4.2)$$

The corresponding principal value of the atomic displacement parameters along the l^{th} principal axis $U_{ii,l}$ is obtained from the variance of the Gaussian distribution,

$$U_{ii,l} = \Sigma_{i,l}^2\langle \delta \rangle^2 \quad (4.3)$$

Generation of amorphous structures. To model the amorphous structure of AZD5718, we carried out MD simulations on periodic amorphous cells with a variable number of water molecules. The atomic positions of a single AZD5718 molecule extracted from the crystal structure determined via single-crystal XRD were first optimised at the B3LYP-D3/6-31G(d,p)^{95, 99, 104, 519-522} level of theory in gas phase using the Gaussian 09 revision D.01 program.⁵²³ Optimised coordinates and CHELPG charges were extracted from the optimisation and used as input to generate amorphous cells. Materials Studio⁵³⁷ together with the COMPASS-II⁵³⁸ force field were used to create cubic amorphous cells of 128 molecules of AZD5718. Five cells of each water content; 0, 0.5, 1.0 and 2.0% (w/w, 0, 16, 32 and 65 water molecules in each cell, respectively), and two cells of 4% water (w/w, 132 water molecules in each cell) were generated. Geometries were optimised during the construction. The mean initial cell volumes were 73,004, 73,372, 73,740, 74,500, and 76,042 Å³ for the 0, 0.5, 1, 2 and 4% water simulations, respectively.

The optimised coordinates and CHELPG charges of AZD5718 were used as input to generate OPLS_2005^{539, 540} force field parameters using the Schrödinger ffile_server.⁵⁴¹ The “ffconv.py” tool was used to convert the topology into GROMACS format.⁵⁴² Water was treated using the TIP3P model in the MD simulations.⁵⁴³

Molecular dynamics simulation of amorphous structures. The GROMACS program (version 2016.4)^{544, 545} was used for all MD simulations throughout the study. The systems were initially equilibrated for 1 ns using the canonical (NVT) ensemble at 298 K. The temperature was held constant using a modified Berendsen thermostat with velocity-rescaling with a coupling constant of 0.1 ps.⁵⁴⁶ A second equilibration was carried out for 10 ns using the isothermal-isobaric ensemble (NPT) at 298 K and 1 bar where the temperature and pressure were held constant using the velocity-rescaling thermostat with a coupling constant of 0.1 ps and a Berendsen barostat with a coupling constant of 1 ps.^{546, 547} Production simulations were carried out for 600 ns using the NPT ensemble at 298 K and 1 bar where the temperature and pressure were held constant using the velocity-rescaling thermostat⁵⁴⁶ with a coupling constant of 0.1 ps and the Parrinello-Rahman barostat with a coupling constant of 4 ps.^{548, 549} A particle mesh Ewald scheme^{550, 551} was used to compute the electrostatic interactions with a 10 Å cutoff in real space. The same cutoff was used for van der Waals interactions, with long-range dispersion correction applied to both energy and pressure. Bond lengths to hydrogens were constrained using the LINCS algorithm.⁵⁵² System trajectories were collected every 10 ps. All simulations were performed using a time step of 2 fs. Models of the amorphous structure were obtained by extracting 1,001 evenly spaced snapshots from the last 100 ns of each MD simulation, corresponding to 100 ps time steps between the extracted snapshots.

Chemical shift predictions and hydrogen bonding motifs in amorphous structures. The predicted shieldings σ_{pred} obtained using ShiftML were converted to chemical shifts δ_{pred} through the relationship given in **Equation 4.1** and where σ_{ref} and b were determined by maximising the cosine similarity between the simulated spectra, obtained by summing Lorentzian functions with a 0.3 ppm linewidth centred on the predicted shifts, and the experimental spectra. For the crystalline compound, the regression parameters were found to be $b = -0.91$ and $\sigma_{\text{ref}} = 27.9$ ppm. For the amorphous form, the regression was only performed on the 4% water simulations, and applied to all other water contents. The obtained parameters are $b = -0.99$ and $\sigma_{\text{ref}} = 30.9$ ppm.

SUPPLEMENTARY DISCUSSION

Chemical shift assignment. The ^1H , ^{13}C and ^{15}N resonances of AZD5718 (**Figure 4.1E**) were assigned using one-dimensional proton, carbon and nitrogen MAS NMR experiments (**Figure 4.1A-C**), as well as two-dimensional refocused ^{13}C - ^{13}C INADEQUATE and ^1H - ^{13}C HETCOR experiments (**Figure 4.1D-E**). The INADEQUATE spectrum (recorded only for the crystalline form) provides the covalent connectivities between carbon atoms, indicated by red lines in **Figure 4.1D**. The HETCOR spectrum (**Figure 4.1E**) correlates chemical shifts of bonded carbon and hydrogen nuclei.

Chemical shift assignments of ^1H , ^{13}C and ^{15}N nuclei are given in **Table 4.6**. The two protons attached to each carbon in aliphatic rings (labelled 15-18, 26 and 27 in **Figure 4.1E**) are not equivalent, thus two values of ^1H chemical shifts are reported for those nuclei.

Table 4.6. Chemical shift assignment of AZD5718. The values for inequivalent protons attached to the same carbon are indicated by a comma, and ambiguous assignments are denoted by a slash. Ambiguous assignments of carbons that were resolved using the computed shifts of structure #1 are indicated by a star.

Label	¹ H chemical shift [ppm]	¹³ C chemical shift [ppm]	¹⁵ N chemical shift [ppm]
1	1.2	11.1	-
2	-	141.5	-
3	5.8	102.3	-
4	-	149.8	-
5	-	-	295.4
6	10.6	-	205.6
7	-	139.5	-
8	6.9/7.3	123.9/125.3	-
9	6.7 / 7 / 7.6	130.1/130.8	-
10	-	133.3	-
11	6.7 / 7 / 7.6	130.1/130.8	-
12	6.9/7.3	123.9/125.3	-
13	-	201.1	-
14	3.9	46.3	-
15	0.0, 1.7	31.2	-
16	-0.5, 0.8	26.6	-
17	-0.5, 0.8	26.0	-
18	1.6, 1.6	29.2	-
19	1.6	49.8	-
20	-	174	-
21	7.7	-	118.6
22	-	125.8	-
23	6.7 / 7 / 7.6	130.8	-
24	-	-	307.2
25	-	-	194.9
26	1.7, 2.7	43.5*	-
27	1.9, 2.7	40.1*	-
28	6.9	-	105.4
29	-	161.8	-
30	-	119.7	-

Comparison of the structures determined via X-ray diffraction and NMR crystallography. The crystal structure determined using single-crystal X-ray diffraction (**Figure 4.7**) was compared to the structure obtained through NMR crystallography. The superposition of the two structures is shown in **Figure 4.8**. The two structures were found to be highly similar except for the conformation of the bicyclo ring (on the left of **Figure 4.8**).

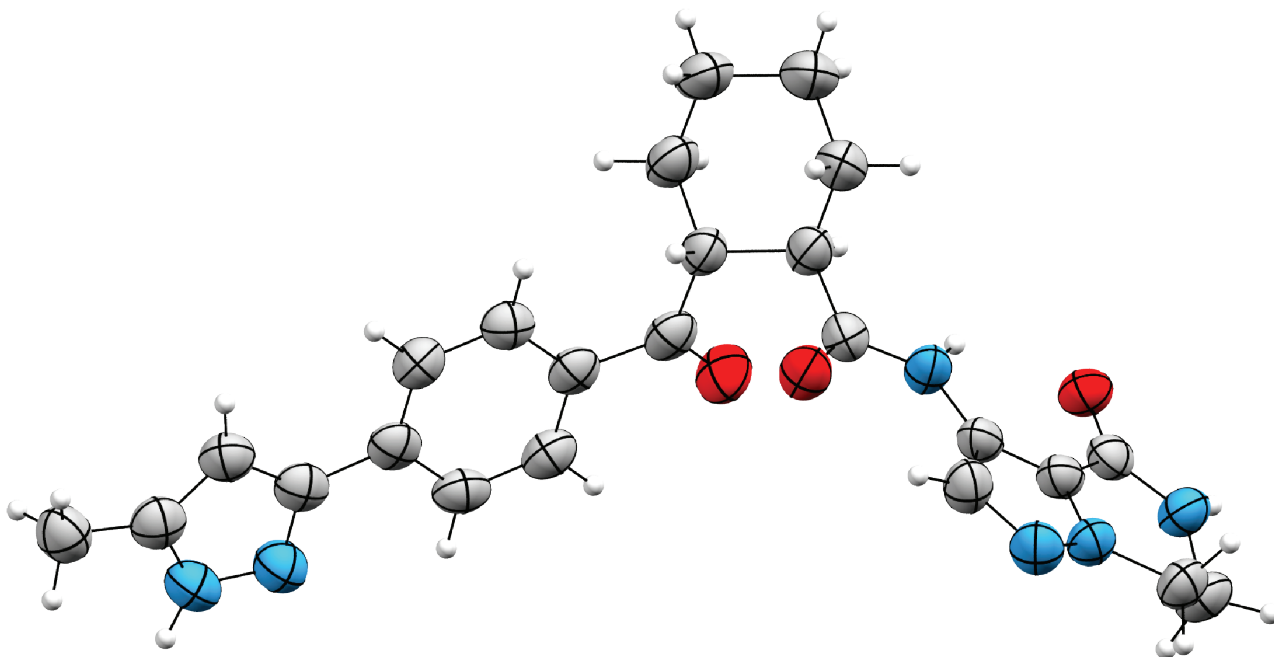


Figure 4.7. Positional uncertainty of the X-ray determined structure of AZD5718. ORTEP plot of the heavy atom ADP tensors for the crystal structure of AZD5718 determined using single crystal X-ray diffraction, drawn at the 90% probability level.

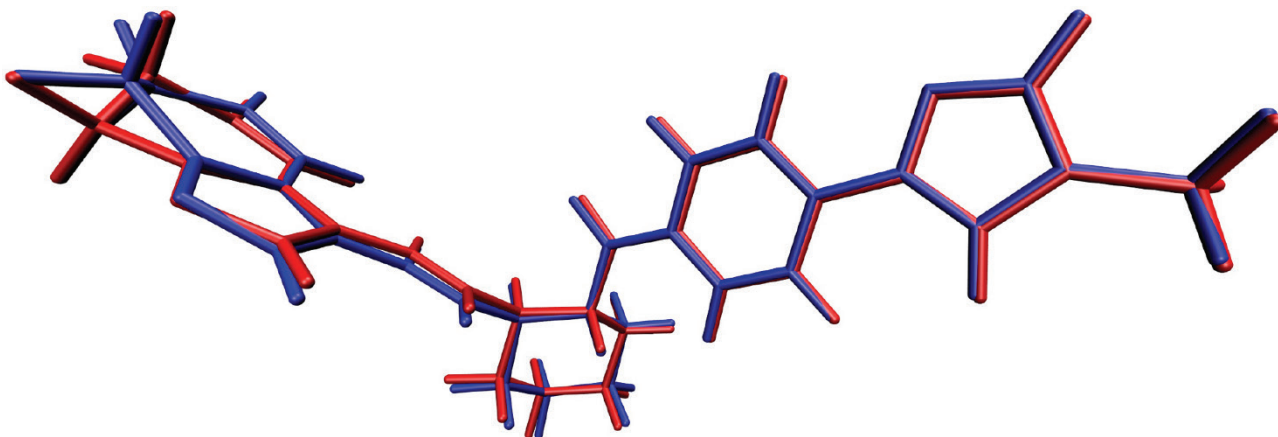


Figure 4.8. Similarity between the XRD and NMR crystallography structures. Comparison between the structure of AZD5718 determined using X-ray diffraction (red) and NMR crystallography (blue).

Simulated spectra of AZD5718 amorphous MD simulations with different water contents. The simulated spectrum for each water content was computed by summing Lorentzian functions centred at the predicted shifts, and with a width of 0.3 ppm. The parameters for the conversion from shielding to shift were extracted by comparing the 4.0% water simulated spectrum with the experimental spectrum, and were applied to all simulations of different water contents. The spectra were normalised such that their maximum is one. Although the experimental peak observed at 11.8 ppm was not observed in the simulated spectra, a larger population of the shifts above 11 ppm was observed with increasing water content (**Figure 4.9**).

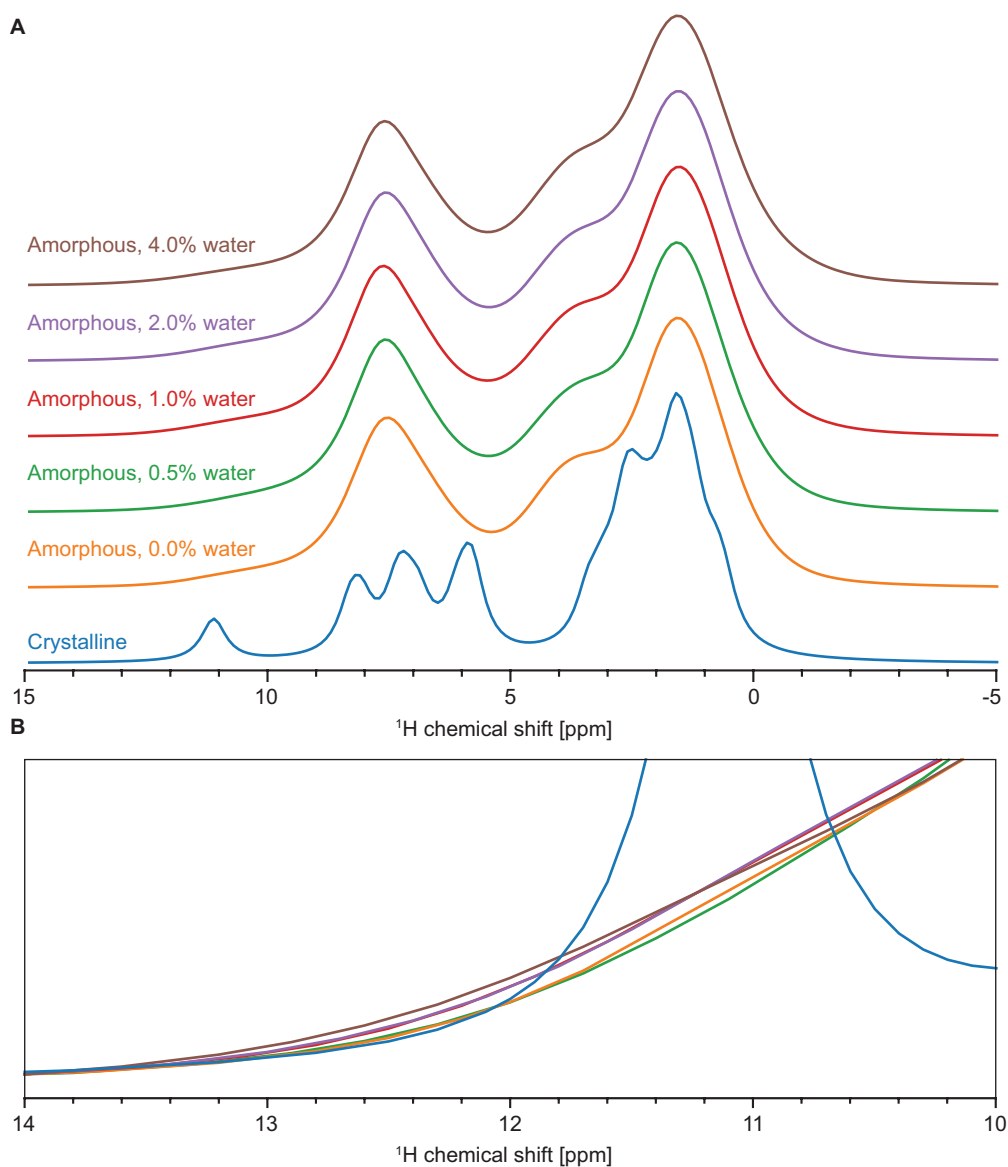


Figure 4.9. Effect of the water content on simulated ^1H NMR spectrum. (A) Simulated ^1H NMR spectra of crystalline and amorphous AZD5718. (B) Close-up view of the spectra in the region between 10 and 14 ppm.

4.3 Atomic-level structure determination of amorphous molecular solids by NMR

This section has been adapted with permission from: Cordova, M.; Moutzouri, P.; Nilsson Lill, S. O.; Cousen, A.; Kearns, M.; Norberg, S. T.; Svensk Ankarberg, A.; McCabe, J.; Pinon, A. C.; Schantz, S.; Emsley, L., Atomic-Level Structure Determination of Amorphous Molecular Solids by NMR. *In press* **2023**.

My contribution was to develop and apply the method and to analyse the results obtained. I also wrote the manuscript, with contributions of all other authors.

4.3.1 Introduction

AZD4625 is a covalent allosteric inhibitor of the mutant GTPase KRAS^{G12C}, and is a clinical development candidate for the treatment of KRAS^{G12C} positive tumors.^{553, 554} In this section we determine the complete ensemble atomic-level structure of the amorphous drug AZD4625 through the combination of DNP-enhanced solid-state NMR, molecular dynamics and machine learned chemical shifts. To do this we introduce a general approach that integrates multiple chemical shifts and includes the experimental spread of chemical shift distributions in NMR spectra of molecular solids, that we use to select an ensemble of local molecular environments that best match the chemical shift distributions in the measured spectra. This process is applied to over one million molecules from MD simulations, for which we predict chemical shifts. From an analysis of the extracted ensemble of local molecular environments in best agreement with the experiments, we identify key intermolecular interactions and conformations present in the amorphous sample. The local atomic environments determined by NMR were found to accurately reproduce the radial distribution function measured for the sample by powder X-Ray diffraction, and to correspond to energetically favourable local structures.

4.3.2 Methods

Synthesis. The synthesis of AZD4625 is described in Ref. 554. The amorphous AZD4625 solid was precipitated from 2-methyltetrahydrofuran (2-MeTHF) and n-heptane. Crude API was initially dissolved in 2-MeTHF, the solution of which was charged directly to n-heptane at 18°C. The precipitate was isolated under vacuum and dried from 25–70°C.

X-ray diffraction experiments. Synchrotron X-ray PDF data were collected on the I15-1 beamline at Diamond Light Source, UK. Powdered samples were contained within a 1 mm inner diameter polyimide capillary with a 0.025 mm wall thickness and spun perpendicular to the beam during data collection. An empty capillary was also collected for background subtraction. Scattering data were collected at an incident X-ray energy of 76.69 keV with one Perkin Elmer XRD4343CT area detector placed close to the sample (~200 mm) for PDF data and a second Perkin Elmer XRD1611CP3 area detector was placed further from the sample (~850 mm) for higher resolution Bragg data. The precise detector geometries were calculated using DAWN⁵⁵⁵ from data collected on a crystalline standard (NIST SRM640c). Total data collection times were 30 minutes for the PDF data and 2 minutes for Bragg data. 2D scattering data were corrected for polarisation, solid angle and detector thickness prior to integration to 1D using DAWN.⁵⁵⁵ The GudrunX program was then used to perform container background, multiple scattering, Compton scattering and absorption corrections on data in the range $0.3 \leq Q \leq 26 \text{ \AA}^{-1}$, prior to Fourier transform to produce the PDF.⁵⁵⁶

NMR experiments. Experiments were carried out using either room temperature ultra-fast MAS techniques that enhance ¹H spectral resolution or Dynamic Nuclear Polarisation (DNP) approaches that enhance the sensitivity of NMR signals. DNP is performed at temperatures of ~100 K and relies on the transfer of high electron spin polarisation, typically from exogenously added solutions of organic radicals, to nuclei of interest upon microwave irradiation.^{413, 491, 509, 516}

The DNP-enhanced NMR experiments were carried out on commercial Bruker Avance Neo NMR spectrometers at a nominal field strength of 9.40 T equipped with either a 264 GHz klystron or a 263 GHz gyrotron microwave source and a 3.2 mm LTMAS DNP probe in a ¹H/¹³C/¹⁵N configuration which was cooled to about 100 K before sample insertion. The DNP sample was packed into a 3.2 mm sapphire rotor, plugged with a Teflon insert, and topped with a zirconia drive cap. Prior to packing, the powder sample of the amorphous form of AZD4625 was ground by hand in a pestle and mortar and then impregnated^{413, 491, 509, 516} with a 20 mM solution of the AMUPol biradical⁵¹⁴ dissolved in a mixture of H₂O:D₂O:¹²C-glycerol (10:30:60 v/v). A DNP enhancement of the drug of a factor 6–8 was achieved, measured as the ratio of the (¹H)¹³C cross-polarisation (CP) signal intensity between spectra acquired with and without microwaves. While this is a modest enhancement, it was sufficient to enable the acquisition of the natural abundance ¹³C-¹³C INADEQUATE experiments described below. DNP spectra were acquired at MAS rates of 8 or 10 kHz.

The room temperature NMR experiments were performed on a dry sample of the powder at a MAS rate of 100 kHz, using a Bruker 0.7 mm room temperature HCN CP-MAS probe at a magnetic field of 21.1 T. A States-TPPI acquisition scheme was used to obtain phase-sensitive two-dimensional spectra. The ^1H and ^{13}C chemical shifts were referenced to literature values. More experimental details and a link to the raw NMR data can be found in **Appendix VIII**.

Chemical shift assignment. The ^1H and ^{13}C resonances of the amorphous form of AZD4625 (**Figure 4.10A**) were assigned using one-dimensional ^1H and ^{13}C MAS NMR experiments, ^{13}C CPPI spectral editing,⁴⁰⁴ (**Figure 4.10B-D, F**), in combination with two-dimensional ^1H - ^1H , ^{13}C - ^{13}C , and ^1H - ^{13}C correlation spectra. The ^1H - ^1H DQ/SQ (**Figure 4.10E**) spectrum provides through-space dipolar correlations between protons, the natural abundance DNP-enhanced refocused ^{13}C - ^{13}C INADEQUATE⁴¹³ (**Figure 4.10H**) provides the covalent connectivities between carbon atoms, and the short- and long-range ^1H - ^{13}C DNP-enhanced DUMBO-HETCOR experiments (**Figure 4.10G, I**), provide ^1H - ^{13}C heteronuclear shift correlations. A DNP-enhanced natural abundance ^{13}C - ^{13}C INADEQUATE spectrum recorded for a crystalline form was also used to guide the assignment (**Figure 4.16**). The chemical shift assignments obtained from an analysis of these spectra for the ^1H and ^{13}C nuclei are given in **Table 4.12**. The chemical shift of C1 was not taken into consideration in the subsequent analysis due to a high uncertainty in the assignment.

Molecular dynamics simulation of AZD4625. The amorphous structure of AZD4625 was modelled by carrying out MD simulations with the OPLS4 force-field⁵⁵⁷ in Desmond^{558, 559} on periodic amorphous cells containing 128 molecules. Eight different amorphous cell simulations were generated and evaluated using Materials Studio.⁵⁶⁰ After equilibration for 1 ns using the canonical NVT ensemble first at 100 K and then at 298 K followed by 22 ns using the isothermal-isobaric ensemble (NPT) at 298 K and 1 bar, production simulations were carried out for 500 ns using the NPT ensemble at 298 K and 1 bar. Snapshots of each MD simulation were extracted every 100 ps and input directly to ShiftML2³⁶⁵ for ^1H and ^{13}C chemical shift predictions. The chemical shielding values were converted to chemical shifts using offsets of 30.78 and 170.04 for ^1H and ^{13}C , respectively. Further information about the MD simulations is given in **Appendix VIII**.

Selection of local molecular environments. Local molecular environments, comprising a central molecule and all other molecules having at least one atom within 7 Å from any atomic site in the central molecule, were extracted from the MD snapshots (1,025,280 environments in total) and selected based on the probability of the molecule at the centre of each environment to match the experimental distributions of chemical shifts. Considering one atomic site a_i in AZD4625, we describe the associated distribution of experimental chemical shifts as a Gaussian function centered on the chemical shift experimentally measured, δ_{exp,a_i} , and with a width given by the linewidth of the peaks observed in the spectra, σ_{exp,a_i} . Based on the measurement of the linewidths in the resolved peaks in the spectra of **Figure 4.10**, here we obtained widths between 2 and 6 ppm for the ^{13}C resonances, and 0.6 and 1 ppm for the ^1H resonances, except for the OH proton for which we obtained a width of 1.8 ppm. The centres and widths of the experimental chemical shift distributions are given in **Table 4.12** and **Figures 4.17-4.20**.

The chemical shift $\delta_{\text{pred},a_i^{(j)}}$ and uncertainty $\sigma_{\text{pred},a_i^{(j)}}$ predicted using ShiftML2 for that atomic site $a_i^{(j)}$ in a molecule j within a given MD snapshot can similarly be described as a Gaussian function centered on the shift prediction and with a width given by the prediction uncertainty. We then define the probability that the computed shift is within the experimental distribution of chemical shift with the two-tailed p-value resulting from the Z-score computed between the two Gaussians,

$$Z_{a_i^{(j)}} = \frac{|\delta_{\text{exp},a_i} - \delta_{\text{pred},a_i^{(j)}}|}{\sqrt{\sigma_{\text{exp},a_i}^2 + \sigma_{\text{pred},a_i^{(j)}}^2}}. \quad (4.4)$$

The p-value $p_{\text{val}}(Z_{a_i^{(j)}})$ thus corresponds to the probability that the computed shift is drawn from the experimental distribution of chemical shift for that atomic site,

$$p_{\text{val}}(Z_{a_i^{(j)}}) = \sqrt{\frac{2}{\pi}} \cdot \int_{Z_{a_i^{(j)}}}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx. \quad (4.5)$$

We note that the p-value corresponds to the null hypothesis, which is here that the shift is drawn from the experimental distribution. A large p-value thus indicates a better correspondence between the predicted shift and experimental distribution. To obtain the probability that the computed shift corresponds to the experimental distribution of shifts, we divide the p-value obtained by the prediction uncertainty divided by the first quartile of all predicted uncertainties obtained for that atomic site in all molecules of all MD snapshots, $\sigma_{\text{pred},a_i}^0$, capped to a minimum value of 1. This step was done in order to prevent chemical shifts predicted with very high uncertainty, thus where the shift prediction is unreliable, from being artificially associated with a high probability of corresponding to the experimental distribution.

$$p_{a_i^{(j)}} = \frac{p_{\text{val}}(Z_{a_i^{(j)}})}{\max\left(1, \frac{\sigma_{\text{pred},a_i^{(j)}}}{\sigma_{\text{pred},a_i}^0}\right)}. \quad (4.6)$$

The probability p_j that a given molecular environment j within an MD snapshot corresponds to the experimental spectrum was then evaluated as the geometric mean of the probabilities obtained using **Equation 4.6** for protons and carbons in the molecule, (except here for the protons and carbon labelled 1 in **Figure 4.10A**, due to the high uncertainty in the assignment of that carbon). This probability was computed for all local environments in all MD snapshots.

$$p_j = \left(\prod_i^n p_{a_i^{(j)}}\right)^{\frac{1}{n}}. \quad (4.7)$$

The selection of the ensemble of local molecular environments most compatible with the experimental spectra, that we refer to as the NMR ensemble, was then performed by selecting all the environments having an overall probability p_j above 0.33, corresponding to about 1% of all local molecular environments present in the MD snapshots (10,107 environments). We note that the cutoff value of 0.33 was chosen as a balance between the maximisation of the overlap and minimisation of the Jensen-Shannon divergence⁵⁶¹ with the experimental shift distributions, and the selection of a large enough ensemble to describe the amorphous compound (see **Figure 4.21**).

In addition, 1,000 local molecular environments were randomly selected from each MD simulation to construct a random ensemble for comparison with the experimentally determined ensemble.

Computation of formation energies of local molecular environments. The formation energy of local molecular environments was computed as the energy difference between the environments (all molecules with at least one atom within 7 Å from any atom of the central molecule) with and without the central molecule. This energy thus includes both the intermolecular interactions and conformational energy of the central molecule. The energies were computed using the DFTB-D3H5 semiempirical level of theory using the 3ob-3-1 parameter set and the DFTB+ software version 22.2.^{325, 326, 349-352, 562}

Identification of hydrogen bonds in local molecular environments. Hydrogen bonds involving the OH proton of the central molecule in each local molecular environment were identified by defining hydrogen bonds as O-H...X motifs (X = O, N) with an O-H-X angle above 130° and H-X distance shorter than 2.5 Å.

Three-dimensional atomic density maps. The three-dimensional atomic density maps were constructed by aligning the NMR ensemble of local molecular environments and a second randomly selected ensemble on given atoms in the central molecule. This was done by minimising the root-mean-square displacement (RMSD) between the positions of the atoms used for the alignment in the central molecule of the different molecular environments. Three-dimensional atomic density maps were then generated by summing three-dimensional Gaussian functions with a width $\sigma = 0.5$ Å placed at the atomic positions r_{a_i} of the aligned local environments, divided by the number of environments aligned,

$$G(\vec{r}) = \frac{1}{N_{\text{env}}} \sum_i^{N_{\text{env}}} \sum_{a_i \in i} \exp\left(-\frac{\|\vec{r} - \vec{r}_{a_i}\|^2}{2\sigma^2}\right). \quad (4.8)$$

Individual atomic density maps were constructed for each element present in the set of aligned environments. The Gaussian functions were not normalised, and this leads to a value of 1 at a given position if an atom of a given element is found at that position in all environments. Each atomic density map was evaluated on a 31x31x31 cubic grid centered at the aligned atomic sites and with 12 Å sides. This corresponds to a spatial sampling of 0.4 Å.

4.3.3 Results and Discussion

Figure 4.10 shows the chemical structure of AZD4625 and the labelling scheme used here, as well as the experimental 1D and 2D NMR spectra obtained for the amorphous form of AZD4625. The spectra display broad linewidths, typical of disordered systems. This highlights the need for multi-dimensional experiments in order to obtain a confident assignment, by spreading the signals over multiple dimensions. With this set of spectra, the ^1H and ^{13}C chemical shifts obtained were assigned as described in **Section 4.3.2**, leading to the assignments given in **Table 4.12**. By fitting Gaussian functions to resolved peaks in the 1D ^1H and ^{13}C MAS spectra, and 2D ^1H - ^1H DQ/SQ spectrum, we obtained linewidths between 2 and 6 ppm for ^{13}C , 0.6 and 1 ppm for C-H protons, and 1.8 ppm for the OH proton (see **Table 4.12** and **Figures 4.17-4.20**). Here, we assume Gaussian shapes for all experimental distributions of chemical shifts. The extracted experimental chemical shift distributions will then serve as the basis to score molecular environments as described in **Section 4.3.2**. No crystalline form of pure AZD4625 has previously been reported.

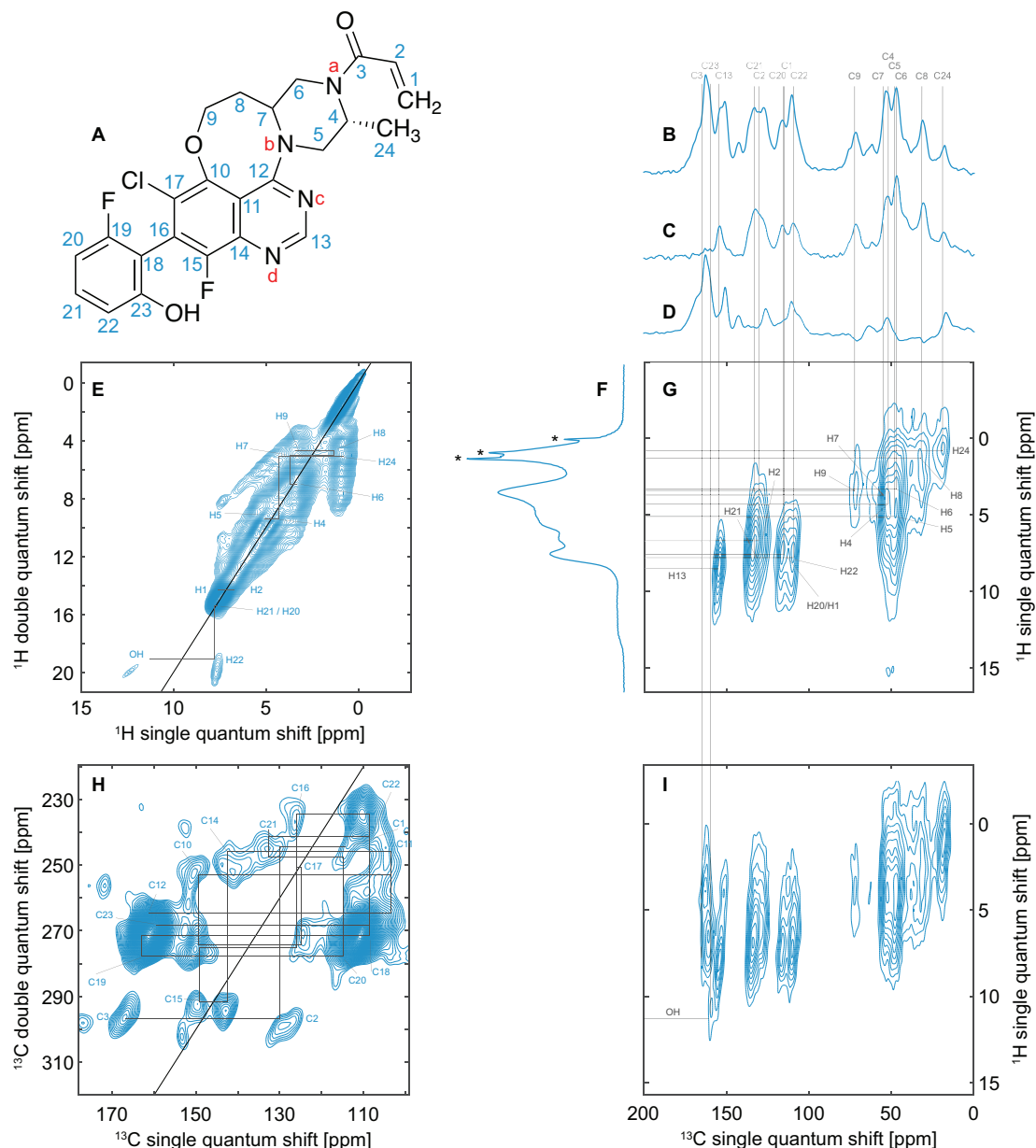


Figure 4.10. NMR spectra of the amorphous form of AZD462 used for chemical shift assignment. (A) Chemical structure of AZD4625 and carbon (blue numbers) and nitrogen (red letters) labelling schemes used here. 1D (B-D) DNP-enhanced ^{13}C CPMAS spectra without (B, C) and with (D) CPPI spectral editing. (F) 1D ^1H 100 kHz MAS spectrum. 2D (E) ^1H - ^1H DQ/SQ, (G, I) DNP-enhanced ^1H - ^{13}C DUMBO-HETCOR and (H) DNP-enhanced ^{13}C - ^{13}C INADEQUATE spectra of amorphous AZD4625. In (D), $-\text{CH}_2$ groups appear negative, $-\text{CH}$ groups disappear and $-\text{C}$ and $-\text{CH}_3$ groups retain a positive intensity. The grey lines indicate correlated peaks or ^{13}C chemical shifts of protonated carbon species. The stars in (F) indicate artifacts due to mobile impurities in the rotor.

To generate a broad ensemble of possible structures, eight MD simulations were carried out with cells containing 128 molecules of AZD4625, randomly initialised in order to model the amorphous system, as described in **Section 4.3.2**. Chemical shift predictions performed using ShiftML2 were then compared with the experimental values obtained for ^1H and ^{13}C (excluding the protons and carbon labelled 1 in **Figure 4.10A** due to the ambiguity in their assignment). A total of 1,025,280 molecular environments, each comprising a central molecule and all molecules that have at least one atom within 7 Å from any atom of the central molecule (see **Section 4.3.2**), were extracted from the MD snapshots. For each atomic site in the central molecule of a molecular environment, we compute the probability that the predicted shift is drawn from the corresponding experimental chemical shift distribution. The probabilities across all atomic sites are then combined into a global probability that the local molecular environment matches the NMR experiments. More details are given in **Section 4.3.2**. **Figure 4.11A** shows the root-mean-square error (RMSE) between ^1H and ^{13}C chemical shifts computed for all AZD4625 molecules in each of the 8,010 snapshots taken from the MD trajectories, as well as the calculated probability that the local molecular environment of each molecule is consistent with the NMR experiments. This includes the computation of chemical shifts for over a million molecules. As expected, higher probability is correlated with lower ^1H and ^{13}C shift RMSE, but it is very important to note that the RMSEs only considers the difference between the centre of the experimental distributions of shifts, and the corresponding chemical shift prediction for each atomic site, while the probability calculated using **Equations 4.4-4.7** also takes into account the width of the experimental distributions as well as the prediction uncertainty, providing an improved picture of the compatibility of a given local molecular environment with the experiments. The histogram of all probabilities of local molecular environments (p_j) to match the experiments is shown in **Figure 4.11B**. Here, we selected the 1% of local molecular environments in best agreement with experiment to construct the NMR ensemble, which corresponds to probabilities above 33%, as indicated by the dashed vertical line in **Figure 4.11B**.

Here, we independently select molecular environments compatible with the NMR experiments. The generation of environments through the MD simulations is inherently biased by the force field used and the starting configurations. The selection of the subset that best matches the experimental data does not aim here to reproduce the exact experimental ensemble of molecular environments in the sample (as is done, e.g., in NMR studies of intrinsically disordered proteins⁵⁶³⁻⁵⁶⁵), but here it provides an additional bias in order to identify systematic structural differences from the ensemble generated by MD, as seen below.

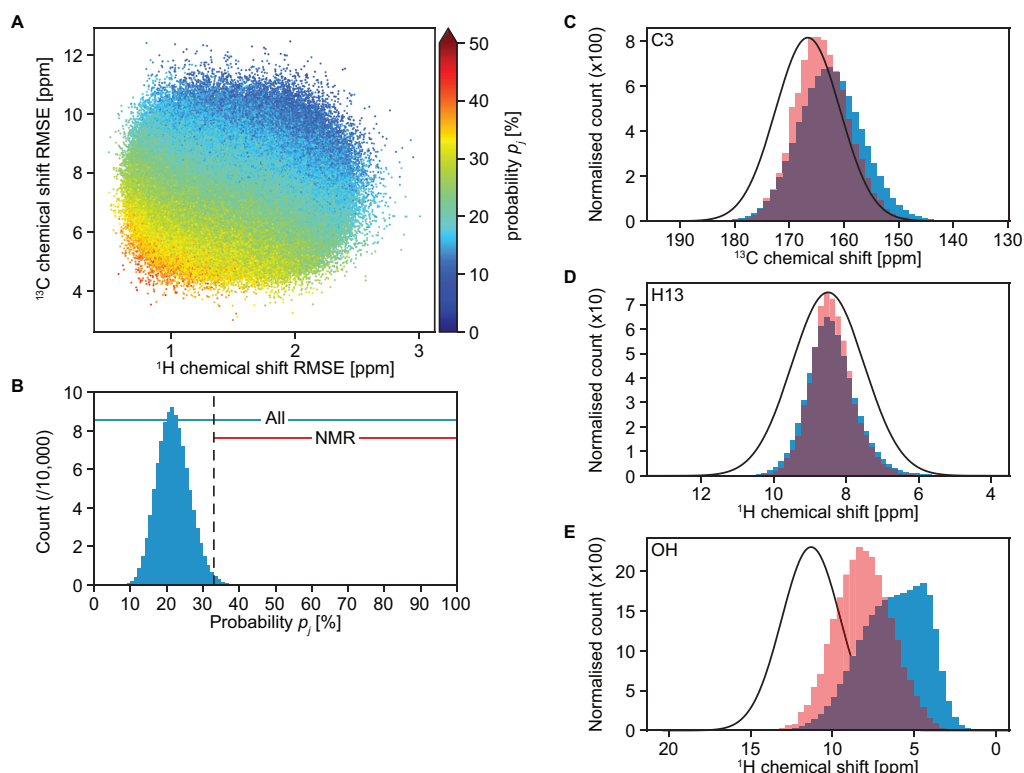


Figure 4.11. Ensemble structure determination. (A) A comparison of ^1H and ^{13}C chemical shift RMSEs for each molecule in the MD snapshots, coloured according to its probability to be simultaneously compatible with the experimental shift distributions for all assigned atoms (as described by **Equations 4.4-4.7**). (B) Histogram of the probabilities of all molecules in the MD snapshots to be compatible with the experimental shift distributions. The dashed line indicates the probability threshold used to select local molecular environments. The ranges of probabilities included in the whole and NMR ensembles are indicated above the histogram. Examples of the predicted chemical shift distributions for the (C) carbon labelled 3, (D) proton labelled 13 and (E) OH proton in all molecular environments (blue) in the MD snapshots and in the NMR ensemble (red), compared to the corresponding experimentally measured distributions (black). Equivalent figures for all the other assigned atoms are given in **Figures 4.22-4.26**.

Figure 4.11C-E shows the histograms of chemical shifts computed for carbon labelled 3, proton labelled 13 and of the OH proton for all AZD4625 molecules in the MD trajectories as compared to those from the NMR ensemble. These examples are taken to illustrate the typical changes of chemical shift distributions seen upon selection of local atomic environments. The distributions for all other protons and carbons considered are given in **Figures 4.22-4.26**. The distribution of predicted shifts for carbon labelled 3 (**Figure 4.11C**) was found to be significantly closer to the experimental distribution of shifts upon selection of local molecular environments, suggesting that this chemical shift does discriminate between the structures. In contrast, for example, the distribution of predicted shifts for the proton labelled 13 (**Figure 4.11D**), which already displays a large overlap with the corresponding experimental distribution of shifts, does not display a significant change upon selection of local molecular environments. Then we note that the distribution of predicted chemical shifts for the OH proton (**Figure 4.11E**) displays a large difference after the selection of local molecular environments, again suggesting that this shift is a powerful discriminator. However, even after selection of the best match structures, the overlap with the predicted distribution is not perfect. We attribute this to the significant proportion of OH protons weakly bonded to hydrogen bond acceptors in the MD trajectories (see **Figure 4.27**). This effect may also be due to bias in the shift predictions. Importantly, we also note that the best match selection does not critically depend on any single shift, but is the result of the joint match to all the shifts in the molecule.

Figure 4.12 shows the analysis of structural properties in the set of best match molecular environments, compared to all molecular environments present in all MD snapshots. As seen in **Figure 4.12A**, the selection of local molecular environments compatible with the NMR experiments promotes hydrogen bonds, in particular with the oxygen labelled 3 and the nitrogen labelled c. Accordingly, the proportion of OH protons not forming hydrogen bonds is significantly reduced in the set of selected local molecular environments. Hydrogen bonding to nitrogen was found to generally lead to further deshielding of the OH proton compared to hydrogen bond to an oxygen, as seen in **Figure 4.27**.

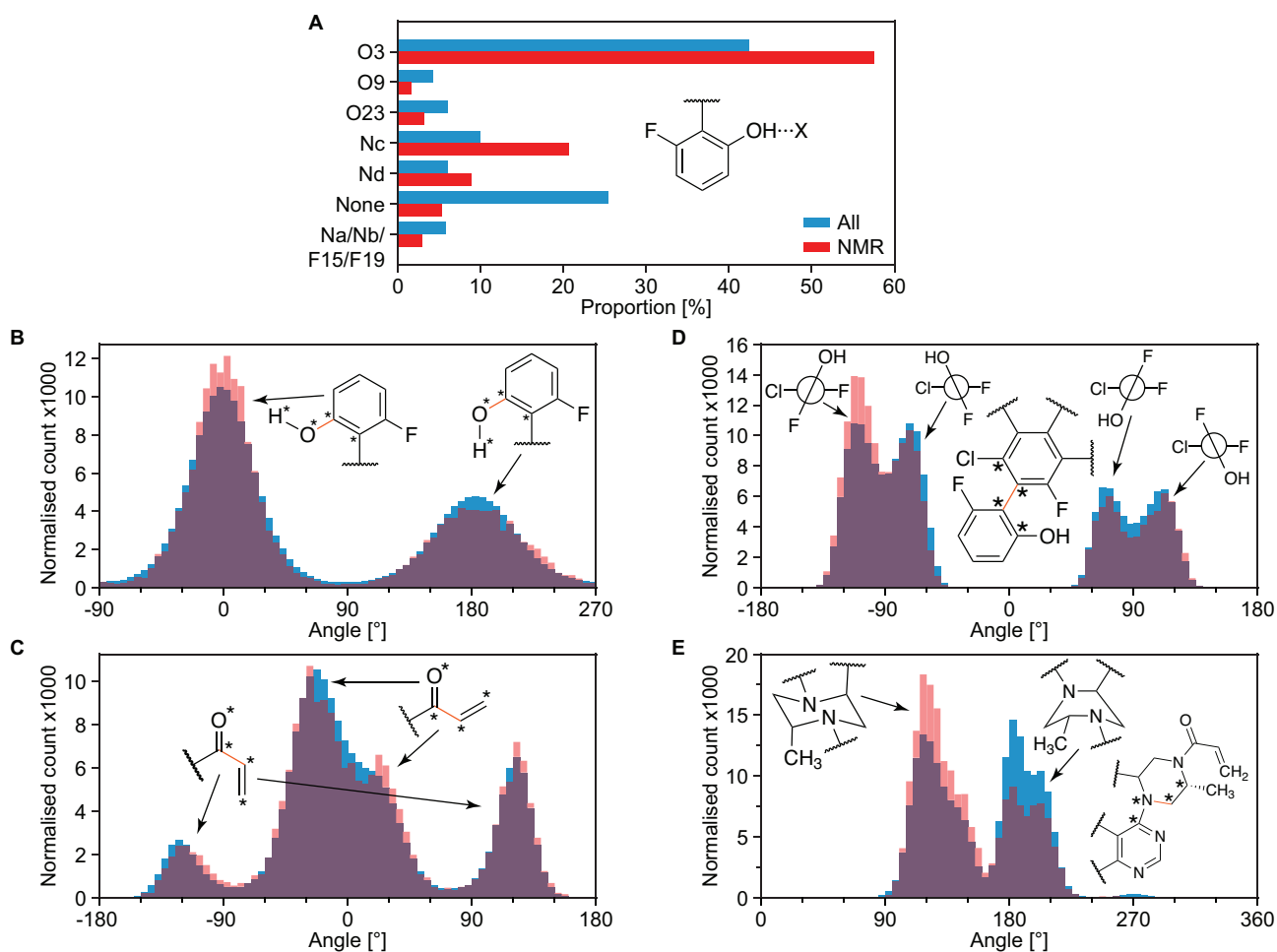


Figure 4.12. Structural properties of the amorphous form of AZD4625. **(A)** Proportions of different hydrogen bond acceptors bonded to the OH group of AZD4625 in all local molecular environments (blue) and in the NMR ensemble (red). Histogram of dihedral angles for the **(B)** OH group, **(C)** enone, **(D)** aromatic planes and **(E)** aliphatic ring in all molecules (blue) and in the NMR ensemble (red). In **(B)**, **(C)**, **(D)** and **(E)**, the rotatable bond associated with the dihedral angle is drawn in orange. Stars indicate the atoms used for the computation of the dihedral angle.

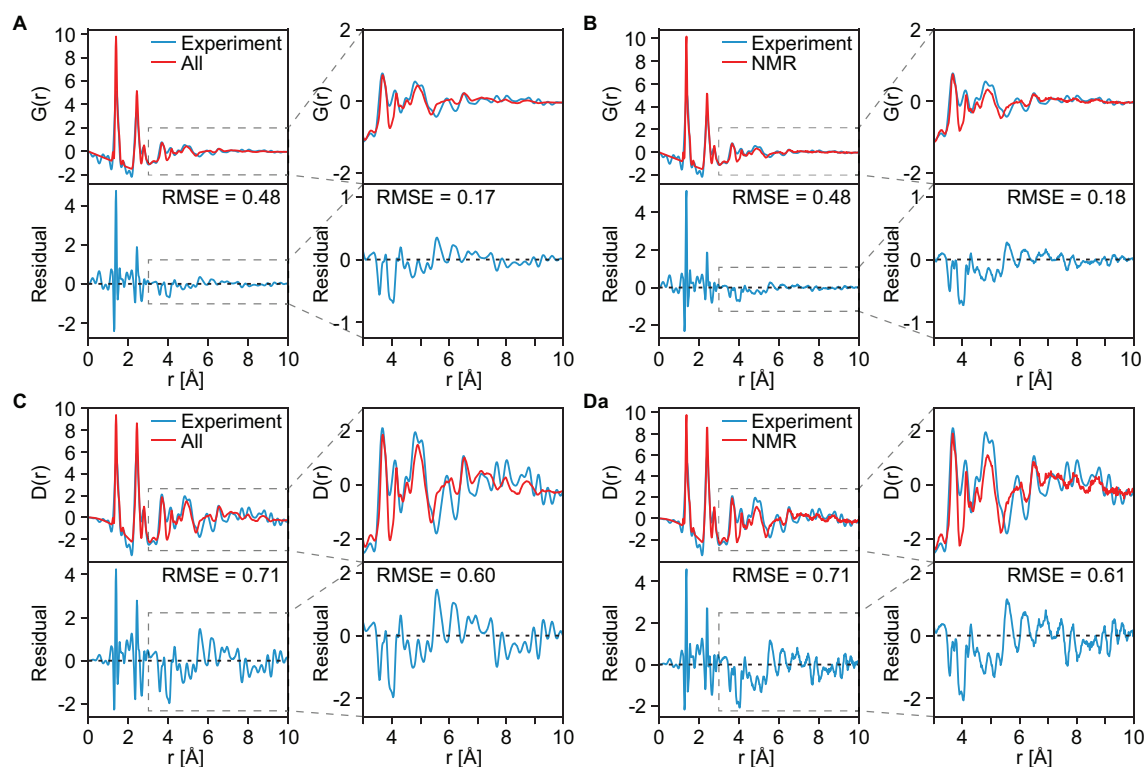


Figure 4.13. Radial distribution functions. The total radial distribution function $G(r)$ (A, B) and the differential correlation function $D(r)$ (C, D) measured from powder X-Ray diffraction (see Section 4.3.2) (blue) and simulated (red) using (A, C) all molecules and (B, D) the best match ensemble by NMR. The lower panels show the residual between the experiment and simulations in each case, along with the RMSE obtained. The plots on the right of each panel shows the range between 3 and 10 Å and the RMSE in the corresponding range.

Preferred conformations of AZD4625 can be extracted from the NMR ensemble. **Figure 4.12B** shows that the position of the OH proton is generally preferred to be pointing away from the body of the molecule, and that this trend is slightly reinforced in the NMR ensemble. Similarly, the Z conformation of the enone group is found to be preferred, and that preference is retained in the NMR ensemble (**Figure 4.12C**). The conformation yielding dihedral angles between the aromatic planes from -120 to -60° were found to be promoted in the NMR ensemble (**Figure 4.12D**). We note that for this case, five of the eight MD simulations carried out started with a dihedral angle around -90° and three of them started with an angle around 90° , which explains the difference in the height of the distributions for positive and negative values in all molecules from the MD snapshots (more details are given in **Appendix VIII**). The chair conformation of the aromatic 6-membered ring was also found to be promoted by the NMR selection of local molecular environments compared to the boat conformation that was also observed in the MD simulations (**Figure 4.12E**).

It is interesting to compare the total radial distribution function $G(r)$ and differential correlation function $D(r)$ obtained from the ensembles before and after selection of local molecular environments with the functions obtained experimentally by powder X-Ray diffraction (**Figure 4.13**). The MD trajectories were found to accurately reproduce the experimental data, with the largest differences found in the two peaks at 1.4 and 2.4 Å. This can be attributed to differences in bond lengths between the MD simulations and in the sample. Importantly, the features at distances above 3 Å are correctly captured by the MD simulation. The selection of local molecular environments was not found to significantly change the similarity between the simulated and experimental $G(r)$ or $D(r)$. This result highlights that the scattering data is unable to sensitively discriminate between ensembles of local molecular environments in the samples studied here.

Figure 4.14 shows the predicted formation energies of molecules of AZD4625 with their local environment, including the formation energy of the central molecule (as described in Section 4.3.2). This is a measure of the stabilisation of the molecules by their environment. On average, the local environments in the NMR ensemble were found to result in the stabilisation of the central molecule by 8.7 ± 0.7 kJ/mol as compared to random local molecular environments extracted from the MD simulations (**Figure 4.14A**). This result suggests that the selection of molecular environments, based purely on NMR chemical shifts, also led to the selection of energetically favourable local molecular environments. **Figure 4.14B** shows that hydrogen bonding of the OH proton of a central molecule to either oxygen labelled 3 or nitrogen labelled d leads to enhanced stabilisation of the central molecule by its whole environment. This also corroborates the increase in hydrogen bonds formed with these two atoms in the NMR ensemble of molecular environments discussed above (**Figure 4.12A**).

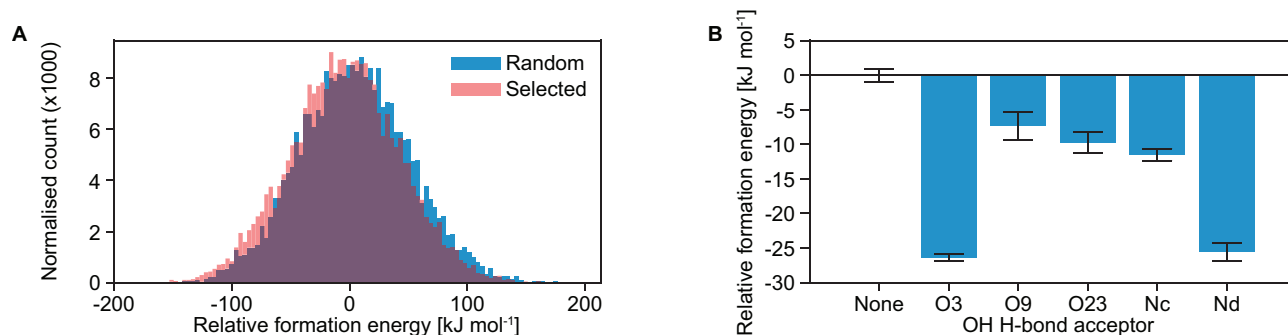


Figure 4.14. Formation energies. (A) Relative formation energies of intermolecular complexes of 8,000 randomly selected molecules (blue) and the molecules from the NMR ensemble (red). The zero is set to be the mean formation energy of all intermolecular complexes. (B) Relative formation energies of the local molecular environments in the NMR ensemble for different hydrogen bond acceptors bonded to the OH proton. The zero is set to be the mean formation energy of intermolecular complexes where no hydrogen bonding acceptor is bonded to the OH proton of the central molecule. Formation energies were computed as the difference in energy between a molecular environment (all molecules with at least one atom within 7 Å from any atom of the central molecule) with and without the central molecule, thus contains both intermolecular interactions and conformational energy of the central molecule.

A set of 20 randomly selected central molecules from the NMR ensemble is shown in **Figure 4.15A**. This highlights the structural flexibility of AZD4625 in the amorphous state. **Figure 4.15B** shows three-dimensional atomic density maps around the OH proton in the NMR (left panel) and the random (middle panel) local molecular environments, as well as the difference between the two atomic density maps (right panel). As expected from **Figure 4.12A** and **Figure 4.14B**, hydrogen bonding towards oxygen and nitrogen atoms is promoted by the selection of local molecular environments. This is highlighted by the contours representing nitrogen and oxygen atomic densities in the rightmost panel in **Figure 4.15B**. This suggests that these interactions are critical to stabilise the structure of amorphous AZD4625. **Figure 4.15C** shows similar atomic density maps, aligned around the methyl group of AZD4625. The difference between atomic density maps highlights the preferred conformation of the 6- and 8-membered aliphatic rings.

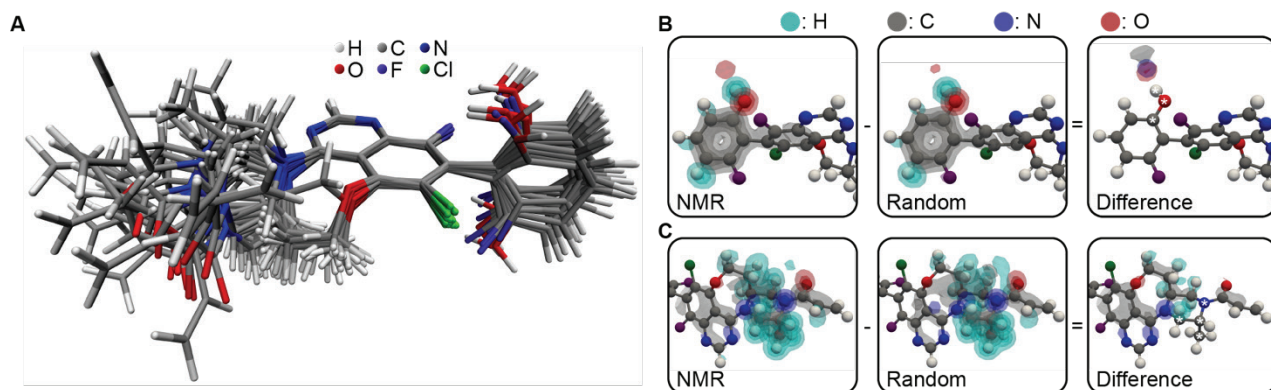


Figure 4.15. Structures representative of the molecular conformations present in the amorphous form of AZD4625. (A) Superposition of 20 molecules of AZD4625 randomly selected from the NMR ensemble. Three-dimensional atomic density maps in NMR-selected and random molecular environments aligned around (B) the OH and (C) the methyl groups. The difference between the 3D maps for the selected and random molecular environments are shown on the right panels, where the atoms aligned are indicated by asterisks in the difference maps. 3D contours are drawn at levels of 0.2, 0.4, 0.6 and 0.8 for the atomic density maps and 0.05, 0.1, 0.15 and 0.2 for the difference maps. The conformation of the molecule displayed along with the atomic density map was chosen such that the various dihedral angles best correspond to the maxima of the distributions for selected local molecular environments in **Figure 4.12B-E**. The three-dimensional contours in the rightmost panels in (B) and (C) highlight the overall structural features promoted by the NMR-based selection.

4.3.4 Conclusion

In this section we have determined the ensemble atomic-level structure of the amorphous form of AZD4625 by combining solid-state NMR experiments with MD simulations and prediction of chemical shifts for over one million AZD4625 molecules in the MD trajectories. Importantly, no crystalline structure of the pure compound has previously been reported.

Local molecular environments compatible with the measured NMR spectra measured were selected through a general approach that integrates multiple chemical shifts, and includes the spread of chemical shift distributions in the experimental spectra as well as the uncertainty of the chemical shift predictions. We expect that the method presented here can be straightforwardly applied to determine the structure of any molecular solid.

The local atomic environments determined by NMR were found to accurately reproduce the radial distribution function measured for the sample by powder X-Ray diffraction. The NMR ensemble was also found to lead to an overall stabilisation of the selected molecules by their environment, observed using approximated calculations of molecular cluster energies.

The ensemble of selected local molecular environments highlights key structural properties in the amorphous sample that play a critical role in the structure and stabilisation of the material in its amorphous form.

4.3.5 Appendix VIII

DATA AVAILABILITY

The NMR raw data are available from the Materialscloud repository <https://doi.org/10.24435/materialscloud:gk-51> in JCAMP-DX version 6.0 standard format and original TopSpin format, as well as the input files for the MD simulations, the MD snapshots extracted, formation energies of intermolecular complexes, and all scripts used to perform the data analysis. All data and scripts are available under the license CC-BY-4.0 (Creative Commons Attribution-ShareAlike 4.0 International).

EXPERIMENTAL DETAILS

1D ^1H MAS experiment. A one rotor period rotor-synchronised spin echo sequence, for background suppression, was used for acquisition. Pre-saturation was applied prior to excitation. No weighting function was applied upon processing.

Table 4.7. Experimental details of the 1D ^1H MAS experiment.

MAS rate (kHz)	VT (K)	^1H 90° RF (kHz)	d1(s)	TD	SW (kHz)	SI
100	278	312.5	5	4096	227.3	8192

1D DNP-enhanced ^{13}C CPMAS experiments. A conventional cross-polarisation (CP)⁴⁰⁸ sequence was used for the acquisition of the spectra presented in **Figure 4.10B-C** with a contact time of 2.5 and 0.1 ms respectively. The short contact time promotes the detections of protonated carbon atoms. Pre-saturation was applied prior to excitation. For **Figure 4.10D** an editing experiment (CPPI)⁴⁰⁴ that relies on phase inversion was used for acquisition. In this sequence a second short (40 μs) cross-polarisation block right after the initial one inverts $-\text{CH}_2$ groups whereas nulls $-\text{CH}$ groups and retains $-\text{C}$ and $-\text{CH}_3$ groups with a positive intensity. For all spectra spin-al-64 ^1H decoupling⁴⁰⁹ with an rf of 71.4 kHz was applied during acquisition. A Lorentzian line broadening of 150 Hz was applied upon processing.

Table 4.8. Experimental details of the 1D ^{13}C CPMAS experiments.

Experiment	MAS rate (kHz)	VT (K)	^1H 90° RF (kHz)	d1(s)	TD	SW (kHz)	SI	Contact Power (kHz), $^1\text{H}/^{13}\text{C}$	Contact time (ms)
Figure 4.10B	10	100	71.4	2	2048	100	8192	54 / 70	2.5
Figure 4.10C	10	100	71.4	2	2048	100	8192	54/70	0.1
Figure 4.10D	10	100	71.4	3	988	100	8192	54/83.3	2.5

2D DNP-enhanced ^{13}C - ^{13}C INADEQUATE experiment. A rotor-synchronised J-based refocused INADEQUATE^{369, 411} sequence was used for acquisition. A States-TPPI acquisition scheme⁵¹⁸ was used to obtain phase-sensitive two-dimensional spectra. Spinal-64 ^1H decoupling⁴⁰⁹ with an rf of 83.3 kHz was applied only during acquisition. A Lorentzian line broadening of 300 Hz was applied upon processing to both dimensions.

Table 4.9. Experimental details of the 2D ^{13}C - ^{13}C INADEQUATE experiment.

MAS rate (kHz)	VT (K)	90° RF amplitude (kHz)	d1(s)	Number of FID points, F2/F1	SW (kHz), F2/F1	Size of real spectrum, F2/F1	J evolution time (μs)	Contact Power (kHz), $^1\text{H}/^{13}\text{C}$	Contact time (ms)
10	100	83.3	3	806/40	81.5/25	2048/256	8	71.5 / 59.8	4

2D ^1H - ^1H DQ/SQ experiment. An eight-rotor period rotor-synchronised BABAXy16³⁶⁷ sequence was used for acquisition. Pre-saturation was also applied prior to excitation. A States-TPPI acquisition scheme⁵¹⁸ was used to obtain phase-sensitive two-dimensional spectra. A Lorentzian line broadening of 100 Hz was applied upon processing to both dimensions.

Table 4.10. Experimental details of the 2D ^1H - ^1H DQ/SQ experiment.

MAS rate (kHz)	VT (K)	90° RF amplitude (kHz)	d1(s)	Number of FID points, F2/F1	SW (kHz), F2/F1	Size of real spectrum, F2/F1	DQ recoupling time (μs)
100	278	312.5	5	9090/200	227.2/33.3	16384/256	80

2D DNP-enhanced ^1H - ^{13}C HETCOR experiment. A DUMBO-HETCOR⁴¹⁰ sequence was used for the acquisition. An eDUMBO-122⁴¹⁰ element (32 μs at 71.5 kHz), applied during t_1 , increases the resolution of the indirect dimension by averaging and therefore decoupling ^1H - ^1H homonuclear dipolar couplings. The rescaling of the indirect dimension, caused by the application of the DUMBO element, was done with the aid of the 1D 100 kHz ^1H MAS spectrum. A States-TPPI acquisition scheme⁵¹⁸ was used to obtain phase-sensitive two-dimensional spectra. Spinal-64 ^1H decoupling⁴⁰⁹ with an rf of 83.3 kHz was applied during acquisition. A Lorentzian line broadening of 200 Hz was applied upon processing to both dimensions.

Table 4.11. Experimental details of the 2D ^1H - ^{13}C HETCOR experiment.

Experiment	MAS rate (kHz)	VT (K)	90° RF amplitude (kHz)	d1(s)	Number of FID points, F2/F1	SW (kHz), F2/F1	Size of real spectrum F2/F1	Contact Power (kHz), $^1\text{H}/^{13}\text{C}$	Contact time (ms)
Figure 4.10H	10	100	83.3	2	1024/128	100/52	8192/1024	83.3/59.8	0.1
Figure 4.10I	10	100	83.3	2	1024/28	100/47	8192/1024	83.3/59.8	0.5

Chemical shift assignment of amorphous AZD4625.

Table 4.12. ^1H and ^{13}C chemical shifts and widths (Gaussian σ) of amorphous AZD4625. ^a Indicates widths that represent several overlapping resonances, thus should be considered as upper bounds to the linewidths. The ^1H and ^{13}C resonances were assigned using the experimental spectra of **Figure 4.10**. Due to the amorphous character of AZD4625, the acquired spectra have broad lineshapes which reduce spectral resolution and often obscure the identification of peak maxima. However, we believe that the assignment presented here is accurate enough to be used for our further analysis. The assignment of C1 is uncertain due to the low signal-to-noise ratio of the INADEQUATE spectrum of the amorphous AZD4625. The assignment of the aliphatic carbon atoms was performed using the 1D CPMAS spectra and the INADEQUATE spectrum of a crystalline form (shown in **Figure 4.16**). The carbon chemical shifts were referenced using glycerol and the proton chemical shifts using L-histidine hydrochloride monohydrate.

Label	^1H Chemical Shift / Width (ppm)	^{13}C Chemical Shift / Width (ppm)
1	7.6/-	114/-
2	6.7/1.0 ^a	129.7/5.5 ^a
3	-	166.6/5.9 ^a
4	4.3/1.0 ^a	50.8/3.1
5	5.1/1.0 ^a	46.9/2.3
6	3.3/0.8	45.8/2.6
7	3.7/1.0 ^a	54/2.7
8	1.3/1.0 ^a	30.3/2.0
9	3.4/0.7	71.3/3.4
10	-	149.3/2.0
11	-	103.3/5.0 ^a
12	-	161/3.9 ^a
13	8.5/1.0 ^a	154/3.7 ^a
14	-	142.2/2.0
15	-	149/2.0
16	-	125.7/2.5
17	-	124.7/3.4 ^a
18	-	108.4/3.6 ^a
19	-	162.7/2.6
20	7.6/0.6	114.7/6.0 ^a
21	7.6/0.6	132.4/5.4 ^a
22	7.8/0.6	108.4/3.0
23	-	159.4/3.4 ^a
24	0.8/0.6	17.7/2.3
OH	11.3/1.8	-

Computational details for the MD simulations. To model the amorphous structure of AZD4625, we carried out MD simulations on periodic amorphous cells. The atomic positions of a single molecule were first optimised at the B3LYP-D3/6-31G(d,p) level of theory^{95, 99, 104, 519, 566, 567} in gas phase using the Gaussian 16 revision C.01 program.⁵³⁶ Optimised coordinates and CHELPG charges⁵²⁵ were extracted from the optimisation and used as input to generate amorphous cells. Materials Studio⁵⁶⁰ together with the COM-PASS-III force field⁵⁶⁸ were used to create cubic amorphous cells (43*43*43 Å) of 128 molecules placed randomly and with identical conformations in eight replicates. These multiple replicas allow the generation of a diverse set of structures. PDB files of the amorphous cells were saved as input for the MD-step. The Desmond program (Schrödinger 2021-4)⁵⁵⁸ was used for all MD simulations throughout the study employing the OPLS4 force field.⁵⁵⁷ The systems were initially equilibrated for 1 ns using the canonical (NVT) ensemble first at 100 K and then at 298 K. The temperature was held constant using a Nosé-Hoover chain thermostat^{569, 570} with a relaxation time of 1.0 ps. A second equilibration was carried out for 22 ns using the isothermal-isobaric ensemble (NPT) at 298 K and 1 bar where the temperature and pressure were held constant using the coupled Martyna-Tobias-Klein method⁵⁷¹ with a relaxation time of 1.0 ps. Production simulations were carried out for 500 ns using the NPT ensemble at 298 K and 1 bar with the same settings as in the second equilibration. Electrostatic interactions were included with a 9 Å cutoff. Trajectories were collected every 100 ps. Models of the amorphous structure were obtained by extracting evenly spaced snapshots from the last 100 ns of each MD simulation. Since no inversion of the aromatic ring containing the OH group was observed during the MD simulations, five simulations were carried out with a starting angle between the aromatic planes around -90° and three were run with a starting angle around 90°. This explains the 5:3 ratio of negative and positive angles in **Figure 4.12D**. The eight simulations (corresponding to 1,025,280 molecular environments) were assumed to fully sample the conformational and noncovalent interaction space of the molecule in the amorphous phase.

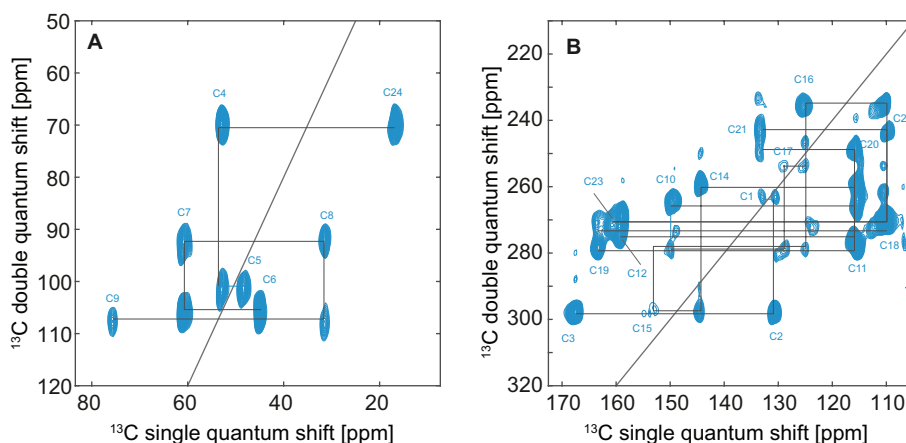


Figure 4.16. ^{13}C - ^{13}C INADEQUATE spectra of a crystalline form of AZD4625. In (A) the aliphatic and in (B) the aromatic regions are plotted.

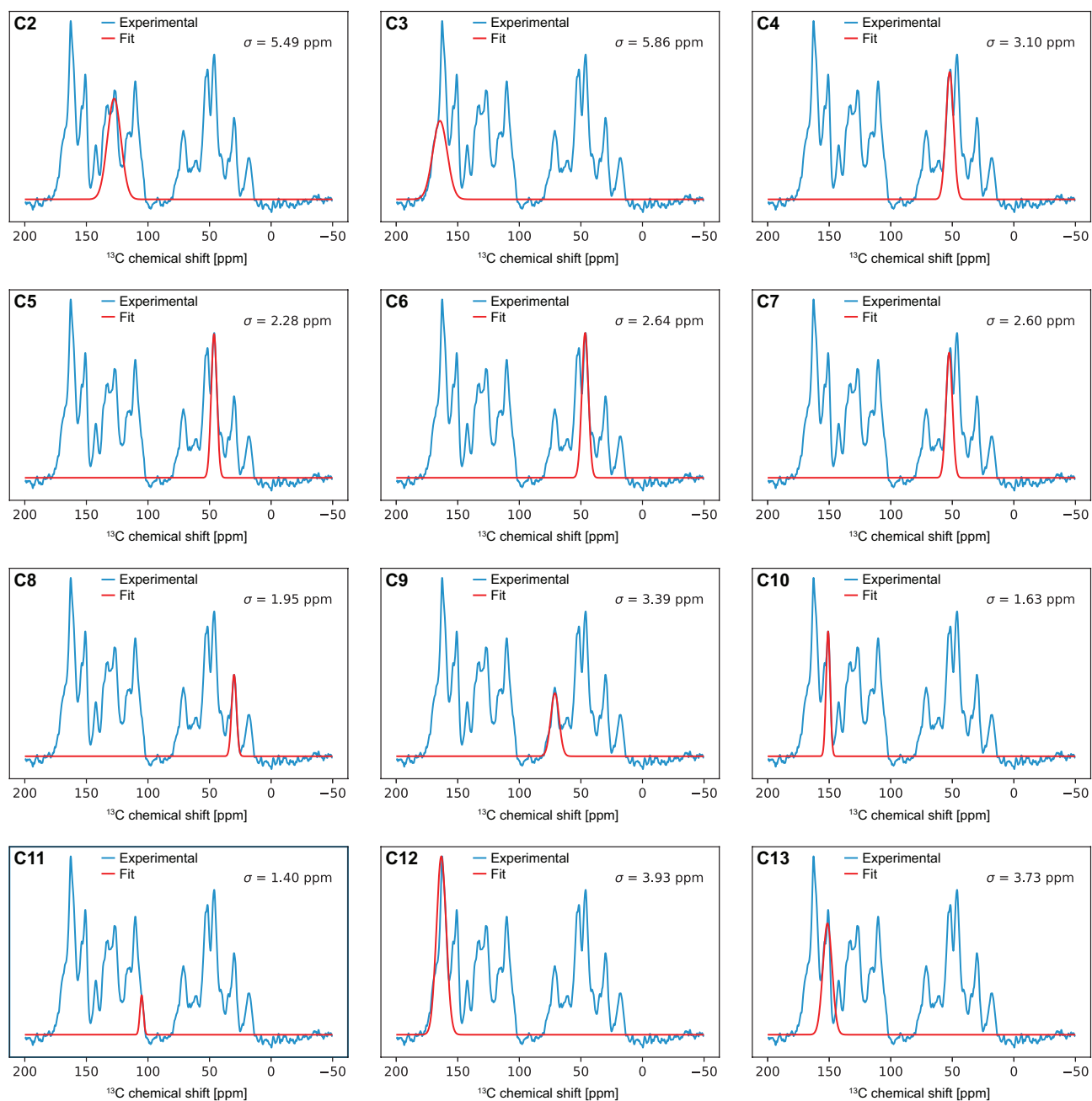


Figure 4.17. Fits of individual resonances of C2-C13 from the 1D ^{13}C CPMAS NMR spectrum of amorphous AZD4625.

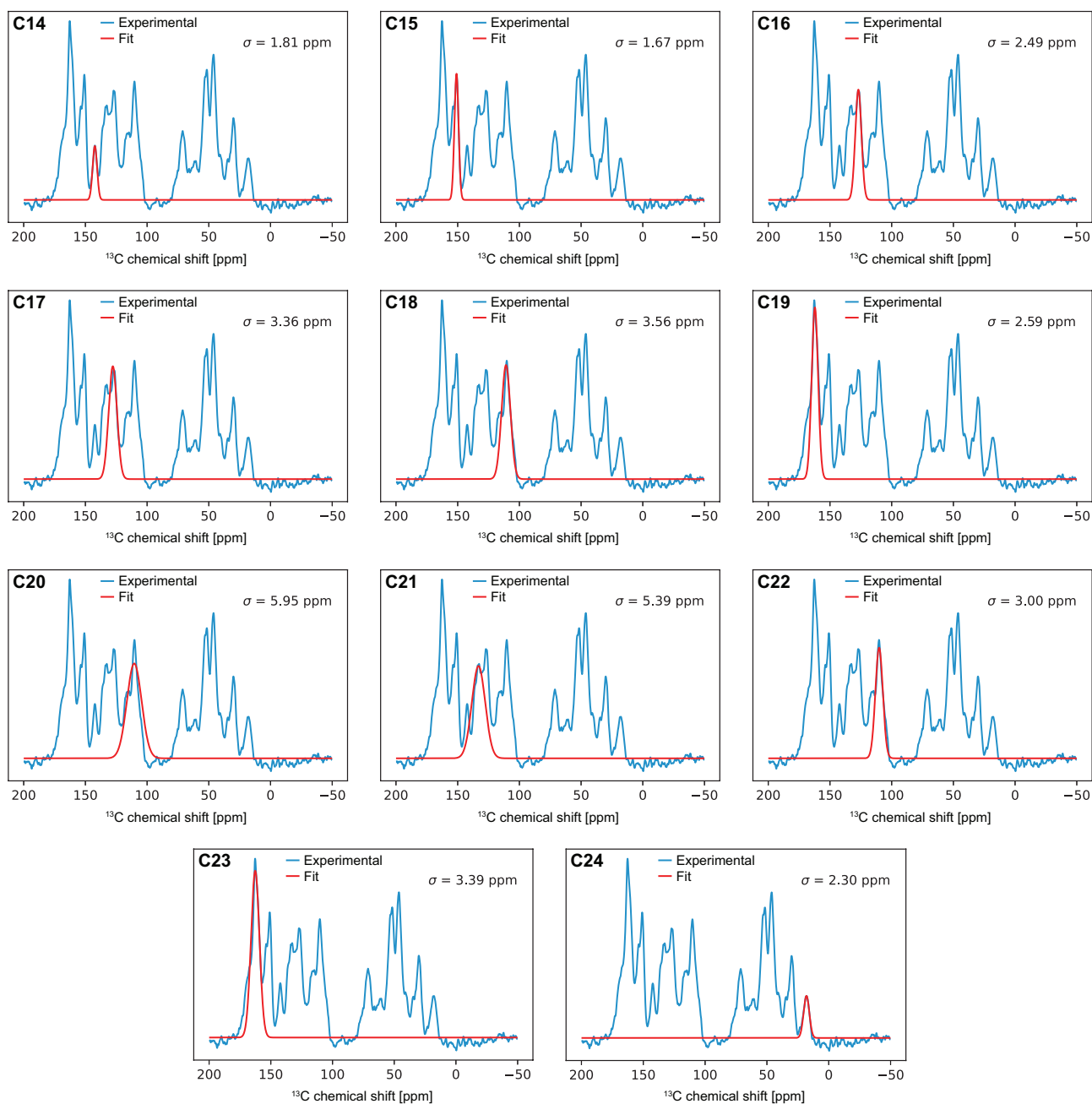


Figure 4.18. Fits of individual resonances of C14-C24 from the 1D ^{13}C CPMAS NMR spectrum of amorphous AZD4625.

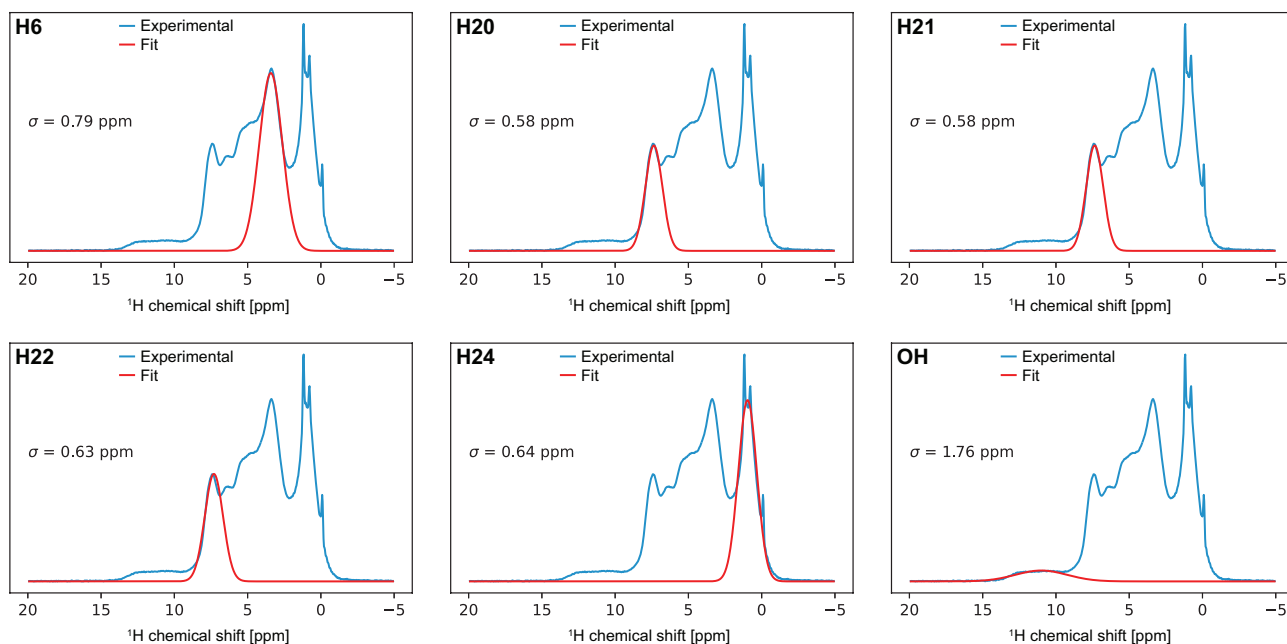


Figure 4.19. Fits of individual resonances of proton sites from the 1D ^1H MAS NMR spectrum of amorphous AZD4625.

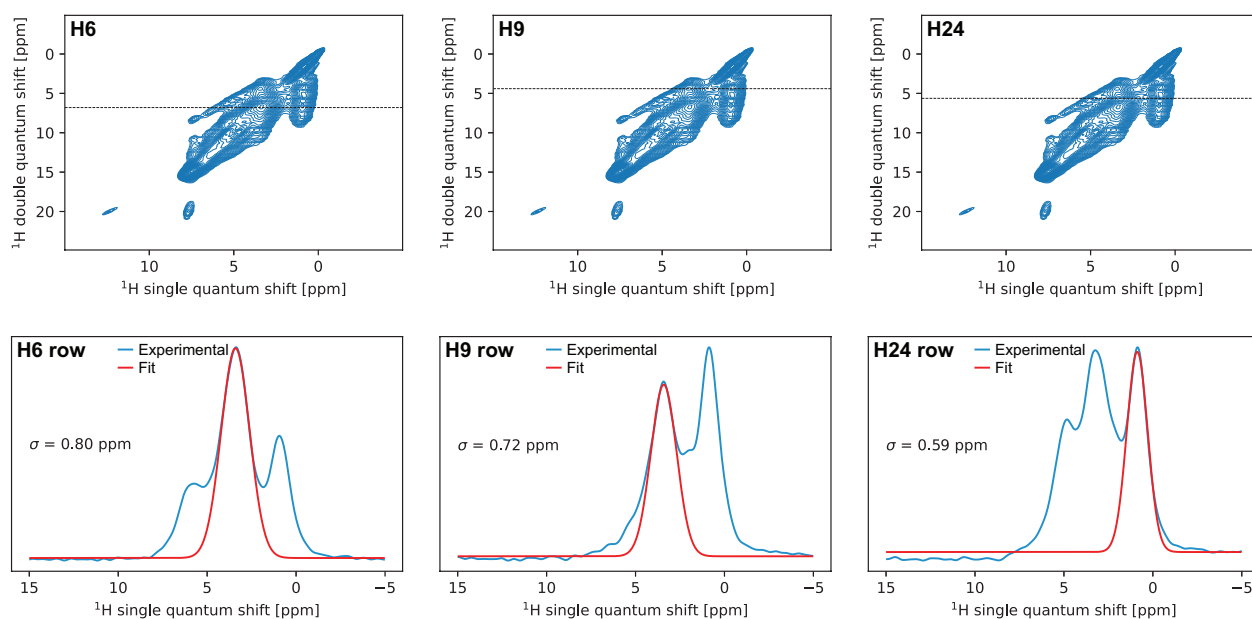


Figure 4.20. Fits of individual proton resonances (bottom panels) from rows extracted from the 2D ^1H - ^1H DQ/SQ MAS NMR spectrum of amorphous AZD4625 (top panels, indicated by dashed black lines).

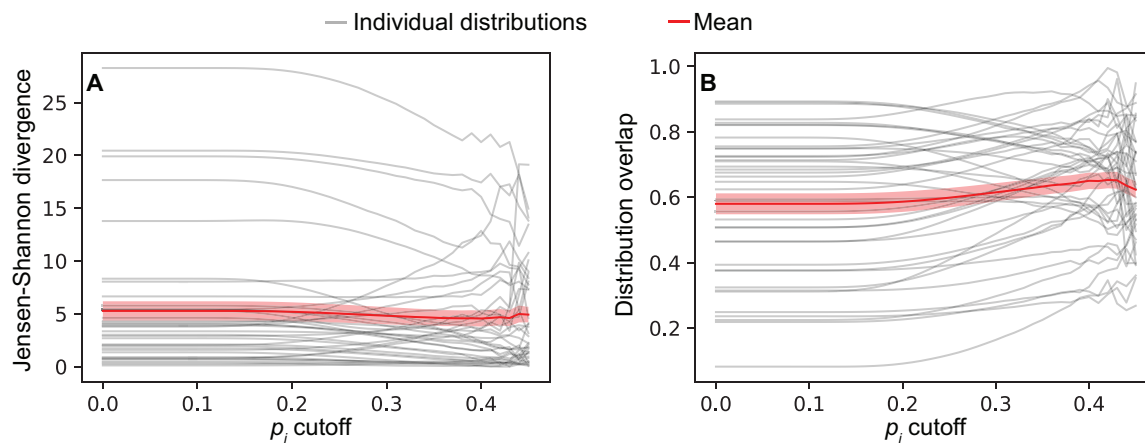


Figure 4.21. (A) Jensen-Shannon divergence and (B) overlap between experimental chemical shift distributions and those obtained from the NMR ensemble. The overlap is defined as the integral under the point-wise minimum between the experimental and NMR-selected shift distribution, where each distribution is assumed to be Gaussian and with an integral of one.

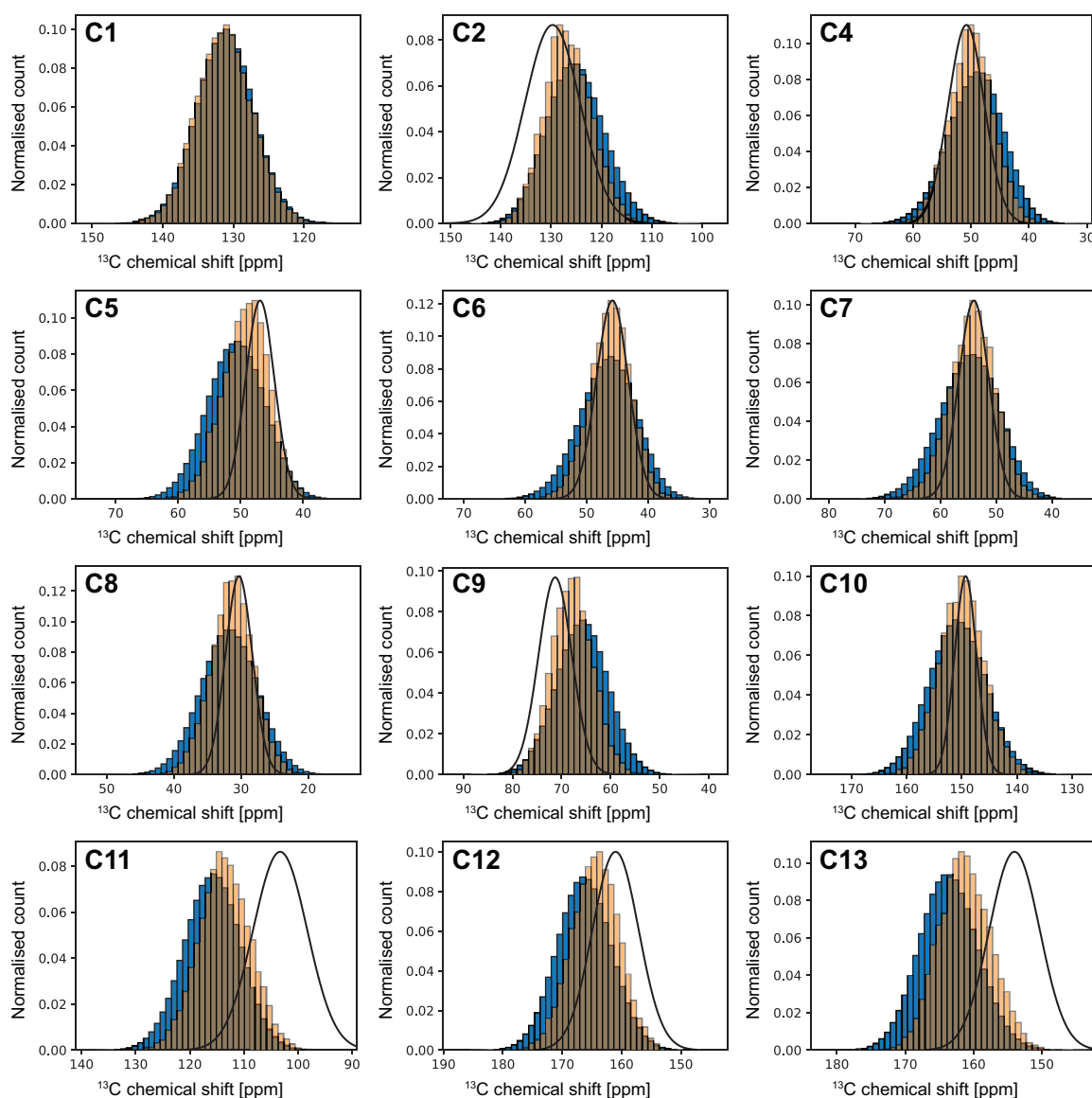


Figure 4.22. Histograms of chemical shifts for individual carbons in the MD (blue) and NMR (orange) ensembles, compared to the experimental distributions when determined (black lines).

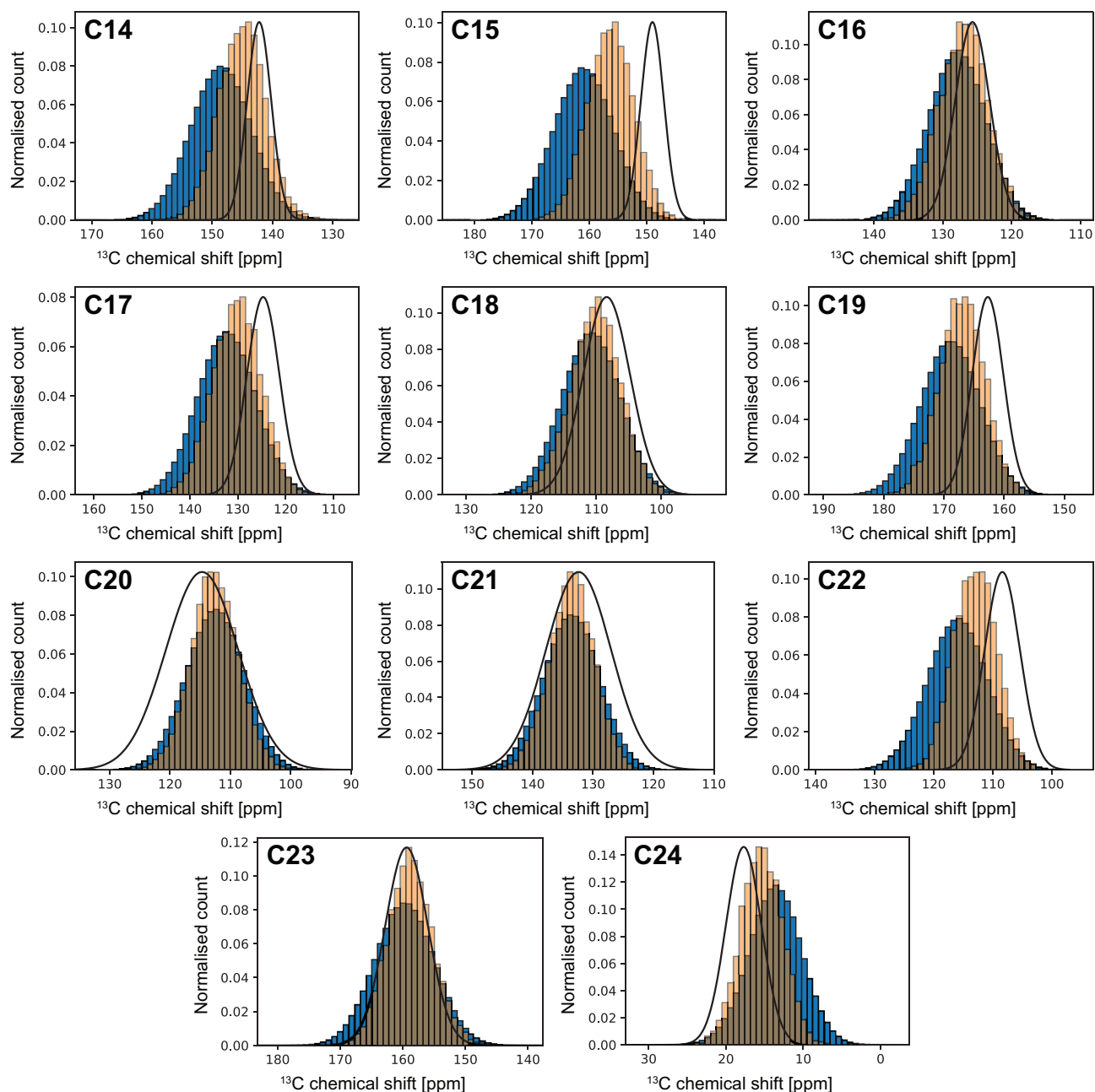


Figure 4.23. Histograms of chemical shifts for individual carbons in the MD (blue) and NMR (orange) ensembles, compared to the experimental distributions when determined (black lines).

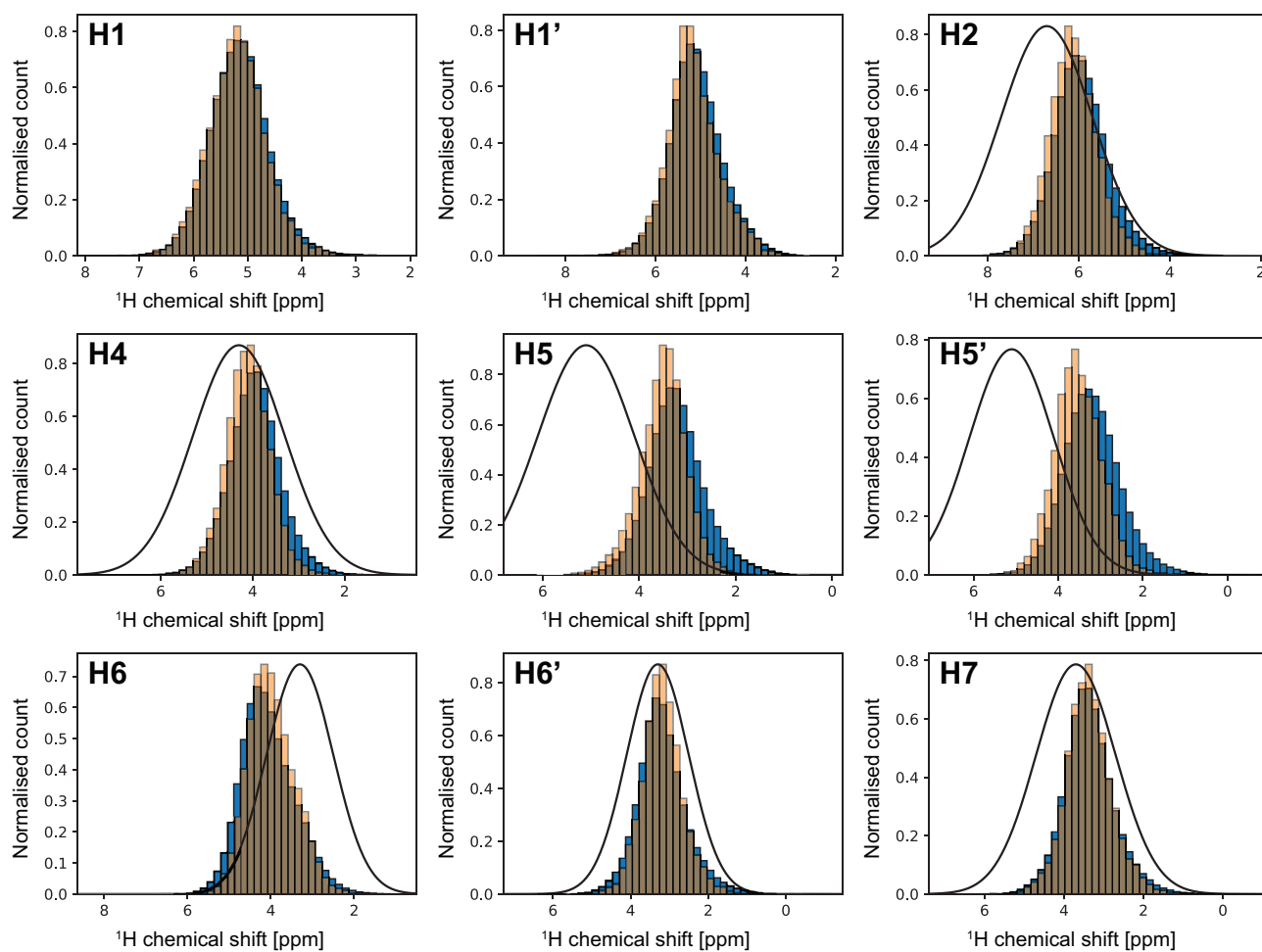


Figure 4.24. Histograms of chemical shifts for individual protons in the MD (blue) and NMR (orange) ensembles, compared to the experimental distributions when determined (black lines).

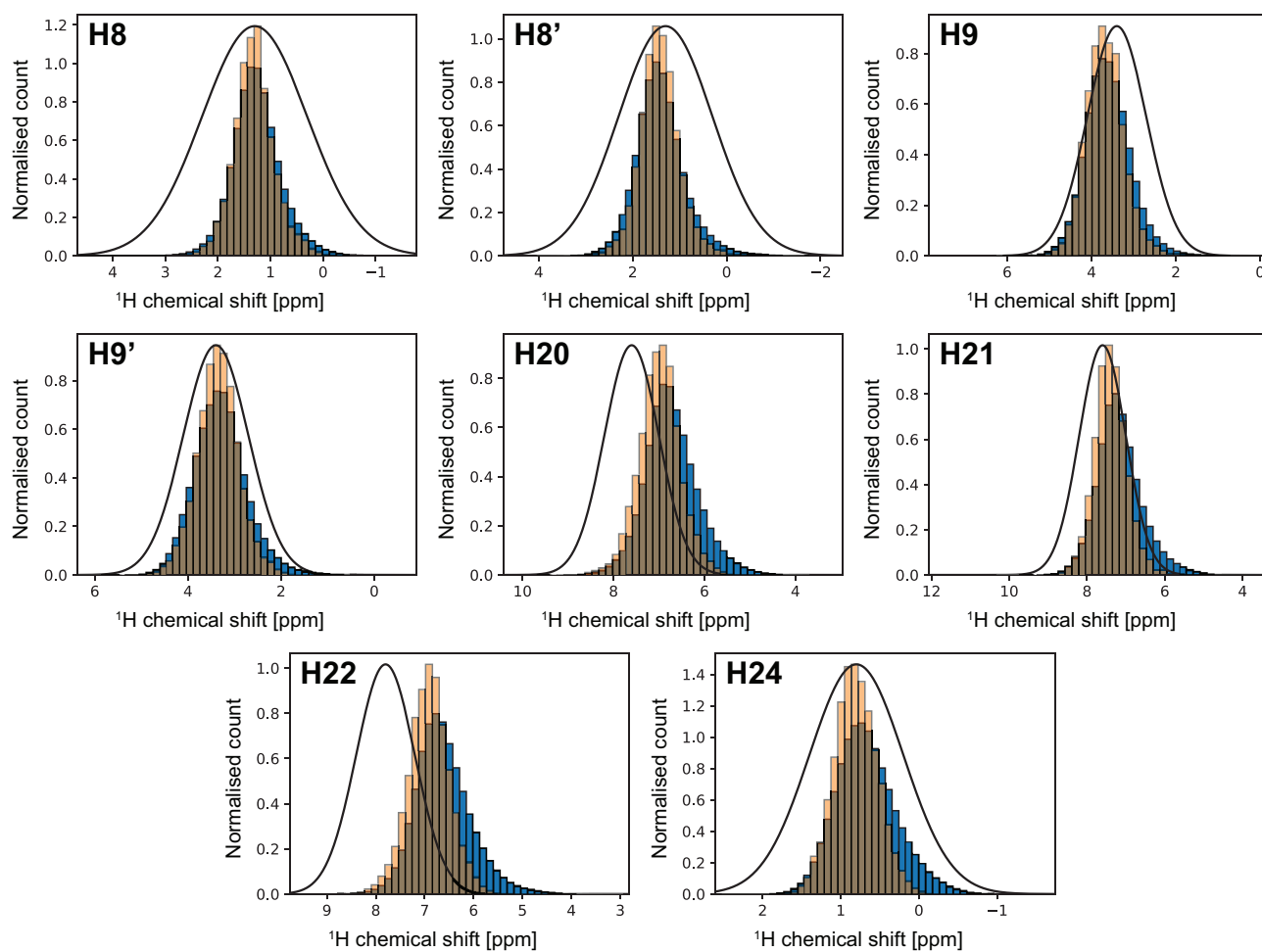


Figure 4.25. Histograms of chemical shifts for individual protons in the MD (blue) and NMR (orange) ensembles, compared to the experimental distributions when determined (black lines).

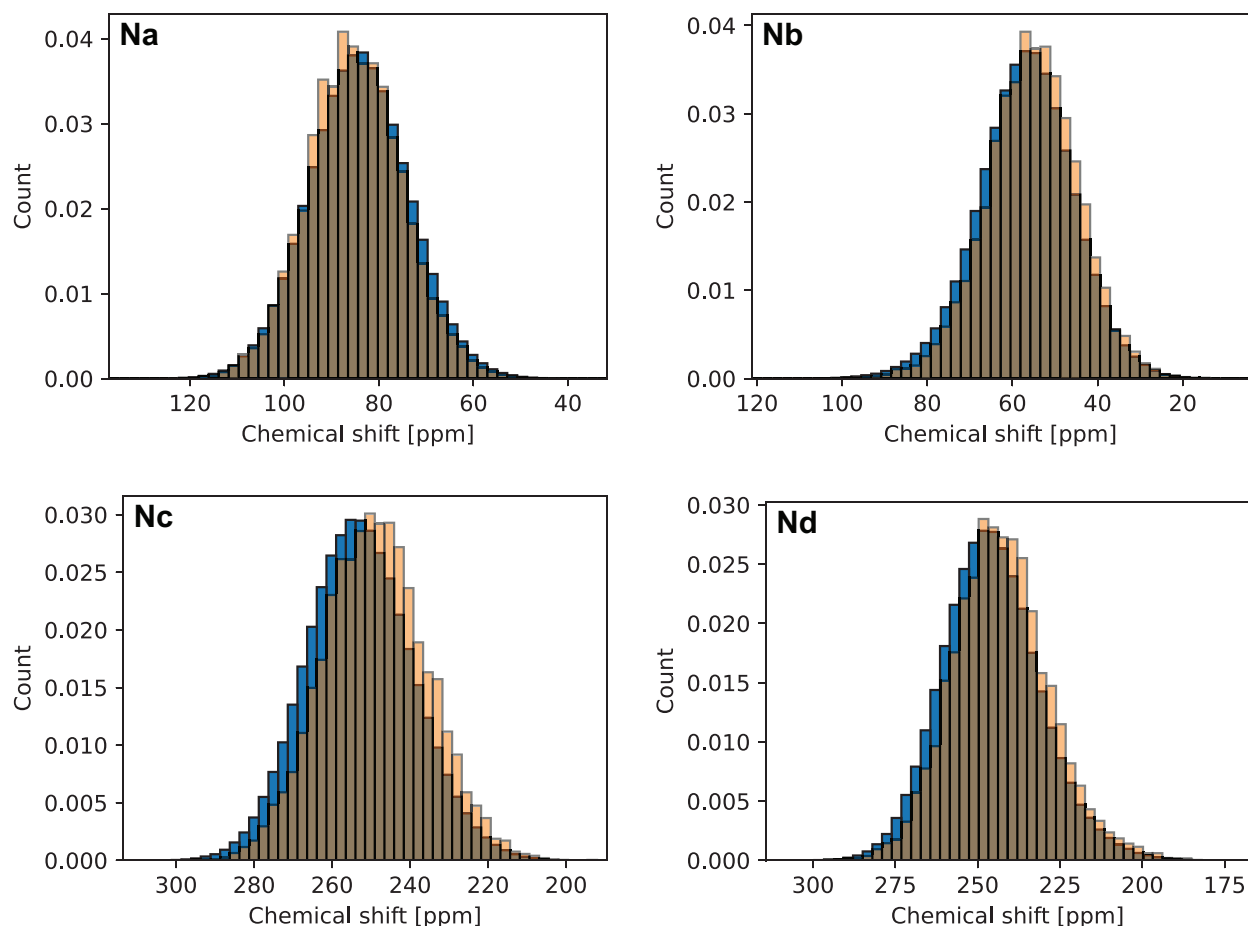


Figure 4.26. Histograms of chemical shifts for individual nitrogens in the MD (blue) and NMR (orange) ensembles.

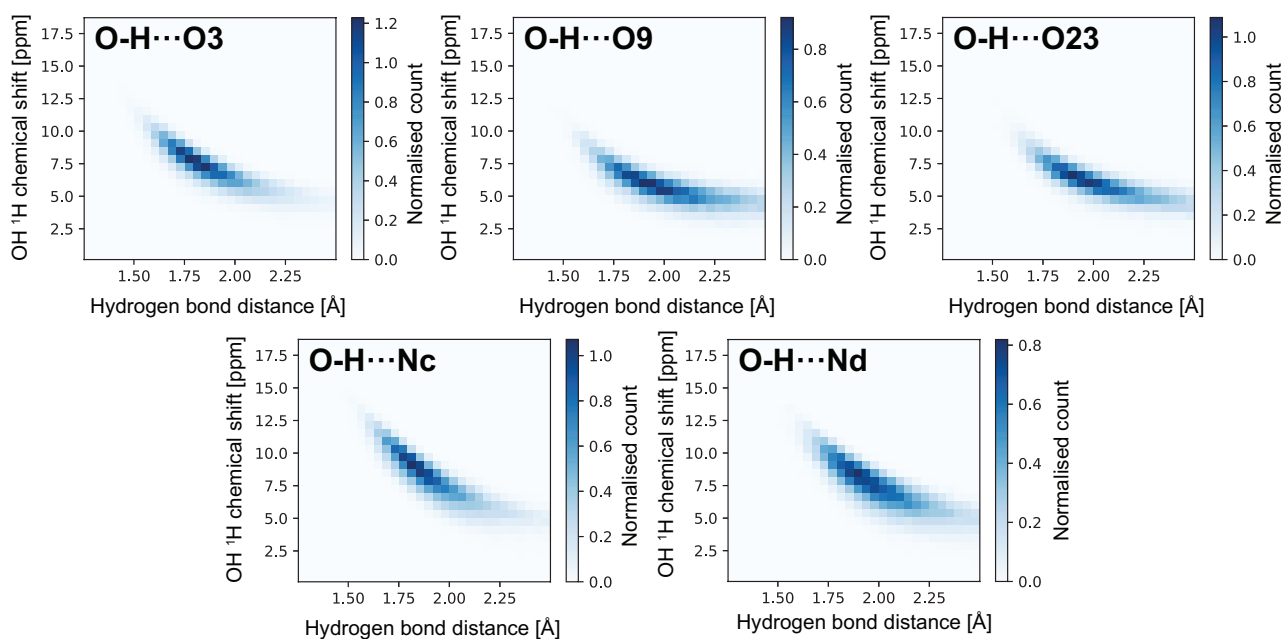


Figure 4.27. Two-dimensional histograms of hydrogen bonding (H...X) distances and OH proton chemical shifts for different hydrogen bond acceptors X.

Chapter 5 Conclusion

5.1 Results achieved

In summary, this thesis presents how NMR crystallography can be accelerated using machine learning. It also shows how the atomic-level structure of amorphous materials can be determined by using a combination of solid-state NMR experiments, molecular dynamics simulations, and machine-learned chemical shifts.

Determining the atomic-level structure of (micro)crystalline molecular solids using NMR chemical shifts requires the combined use of solid-state NMR experiments, crystal structure prediction (CSP) protocols, and DFT chemical shift computations. A machine learning model of chemical shifts was presented as an alternative to DFT computations for solids containing up to 12 elements and finite-temperature or distorted structures, allowing chemical shifts to be obtained in seconds and with DFT-level accuracy for typical crystal structures, and enabling the computation of chemical shifts for large ensembles of large structures. Incorporating machine-learned chemical shifts in CSP procedures was shown to improve the generation of candidate crystal structures by targeting the experimentally measured chemical shifts. In addition, ShiftML enabled the prediction of chemical shifts for a large database of crystal structures, which in turn was found to allow the identification of preferred intermolecular interactions in crystal structures directly from the chemical structure of a molecule and its assigned chemical shifts. This can in turn be used to rank candidate crystal structures, and could enable the construction of structural constraints to further accelerate CSP procedures.

Improvements of the measurement and assignment of chemical shifts were also presented in order to speed up and increase the robustness of these important steps in NMR studies. Using the database of chemical shifts predicted by ShiftML, a probabilistic framework to assign experimental chemical shifts to their associated atomic sites using a database of chemical shifts predicted by ShiftML for a large number of crystal structures was presented. This allows the determination of the assignment without prior knowledge of the three-dimensional structure of the material, and is typically able to confidently assign most chemical shifts, with only a few ambiguities remaining. Then, a convolutional LSTM neural network able to increase the spectral resolution of ^1H spectra by removing MAS-dependent broadenings and shifts was introduced. This allows a more confident measurement of chemical shifts compared to experimental MAS spectra measured at the fastest rates available. The model was also adapted to two-dimensional ^1H - ^1H correlation spectra. Overall, these methods are able to help provide accurate chemical shift measurements and assignments directly from simple and highly sensitive experiments.

Finally, structure determination of amorphous molecular solids was also shown to be possible through a combination of solid-state NMR experiments, molecular dynamics simulation and machine-learned chemical shifts. Comparing chemical shifts predicted for MD structures to experimental values allows the determination of an ensemble of local molecular structures compatible with experiments, which can be analysed to identify preferred molecular configurations and intermolecular interactions in the material under study. This resulted in a general method to determine the atomic-level structure of amorphous molecular solids by NMR by simultaneously considering experimental and computed shifts from multiple atomic sites.

5.2 Future development

Several improvements to the methods presented here could greatly contribute to the improvement and democratisation of structure determination of molecular solids by NMR crystallography.

Developing more accurate machine learning models of chemical shifts would improve the analyses using ShiftML presented in this thesis. In this direction, models able to predict shifts with an accuracy beyond the PBE level of theory typically used for GIPAW chemical shift computations would allow both more confident and faster NMR crystallography compared to procedures using GIPAW. The main challenge to constructing such models is the availability of training data. While ideally experimental chemical shifts and structures should be used as training data to avoid any bias from the method used to compute chemical shifts, this is currently unrealistic due to the lack of a large, centralised database of experimental chemical shifts of molecular solids. In addition, experimental shifts would require thorough and accurate referencing to avoid unwanted noise in the data. An alternative approach to constructing more accurate models of chemical shifts is the computation of training data using higher levels of theory, however approximations may be required in order to enable the computation of the training data using currently available computational resources.³⁰²

NMR crystallography would also benefit from the computation of time-averaged chemical shifts obtained by path-integral molecular dynamics (PIMD) trajectories of candidate crystal structures. This process includes the effect of molecular dynamics and nuclear quantum effects into the chemical shifts obtained, and has been shown to lead to more accurate chemical shifts in better agreement with experimental values.^{303, 306} Coupling PIMD with general-purpose machine learning models of chemical shifts and machine learning potentials could enable routine evaluation of PIMD-averaged chemical shifts in order to improve the description of thermal and quantum fluctuations influencing experimental chemical shifts.³⁰³

ShiftML directly translates atomic-level structures into chemical shifts. This allows NMR crystallography through the generation of candidate crystal structures, followed by chemical shift computation and comparison with experiments. A more direct process to determine structures from chemical shifts would greatly accelerate NMR crystallography. In particular, constructing a model able to predict structures directly from experimental shifts would lead to a paradigm shift in NMR crystallography, similar to the recent introduction of AlphaFold in the field of protein structure determination, that is able to determine protein structures directly from the sequence of amino acids.^{572, 573} However, several challenges including the determination of the space group and unit cell parameters prevent the straightforward application of models similar to AlphaFold to determine the structure of molecular solids.

Finally, the determination of structural ensembles that quantitatively reproduce the spectra measured for amorphous materials would significantly improve their structure determination. In that regard, adapting methods to determine the conformational ensembles of intrinsically disordered proteins (IDPs)⁵⁶³⁻⁵⁶⁵ to molecular solids could be an important step forward.

Bibliography

1. Anderson, A. C., The Process of Structure-Based Drug Design. *Chemistry & Biology* **2003**, *10* (9), 787-797.
2. Lionta, E.; Spyrou, G.; Vassilatis, D. K.; Cournia, Z., Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Current Topics in Medicinal Chemistry* **2014**, *14* (16), 1923-1938.
3. Batool, M.; Ahmad, B.; Choi, S., A Structure-Based Drug Discovery Paradigm. *International Journal of Molecular Sciences* **2019**, *20* (11), 2783.
4. Quarti, C.; Mosconi, E.; Ball, J. M.; D'Innocenzo, V.; Tao, C.; Pathak, S.; Snaith, H. J.; Petrozza, A.; De Angelis, F., Structural and optical properties of methylammonium lead iodide across the tetragonal to cubic phase transition: implications for perovskite solar cells. *Energy & Environmental Science* **2016**, *9* (1), 155-163.
5. Wang, L.; Huang, L.; Tan, W. C.; Feng, X.; Chen, L.; Huang, X.; Ang, K.-W., 2D Photovoltaic Devices: Progress and Prospects. *Small Methods* **2018**, *2* (3), 1700294.
6. Alharbi, E. A.; Alyamani, A. Y.; Kubicki, D. J.; Uhl, A. R.; Walder, B. J.; Alanazi, A. Q.; Luo, J.; Burgos-Caminal, A.; Albadri, A.; Albrithen, H.; Alotaibi, M. H.; Moser, J.-E.; Zakeeruddin, S. M.; Giordano, F.; Emsley, L.; Grätzel, M., Atomic-level passivation mechanism of ammonium salts enabling highly efficient perovskite solar cells. *Nature Communications* **2019**, *10* (1), 3008.
7. Bloom, J.; Meyer, M.; Meinhold, P.; Otey, C.; Macmillan, D.; Arnold, F., Evolving strategies for enzyme engineering. *Current Opinion in Structural Biology* **2005**, *15* (4), 447-452.
8. Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. N., Computational Enzyme Design. *Angewandte Chemie International Edition* **2013**, *52* (22), 5700-5725.
9. Zhu, X.; Guo, Q.; Sun, Y.; Chen, S.; Wang, J.-Q.; Wu, M.; Fu, W.; Tang, Y.; Duan, X.; Chen, D.; Wan, Y., Optimising surface d charge of AuPd nanoalloy catalysts for enhanced catalytic activity. *Nature Communications* **2019**, *10* (1), 1428.
10. Zhang, T.; Walsh, A. G.; Yu, J.; Zhang, P., Single-atom alloy catalysts: structural analysis, electronic properties and catalytic activities. *Chemical Society Reviews* **2021**, *50* (1), 569-588.
11. Bragg, W. L., The structure of some crystals as indicated by their diffraction of X-rays. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **1913**, *89* (610), 248-277.
12. Dickinson, R. G.; Raymond, A. L., The crystal structure of hexamethylene-tetramine. *Journal of the American Chemical Society* **1923**, *45* (1), 22-29.
13. Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C., A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **1958**, *181* (4610), 662-666.
14. Smyth, M. S.; Martin, J. H. J., x Ray crystallography. *Molecular Pathology* **2000**, *53* (1), 8-14.
15. Ma, T.; Kapustin, E. A.; Yin, S. X.; Liang, L.; Zhou, Z.; Niu, J.; Li, L.-H.; Wang, Y.; Su, J.; Li, J.; Wang, X.; Wang, W. D.; Wang, W.; Sun, J.; Yaghi, O. M., Single-crystal x-ray diffraction structures of covalent organic frameworks. *Science* **2018**, *361* (6397), 48-52.
16. Rietveld, H. M., A profile refinement method for nuclear and magnetic structures. *Journal of Applied Crystallography* **1969**, *2* (2), 65-71.
17. Martí-Rujas, J., Structural elucidation of microcrystalline MOFs from powder X-ray diffraction. *Dalton Transactions* **2020**, *49* (40), 13897-13916.
18. Meden, A.; Radosavljevic Evans, I., Structure determination from powder diffraction data: past, present and future challenges. *Crystal Research and Technology* **2015**, *50* (9-10), 747-758.
19. Belik, A. A.; Iikubo, S.; Kodama, K.; Igawa, N.; Shamoto, S.-i.; Niitaka, S.; Azuma, M.; Shimakawa, Y.; Takano, M.; Izumi, F.; Takayama-Muromachi, E., Neutron Powder Diffraction Study on the Crystal and Magnetic Structures of BiCoO₃. *Chemistry of Materials* **2006**, *18* (3), 798-803.
20. Lutterotti, L.; Matthies, S.; Wenk, H. R.; Schultz, A. S.; Richardson, J. W., Combined texture and structure analysis of deformed limestone from time-of-flight neutron diffraction spectra. *Journal of Applied Physics* **1997**, *81* (2), 594-600.
21. Choi, C. S.; Boutin, H. P., A study of the crystal structure of β -cyclotetramethylene tetranitramine by neutron diffraction. *Acta Crystallographica Section B Structural Crystallography and Crystal Chemistry* **1970**, *26* (9), 1235-1240.
22. Soper, A. K., The Radial Distribution Functions of Water as Derived from Radiation Total Scattering Experiments: Is There Anything We Can Say for Sure? *ISRN Physical Chemistry* **2013**, *2013*, 1-67.
23. Billinge, S. J. L.; Kanatzidis, M. G., Beyond crystallography: the study of disorder, nanocrystallinity and crystallographically challenged materials with pair distribution functions. *Chemical Communications* **2004**, (7).
24. Billinge, S. J. L.; Dykhne, T.; Juhás, P.; Božin, E.; Taylor, R.; Florence, A. J.; Shankland, K., Characterisation of amorphous and nanocrystalline molecular materials by total scattering. *CrystEngComm* **2010**, *12* (5), 1366-1368.
25. Finney, J. L.; Hallbrucker, A.; Kohl, I.; Soper, A. K.; Bowron, D. T., Structures of High and Low Density Amorphous Ice by Neutron Diffraction. *Physical Review Letters* **2002**, *88* (22), 225503.

26. Weirich, T. E.; Zou, X.; Ramlau, R.; Simon, A.; Cascarano, G. L.; Giacobazzo, C.; Hovmöller, S., Structures of nanometre-size crystals determined from selected-area electron diffraction data. *Acta Crystallographica Section A Foundations of Crystallography* **2000**, *56* (1), 29-35.
27. Zaefferer, S., New developments of computer-aided crystallographic analysis in transmission electron microscopy. *Journal of Applied Crystallography* **2000**, *33* (1), 10-25.
28. Van Aert, S.; De Backer, A.; Martinez, G. T.; den Dekker, A. J.; Van Dyck, D.; Bals, S.; Van Tendeloo, G., Advanced electron crystallography through model-based imaging. *IUCr* **2016**, *3* (1), 71-83.
29. De Backer, A.; Martinez, G. T.; MacArthur, K. E.; Jones, L.; Béch , A.; Nellist, P. D.; Van Aert, S., Dose limited reliability of quantitative annular dark field scanning transmission electron microscopy for nano-particle atom-counting. *Ultramicroscopy* **2015**, *151*, 56-61.
30. Dycus, J. H.; Harris, J. S.; Sang, X.; Fancher, C. M.; Findlay, S. D.; Oni, A. A.; Chan, T.-t. E.; Koch, C. C.; Jones, J. L.; Allen, L. J.; Irving, D. L.; LeBeau, J. M., Accurate Nanoscale Crystallography in Real-Space Using Scanning Transmission Electron Microscopy. *Microscopy and Microanalysis* **2015**, *21* (4), 946-952.
31. Nakane, T.; Kotecha, A.; Sente, A.; McMullan, G.; Masiulis, S.; Brown, P. M. G. E.; Grigoras, I. T.; Malinauskait , L.; Malinauskas, T.; Miehl , J.; Uchański, T.; Yu, L.; Karia, D.; Pechnikova, E. V.; de Jong, E.; Keizer, J.; Bischoff, M.; McCormack, J.; Tiemeijer, P.; Hardwick, S. W.; Chirgadze, D. Y.; Murshudov, G.; Aricescu, A. R.; Scheres, S. H. W., Single-particle cryo-EM at atomic resolution. *Nature* **2020**, *587* (7832), 152-156.
32. Yip, K. M.; Fischer, N.; Paknia, E.; Chari, A.; Stark, H., Atomic-resolution protein structure determination by cryo-EM. *Nature* **2020**, *587* (7832), 157-161.
33. Jones, C. G.; Martynowycz, M. W.; Hattne, J.; Fulton, T. J.; Stoltz, B. M.; Rodriguez, J. A.; Nelson, H. M.; Gonen, T., The CryoEM Method MicroED as a Powerful Tool for Small Molecule Structure Determination. *ACS Central Science* **2018**, *4* (11), 1587-1592.
34. Morris, G. A., Modern NMR techniques for structure elucidation. *Magnetic Resonance in Chemistry* **1986**, *24* (5), 371-403.
35. Marcarino, M. O.; Zanardi, M. a. M.; Cicetti, S.; Sarotti, A. M., NMR Calculations with Quantum Methods: Development of New Tools for Structural Elucidation and Beyond. *Accounts of Chemical Research* **2020**, *53* (9), 1922-1932.
36. Lodewyk, M. W.; Soldi, C.; Jones, P. B.; Olmstead, M. M.; Rita, J.; Shaw, J. T.; Tantillo, D. J., The Correct Structure of Aquatolide—Experimental Validation of a Theoretically-Predicted Structural Revision. *Journal of the American Chemical Society* **2012**, *134* (45), 18550-18553.
37. Southern, S. A.; Bryce, D. L., Chapter One - Recent advances in NMR crystallography and polymorphism. In *Annual Reports on NMR Spectroscopy*, Webb, G. A., Ed. Academic Press: 2021; Vol. 102, pp 1-80.
38. Brown, S. P.; Schaller, T.; Seelbach, U. P.; Koziol, F.; Ochsenfeld, C.; Kl rner, F.-G.; Spiess, H. W., Structure and Dynamics of the Host-Guest Complex of a Molecular Tweezer: Coupling Synthesis, Solid-State NMR, and Quantum-Chemical Calculations. *Angewandte Chemie International Edition* **2001**, *40* (4), 717-720.
39. Brown, S. P.; Spiess, H. W., Advanced Solid-State NMR Methods for the Elucidation of Structure and Dynamics of Molecular, Macromolecular, and Supramolecular Systems. *Chemical Reviews* **2001**, *101* (12), 4125-4156.
40. 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. *Journal of the American Chemical Society* **2003**, *125* (19), 5792-5800.
41. Rapp, A.; Schnell, I.; Sebastiani, D.; Brown, S. P.; Percec, V.; Spiess, H. W., Supramolecular Assembly of Dendritic Polymers Elucidated by ¹H and ¹³C Solid-State MAS NMR Spectroscopy. *Journal of the American Chemical Society* **2003**, *125* (43), 13284-13297.
42. Elena, B.; Emsley, L., Powder Crystallography by Proton Solid-State NMR Spectroscopy. *Journal of the American Chemical Society* **2005**, *127* (25), 9140-9146.
43. Harper, J. K.; Barich, D. H.; Heider, E. M.; Grant, D. M.; Franke, R. R.; Johnson, J. H.; Zhang, Y.; Lee, P. L.; Von Dreele, R. B.; Scott, B.; Williams, D.; Ansell, G. B., A Combined Solid-State NMR and X-ray Powder Diffraction Study of a Stable Polymorph of Paclitaxel. *Crystal Growth & Design* **2005**, *5* (5), 1737-1742.
44. 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.
45. Harper, J. K.; Grant, D. M., Enhancing Crystal-Structure Prediction with NMR Tensor Data. *Crystal Growth & Design* **2006**, *6* (10), 2315-2321.
46. Heider, E. M.; Harper, J. K.; Grant, D. M., Structural characterization of an anhydrous polymorph of paclitaxel by solid-state NMR. *Physical Chemistry Chemical Physics* **2007**, *9* (46), 6083-6097.
47. Pickard, C. J.; Salager, E.; Pintacuda, G.; Elena, B.; Emsley, L., Resolving Structures from Powders by NMR Crystallography Using Combined Proton Spin Diffusion and Plane Wave DFT Calculations. *Journal of the American Chemical Society* **2007**, *129* (29), 8932-8933.
48. Schaller, T.; B chele, U. P.; Kl rner, F.-G.; Bl ser, D.; Boese, R.; Brown, S. P.; Spiess, H. W.; Koziol, F.; Kussmann, J.; Ochsenfeld, C., Structure of Molecular Tweezer Complexes in the Solid State: NMR Experiments, X-ray Investigations, and Quantum Chemical Calculations. *Journal of the American Chemical Society* **2007**, *129* (5), 1293-1303.
49. 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 ¹H Solid-State NMR Spectroscopy. *Journal of the American Chemical Society* **2010**, *132* (8), 2564-2566.

50. Salager, E.; Stein, R. S.; Pickard, C. J.; Elena, B.; Emsley, L., Powder NMR crystallography of thymol. *Physical Chemistry Chemical Physics* **2009**, *11* (15), 2610-2621.
51. Dudenko, D.; Kiersnowski, A.; Shu, J.; Pisula, W.; Sebastiani, D.; Spiess, H. W.; Hansen, M. R., A Strategy for Revealing the Packing in Semicrystalline π -Conjugated Polymers: Crystal Structure of Bulk Poly-3-hexyl-thiophene (P3HT). *Angewandte Chemie International Edition* **2012**, *51* (44), 11068-11072.
52. Baías, 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), 8069-8080.
53. Baías, 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. *Journal of the American Chemical Society* **2013**, *135* (46), 17501-17507.
54. Harper, J. K.; Iulucci, R.; Gruber, M.; Kalakewich, K., Refining crystal structures with experimental ^{13}C NMR shift tensors and lattice-including electronic structure methods. *CrystEngComm* **2013**, *15* (43).
55. Hofstetter, A.; Balodis, M.; Paruzzo, F. M.; Widdifield, C. M.; Stevanato, G.; Pinon, A. C.; Bygrave, P. J.; 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**, *141* (42), 16624-16634.
56. Li, M.; Meng, F.; Tsutsumi, Y.; Amoureux, J.-P.; Xu, W.; Lu, X.; Zhang, F.; Su, Y., Understanding Molecular Interactions in Rafoxanide–Povidone Amorphous Solid Dispersions from Ultrafast Magic Angle Spinning NMR. *Molecular Pharmaceutics* **2020**, *17* (6), 2196-2207.
57. Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M., Protein structure determination from NMR chemical shifts. *Proceedings of the National Academy of Sciences* **2007**, *104* (23), 9615-9620.
58. Wüthrich, K., Protein structure determination in solution by NMR spectroscopy. *Journal of Biological Chemistry* **1990**, *265* (36), 22059-22062.
59. Kay, L. E., NMR studies of protein structure and dynamics. *Journal of Magnetic Resonance* **2011**, *213* (2), 477-491.
60. Klein, A.; Rovó, P.; Sakhrani, V. V.; Wang, Y.; Holmes, J. B.; Liu, V.; Skowronek, P.; Kukuk, L.; Vasa, S. K.; Güntert, P.; Mueller, L. J.; Linser, R., Atomic-resolution chemical characterization of (2x)72-kDa tryptophan synthase via four- and five-dimensional ^1H -detected solid-state NMR. *Proceedings of the National Academy of Sciences* **2022**, *119* (4), e2114690119.
61. 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.
62. Fyfe, C. A.; Brouwer, D. H., Optimization, Standardization, and Testing of a New NMR Method for the Determination of Zeolite Host–Organic Guest Crystal Structures. *Journal of the American Chemical Society* **2006**, *128* (36), 11860-11871.
63. Brouwer, D. H., NMR Crystallography of Zeolites: Refinement of an NMR-Solved Crystal Structure Using ab Initio Calculations of ^{29}Si Chemical Shift Tensors. *Journal of the American Chemical Society* **2008**, *130* (20), 6306-6307.
64. Cadars, S.; Brouwer, D. H.; Chmelka, B. F., Probing local structures of siliceous zeolite frameworks by solid-state NMR and first-principles calculations of ^{29}Si – ^{29}Si scalar couplings. *Physical Chemistry Chemical Physics* **2009**, *11* (11), 1825-1837.
65. Brouwer, D. H.; Moudrakovski, I. L.; Darton, R. J.; Morris, R. E., Comparing quantum-chemical calculation methods for structural investigation of zeolite crystal structures by solid-state NMR spectroscopy. *Magnetic Resonance in Chemistry* **2010**, *48* (S1), S113-S121.
66. 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.
67. Kunhi Mohamed, A.; Moutzouri, P.; Berruyer, P.; Walder, B. J.; Siramanont, J.; Harris, M.; Negroni, M.; Galmarini, S. C.; Parker, S. C.; Scrivener, K. L.; Emsley, L.; Bowen, P., The Atomic-Level Structure of Cementitious Calcium Aluminate Silicate Hydrate. *Journal of the American Chemical Society* **2020**, *142* (25), 11060-11071.
68. Bamine, T.; Boivin, E.; Boucher, F.; Messinger, R. J.; Salager, E.; Deschamps, M.; Masquelier, C.; Croguennec, L.; Ménétrier, M.; Carlier, D., Understanding Local Defects in Li-Ion Battery Electrodes through Combined DFT/NMR Studies: Application to LiVPO₄F. *The Journal of Physical Chemistry C* **2017**, *121* (6), 3219-3227.
69. Datta, K.; Caiazzo, A.; Hope, M. A.; Li, J.; Mishra, A.; Cordova, M.; Chen, Z.; Emsley, L.; Wienk, M. M.; Janssen, R. A. J., Light-Induced Halide Segregation in 2D and Quasi-2D Mixed-Halide Perovskites. *ACS Energy Letters* **2023**, *8* (4), 1662-1670.
70. Loiseau, T.; Lecroq, L.; Volkringer, C.; Marrot, J.; Férey, G.; Haouas, M.; Taulelle, F.; Bourrelly, S.; Llewellyn, P. L.; Latroche, M., MIL-96, a Porous Aluminum Trimesate 3D Structure Constructed from a Hexagonal Network of 18-Membered Rings and μ_3 -Oxo-Centered Trinuclear Units. *Journal of the American Chemical Society* **2006**, *128* (31), 10223-10230.
71. Brouwer, D. H., Structure solution of network materials by solid-state NMR without knowledge of the crystallographic space group. *Solid State Nuclear Magnetic Resonance* **2013**, *51-52*, 37-45.
72. 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.
73. Ashbrook, S. E.; McKay, D., Combining solid-state NMR spectroscopy with first-principles calculations – a guide to NMR crystallography. *Chemical Communications* **2016**, *52* (45), 7186-7204.
74. Valla, M.; Rossini, A. J.; Caillot, M.; Chizallet, C.; Raybaud, P.; Digne, M.; Chaumonnot, A.; Lesage, A.; Emsley, L.; van Bokhoven, J. A.; Copéret, C., Atomic Description of the Interface between Silica and Alumina in Aluminosilicates through Dynamic

Nuclear Polarization Surface-Enhanced NMR Spectroscopy and First-Principles Calculations. *Journal of the American Chemical Society* **2015**, *137* (33), 10710-10719.

75. Harper, A. F.; Emge, S. P.; Magusin, P. C. M. M.; Grey, C. P.; Morris, A. J., Modelling amorphous materials via a joint solid-state NMR and X-ray absorption spectroscopy and DFT approach: application to alumina. *Chemical Science* **2023**, *14* (5), 1155-1167.
76. Martineau, C., NMR crystallography: Applications to inorganic materials. *Solid State Nuclear Magnetic Resonance* **2014**, *63-64*, 1-12.
77. Keeler, J., *Understanding NMR spectroscopy*. John Wiley & Sons: 2010.
78. Levitt, M. H., *Spin dynamics: basics of nuclear magnetic resonance*. John Wiley & Sons: 2013.
79. Hore, P. J., *Nuclear magnetic resonance*. Oxford University Press, USA: 2015.
80. Ernst, R. R.; Bodenhausen, G.; Wokaun, A., *Principles of nuclear magnetic resonance in one and two dimensions*. Clarendon Press: Oxford, 1987.
81. Lowe, I. J., Free Induction Decays of Rotating Solids. *Physical Review Letters* **1959**, *2* (7), 285-287.
82. Andrew, E. R.; Bradbury, A.; Eades, R. G., Removal of Dipolar Broadening of Nuclear Magnetic Resonance Spectra of Solids by Specimen Rotation. *Nature* **1959**, *183*, 1802-1803.
83. Haeberlen, U.; Waugh, J. S., Coherent Averaging Effects in Magnetic Resonance. *Physical Review* **1968**, *175* (2), 453-467.
84. Maricq, M. M.; Waugh, J. S., NMR in rotating solids. *The Journal of Chemical Physics* **1979**, *70* (7), 3300-3316.
85. Simões de Almeida, B.; Moutzouri, P.; Stevanato, G.; Emsley, L., Theory and simulations of homonuclear three-spin systems in rotating solids. *The Journal of Chemical Physics* **2021**, *155* (8), 084201.
86. Schrödinger, E., An undulatory theory of the mechanics of atoms and molecules. *Physical Review* **1926**, *28* (6), 1049-1070.
87. Hartree, D. R., The wave mechanics of an atom with a non-Coulomb central field Part I theory and methods. *Proceedings of the Cambridge Philosophical Society* **1928**, *24* (1), 89-110.
88. Slater, J. C., The Self Consistent Field and the Structure of Atoms. *Physical Review* **1928**, *32* (3), 339-348.
89. Fock, V., Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Zeitschrift für Physik* **1930**, *61*, 126-148.
90. Møller, C.; Plesset, M. S., Note on an Approximation Treatment for Many-Electron Systems. *Physical Review* **1934**, *46* (7), 618-622.
91. Purvis, G. D.; Bartlett, R. J., A full coupled-cluster singles and doubles model: The inclusion of disconnected triples. *The Journal of Chemical Physics* **1982**, *76* (4), 1910-1918.
92. Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M., A fifth-order perturbation comparison of electron correlation theories. *Chemical Physics Letters* **1989**, *157* (6), 479-483.
93. Hohenberg, P.; Kohn, W., Inhomogeneous Electron Gas. *Physical Review* **1964**, *136* (3B), B864-B871.
94. Cohen, A. J.; Mori-Sánchez, P.; Yang, W., Challenges for Density Functional Theory. *Chemical Reviews* **2011**, *112* (1), 289-320.
95. Vosko, S. H.; Wilk, L.; Nusair, M., Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of Physics* **1980**, *58* (8), 1200-1211.
96. Perdew, J. P.; Wang, Y., Accurate and simple analytic representation of the electron-gas correlation energy. *Physical Review B* **1992**, *45* (23), 13244-13249.
97. Perdew, J. P.; Burke, K.; Ernzerhof, M., Generalized Gradient Approximation Made Simple. *Physical Review Letters* **1996**, *77* (18), 3865-3868.
98. Perdew, J. P., Density-Functional Approximation for the Correlation-Energy of the Inhomogeneous Electron-Gas. *Physical Review B* **1986**, *33* (12), 8822-8824.
99. Becke, A. D., Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A* **1988**, *38* (6), 3098-3100.
100. Zhao, Y.; Truhlar, D. G., A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *The Journal of Chemical Physics* **2006**, *125* (19), 194101.
101. Zhao, Y.; Truhlar, D. G., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts* **2007**, *120*, 215-241.
102. Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E., Climbing the Density Functional Ladder: Nonempirical Meta-Generalized Gradient Approximation Designed for Molecules and Solids. *Physical Review Letters* **2003**, *91* (14), 146401.
103. Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P., Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *The Journal of Chemical Physics* **2003**, *119* (23), 12129-12137.
104. Lee, C.; Yang, W.; Parr, R. G., Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B* **1988**, *37* (2), 785-789.
105. Perdew, J. P.; Ernzerhof, M.; Burke, K., Rationale for mixing exact exchange with density functional approximations. *The Journal of Chemical Physics* **1996**, *105* (22), 9982-9985.
106. Adamo, C.; Barone, V., Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of Chemical Physics* **1999**, *110* (13), 6158-6170.
107. Slater, J. C., Atomic Shielding Constants. *Physical Review* **1930**, *36* (1), 57-64.
108. Hellmann, H., A New Approximation Method in the Problem of Many Electrons. *The Journal of Chemical Physics* **1935**, *3* (1), 61.

109. Hamann, D. R.; Schlüter, M.; Chiang, C., Norm-Conserving Pseudopotentials. *Physical Review Letters* **1979**, *43* (20), 1494-1497.
110. Schwerdtfeger, P., The Pseudopotential Approximation in Electronic Structure Theory. *ChemPhysChem* **2011**, *12* (17), 3143-3155.
111. Blöchl, P. E., Projector augmented-wave method. *Physical Review B* **1994**, *50* (24), 17953-17979.
112. London, F., Théorie quantique des courants interatomiques dans les combinaisons aromatiques. *Journal de Physique et le Radium* **1937**, *8* (10), 397-409.
113. Ditchfield, R., Self-Consistent Perturbation Theory of Diamagnetism. 1. Gauge-Invariant LCAO Method for N.M.R. Chemical Shifts. *Molecular Physics* **1974**, *27* (4), 789-807.
114. McWeeny, R., Perturbation Theory for the Fock-Dirac Density Matrix. *Physical Review* **1962**, *126* (3), 1028-1034.
115. 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.
116. Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. J., A comparison of models for calculating nuclear magnetic resonance shielding tensors. *The Journal of Chemical Physics* **1996**, *104* (14), 5497-5509.
117. Pickard, C. J.; Mauri, F., All-electron magnetic response with pseudopotentials: NMR chemical shifts. *Physical Review B* **2001**, *63* (24), 245101.
118. Yates, J. R.; Pickard, C. J.; Mauri, F., Calculation of NMR chemical shifts for extended systems using ultrasoft pseudopotentials. *Physical Review B* **2007**, *76* (2), 024401.
119. Hinton, J. F.; Guthrie, P.; Pulay, P.; Wolinski, K., Ab initio quantum mechanical calculation of the chemical shift anisotropy of the hydrogen atom in the (H₂O)₁₇ water cluster. *Journal of the American Chemical Society* **2002**, *114* (5), 1604-1605.
120. Koller, H.; Engelhardt, G.; Kentgens, A. P. M.; Sauer, J., ²³Na NMR Spectroscopy of Solids: Interpretation of Quadrupole Interaction Parameters and Chemical Shifts. *The Journal of Physical Chemistry* **2002**, *98* (6), 1544-1551.
121. Holmes, S. T.; Iuliucci, R. J.; Mueller, K. T.; Dybowski, C., Density functional investigation of intermolecular effects on C-13 NMR chemical-shielding tensors modeled with molecular clusters. *Journal of Chemical Physics* **2014**, *141* (16), 164121.
122. 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.
123. Jacob, C. R.; Visscher, L., Calculation of nuclear magnetic resonance shieldings using frozen-density embedding. *The Journal of Chemical Physics* **2006**, *125* (19).
124. Hartman, J. D.; Balaji, A.; Beran, G. J. O., Improved Electrostatic Embedding for Fragment-Based Chemical Shift Calculations in Molecular Crystals. *Journal of Chemical Theory and Computation* **2017**, *13* (12), 6043-6051.
125. Gao, Q.; Yokojima, S.; Fedorov, D. G.; Kitaura, K.; Sakurai, M.; Nakamura, S., Fragment-Molecular-Orbital-Method-Based ab Initio NMR Chemical-Shift Calculations for Large Molecular Systems. *Journal of Chemical Theory and Computation* **2010**, *6* (4), 1428-1444.
126. 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.
127. Hartman, J. D.; Kudla, R. A.; Day, G. M.; Mueller, L. J.; Beran, G. J. O., Benchmark fragment-based ¹H, ¹³C, ¹⁵N and ¹⁷O chemical shift predictions in molecular crystals. *Physical Chemistry Chemical Physics* **2016**, *18* (31), 21686-21709.
128. Hartman, J. D.; Monaco, S.; Schatschneider, B.; Beran, G. J. O., Fragment-based ¹³C nuclear magnetic resonance chemical shift predictions in molecular crystals: An alternative to planewave methods. *The Journal of Chemical Physics* **2015**, *143* (10), 102809.
129. Perras, F. A.; Bryce, D. L., Multinuclear Magnetic Resonance Crystallographic Structure Refinement and Cross-Validation Using Experimental and Computed Electric Field Gradients: Application to Na₂Al₂B₂O₇. *The Journal of Physical Chemistry C* **2012**, *116* (36), 19472-19482.
130. Romao, C. P.; Perras, F. A.; Werner-Zwanziger, U.; Lussier, J. A.; Miller, K. J.; Calahoo, C. M.; Zwanziger, J. W.; Bieringer, M.; Marinkovic, B. A.; Bryce, D. L.; White, M. A., Zero Thermal Expansion in ZrMgMo₃O₁₂: NMR Crystallography Reveals Origins of Thermoelastic Properties. *Chemistry of Materials* **2015**, *27* (7), 2633-2646.
131. de Dios, A. C.; Pearson, J. G.; Oldfield, E., Secondary and Tertiary Structural Effects on Protein NMR Chemical Shifts: an ab Initio Approach. *Science* **1993**, *260* (5113), 1491-1496.
132. Facelli, J. C.; Grant, D. M., Determination of molecular symmetry in crystalline naphthalene using solid-state NMR. *Nature* **1993**, *365* (6444), 325-327.
133. Brown, S. P., Recent Advances in Solid-State MAS NMR Methodology for Probing Structure and Dynamics in Polymeric and Supramolecular Systems. *Macromolecular Rapid Communications* **2009**, *30* (9-10), 688-716.
134. Yates, J. R.; Pickard, C. J.; Payne, M. C.; Dupree, R.; Profeta, M.; Mauri, F., Theoretical investigation of oxygen-17 NMR shielding and electric field gradients in glutamic acid polymorphs. *Journal of Physical Chemistry A* **2004**, *108* (28), 6032-6037.
135. Yates, J. R.; Dobbins, S. E.; Pickard, C. J.; Mauri, F.; Ghi, P. Y.; Harris, R. K., A combined first principles computational and solid-state NMR study of a molecular crystal: flurbiprofen. *Physical Chemistry Chemical Physics* **2005**, *7* (7), 1402-1407.
136. 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* (S1), S103-S112.
137. Webber, A. L.; Elena, B.; Griffin, J. M.; Yates, J. R.; Pham, T. N.; Mauri, F.; Pickard, C. J.; Gil, A. M.; Stein, R.; Lesage, A.; Emsley, L.; Brown, S. P., Complete ¹H resonance assignment of β-maltose from ¹H-¹H DQ-SQ CRAMPS and ¹H (DQ-DUMBO)-¹³C

- SQ refocused INEPT 2D solid-state NMR spectra and first principles GIPAW calculations. *Physical Chemistry Chemical Physics* **2010**, 12 (26), 6970-6983.
138. Harris, R. K.; Joyce, S. A.; Pickard, C. J.; Cadars, S.; Emsley, L., Assigning carbon-13 NMR spectra to crystal structures by the INADEQUATE pulse sequence and first principles computation: a case study of two forms of testosterone. *Phys. Chem. Chem. Phys.* **2006**, 8 (1), 137-143.
139. Mifsud, N.; Elena, B.; Pickard, C. J.; Lesage, A.; Emsley, L., Assigning powders to crystal structures by high-resolution 1H–1H double quantum and 1H–13C J-INEPT solid-state NMR spectroscopy and first principles computation. A case study of penicillin G. *Physical Chemistry Chemical Physics* **2006**, 8 (29), 3418-3422.
140. Ashbrook, S. E.; Le Pollès, L.; Pickard, C. J.; Berry, A. J.; Wimperis, S.; Farnan, I., First-principles calculations of solid-state 17O and 29Si NMR spectra of Mg2SiO4 polymorphs. *Phys. Chem. Chem. Phys.* **2007**, 9 (13), 1587-1598.
141. Yazawa, K.; Suzuki, F.; Nishiyama, Y.; Ohata, T.; Aoki, A.; Nishimura, K.; Kaji, H.; Shimizu, T.; Asakura, T., Determination of accurate 1H positions of an alanine tripeptide with anti-parallel and parallel β -sheet structures by high resolution 1H solid state NMR and GIPAW chemical shift calculation. *Chemical Communications* **2012**, 48 (91), 11199-11201.
142. Asakura, T.; Yazawa, K.; Horiguchi, K.; Suzuki, F.; Nishiyama, Y.; Nishimura, K.; Kaji, H., Difference in the structures of alanine tri- and tetra-peptides with antiparallel β -sheet assessed by X-ray diffraction, solid-state NMR and chemical shift calculations by GIPAW. *Biopolymers* **2014**, 101 (1), 13-20.
143. Tatton, A. S.; Blade, H.; Brown, S. P.; Hodgkinson, P.; Hughes, L. P.; Nilsson Lill, S. O.; Yates, J. R., Improving Confidence in Crystal Structure Solutions Using NMR Crystallography: The Case of β -Piroxicam. *Crystal Growth & Design* **2018**, 18 (6), 3339-3351.
144. Widdifield, C. M.; Robson, H.; Hodgkinson, P., Furosemide's one little hydrogen atom: NMR crystallography structure verification of powdered molecular organics. *Chemical Communications* **2016**, 52 (40), 6685-6688.
145. Woińska, M.; Grabowsky, S.; Dominiak, P. M.; Woźniak, K.; Jayatilaka, D., Hydrogen atoms can be located accurately and precisely by x-ray crystallography. *Science Advances* **2016**, 2 (5), e1600192.
146. Dračínský, M.; Jansa, P.; Ahonen, K.; Buděšínský, M., Tautomerism and the Protonation/Deprotonation of Isocytosine in Liquid- and Solid-States Studied by NMR Spectroscopy and Theoretical Calculations. *European Journal of Organic Chemistry* **2011**, 2011 (8), 1544-1551.
147. Osmiałowski, B.; Kolehmainen, E.; Ikonen, S.; Ahonen, K.; Löfman, M., NMR crystallography of 2-acylamino-6-[1H]-pyridones: Solid-state NMR, GIPAW computational, and single crystal X-ray diffraction studies. *Journal of Molecular Structure* **2011**, 1006 (1-3), 678-683.
148. Bártová, K.; Císařová, I.; Lyčka, A.; Dračínský, M., Tautomerism of azo dyes in the solid state studied by 15N, 14N, 13C and 1H NMR spectroscopy, X-ray diffraction and quantum-chemical calculations. *Dyes and Pigments* **2020**, 178, 108342.
149. Marín-Luna, M.; Claramunt, R. M.; Nieto, C. I.; Alkorta, I.; Elguero, J.; Reviriego, F., A theoretical NMR study of polymorphism in crystal structures of azoles and benzazoles. *Magnetic Resonance in Chemistry* **2019**, 57 (6), 275-284.
150. Gumbert, S. D.; Körbitzer, M.; Alig, E.; Schmidt, M. U.; Chierotti, M. R.; Gobetto, R.; Li, X.; van de Streek, J., Crystal structure and tautomerism of Pigment Yellow 138 determined by X-ray powder diffraction and solid-state NMR. *Dyes and Pigments* **2016**, 131, 364-372.
151. Hodgkinson, P., NMR crystallography of molecular organics. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2020**, 118-119, 10-53.
152. Chierotti, M. R.; Gaglioti, K.; Gobetto, R.; Braga, D.; Grepioni, F.; Maini, L., From molecular crystals to salt co-crystals of barbituric acid via the carbonate ion and an improvement of the solid state properties. *CrystEngComm* **2013**, 15 (37), 7598-7605.
153. Bernasconi, D.; Bordignon, S.; Rossi, F.; Priola, E.; Nervi, C.; Gobetto, R.; Voinovich, D.; Hasa, D.; Duong, N. T.; Nishiyama, Y.; Chierotti, M. R., Selective Synthesis of a Salt and a Cocrystal of the Ethionamide–Salicylic Acid System. *Crystal Growth & Design* **2019**, 20 (2), 906-915.
154. Rossi, F.; Cerreia Vioglio, P.; Bordignon, S.; Giorgio, V.; Nervi, C.; Priola, E.; Gobetto, R.; Yazawa, K.; Chierotti, M. R., Unraveling the Hydrogen Bond Network in a Theophylline–Pyridoxine Salt Cocrystal by a Combined X-ray Diffraction, Solid-State NMR, and Computational Approach. *Crystal Growth & Design* **2018**, 18 (4), 2225-2233.
155. Bravetti, F.; Russo, R. E.; Bordignon, S.; Gallo, A.; Rossi, F.; Nervi, C.; Gobetto, R.; Chierotti, M. R., Zwitterionic or Not? Fast and Reliable Structure Determination by Combining Crystal Structure Prediction and Solid-State NMR. *Molecules* **2023**, 28 (4).
156. 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. *Physical Chemistry Chemical Physics* **2017**, 19 (25), 16650-16661.
157. Rehman, Z.; Franks, W. T.; Nguyen, B.; Schmidt, H. F.; Scrivens, G.; Brown, S. P., Discovering the Solid-State Secrets of Lorlatinib by NMR Crystallography: To Hydrogen Bond or not to Hydrogen Bond. *Journal of Pharmaceutical Sciences* **2023**, 112 (7), 1915-1928.
158. Woodley, S. M.; Catlow, R., Crystal structure prediction from first principles. *Nature Materials* **2008**, 7 (12), 937-946.
159. Day, G. M.; S. Motherwell, W. D.; 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.
160. Mooij, W. T. M.; van Eijck, B. P.; Price, S. L.; Verwer, P.; Kroon, J., Crystal structure predictions for acetic acid. *Journal of Computational Chemistry* **1998**, 19 (4), 459-474.
161. Schneider, E.; Vogt, L.; Tuckerman, M. E., Exploring polymorphism of benzene and naphthalene with free energy based enhanced molecular dynamics. *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials* **2016**, 72 (4), 542-550.

162. Francia, N. F.; Price, L. S.; Nyman, J.; Price, S. L.; Salvalaglio, M., Systematic Finite-Temperature Reduction of Crystal Energy Landscapes. *Crystal Growth & Design* **2020**, *20* (10), 6847-6862.
163. Francia, N. F.; Price, L. S.; Salvalaglio, M., Reducing crystal structure overprediction of ibuprofen with large scale molecular dynamics simulations. *CrystEngComm* **2021**, *23* (33), 5575-5584.
164. Smalley, C. J. H.; Hoskyns, H. E.; Hughes, C. E.; Johnstone, D. N.; Willhammar, T.; Young, M. T.; Pickard, C. J.; Logsdail, A. J.; Midgley, P. A.; Harris, K. D. M., A structure determination protocol based on combined analysis of 3D-ED data, powder XRD data, solid-state NMR data and DFT-D calculations reveals the structure of a new polymorph of l-tyrosine. *Chemical Science* **2022**, *13* (18), 5277-5288.
165. Lai, J.; Niks, D.; Wang, Y.; Domratcheva, T.; Barends, T. R.; Schwarz, F.; Olsen, R. A.; Elliott, D. W.; Fatmi, M. Q.; Chang, C. E.; 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.
166. Hughes, C. E.; Reddy, G. N. M.; Masiero, S.; Brown, S. P.; Williams, P. A.; Harris, K. D. M., Determination of a complex crystal structure in the absence of single crystals: analysis of powder X-ray diffraction data, guided by solid-state NMR and periodic DFT calculations, reveals a new 2'-deoxyguanosine structural motif. *Chemical Science* **2017**, *8* (5), 3971-3979.
167. Rajeswaran, M.; Blanton, T. N.; Zumbulyadis, N.; Giesen, D. J.; Conesa-Moratilla, C.; Mixture, S. T.; Stephens, P. W.; Huq, A., Three-Dimensional Structure Determination of N-(p-Tolyl)-dodecylsulfonamide from Powder Diffraction Data and Validation of Structure Using Solid-State NMR Spectroscopy. *Journal of the American Chemical Society* **2002**, *124* (48), 14450-14459.
168. 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. *Crystal Growth & Design* **2015**, *15* (8), 3674-3683.
169. 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. *Crystal Growth & Design* **2016**, *16* (4), 1798-1804.
170. Dudek, M. K.; Paluch, P.; Śniechowska, J.; Nartowski, K. P.; Day, G. M.; Potrzebowski, M. J., Crystal structure determination of an elusive methanol solvate – hydrate of catechin using crystal structure prediction and NMR crystallography. *CrystEngComm* **2020**, *22* (30), 4969-4981.
171. Dudek, M. K.; Paluch, P.; Pindelska, E., Crystal structures of two furazidin polymorphs revealed by a joint effort of crystal structure prediction and NMR crystallography. *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials* **2020**, *76* (3), 322-335.
172. Brus, J.; Czernek, J.; Kobera, L.; Urbanova, M.; Abbrent, S.; Husak, M., Predicting the Crystal Structure of Decitabine by Powder NMR Crystallography: Influence of Long-Range Molecular Packing Symmetry on NMR Parameters. *Crystal Growth & Design* **2016**, *16* (12), 7102-7111.
173. Baías, 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.
174. 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. *Chemical Science* **2016**, *7* (7), 4379-4390.
175. Hofstetter, A.; Emsley, L., Positional Variance in NMR Crystallography. *Journal of the American Chemical Society* **2017**, *139* (7), 2573-2576.
176. Engel, E. A.; Anelli, A.; Hofstetter, A.; Paruzzo, F.; Emsley, L.; Ceriotti, M., A Bayesian approach to NMR crystal structure determination. *Physical Chemistry Chemical Physics* **2019**, *21* (42), 23385-23400.
177. Cordova, M.; Balodis, M.; Hofstetter, A.; Paruzzo, F.; Nilsson Lill, S. O.; Eriksson, E. S. E.; Berruyer, P.; Simões de Almeida, B.; Quayle, M. J.; Norberg, S. T.; Svensk Ankarberg, A.; Schantz, S.; Emsley, L., Structure determination of an amorphous drug through large-scale NMR predictions. *Nature Communications* **2021**, *12* (1), 2964.
178. Burnett, M. N.; Johnson, C. K. *ORTEP-III: Oak Ridge thermal ellipsoid plot program for crystal structure illustrations*; Oak ridge national laboratory report ORNL-6895, Tennessee: 1996.
179. Dunitz, J. D.; Schomaker, V.; Trueblood, K. N., Interpretation of Atomic Displacement Parameters from Diffraction Studies of Crystals. *Journal of Physical Chemistry* **1988**, *92* (4), 856-867.
180. Morales-Melgares, A.; Casar, Z.; Moutzouri, P.; Venkatesh, A.; Cordova, M.; Mohamed, A. K.; Scrivener, K. L.; Bowen, P.; Emsley, L., Atomic-Level Structure of Zinc-Modified Cementitious Calcium Silicate Hydrate. *Journal of the American Chemical Society* **2022**, *144* (50), 22915-22924.
181. Scrivener, K. L.; Kirkpatrick, R. J., Innovation in use and research on cementitious material. *Cement and Concrete Research* **2008**, *38* (2), 128-136.
182. Megat Johari, M. A.; Brooks, J. J.; Kabir, S.; Rivard, P., Influence of supplementary cementitious materials on engineering properties of high strength concrete. *Construction and Building Materials* **2011**, *25* (5), 2639-2648.
183. Juenger, M. C. G.; Siddique, R., Recent advances in understanding the role of supplementary cementitious materials in concrete. *Cement and Concrete Research* **2015**, *78*, 71-80.
184. Stephan, D.; Maleki, H.; Knöfel, D.; Eber, B.; Härdtl, R., Influence of Cr, Ni, and Zn on the properties of pure clinker phases. *Cement and Concrete Research* **1999**, *29* (4), 545-552.
185. Li, X.; Scrivener, K. L., Impact of ZnO on C3S hydration and C-S-H morphology at early ages. *Cement and Concrete Research* **2022**, *154*, 106734.

186. Odler, I.; Abdul-Maula, S., Polymorphism and Hydration of Tricalcium Silicate Doped With ZnO. *Journal of the American Ceramic Society* **1983**, *66* (1), 1-4.
187. Bazzoni, A.; Ma, S.; Wang, Q.; Shen, X.; Cantoni, M.; Scrivener, K. L.; Scherer, G., The Effect of Magnesium and Zinc Ions on the Hydration Kinetics of C3S. *Journal of the American Ceramic Society* **2014**, *97* (11), 3684-3693.
188. Kunhi Mohamed, A.; Parker, S. C.; Bowen, P.; Galmarini, S., An atomistic building block description of C-S-H - Towards a realistic C-S-H model. *Cement and Concrete Research* **2018**, *107*, 221-235.
189. Klaska, K. H.; Eck, J. C.; Pohl, D., New investigation of willemite. *Acta Crystallographica Section B Structural Crystallography and Crystal Chemistry* **1978**, *34* (11), 3324-3325.
190. Hill, R. J.; Gibbs, G. V.; Craig, J. R., A neutron-diffraction study of hemimorphite. *Zeitschrift für Kristallographie - Crystalline Materials* **1978**, *146* (1-6), 241-260.
191. Gard, J. A.; Taylor, H. F. W., The crystal structure of foshagite. *Acta Crystallographica* **1960**, *13* (10), 785-793.
192. Smith, G. S.; Alexander, L. E., Refinement of the atomic parameters of α -quartz. *Acta Crystallographica* **1963**, *16* (6), 462-471.
193. Chandrappa, G. T.; Ghosh, S.; Patil, K. C., Synthesis and Properties of Willemite, Zn_2SiO_4 , and $\text{M}^{2+}:\text{Zn}_2\text{SiO}_4$ (M = Co and Ni). *Journal of Materials Synthesis and Processing* **1999**, *7* (5), 273-279.
194. Lippmaa, E.; Magi, M.; Samoson, A.; Engelhardt, G.; Grimmer, A. R., Structural Studies of Silicates by Solid-State High-Resolution Si-29 Nmr. *Journal of the American Chemical Society* **1980**, *102* (15), 4889-4893.
195. Fyfe, C. A.; Grondy, H.; Feng, Y.; Kokotailo, G. T., Natural-abundance two-dimensional silicon-29 MAS NMR investigation of the three-dimensional bonding connectivities in the zeolite catalyst ZSM-5. *Journal of the American Chemical Society* **2002**, *124* (24), 8812-8820.
196. 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.
197. Hope, M. A.; Nakamura, T.; Ahlawat, P.; Mishra, A.; Cordova, M.; Jahanbakhshi, F.; Mladenovic, M.; Runjhun, R.; Merten, L.; Hinderhofer, A.; Carlsen, B. I.; Kubicki, D. J.; Gershoni-Poranne, R.; Schneeberger, T.; Carbone, L. C.; Liu, Y. H.; Zakeeruddin, S. M.; Lewinski, J.; Hagfeldt, A.; Schreiber, F.; Rothlisberger, U.; Grätzel, M.; Milic, J. V.; Emsley, L., Nanoscale Phase Segregation in Supramolecular pi-Templating for Hybrid Perovskite Photovoltaics from NMR Crystallography. *Journal of the American Chemical Society* **2021**, *143* (3), 1529-1538.
198. Jena, A. K.; Kulkarni, A.; Miyasaka, T., Halide Perovskite Photovoltaics: Background, Status, and Future Prospects. *Chemical Reviews* **2019**, *119* (5), 3036-3103.
199. Kim, H.-S.; Lee, C.-R.; Im, J.-H.; Lee, K.-B.; Moehl, T.; Marchioro, A.; Moon, S.-J.; Humphry-Baker, R.; Yum, J.-H.; Moser, J. E.; Grätzel, M.; Park, N.-G., Lead Iodide Perovskite Sensitized All-Solid-State Submicron Thin Film Mesoscopic Solar Cell with Efficiency Exceeding 9%. *Scientific Reports* **2012**, *2* (1), 591.
200. Eperon, G. E.; Stranks, S. D.; Menelaou, C.; Johnston, M. B.; Herz, L. M.; Snaith, H. J., Formamidinium lead trihalide: a broadly tunable perovskite for efficient planar heterojunction solar cells. *Energy & Environmental Science* **2014**, *7* (3), 982.
201. Grätzel, M., The Rise of Highly Efficient and Stable Perovskite Solar Cells. *Accounts of Chemical Research* **2017**, *50* (3), 487-491.
202. Wang, R.; Mujahid, M.; Duan, Y.; Wang, Z. K.; Xue, J.; Yang, Y., A Review of Perovskites Solar Cell Stability. *Advanced Functional Materials* **2019**, *29* (47), 1808843.
203. Ruiz-Preciado, M. A.; Kubicki, D. J.; Hofstetter, A.; McGovern, L.; Futscher, M. H.; Ummadisingu, A.; Gershoni-Poranne, R.; Zakeeruddin, S. M.; Ehrler, B.; Emsley, L.; Milić, 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. *Journal of the American Chemical Society* **2020**, *142* (3), 1645-1654.
204. Szell, P. M. J.; Gabriel, S. A.; Gill, R. D. D.; Wan, S. Y. H.; 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 Crystallographica Section C Structural Chemistry* **2017**, *73* (3), 157-167.
205. Viger-Gravel, J.; Avalos, C. E.; Kubicki, D. J.; Gajan, D.; Lelli, M.; Ouari, O.; Lesage, A.; Emsley, L., ^{19}F Magic Angle Spinning Dynamic Nuclear Polarization Enhanced NMR Spectroscopy. *Angewandte Chemie International Edition* **2019**, *58* (22), 7249-7253.
206. 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 ^{19}F NMR of octafluoronaphthalene. *Physical Chemistry Chemical Physics* **2007**, *9* (19), 2389-2396.
207. 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.
208. Vícha, J.; Švec, P.; Růžicková, Z.; Samsonov, M. A.; Bártová, K.; Růžicka, A.; Straka, M.; Dračinský, M., Experimental and Theoretical Evidence of Spin-Orbit Heavy Atom on the Light Atom ^1H NMR Chemical Shifts Induced through $\text{H}\cdots\text{I}$ -Hydrogen Bond. *Chemistry – A European Journal* **2020**, *26* (40), 8698-8702.
209. Vícha, J.; Novotný, J.; Komorovsky, S.; Straka, M.; Kaupp, M.; Marek, R., Relativistic Heavy-Neighbor-Atom Effects on NMR Shifts: Concepts and Trends Across the Periodic Table. *Chemical Reviews* **2020**, *120* (15), 7065-7103.
210. Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A., Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters* **2012**, *108* (5), 058301.
211. Misra, M.; Andrienko, D.; Baumeier, B.; Faulon, J.-L.; von Lilienfeld, O. A., Toward Quantitative Structure-Property Relationships for Charge Transfer Rates of Polycyclic Aromatic Hydrocarbons. *Journal of Chemical Theory and Computation* **2011**, *7* (8), 2549-2555.

212. Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Anatole von Lilienfeld, O., Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics* **2013**, *15* (9), 095003.
213. Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R., Accelerating materials property predictions using machine learning. *Scientific Reports* **2013**, *3* (1), 2810.
214. Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C., Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B* **2014**, *89* (9), 094104.
215. Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U., How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Physical Review B* **2014**, *89* (20), 205118.
216. Janet, J. P.; Chan, L.; Kulik, H. J., Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *The Journal of Physical Chemistry Letters* **2018**, *9* (5), 1064-1071.
217. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A., Big Data Meets Quantum Chemistry Approximations: The Δ -Machine Learning Approach. *Journal of Chemical Theory and Computation* **2015**, *11* (5), 2087-2096.
218. Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K., Quantum chemical accuracy from density functional approximations via machine learning. *Nature Communications* **2020**, *11*, 5223.
219. Ruth, M.; Gerbig, D.; Schreiner, P. R., Machine Learning of Coupled Cluster (T)-Energy Corrections via Delta (Δ)-Learning. *Journal of Chemical Theory and Computation* **2022**, *18* (8), 4846-4855.
220. Hu, L.; Wang, X.; Wong, L.; Chen, G., Combined first-principles calculation and neural-network correction approach for heat of formation. *The Journal of Chemical Physics* **2003**, *119* (22), 11501-11507.
221. Kayala, M. A.; Baldi, P., ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level Using Machine Learning. *Journal of Chemical Information and Modeling* **2012**, *52* (10), 2526-2540.
222. Ziletti, A.; Kumar, D.; Scheffler, M.; Ghiringhelli, L. M., Insightful classification of crystal structures using deep learning. *Nature Communications* **2018**, *9* (1), 2775.
223. Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G., Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **2018**, *360* (6385), 186-190.
224. Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A., Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* **2019**, *5* (9), 1572-1583.
225. Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T., “Found in Translation”: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical Science* **2018**, *9* (28), 6091-6098.
226. Saunders, C.; Gammernan, A.; Vovk, V., Ridge Regression Learning Algorithm in Dual Variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc.: 1998; pp 515–521.
227. Cortes, C.; Vapnik, V., Support-vector networks. *Machine Learning* **1995**, *20* (3), 273-297.
228. Tin Kam, H., Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995; pp 278-282.
229. Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J., *Classification And Regression Trees*. Routledge: New York, 1984.
230. Friedman, J. H., Greedy function approximation: A gradient boosting machine. *Annals of Statistics* **2001**, *29* (5), 1189-1232.
231. Friedman, J. H., Stochastic gradient boosting. *Computational Statistics & Data Analysis* **2002**, *38* (4), 367-378.
232. Macukow, B., Neural Networks – State of Art, Brief History, Basic Models and Architecture. In *Computer Information Systems and Industrial Management*, 2016; pp 3-14.
233. Rumelhart, D. E.; Hinton, G. E.; Williams, R. J., Learning representations by back-propagating errors. *Nature* **1986**, *323* (6088), 533-536.
234. Bartók, A. P.; Kondor, R.; Csányi, G., On representing chemical environments. *Physical Review B* **2013**, *87* (18), 019902.
235. Huang, B.; von Lilienfeld, O. A., Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nature Chemistry* **2020**, *12* (10), 945-951.
236. 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.
237. Meiler, J.; Maier, W.; Will, M.; Meusinger, R., Using Neural Networks for ¹³C NMR Chemical Shift Prediction—Comparison with Traditional Methods. *Journal of Magnetic Resonance* **2002**, *157* (2), 242-252.
238. Aires-de-Sousa, J.; Hemmer, M. C.; Gasteiger, J., Prediction of ¹H NMR Chemical Shifts Using Neural Networks. *Analytical Chemistry* **2001**, *74* (1), 80-90.
239. Jonas, E.; Kuhn, S.; Schlörer, N., Prediction of chemical shift in NMR: A review. *Magnetic Resonance in Chemistry* **2021**, *60* (11), 1021-1031.
240. Da Costa, F. B.; Binev, Y.; Gasteiger, J.; Aires-De-Sousa, J., Structure-based predictions of H-1 NMR chemical shifts of sesquiterpene lactones using neural networks. *Tetrahedron Letters* **2004**, *45* (37), 6931-6935.
241. Binev, Y.; Aires-De-Sousa, J., Structure-based predictions of H-1 NMR chemical shifts using feed-forward neural networks. *Journal of Chemical Information and Computer Sciences* **2004**, *44* (3), 940-945.
242. Loss, A.; Stenutz, R.; Schwarzer, E.; von der Lieth, C. W., GlyNest and CASPER: two independent approaches to estimate ¹H and ¹³C NMR shifts of glycans available through a common web-interface. *Nucleic Acids Research* **2006**, *34* (suppl_2), W733-W737.

243. Ksenofontov, A. A.; Isaev, Y. I.; Lukanov, M. M.; Makarov, D. M.; Eventova, V. A.; Khodov, I. A.; Berezin, M. B., Accurate prediction of ¹¹B NMR chemical shift of BODIPYs via machine learning. *Physical Chemistry Chemical Physics* **2023**, *25* (13), 9472-9481.
244. Gao, P.; Zhang, J.; Peng, Q.; Zhang, J.; Glezakou, V.-A., General Protocol for the Accurate Prediction of Molecular ¹³C/¹H NMR Chemical Shifts via Machine Learning Augmented DFT. *Journal of Chemical Information and Modeling* **2020**, *60* (8), 3746-3754.
245. Unzueta, P. A.; Greenwell, C. S.; Beran, G. J. O., Predicting Density Functional Theory-Quality Nuclear Magnetic Resonance Chemical Shifts via Δ -Machine Learning. *Journal of Chemical Theory and Computation* **2021**, *17* (2), 826-840.
246. Spera, S.; Bax, A., Empirical correlation between protein backbone conformation and C.alpha. and C.beta. ¹³C nuclear magnetic resonance chemical shifts. *Journal of the American Chemical Society* **2002**, *113* (14), 5490-5492.
247. Meiler, J., PROSHIFT: Protein chemical shift prediction using artificial neural networks. *Journal of Biomolecular NMR* **2003**, *26* (1), 25-37.
248. Neal, S.; Nip, A. M.; Zhang, H.; Wishart, D. S., Rapid and accurate calculation of protein ¹H, ¹³C and ¹⁵N chemical shifts. *Journal of Biomolecular NMR* **2003**, *26* (3), 215-240.
249. Shen, Y.; Bax, A., Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *Journal of Biomolecular NMR* **2007**, *38* (4), 289-302.
250. Xu, X.-P.; Case, D. A., Automated prediction of ¹⁵N, ¹³C α , ¹³C β and ¹³C' chemical shifts in proteins using a density functional database. *Journal of Biomolecular NMR* **2001**, *21* (4), 321-333.
251. Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S., SHIFTX2: significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR* **2011**, *50* (1), 43-57.
252. Shen, Y.; Bax, A., SPARTA plus : a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *Journal of Biomolecular NMR* **2010**, *48* (1), 13-22.
253. Yang, Z.; Chakraborty, M.; White, A. D., Predicting chemical shifts with graph neural networks. *Chemical Science* **2021**, *12* (32), 10802-10809.
254. 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. *Journal of Chemical Theory and Computation* **2016**, *12* (2), 765-773.
255. Chaker, Z.; Salanne, M.; Delaye, J.-M.; Charpentier, T., NMR shifts in aluminosilicate glasses via machine learning. *Physical Chemistry Chemical Physics* **2019**, *21* (39), 21709-21725.
256. Ohkubo, T.; Takei, A.; Tachi, Y.; Fukatsu, Y.; Deguchi, K.; Ohki, S.; Shimizu, T., New Approach To Understanding the Experimental ¹³³Cs NMR Chemical Shift of Clay Minerals via Machine Learning and DFT-GIPAW Calculations. *The Journal of Physical Chemistry A* **2023**, *127* (4), 973-986.
257. Gaumard, R.; Dragún, D.; Pedroza-Montero, J. N.; Alonso, B.; Guesmi, H.; Malkin Ondík, I.; Mineva, T., Regression Machine Learning Models Used to Predict DFT-Computed NMR Parameters of Zeolites. *Computation* **2022**, *10* (5), 74.
258. Dawson, D. M.; Seymour, V. R.; Ashbrook, S. E., Effects of Extraframework Species on the Structure-Based Prediction of ³¹P Isotropic Chemical Shifts of Aluminophosphates. *The Journal of Physical Chemistry C* **2017**, *121* (50), 28065-28076.
259. Gerrard, W.; Bratholm, L. A.; Packer, M. J.; Mulholland, A. J.; Glowacki, D. R.; Butts, C. P., IMPRESSION – prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chemical Science* **2020**, *11* (2), 508-515.
260. Guan, Y.; Shree Sowndarya, S. V.; Gallegos, L. C.; St. John, P. C.; Paton, R. S., Real-time prediction of ¹H and ¹³C chemical shifts with DFT accuracy using a 3D graph neural network. *Chemical Science* **2021**, *12* (36), 12012-12026.
261. 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.
262. Liu, S.; Li, J.; Bennett, K. C.; Ganoe, B.; Stauch, T.; Head-Gordon, M.; Hexemer, A.; Ushizima, D.; Head-Gordon, T., Multiresolution 3D-DenseNet for Chemical Shift Prediction in NMR Crystallography. *The Journal of Physical Chemistry Letters* **2019**, *10* (16), 4558-4565.
263. Dailey, B. P.; Shoolery, J. N., The Electron Withdrawal Power of Substituent Groups. *Journal of the American Chemical Society* **1955**, *77* (15), 3977-3981.
264. Paul, E. G.; Grant, D. M., Additivity Relationships in Carbon-13 Chemical Shift Data for Linear Alkanes. *Journal of the American Chemical Society* **1963**, *85* (11), 1701-1702.
265. Pretsch, E.; Bühlmann, P.; Affolter, C.; Pretsch, E.; Bühlmann, P.; Affolter, C., *Structure determination of organic compounds*. Springer: 2000.
266. Bremser, W., Hose - Novel Substructure Code. *Analytica Chimica Acta-Computer Techniques and Optimization* **1978**, *2* (4), 355-365.
267. Shen, Y.; Delaglio, F.; Cornilescu, G.; Bax, A., TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *Journal of Biomolecular NMR* **2009**, *44* (4), 213-223.
268. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Research* **2000**, *28* (1), 235-242.
269. Hoch, J. C.; Baskaran, K.; Burr, H.; Chin, J.; Eghbalnia, Hamid R.; Fujiwara, T.; Gryk, Michael R.; Iwata, T.; Kojima, C.; Kurisu, G.; Maziuk, D.; Miyanoiri, Y.; Wedell, Jonathan R.; Wilburn, C.; Yao, H.; Yokochi, M., Biological Magnetic Resonance Data Bank. *Nucleic Acids Research* **2023**, *51* (D1), D368-D376.

270. Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Kent Wenger, R.; Yao, H.; Markley, J. L., BioMagResBank. *Nucleic Acids Research* **2007**, *36*, D402-D408.
271. Fukushima, K.; Miyake, S., Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition* **1982**, *15* (6), 455-469.
272. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P., Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **1998**, *86* (11), 2278-2324.
273. Vaillant, R.; Monroq, C.; LeCun, Y., Original approach for the localisation of objects in images. *IEE Proceedings - Vision, Image, and Signal Processing* **1994**, *141* (4), 245-250.
274. Lawrence, S.; Giles, C. L.; Ah Chung, T.; Back, A. D., Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks* **1997**, *8* (1), 98-113.
275. LeCun, Y.; Bengio, Y.; Hinton, G., Deep learning. *Nature* **2015**, *521* (7553), 436-444.
276. Lee, H. H.; Kim, H., Intact metabolite spectrum mining by deep learning in proton magnetic resonance spectroscopy of the brain. *Magnetic Resonance in Medicine* **2019**, *82* (1), 33-48.
277. Yang, Z.; Zheng, X.; Gao, X.; Zeng, Q.; Yang, C.; Luo, J.; Zhan, C.; Lin, Y., Deep Learning Methodology for Obtaining Ultraclean Pure Shift Proton Nuclear Magnetic Resonance Spectra. *The Journal of Physical Chemistry Letters* **2023**, *14* (14), 3397-3402.
278. Wu, K.; Luo, J.; Zeng, Q.; Dong, X.; Chen, J.; Zhan, C.; Chen, Z.; Lin, Y., Improvement in Signal-to-Noise Ratio of Liquid-State NMR Spectroscopy via a Deep Neural Network DN-Unet. *Analytical Chemistry* **2020**, *93* (3), 1377-1382.
279. Schmid, N.; Bruderer, S.; Paruzzo, F.; Fischetti, G.; Toscano, G.; Graf, D.; Fey, M.; Henrici, A.; Ziebart, V.; Heitmann, B.; Grabner, H.; Wegner, J. D.; Sigel, R. K. O.; Wilhelm, D., Deconvolution of 1D NMR spectra: A deep learning-based approach. *Journal of Magnetic Resonance* **2023**, *347*, 107357.
280. Klukowski, P.; Riek, R.; Güntert, P., Rapid protein assignments and structures from raw NMR spectra with the deep learning technique ARTINA. *Nature Communications* **2022**, *13*, 6151.
281. Li, D.-W.; Hansen, A. L.; Yuan, C.; Bruschweiler-Li, L.; Bruschweiler, R., DEEP picker is a deep neural network for accurate deconvolution of complex two-dimensional NMR spectra. *Nature Communications* **2021**, *12*, 5229.
282. Li, D.-W.; Bruschweiler-Li, L.; Hansen, A. L.; Bruschweiler, R., DEEP Picker1D and Voigt Fitter1D: a versatile tool set for the automated quantitative spectral deconvolution of complex 1D-NMR spectra. *Magnetic Resonance* **2023**, *4*, 19-26.
283. Karunanithy, G.; Hansen, D. F., FID-Net: A versatile deep neural network architecture for NMR spectral reconstruction and virtual decoupling. *Journal of Biomolecular NMR* **2021**, *75*, 179-191.
284. Karunanithy, G.; Mackenzie, H. W.; Hansen, D. F., Virtual Homonuclear Decoupling in Direct Detection Nuclear Magnetic Resonance Experiments Using Deep Neural Networks. *Journal of the American Chemical Society* **2021**, *143* (41), 16935-16942.
285. Qu, X.; Huang, Y.; Lu, H.; Qiu, T.; Guo, D.; Agback, T.; Orekhov, V.; Chen, Z., Accelerated Nuclear Magnetic Resonance Spectroscopy with Deep Learning. *Angewandte Chemie International Edition* **2020**, *59* (26), 10297-10300.
286. Luo, J.; Zeng, Q.; Wu, K.; Lin, Y., Fast reconstruction of non-uniform sampling multidimensional NMR spectroscopy via a deep neural network. *Journal of Magnetic Resonance* **2020**, *317*, 106772.
287. Hubel, D. H.; Wiesel, T. N., Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology* **1962**, *160* (1), 106-154.
288. Felleman, D. J.; Van Essen, D. C., Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex* **1991**, *1* (1), 1-47.
289. Bethge, M.; Cadieu, C. F.; Hong, H.; Yamins, D. L. K.; Pinto, N.; Ardila, D.; Solomon, E. A.; Majaj, N. J.; DiCarlo, J. J., Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology* **2014**, *10* (12).
290. Hochreiter, S.; Schmidhuber, J., Long short-term memory. *Neural Computation* **1997**, *9* (8), 1735-1780.
291. Hansen, D. F., Using Deep Neural Networks to Reconstruct Non-uniformly Sampled NMR Spectra. *Journal of Biomolecular NMR* **2019**, *73*, 577-585.
292. Becker, M.; Lehmkuhl, S.; Kesselheim, S.; Korvink, J. G.; Jouda, M., Acquisitions with random shim values enhance AI-driven NMR shimming. *Journal of Magnetic Resonance* **2022**, *345*, 107323.
293. Sundermeyer, M.; Schluter, R.; Ney, H., LSTM Neural Networks for Language Modeling. *13th Annual Conference of the International Speech Communication Association 2012 (Interspeech 2012)* **2012**, 1-3, 194-197.
294. Graves, A.; Jaitly, N.; Mohamed, A.-r., Hybrid speech recognition with Deep Bidirectional LSTM. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013; pp 273-278.
295. Worswick, S. G.; Spencer, J. A.; Jeschke, G.; Kuprov, I., Deep neural network processing of DEER data. *Science Advances* **2018**, *4* (8), eaat5218.
296. Keeley, J.; Choudhury, T.; Galazzo, L.; Bordignon, E.; Feintuch, A.; Goldfarb, D.; Russell, H.; Taylor, M. J.; Lovett, J. E.; Eggeling, A.; Fábregas Ibáñez, L.; Keller, K.; Yulikov, M.; Jeschke, G.; Kuprov, I., Neural networks in pulsed dipolar spectroscopy: A practical guide. *Journal of Magnetic Resonance* **2022**, *338*, 107186.
297. Brog, J.-P.; Chanez, C.-L.; Crochet, A.; Fromm, K. M., Polymorphism, what it is and how to identify it: a systematic review. *RSC Advances* **2013**, *3* (38), 16905-16931.
298. Santos, O.; Freitas, J.; Cazedey, E.; Araújo, M.; Dorigueto, A., Structure, Solubility and Stability of Orbifloxacin Crystal Forms: Hemihydrate versus Anhydrate. *Molecules* **2016**, *21* (3), 328.

299. Lee, A. Y.; Erdemir, D.; Myerson, A. S., Crystal Polymorphism in Chemical Process Development. *Annual Review of Chemical and Biomolecular Engineering* **2011**, *2* (1), 259-280.
300. Pudipeddi, M.; Serajuddin, A. T. M., Trends in Solubility of Polymorphs. *Journal of Pharmaceutical Sciences* **2005**, *94* (5), 929-939.
301. Dračinský, M.; Möller, H. M.; Exner, T. E., Conformational Sampling by Ab Initio Molecular Dynamics Simulations Improves NMR Chemical Shift Predictions. *Journal of Chemical Theory and Computation* **2013**, *9* (8), 3806-3815.
302. Dračinský, 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**, *21* (27), 14992-15000.
303. Engel, E. A.; Kapil, V.; Ceriotti, M., Importance of Nuclear Quantum Effects for NMR Crystallography. *The Journal of Physical Chemistry Letters* **2021**, *12* (32), 7701-7707.
304. 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 & Design* **2016**, *16* (11), 6479-6493.
305. Beran, G. J. O., Calculating Nuclear Magnetic Resonance Chemical Shifts from Density Functional Theory: A Primer. *Emagres* **2019**, *8* (3), 215-226.
306. Dračinský, M.; Vícha, J.; Bártoňová, K.; Hodgkinson, P., Towards Accurate Predictions of Proton NMR Spectroscopic Parameters in Molecular Solids. *ChemPhysChem* **2020**, *21* (18), 2075-2083.
307. 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* (S1), S174-S186.
308. Reif, B.; Ashbrook, S. E.; Emsley, L.; Hong, M., Solid-state NMR spectroscopy. *Nature Reviews Methods Primers* **2021**, *1*, 2.
309. Gupta, A.; Chakraborty, S.; Ramakrishnan, R., Revving up ¹³C NMR shielding predictions across chemical space: benchmarks for atoms-in-molecules kernel machine learning with new data for 134 kilo molecules. *Machine Learning: Science and Technology* **2021**, *2* (3).
310. Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O. A., Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *The Journal of Physical Chemistry Letters* **2015**, *6* (16), 3309-3313.
311. Cobas, C., NMR signal processing, prediction, and structure verification with machine learning techniques. *Magnetic Resonance in Chemistry* **2020**, *58* (6), 512-519.
312. 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), 171-179.
313. 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.
314. 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 & Design* **2013**, *13* (6), 2390-2395.
315. Wang, Y.; Lv, J.; Zhu, L.; Ma, Y., Crystal structure prediction via particle-swarm optimization. *Physical Review B* **2010**, *82* (9), 094116.
316. 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.
317. 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.
318. Curtis, F.; Li, X.; Rose, T.; Vázquez-Mayagoitia, Á.; Bhattacharya, S.; Ghiringhelli, L. M.; Marom, N., GATOR: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction. *Journal of Chemical Theory and Computation* **2018**, *14* (4), 2246-2264.
319. Karfunkel, H. R.; Gdanitz, R. J., Ab Initio prediction of possible crystal structures for general organic molecules. *Journal of Computational Chemistry* **1992**, *13* (10), 1171-1183.
320. Bazterra, V. E.; Ferraro, M. B.; Facelli, J. C., Modified genetic algorithm to model crystal structures. I. Benzene, naphthalene and anthracene. *The Journal of Chemical Physics* **2002**, *116* (14), 5984-5991.
321. Zhu, Q.; Oganov, A. R.; Glass, C. W.; Stokes, H. T., Constrained evolutionary algorithm for structure prediction of molecular crystals: methodology and applications. *Acta Crystallographica Section B Structural Science* **2012**, *68* (3), 215-226.
322. Zilka, M.; Dudenko, D. V.; Hughes, C. E.; Williams, P. A.; Sturniolo, S.; Franks, W. T.; Pickard, C. J.; Yates, J. R.; Harris, K. D. M.; Brown, S. P., Ab initio random structure searching of organic molecular solids: assessment and validation against experimental data. *Physical Chemistry Chemical Physics* **2017**, *19* (38), 25949-25960.
323. Yang, S.; Day, G. M., Exploration and Optimization in Crystal Structure Prediction: Combining Basin Hopping with Quasi-Random Sampling. *Journal of Chemical Theory and Computation* **2021**, *17* (3), 1988-1999.
324. Wood, P. A.; Olsson, T. S. G.; Cole, J. C.; Cottrell, S. J.; Feeder, N.; Galek, P. T. A.; Groom, C. R.; Pidcock, E., Evaluation of molecular crystal structures using Full Interaction Maps. *CrystEngComm* **2013**, *15* (1), 65-72.
325. Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G., Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Physical Review B* **1998**, *58* (11), 7260-7268.
326. Gaus, M.; Cui, Q.; Elstner, M., DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *Journal of Chemical Theory and Computation* **2011**, *7* (4), 931-948.
327. Eldar, Y.; Lindenbaum, M.; Porat, M.; Zeevi, Y. Y., The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing* **1997**, *6* (9), 1305-1315.

328. Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; Dal Corso, A.; de Gironcoli, S.; Fabris, S.; Fratesi, G.; Gebauer, R.; Gerstmann, U.; Gougoussis, C.; Kokalj, A.; Lazzeri, M.; Martin-Samos, L.; Marzari, N.; Mauri, F.; Mazzarello, R.; Paolini, S.; 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.
329. 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.; Küçükbenli, E.; Lazzeri, M.; Marsili, M.; Marzari, N.; Mauri, F.; Nguyen, N. L.; Nguyen, H. V.; Otero-de-la-Roza, A.; Paulatto, L.; Poncè, 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. *Journal of Physics: Condensed Matter* **2017**, *29* (46), 465901.
330. Grimme, S., Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry* **2006**, *27* (15), 1787-1799.
331. Barone, V.; Casarin, M.; Forrer, D.; Pavone, M.; Sami, M.; Vittadini, A., Role and effective treatment of dispersive forces in materials: Polyethylene and graphite crystals as test cases. *Journal of Computational Chemistry* **2009**, *30* (6), 934-939.
332. Dal Corso, A., Pseudopotentials periodic table: From H to Pu. *Computational Materials Science* **2014**, *95*, 337-350.
333. Kresse, G.; Joubert, D., From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B* **1999**, *59* (3), 1758-1775.
334. 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.
335. Kapil, V.; Rossi, M.; Marsalek, O.; Petraglia, R.; Litman, Y.; Spura, T.; Cheng, B.; Cuzzocrea, A.; Meißner, R. H.; Wilkins, D. M.; Helfrecht, B. A.; Juda, P.; Bienvenue, S. P.; Fang, W.; Kessler, J.; Poltavsky, I.; Vandenbrande, S.; Wieme, J.; Corminboeuf, C.; Kühne, T. D.; Manolopoulos, D. E.; Markland, T. E.; Richardson, J. O.; Tkatchenko, A.; Tribello, G. A.; Van Speybroeck, V.; Ceriotti, M., i-PI 2.0: A universal force engine for advanced molecular simulations. *Computer Physics Communications* **2019**, *236*, 214-223.
336. Ceriotti, M.; Bussi, G.; Parrinello, M., Langevin Equation with Colored Noise for Constant-Temperature Molecular Dynamics Simulations. *Physical Review Letters* **2009**, *102* (2), 020601.
337. Ceriotti, M.; Bussi, G.; Parrinello, M., Colored-Noise Thermostats à la Carte. *Journal of Chemical Theory and Computation* **2010**, *6* (4), 1170-1180.
338. Monkhorst, H. J.; Pack, J. D., Special points for Brillouin-zone integrations. *Physical Review B* **1976**, *13* (12), 5188-5192.
339. 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.
340. Musil, F.; Veit, M.; Goscinski, A.; Fraux, G.; Willatt, M. J.; Stricker, M.; Junge, T.; Ceriotti, M., Efficient implementation of atom-density representations. *The Journal of Chemical Physics* **2021**, *154* (11), 114109.
341. Goscinski, A.; Musil, F.; Pozdnyakov, S.; Nigam, J.; Ceriotti, M., Optimal radial basis for density-based atomic representations. *The Journal of Chemical Physics* **2021**, *155* (10), 104106.
342. RDKit: Open-source cheminformatics, version 2022.03.4; <https://www.rdkit.org>.
343. Imbalzano, G.; Zhuang, Y.; Kapil, V.; Rossi, K.; Engel, E. A.; Grasselli, F.; Ceriotti, M., Uncertainty estimation for molecular dynamics and sampling. *The Journal of Chemical Physics* **2021**, *154* (7), 074102.
344. Balodis, M.; Cordova, M.; Hofstetter, A.; Day, G. M.; Emsley, L., De Novo Crystal Structure Determination from Machine Learned Chemical Shifts. *Journal of the American Chemical Society* **2022**, *144* (16), 7215-7223.
345. Maruyoshi, K.; Iuga, D.; Antzutkin, O. N.; Alhalaweh, A.; Velaga, S. P.; Brown, S. P., Identifying the intermolecular hydrogen-bonding supramolecular synthons in an indomethacin–nicotinamide cocrystal by solid-state NMR. *Chemical Communications* **2012**, *48* (88), 10844-10846.
346. Gervais, C.; Profeta, M.; Lafond, V.; Bonhomme, C.; Azaïs, T.; Mutin, H.; Pickard, C. J.; Mauri, F.; Babonneau, F., Combined ab initio computational and experimental multinuclear solid-state magnetic resonance study of phenylphosphonic acid. *Magnetic Resonance in Chemistry* **2004**, *42* (5), 445-452.
347. 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.
348. Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E., Array programming with NumPy. *Nature* **2020**, *585* (7825), 357-362.
349. Hourahine, B.; Aradi, B.; Blum, V.; Bonafé, F.; Buccheri, A.; Camacho, C.; Cevallos, C.; Deshayé, M. Y.; Dumitrică, T.; Dominguez, A.; Ehlert, S.; Elstner, M.; van der Heide, T.; Hermann, J.; Irle, S.; Kranz, J. J.; Köhler, C.; Kowalczyk, T.; Kubař, T.; Lee, I. S.; Lutscher, V.; Maurer, R. J.; Min, S. K.; Mitchell, I.; Negre, C.; Niehaus, T. A.; Niklasson, A. M. N.; Page, A. J.; Pecchia, A.; Penazzi, G.; Persson, M. P.; Řezáč, J.; Sánchez, C. G.; Sternberg, M.; Stöhr, M.; Stuckenberg, F.; Tkatchenko, A.; Yu, V. W. z.; Frauenheim, T., DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *The Journal of Chemical Physics* **2020**, *152* (12), 124101.
350. Gaus, M.; Goez, A.; Elstner, M., Parametrization and Benchmark of DFTB3 for Organic Molecules. *Journal of Chemical Theory and Computation* **2013**, *9* (1), 338-354.

351. Řezáč, J., Empirical Self-Consistent Correction for the Description of Hydrogen Bonds in DFTB3. *Journal of Chemical Theory and Computation* **2017**, *13* (10), 4804-4817.
352. Aradi, B.; Hourahine, B.; Frauenheim, T., DFTB+, a Sparse Matrix-Based Implementation of the DFTB Method. *The Journal of Physical Chemistry A* **2007**, *111* (26), 5678-5684.
353. Chisholm, J. A.; Motherwell, S., COMPACT: a program for identifying crystal structure similarity using distances. *Journal of Applied Crystallography* **2005**, *38* (1), 228-231.
354. Nyman, J.; Day, G. M., Static and lattice vibrational energy differences between polymorphs. *CrystEngComm* **2015**, *17* (28), 5154-5165.
355. Iuzzolino, L.; McCabe, P.; Price, Sarah L.; Brandenburg, J. G., Crystal structure prediction of flexible pharmaceutical-like molecules: density functional tight-binding as an intermediate optimisation method and for free energy estimation. *Faraday Discussions* **2018**, *211*, 275-296.
356. 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 Jr, 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. 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. 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.
357. Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P., Optimization by Simulated Annealing. *Science* **1983**, *220* (4598), 671-680.
358. Cordova, M.; Balodis, M.; Simões de Almeida, B.; Ceriotti, M.; Emsley, L., Bayesian probabilistic assignment of chemical shifts in organic solids. *Science Advances* **2021**, *7* (48), eabk2341.
359. Sebastiani, D., Current Densities and Nucleus-Independent Chemical Shift Maps from Reciprocal-Space Density Functional Perturbation Theory Calculations. *ChemPhysChem* **2006**, *7* (1), 164-175.
360. Chen, Z.; Wannere, C. S.; Corminboeuf, C.; Puchta, R.; Schleyer, P. v. R., Nucleus-Independent Chemical Shifts (NICS) as an Aromaticity Criterion. *Chemical Reviews* **2005**, *105* (10), 3842-3888.
361. Schleyer, P. v. R.; Maerker, C.; Dransfeld, A.; Jiao, H.; van Eikema Hommes, N. J. R., Nucleus-Independent Chemical Shifts: A Simple and Efficient Aromaticity Probe. *Journal of the American Chemical Society* **1996**, *118* (26), 6317-6318.
362. Schmidt, J.; Hoffmann, A.; Spiess, H. W.; Sebastiani, D., Bulk Chemical Shifts in Hydrogen-Bonded Systems from First-Principles Calculations and Solid-State-NMR. *The Journal of Physical Chemistry B* **2006**, *110* (46), 23204-23210.
363. Zilka, M.; Sturniolo, S.; Brown, S. P.; Yates, J. R., Visualising crystal packing interactions in solid-state NMR: Concepts and applications. *The Journal of Chemical Physics* **2017**, *147* (14).
364. Miclaus, M.; Grosu, I.-G.; Filip, X.; Tripon, C.; Filip, C., Optimizing structure determination from powders of crystalline organic solids with high molecular flexibility: the case of lisinopril dihydrate. *CrystEngComm* **2014**, *16* (3), 299-303.
365. Cordova, M.; Engel, E. A.; Stefaniuk, A.; Paruzzo, F.; Hofstetter, A.; Ceriotti, M.; Emsley, L., A Machine Learning Model of Chemical Shifts for Chemically and Structurally Diverse Molecular Solids. *Journal of Physical Chemistry C* **2022**, *126* (39), 16710-16720.
366. Feike, M.; Demco, D. E.; Graf, R.; Gottwald, J.; Hafner, S.; Spiess, H. W., Broadband Multiple-Quantum NMR Spectroscopy. *Journal of Magnetic Resonance, Series A* **1996**, *122* (2), 214-221.
367. Saalwächter, K.; Lange, F.; Matyjaszewski, K.; Huang, C.-F.; Graf, R., BaBa-xy16: Robust and broadband homonuclear DQ recoupling for applications in rigid and soft solids up to the highest MAS frequencies. *Journal of Magnetic Resonance* **2011**, *212* (1), 204-215.
368. Tricot, G.; Trébosc, J.; Pourpoint, F.; Gauvin, R.; Delevoye, L., The D-HMQC MAS-NMR Technique. In *Annual Reports on NMR Spectroscopy*, 2014; Vol. 81, pp 145-184.
369. Lesage, A.; Auger, C.; Caldarelli, S.; Emsley, L., Determination of Through-Bond Carbon-Carbon Connectivities in Solid-State NMR Using the INADEQUATE Experiment. *Journal of the American Chemical Society* **1997**, *119* (33), 7867-7868.
370. Guerry, P.; Herrmann, T., Advances in automated NMR protein structure determination. *Quarterly Reviews of Biophysics* **2011**, *44* (3), 257-309.
371. Schmidt, E.; Güntert, P., A New Algorithm for Reliable and General NMR Resonance Assignment. *Journal of the American Chemical Society* **2012**, *134* (30), 12817-12829.
372. Aeschbacher, T.; Schmidt, E.; Blatter, M.; Maris, C.; Duss, O.; Allain, F. H. T.; Güntert, P.; Schubert, M., Automated and assisted RNA resonance assignment using NMR chemical shift statistics. *Nucleic Acids Research* **2013**, *41* (18), e172.
373. Chávez, M.; Wiegand, T.; Malär, A. A.; Meier, B. H.; Ernst, M., Residual dipolar line width in magic-angle spinning proton solid-state NMR. *Magnetic Resonance* **2021**, *2*, 499-509.
374. Kobayashi, T.; Mao, K.; Paluch, P.; Nowak-Król, A.; Sniechowska, J.; Nishiyama, Y.; Gryko, D. T.; Potrzebowski, M. J.; Pruski, M., Study of Intermolecular Interactions in the Corrole Matrix by Solid-State NMR under 100 kHz MAS and Theoretical Calculations. *Angewandte Chemie International Edition* **2013**, *52* (52), 14108-14111.

375. Agarwal, V.; Penzel, S.; Szekely, K.; Cadalbert, R.; Testori, E.; Oss, A.; Past, J.; Samoson, A.; Ernst, M.; Böckmann, A.; Meier, B. H., De Novo 3D Structure Determination from Sub-milligram Protein Samples by Solid-State 100 kHz MAS NMR Spectroscopy. *Angewandte Chemie International Edition* **2014**, *53* (45), 12253-12256.
376. Nishiyama, Y.; Malon, M.; Ishii, Y.; Ramamoorthy, A., 3D 15N/15N/1H chemical shift correlation experiment utilizing an RFDR-based 1H/1H mixing period at 100kHz MAS. *Journal of Magnetic Resonance* **2014**, *244*, 1-5.
377. Lin, Y.-L.; Cheng, Y.-S.; Ho, C.-I.; Guo, Z.-H.; Huang, S.-J.; Org, M.-L.; Oss, A.; Samoson, A.; Chan, J. C. C., Preparation of fibril nuclei of beta-amyloid peptides in reverse micelles. *Chemical Communications* **2018**, *54* (74), 10459-10462.
378. Penzel, S.; Oss, A.; Org, M.-L.; Samoson, A.; Böckmann, A.; Ernst, M.; Meier, B. H., Spinning faster: protein NMR at MAS frequencies up to 126 kHz. *Journal of Biomolecular NMR* **2019**, *73* (1-2), 19-29.
379. Sternberg, U.; Witter, R.; Kuprov, I.; Lamley, J. M.; Oss, A.; Lewandowski, J. R.; Samoson, A., 1H line width dependence on MAS speed in solid state NMR – Comparison of experiment and simulation. *Journal of Magnetic Resonance* **2018**, *291*, 32-39.
380. Levitt, M. H.; Raleigh, D. P.; Creuzet, F.; Griffin, R. G., Theory and simulations of homonuclear spin pair systems in rotating solids. *The Journal of Chemical Physics* **1990**, *92* (11), 6347-6364.
381. Nakai, T.; McDowell, C. A., Application of Floquet theory to the nuclear magnetic resonance spectra of homonuclear two-spin systems in rotating solids. *The Journal of Chemical Physics* **1992**, *96* (5), 3452-3466.
382. Brown, S. P., Applications of high-resolution 1H solid-state NMR. *Solid State Nuclear Magnetic Resonance* **2012**, *41*, 1-27.
383. Andreas, L. B.; Jaudzems, K.; Stanek, J.; Lalli, D.; Bertarello, A.; Le Marchand, T.; Cala-De Paepe, D.; Kotelovica, S.; Akopjana, I.; Knott, B.; Wegner, S.; Engelke, F.; Lesage, A.; Emsley, L.; Tars, K.; Herrmann, T.; Pintacuda, G., Structure of fully protonated proteins by proton-detected magic-angle spinning NMR. *Proceedings of the National Academy of Sciences* **2016**, *113* (33), 9187-9192.
384. Chevelkov, V.; Rehbein, K.; Diehl, A.; Reif, B., Ultrahigh Resolution in Proton Solid-State NMR Spectroscopy at High Levels of Deuteration. *Angewandte Chemie International Edition* **2006**, *45* (23), 3878-3881.
385. Schledorn, M.; Malär, A. A.; Torosyan, A.; Penzel, S.; Klose, D.; Oss, A.; Org, M. L.; Wang, S.; Lecoq, L.; Cadalbert, R.; Samoson, A.; Böckmann, A.; Meier, B. H., Protein NMR Spectroscopy at 150 kHz Magic-Angle Spinning Continues To Improve Resolution and Mass Sensitivity. *ChemBioChem* **2020**, *21* (17), 2540-2548.
386. Gerstein, B. C.; Pembleton, R. G.; Wilson, R. C.; Ryan, L. M., High resolution NMR in randomly oriented solids with homonuclear dipolar broadening: Combined multiple pulse NMR and magic angle spinning. *The Journal of Chemical Physics* **1977**, *66* (1), 361-362.
387. Gan, Z.; Madhu, P. K.; Amoureux, J.-P.; Trébosc, J.; Lafon, O., A tunable homonuclear dipolar decoupling scheme for high-resolution proton NMR of solids from slow to fast magic-angle spinning. *Chemical Physics Letters* **2011**, *503* (1-3), 167-170.
388. Halse, M. E.; Emsley, L., Improved Phase-Modulated Homonuclear Dipolar Decoupling for Solid-State NMR Spectroscopy from Symmetry Considerations. *The Journal of Physical Chemistry A* **2013**, *117* (25), 5280-5290.
389. Leskes, M.; Steuernagel, S.; Schneider, D.; Madhu, P. K.; Vega, S., Homonuclear dipolar decoupling at magic-angle spinning frequencies up to 65kHz in solid-state nuclear magnetic resonance. *Chemical Physics Letters* **2008**, *466* (1-3), 95-99.
390. Nishiyama, Y.; Lu, X.; Trébosc, J.; Lafon, O.; Gan, Z.; Madhu, P. K.; Amoureux, J.-P., Practical choice of 1H–1H decoupling schemes in through-bond 1H–{X} HMQC experiments at ultra-fast MAS. *Journal of Magnetic Resonance* **2012**, *214*, 151-158.
391. Salager, E.; Dumez, J.-N.; Stein, R. S.; Steuernagel, S.; Lesage, A.; Elena-Herrmann, B.; Emsley, L., Homonuclear dipolar decoupling with very large scaling factors for high-resolution ultrafast magic angle spinning 1H solid-state NMR spectroscopy. *Chemical Physics Letters* **2010**, *498* (1-3), 214-220.
392. Paruzzo, F. M.; Emsley, L., High-resolution 1H NMR of powdered solids by homonuclear dipolar decoupling. *Journal of Magnetic Resonance* **2019**, *309*, 106598.
393. Vinogradov, E.; Madhu, P. K.; Vega, S., High-resolution proton solid-state NMR spectroscopy by phase-modulated Lee–Goldburg experiment. *Chemical Physics Letters* **1999**, *314* (5-6), 443-450.
394. Sakellariou, D.; Lesage, A.; Hodgkinson, P.; Emsley, L., Homonuclear dipolar decoupling in solid-state NMR using continuous phase modulation. *Chemical Physics Letters* **2000**, *319* (3-4), 253-260.
395. Moutzouri, P.; Paruzzo, F. M.; Simões de Almeida, B.; Stevanato, G.; Emsley, L., Homonuclear Decoupling in 1H NMR of Solids by Remote Correlation. *Angewandte Chemie* **2020**, *132* (15), 6294-6297.
396. Moutzouri, P.; Simões de Almeida, B.; Emsley, L., Fast remote correlation experiments for 1H homonuclear decoupling in solids. *Journal of Magnetic Resonance* **2020**, *321*, 106856.
397. Pell, A. J.; Edden, R. A. E.; Keeler, J., Broadband proton-decoupled proton spectra. *Magnetic Resonance in Chemistry* **2007**, *45* (4), 296-316.
398. Oschkinat, H.; Pastore, A.; Pfändler, P.; Bodenhausen, G., Two-dimensional correlation of directly and remotely connected transitions by z-filtered COSY. *Journal of Magnetic Resonance (1969)* **1986**, *69* (3), 559-566.
399. Schnell, I.; Spiess, H. W., High-Resolution 1H NMR Spectroscopy in the Solid State: Very Fast Sample Rotation and Multiple-Quantum Coherences. *Journal of Magnetic Resonance* **2001**, *151* (2), 153-227.
400. Moutzouri, P.; Simões de Almeida, B.; Torodii, D.; Emsley, L., Pure Isotropic Proton Solid State NMR. *Journal of the American Chemical Society* **2021**, *143* (26), 9834-9841.
401. Caravatti, P.; Bodenhausen, G.; Ernst, R. R., Heteronuclear solid-state correlation spectroscopy. *Chemical Physics Letters* **1982**, *89* (5), 363-367.
402. Caravatti, P.; Braunschweiler, L.; Ernst, R. R., Heteronuclear correlation spectroscopy in rotating solids. *Chemical Physics Letters* **1983**, *100* (4), 305-310.

403. Opella, S. J.; Frey, M. H., Selection of nonprotonated carbon resonances in solid-state nuclear magnetic resonance. *Journal of the American Chemical Society* **1979**, *101* (19), 5854-5856.
404. Wu, X. L.; Zilm, K. W., Complete Spectral Editing in CPMAS NMR. *Journal of Magnetic Resonance, Series A* **1993**, *102* (2), 205-213.
405. Wu, X. L.; Burns, S. T.; Zilm, K. W., Spectral Editing in CPMAS NMR. Generating Subspectra Based on Proton Multiplicities. *Journal of Magnetic Resonance, Series A* **1994**, *111* (1), 29-36.
406. Lesage, A.; Sakellariou, D.; Steuernagel, S.; Emsley, L., Carbon-Proton Chemical Shift Correlation in Solid-State NMR by Through-Bond Multiple-Quantum Spectroscopy. *Journal of the American Chemical Society* **1998**, *120* (50), 13194-13201.
407. Lesage, A.; Steuernagel, S.; Emsley, L., Carbon-13 Spectral Editing in Solid-State NMR Using Heteronuclear Scalar Couplings. *Journal of the American Chemical Society* **1998**, *120* (28), 7095-7100.
408. Pines, A.; Gibby, M. G.; Waugh, J. S., Proton-enhanced NMR of dilute spins in solids. *The Journal of Chemical Physics* **1973**, *59* (2), 569-590.
409. Fung, B. M.; Khitrin, A. K.; Ermolaev, K., An Improved Broadband Decoupling Sequence for Liquid Crystals and Solids. *Journal of Magnetic Resonance* **2000**, *142* (1), 97-101.
410. Elena, B.; de Paëpe, G.; Emsley, L., Direct spectral optimisation of proton-proton homonuclear dipolar decoupling in solid-state NMR. *Chemical Physics Letters* **2004**, *398* (4-6), 532-538.
411. Bax, A.; Freeman, R.; Frenkiel, T. A., An NMR technique for tracing out the carbon skeleton of an organic molecule. *Journal of the American Chemical Society* **2002**, *103* (8), 2102-2104.
412. Peersen, O. B.; Wu, X. L.; Kustanovich, I.; Smith, S. O., Variable-Amplitude Cross-Polarization MAS NMR. *Journal of Magnetic Resonance, Series A* **1993**, *104* (3), 334-339.
413. Rossini, A. J.; Zagdoun, A.; Hegner, F.; Schwarzwälder, M.; Gajan, D.; Copéret, C.; Lesage, A.; Emsley, L., Dynamic Nuclear Polarization NMR Spectroscopy of Microcrystalline Solids. *Journal of the American Chemical Society* **2012**, *134* (40), 16899-16908.
414. 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. *Physical Chemistry Chemical Physics* **2010**, *12* (22), 5850-5860.
415. Gaus, M.; Lu, X.; Elstner, M.; Cui, Q., Parameterization of DFTB3/3OB for Sulfur and Phosphorus for Chemical and Biological Applications. *Journal of Chemical Theory and Computation* **2014**, *10* (4), 1518-1537.
416. Rogers, D.; Hahn, M., Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50* (5), 742-754.
417. García-Domenech, R.; Gálvez, J.; de Julián-Ortiz, J. V.; Pogliani, L., Some New Trends in Chemical Graph Theory. *Chemical Reviews* **2008**, *108* (3), 1127-1169.
418. Danishuddin; Khan, A. U., Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discovery Today* **2016**, *21* (8), 1291-1302.
419. Babai, L., Graph isomorphism in quasipolynomial time. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, 2016; pp 684-697.
420. Grohe, M.; Schweitzer, P., The graph isomorphism problem. *Communications of the ACM* **2020**, *63* (11), 128-134.
421. Weisfeiler, B.; Lehman, A., The reduction of a graph to canonical form and the algebra which appears therein. *nti, Series* **1968**, *2* (9), 12-16.
422. Brus, J.; Jegorov, A., Through-Bonds and Through-Space Solid-State NMR Correlations at Natural Isotopic Abundance: Signal Assignment and Structural Study of Simvastatin. *The Journal of Physical Chemistry A* **2004**, *108* (18), 3955-3964.
423. 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 15N-edited and 2D 14N-1H solid-state NMR. *CrystEngComm* **2012**, *14* (8), 2654-2659.
424. Uldry, A.-C.; Griffin, J. M.; Yates, J. R.; Pérez-Torralba, M.; Santa María, M. D.; Webber, A. L.; Beaumont, M. L. L.; Samoson, A.; Claramunt, R. M.; Pickard, C. J.; Brown, S. P., 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.
425. Harris, R. K.; Jackson, P., High-resolution 1H and 13C NMR of solid 2-Aminobenzoic acid. *Journal of Physics and Chemistry of Solids* **1987**, *48* (9), 813-818.
426. 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.
427. Liu, F.; Phung, C. G.; Alderman, D. W.; Grant, D. M., Carbon-13 Chemical Shift Tensors in Methyl Glycosides, Comparing Diffraction and Optimized Structures with Single-Crystal NMR. *Journal of the American Chemical Society* **1996**, *118* (43), 10629-10634.
428. Sherwood, M. H.; Alderman, D. W.; Grant, D. M., Assignment of Carbon-13 Chemical-Shift Tensors in Single-Crystal Sucrose. *Journal of Magnetic Resonance, Series A* **1993**, *104* (2), 132-145.
429. Liu, F.; Phung, C. G.; Alderman, D. W.; Grant, D. M., Analyzing and Assigning Carbon-13 Chemical-Shift Tensors in α -L-Rhamnose Monohydrate Single Crystals. *Journal of Magnetic Resonance, Series A* **1996**, *120* (2), 242-248.
430. Liu, F.; Phung, C. G.; Alderman, D. W.; Grant, D. M., Analyzing and Assigning Carbon-13 Chemical-Shift Tensors in Fructose, Sorbose, and Xylose Single Crystals. *Journal of Magnetic Resonance, Series A* **1996**, *120* (2), 231-241.

431. Malkin, V. G.; Malkina, O. L.; Salahub, D. R., Influence of Intermolecular Interactions on the ^{13}C NMR Shielding Tensor in Solid Alpha-Glycine. *Journal of the American Chemical Society* **1995**, *117* (11), 3294-3295.
432. Naito, A.; Ganapathy, S.; Akasaka, K.; McDowell, C. A., Chemical shielding tensor and ^{13}C – ^{14}N dipolar splitting in single crystals of L-alanine. *The Journal of Chemical Physics* **1981**, *74* (6), 3190-3197.
433. Chen, X.; Zhan, C.-G., First-principles studies of C- 13 NMR chemical shift tensors of amino acids in crystal state. *Journal of Molecular Structure: THEOCHEM* **2004**, *682* (1-3), 73-82.
434. Naito, A.; Ganapathy, S.; Raghunathan, P.; McDowell, C. A., Determination of the ^{14}N quadrupole coupling tensor and the ^{13}C chemical shielding tensors in a single crystal of L-serine monohydrate. *The Journal of Chemical Physics* **1983**, *79* (9), 4173-4182.
435. Naito, A.; McDowell, C. A., Determination of the ^{14}N quadrupole coupling tensors and the ^{13}C chemical shielding tensors in a single crystal of L-asparagine monohydrate. *The Journal of Chemical Physics* **1984**, *81* (11), 4795-4803.
436. Janes, N.; Ganapathy, S.; Oldfield, E., Carbon- 13 chemical shielding tensors in L-threonine. *Journal of Magnetic Resonance (1969)* **1983**, *54* (1), 111-121.
437. Sherwood, M. H.; Facelli, J. C.; Alderman, D. W.; Grant, D. M., Carbon- 13 chemical shift tensors in polycyclic aromatic compounds. 2. Single-crystal study of naphthalene. *Journal of the American Chemical Society* **2002**, *113* (3), 750-753.
438. Iuliucci, R. J.; Facelli, J. C.; Alderman, D. W.; Grant, D. M., Carbon- 13 Chemical Shift Tensors in Polycyclic Aromatic Compounds. 5. Single-Crystal Study of Acenaphthene. *Journal of the American Chemical Society* **2002**, *117* (8), 2336-2343.
439. Portieri, A.; Harris, R. K.; Fletton, R. A.; Lancaster, R. W.; Threlfall, T. L., Effects of polymorphic differences for sulfanilamide, as seen through ^{13}C and ^{15}N solid-state NMR, together with shielding calculations. *Magnetic Resonance in Chemistry* **2004**, *42* (3), 313-320.
440. Stueber, D.; Grant, D. M., ^{13}C and ^{15}N Chemical Shift Tensors in Adenosine, Guanosine Dihydrate, 2'-Deoxythymidine, and Cytidine. *Journal of the American Chemical Society* **2002**, *124* (35), 10539-10551.
441. Sharif, S.; Schagen, D.; Toney, M. D.; Limbach, H.-H., Coupling of Functional Hydrogen Bonds in Pyridoxal-5'-phosphate–Enzyme Model Systems Observed by Solid-State NMR Spectroscopy. *Journal of the American Chemical Society* **2007**, *129* (14), 4440-4455.
442. Wei, Y.; de Dios, A. C.; McDermott, A. E., Solid-State ^{15}N NMR Chemical Shift Anisotropy of Histidines: Experimental and Theoretical Studies of Hydrogen Bonding. *Journal of the American Chemical Society* **1999**, *121* (44), 10389-10394.
443. Gervais, C.; Dupree, R.; Pike, K. J.; Bonhomme, C.; Profeta, M.; Pickard, C. J.; Mauri, F., Combined First-Principles Computational and Experimental Multinuclear Solid-State NMR Investigation of Amino Acids. *The Journal of Physical Chemistry A* **2005**, *109* (31), 6960-6969.
444. Hu, J. Z.; Facelli, J. C.; Alderman, D. W.; Pugmire, R. J.; Grant, D. M., ^{15}N Chemical Shift Tensors in Nucleic Acid Bases. *Journal of the American Chemical Society* **1998**, *120* (38), 9863-9869.
445. Smith, E. D. L.; Hammond, R. B.; Jones, M. J.; Roberts, K. J.; Mitchell, J. B. O.; Price, S. L.; Harris, R. K.; Apperley, D. C.; Cherryman, J. C.; Docherty, R., The Determination of the Crystal Structure of Anhydrous Theophylline by X-ray Powder Diffraction with a Systematic Search Algorithm, Lattice Energy Calculations, and ^{13}C and ^{15}N Solid-State NMR: A Question of Polymorphism in a Given Unit Cell. *The Journal of Physical Chemistry B* **2001**, *105* (24), 5818-5826.
446. Yamada, K.; Dong, S.; Wu, G., Solid-State ^{17}O NMR Investigation of the Carbonyl Oxygen Electric-Field-Gradient Tensor and Chemical Shielding Tensor in Amides. *Journal of the American Chemical Society* **2000**, *122* (47), 11602-11609.
447. Dong, S.; Yamada, K.; Wu, G., Oxygen- 17 Nuclear Magnetic Resonance of Organic Solids. *Zeitschrift für Naturforschung A* **2000**, *55* (1-2), 21-28.
448. Krizhevsky, A.; Sutskever, I.; Hinton, G. E., ImageNet classification with deep convolutional neural networks. *Communications of the ACM* **2017**, *60* (6), 84-90.
449. Liu, J.; Osadchy, M.; Ashton, L.; Foster, M.; Solomon, C. J.; Gibson, S. J., Deep convolutional neural networks for Raman spectrum recognition: a unified solution. *The Analyst* **2017**, *142* (21), 4067-4074.
450. Chen, D.; Wang, Z.; Guo, D.; Orekhov, V.; Qu, X., Review and Prospect: Deep Learning in Nuclear Magnetic Resonance Spectroscopy. *Chemistry – A European Journal* **2020**, *26* (46), 10391-10401.
451. Kong, W.; Dong, Z. Y.; Jia, Y.; Hill, D. J.; Xu, Y.; Zhang, Y., Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network. *IEEE Transactions on Smart Grid* **2019**, *10* (1), 841-851.
452. Jain, V.; Biesinger, M. C.; Linford, M. R., The Gaussian-Lorentzian Sum, Product, and Convolution (Voigt) functions in the context of peak fitting X-ray photoelectron spectroscopy (XPS) narrow scans. *Applied Surface Science* **2018**, *447*, 548-553.
453. Mehring, M., *Principles of high-resolution NMR in solids*. 2nd, rev. and enl. ed.; Springer-Verlag: Berlin ; New York, 1983; p viii, 342 p.
454. Filip, C.; Hafner, S.; Schnell, I.; Demco, D. E.; Spiess, H. W., Solid-state nuclear magnetic resonance spectra of dipolar-coupled multi-spin systems under fast magic angle spinning. *The Journal of Chemical Physics* **1999**, *110*, 423-440.
455. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; Woo, W.-c., Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, 2015; Vol. 28.
456. Gu, Z.; Ridenour, C. F.; Bronnimann, C. E.; Iwashita, T.; McDermott, A., Hydrogen Bonding and Distance Studies of Amino Acids and Peptides Using Solid State 2D ^1H – ^{13}C Heteronuclear Correlation Spectra. *Journal of the American Chemical Society* **1996**, *118* (4), 822-829.
457. Lee, G. S. H.; Taylor, R. C.; Dawson, M.; Kannangara, G. S. K.; Wilson, M. A., High-resolution solid state ^{13}C nuclear magnetic resonance spectra of 3,4-methylenedioxymphetamine hydrochloride and related compounds and their mixtures with lactose. *Solid State Nuclear Magnetic Resonance* **2000**, *16* (4), 225-237.

458. Morimoto, B. H.; Lovell, S.; Kahr, B., Ecstasy: 3,4-Methylenedioxymethamphetamine (MDMA). *Acta Crystallographica Section C Crystal Structure Communications* **1998**, *54* (2), 229-231.
459. Vanderhart, D. L.; Earl, W. L.; Garroway, A. N., Resolution in ^{13}C NMR of organic solids using high-power proton decoupling and magic-angle sample spinning. *Journal of Magnetic Resonance (1969)* **1981**, *44* (2), 361-401.
460. Alla, M.; Lippmaa, E., Resolution limits in magic-angle rotation NMR spectra of polycrystalline solids. *Chemical Physics Letters* **1982**, *87* (1), 30-33.
461. Jeener, J.; Meier, B. H.; Bachmann, P.; Ernst, R. R., Investigation of exchange processes by two-dimensional NMR spectroscopy. *The Journal of Chemical Physics* **1979**, *71* (11), 4546-4553.
462. Meier, B. H., Polarization transfer and spin diffusion in solid-state NMR. In *Advances in Magnetic and Optical Resonance*, 1994; Vol. 18, pp 1-116.
463. Cordova, M.; Moutzouri, P.; Simões de Almeida, B.; Torodii, D.; Emsley, L., Pure Isotropic Proton NMR Spectra in Solids using Deep Learning. *Angewandte Chemie-International Edition* **2023**, *62* (8), e202216607.
464. Klukowski, P.; Augoff, M.; Zięba, M.; Drwal, M.; Gonczarek, A.; Walczak, M. J.; Valencia, A., NMRNet: a deep learning approach to automated peak picking of protein NMR spectra. *Bioinformatics* **2018**, *34* (15), 2590-2597.
465. Cadars, S.; Lesage, A.; Emsley, L., Chemical Shift Correlations in Disordered Solids. *Journal of the American Chemical Society* **2005**, *127* (12), 4466-4476.
466. Sakellariou, D.; Brown, S. P.; Lesage, A.; Hediger, S.; Bardet, M.; Meriles, C. A.; Pines, A.; Emsley, L., High-Resolution NMR Correlation Spectra of Disordered Solids. *Journal of the American Chemical Society* **2003**, *125* (14), 4376-4380.
467. Fabbiani, M.; Al-Nahari, S.; Piveteau, L.; Dib, E.; Veremeienko, V.; Gaje, A.; Dumitrescu, D. G.; Gaveau, P.; Mineva, T.; Massiot, D.; van der Lee, A.; Haines, J.; Alonso, B., Host–Guest Silicalite-1 Zeolites: Correlated Disorder and Phase Transition Inhibition by a Small Guest Modification. *Chemistry of Materials* **2022**, *34* (1), 366-387.
468. Fayon, F.; Le Saout, G.; Emsley, L.; Massiot, D., Through-bond phosphorus–phosphorus connectivities in crystalline and disordered phosphates by solid-state NMR. *Chemical Communications* **2002**, (16), 1702-1703.
469. Corlett, E. K.; Blade, H.; Hughes, L. P.; Sidebottom, P. J.; Walker, D.; Walton, R. I.; Brown, S. P., An XRD and NMR crystallographic investigation of the structure of 2,6-lutidinium hydrogen fumarate. *CrystEngComm* **2019**, *21* (22), 3502-3516.
470. Wu, X.; Hong, Y.-I.; Xu, B.; Nishiyama, Y.; Jiang, W.; Zhu, J.; Zhang, G.; Kitagawa, S.; Horike, S., Perfluoroalkyl-Functionalized Covalent Organic Frameworks with Superhydrophobicity for Anhydrous Proton Conduction. *Journal of the American Chemical Society* **2020**, *142* (33), 14357-14364.
471. Cadars, S.; Layrac, G.; Gérardin, C.; Deschamps, M.; Yates, J. R.; Tichit, D.; Massiot, D., Identification and Quantification of Defects in the Cation Ordering in Mg/Al Layered Double Hydroxides. *Chemistry of Materials* **2011**, *23* (11), 2821-2831.
472. Hanrahan, M. P.; Venkatesh, A.; Carnahan, S. L.; Calahan, J. L.; Lubach, J. W.; Munson, E. J.; Rossini, A. J., Enhancing the resolution of ^1H and ^{13}C solid-state NMR spectra by reduction of anisotropic bulk magnetic susceptibility broadening. *Physical Chemistry Chemical Physics* **2017**, *19* (41), 28153-28162.
473. Schwerk, U.; Michel, D.; Pruski, M., Local Magnetic Field Distribution in a Polycrystalline Sample Exposed to a Strong Magnetic Field. *Journal of Magnetic Resonance, Series A* **1996**, *119* (2), 157-164.
474. Dumez, J.-N.; Butler, M. C.; Salager, E.; Elena-Herrmann, B.; Emsley, L., Ab initio simulation of proton spin diffusion. *Physical Chemistry Chemical Physics* **2010**, *12* (32), 9172-9175.
475. Deschamps, M.; Fayon, F.; Cadars, S.; Rollet, A.-L.; Massiot, D., ^1H and ^{19}F ultra-fast MAS double-quantum single-quantum NMR correlation experiments using three-spin terms of the dipolar homonuclear Hamiltonian. *Physical Chemistry Chemical Physics* **2011**, *13* (17), 8024-8030.
476. Nishiyama, Y.; Agarwal, V.; Zhang, R., Efficient symmetry-based γ -encoded DQ recoupling sequences for suppression of t_1 -noise in solid-state NMR spectroscopy at fast MAS. *Solid State Nuclear Magnetic Resonance* **2021**, *114*, 101734.
477. King, R. D.; Muggleton, S.; Lewis, R. A.; Sternberg, M. J., Drug design by machine learning: the use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proceedings of the National Academy of Sciences* **1992**, *89* (23), 11322-11326.
478. McTigue, M.; Murray, B. W.; Chen, J. H.; Deng, Y.-L.; Solowiej, J.; Kania, R. S., Molecular conformations, interactions, and properties associated with drug efficiency and clinical performance among VEGFR TK inhibitors. *Proceedings of the National Academy of Sciences* **2012**, *109* (45), 18281-18289.
479. Daina, A.; Michelin, O.; Zoete, V., iLOGP: A Simple, Robust, and Efficient Description of n-Octanol/Water Partition Coefficient for Drug Design Using the GB/SA Approach. *Journal of Chemical Information and Modeling* **2014**, *54* (12), 3284-3301.
480. Harris, K. D. M., Powder Diffraction Crystallography of Molecular Solids. In *Advanced X-Ray Crystallography*, 2012; Vol. 315, pp 133-177.
481. Hughes, C. E.; Boughdiri, I.; Bouakkaz, C.; Williams, P. A.; Harris, K. D. M., Elucidating the Crystal Structure of dl-Arginine by Combined Powder X-ray Diffraction Data Analysis and Periodic DFT-D Calculations. *Crystal Growth & Design* **2017**, *18* (1), 42-46.
482. Amstad, E.; Gopinadhan, M.; Holtze, C.; Osuji, C. O.; Brenner, M. P.; Spaepen, F.; Weitz, D. A., Production of amorphous nanoparticles by supersonic spray-drying with a microfluidic nebulator. *Science* **2015**, *349* (6251), 956-960.
483. Descamps, M., Amorphous Pharmaceutical Solids. *Advanced Drug Delivery Reviews* **2016**, *100*, 1-2.
484. Nartowski, K. P.; Malhotra, D.; Hawarden, L. E.; Sibik, J.; Iuga, D.; Zeitler, J. A.; Fábíán, L.; Khimyak, Y. Z., ^{19}F NMR Spectroscopy as a Highly Sensitive Method for the Direct Monitoring of Confined Crystallization within Nanoporous Materials. *Angewandte Chemie International Edition* **2016**, *55* (31), 8904-8908.
485. Rams-Baron, M.; Jachowicz, R.; Boldyreva, E.; Zhou, D.; Jamroz, W.; Paluch, M., Why Amorphous Drugs? In *Amorphous Drugs*, Springer, Cham.: 2018.

486. Suresh, K.; Matzger, A. J., Enhanced Drug Delivery by Dissolution of Amorphous Drug Encapsulated in a Water Unstable Metal–Organic Framework (MOF). *Angewandte Chemie International Edition* **2019**, *58* (47), 16790–16794.
487. Wilson, V. R.; Lou, X.; Osterling, D. J.; Stolarik, D. F.; Jenkins, G. J.; Nichols, B. L. B.; Dong, Y.; Edgar, K. J.; Zhang, G. G. Z.; Taylor, L. S., Amorphous solid dispersions of enzalutamide and novel polysaccharide derivatives: investigation of relationships between polymer structure and performance. *Scientific Reports* **2020**, *10*, 18535.
488. Tong, P.; Zografi, G., Effects of water vapor absorption on the physical and chemical stability of amorphous sodium indomethacin. *AAPS PharmSciTech* **2004**, *5* (2), 26.
489. Surana, R.; Pyne, A.; Suryanarayanan, R., Effect of Aging on the Physical Properties of Amorphous Trehalose. *Pharmaceutical Research* **2004**, *21* (5), 867–874.
490. Indulkar, A. S.; Lou, X.; Zhang, G. G. Z.; Taylor, L. S., Insights into the Dissolution Mechanism of Ritonavir–Copovidone Amorphous Solid Dispersions: Importance of Congruent Release for Enhanced Performance. *Molecular Pharmaceutics* **2019**, *16* (3), 1327–1339.
491. Rossini, A. J.; Widdifield, C. M.; Zagdoun, A.; Lelli, M.; Schwarzwälder, M.; Copéret, C.; Lesage, A.; Emsley, L., Dynamic Nuclear Polarization Enhanced NMR Spectroscopy for Pharmaceutical Formulations. *Journal of the American Chemical Society* **2014**, *136* (6), 2324–2334.
492. Ni, Q. Z.; Yang, F.; Can, T. V.; Sergeyev, I. V.; D’Addio, S. M.; Jawa, S. K.; Li, Y.; Lipert, M. P.; Xu, W.; Williamson, R. T.; Leone, A.; Griffin, R. G.; Su, Y., In Situ Characterization of Pharmaceutical Formulations by Dynamic Nuclear Polarization Enhanced MAS NMR. *The Journal of Physical Chemistry B* **2017**, *121* (34), 8132–8141.
493. Kerber, R. N.; Kermagoret, A.; Callens, E.; Florian, P.; Massiot, D.; Lesage, A.; Copéret, C.; Delbecq, F.; Rozanska, X.; Sautet, P., Nature and Structure of Aluminum Surface Sites Grafted on Silica from a Combination of High-Field Aluminum-27 Solid-State NMR Spectroscopy and First-Principles Calculations. *Journal of the American Chemical Society* **2012**, *134* (15), 6767–6775.
494. Schahl, A.; Gerber, I. C.; Réat, V.; Jolibois, F., Diversity of the Hydrogen Bond Network and Its Impact on NMR Parameters of Amylose B Polymorph: A Study Using Molecular Dynamics and DFT Calculations Within Periodic Boundary Conditions. *The Journal of Physical Chemistry B* **2020**, *125* (1), 158–168.
495. Holmes, J. B.; Liu, V.; Caulkins, B. G.; Hilario, E.; Ghosh, R. K.; Drago, V. N.; Young, R. P.; Romero, J. A.; Gill, A. D.; Bogie, P. M.; Paulino, J.; Wang, X.; Riviere, G.; Bosken, Y. K.; Struppe, J.; Hassan, A.; Guidoulianov, J.; Perrone, B.; Mentink-Vigier, F.; Chang, C.-e. A.; Long, J. R.; Hooley, R. J.; Mueser, T. C.; Dunn, M. F.; Mueller, L. J., Imaging active site chemistry and protonation states: NMR crystallography of the tryptophan synthase α -aminoacrylate intermediate. *Proceedings of the National Academy of Sciences* **2022**, *119* (2), e2109235119.
496. Jose, K. V. J.; Raghavachari, K., Fragment-Based Approach for the Evaluation of NMR Chemical Shifts for Large Biomolecules Incorporating the Effects of the Solvent Environment. *Journal of Chemical Theory and Computation* **2017**, *13* (3), 1147–1158.
497. Gascón, J. A.; Sproviero, E. M.; Batista, V. S., QM/MM Study of the NMR Spectroscopy of the Retinyl Chromophore in Visual Rhodopsin. *Journal of Chemical Theory and Computation* **2005**, *1* (4), 674–685.
498. Jin, X.; Zhu, T.; Zhang, J. Z. H.; He, X., Automated Fragmentation QM/MM Calculation of NMR Chemical Shifts for Protein-Ligand Complexes. *Frontiers in Chemistry* **2018**, *6*.
499. Uluca, B.; Viennet, T.; Petrović, D.; Shaykhalishahi, H.; Weirich, F.; Gönülalan, A.; Strodel, B.; Etzkorn, M.; Hoyer, W.; Heise, H., DNP-Enhanced MAS NMR: A Tool to Snapshot Conformational Ensembles of α -Synuclein in Different States. *Biophysical Journal* **2018**, *114* (7), 1614–1623.
500. Heise, H.; Luca, S.; de Groot, B. L.; Grubmüller, H.; Baldus, M., Probing Conformational Disorder in Neurotensin by Two-Dimensional Solid-State NMR and Comparison to Molecular Dynamics Simulations. *Biophysical Journal* **2005**, *89* (3), 2113–2120.
501. Siemer, A. B., Advances in studying protein disorder with solid-state NMR. *Solid State Nuclear Magnetic Resonance* **2020**, *106*, 101643.
502. Li, J.; Bennett, K. C.; Liu, Y.; Martin, M. V.; Head-Gordon, T., Accurate prediction of chemical shifts for aqueous protein structure on “Real World” data. *Chemical Science* **2020**, *11* (12), 3180–3191.
503. Ericsson, H.; Nelander, K.; Lagerstrom-Fermer, M.; Balendran, C.; Bhat, M.; Chialda, L.; Gan, L.-M.; Heijer, M.; Kjaer, M.; Lambert, J.; Lindstedt, E.-L.; Forsberg, G.-B.; Whatling, C.; Skrtic, S., Initial Clinical Experience with AZD5718, a Novel Once Daily Oral 5-Lipoxygenase Activating Protein Inhibitor. *Clinical and Translational Science* **2018**, *11* (3), 330–338.
504. Pettersen, D.; Broddefalk, J.; Emtenäs, H.; Hayes, M. A.; Lemurell, M.; Swanson, M.; Ulander, J.; Whatling, C.; Amilon, C.; Ericsson, H.; Westin Eriksson, A.; Granberg, K.; Plowright, A. T.; Shamovsky, I.; Dellsén, A.; Sundqvist, M.; Någård, M.; Lindstedt, E.-L., Discovery and Early Clinical Development of an Inhibitor of 5-Lipoxygenase Activating Protein (AZD5718) for Treatment of Coronary Artery Disease. *Journal of Medicinal Chemistry* **2019**, *62* (9), 4312–4324.
505. Hukushima, K.; Nemoto, K., Exchange Monte Carlo Method and Application to Spin Glass Simulations. *Journal of the Physical Society of Japan* **1996**, *65* (6), 1604–1608.
506. Steiner, T., The Hydrogen Bond in the Solid State. *Angewandte Chemie International Edition* **2002**, *41* (1), 48–76.
507. Yuan, X.; Xiang, T.-X.; Anderson, B. D.; Munson, E. J., Hydrogen Bonding Interactions in Amorphous Indomethacin and Its Amorphous Solid Dispersions with Poly(vinylpyrrolidone) and Poly(vinylpyrrolidone-co-vinyl acetate) Studied Using ¹³C Solid-State NMR. *Molecular Pharmaceutics* **2015**, *12* (12), 4518–4528.
508. Melnyk, A.; Junk, M. J. N.; McGehee, M. D.; Chmelka, B. F.; Hansen, M. R.; Andrienko, D., Macroscopic Structural Compositions of π -Conjugated Polymers: Combined Insights from Solid-State NMR and Molecular Dynamics Simulations. *The Journal of Physical Chemistry Letters* **2017**, *8* (17), 4155–4160.

509. Lesage, A.; Lelli, M.; Gajan, D.; Caporini, M. A.; Vitzthum, V.; Miéville, P.; Alauzun, J.; Roussey, A.; Thieuleux, C.; Mehdi, A.; Bodenhausen, G.; Coperet, C.; Emsley, L., Surface Enhanced NMR Spectroscopy by Dynamic Nuclear Polarization. *Journal of the American Chemical Society* **2010**, *132* (44), 15459-15461.
510. Zagdoun, A.; Rossini, A. J.; Gajan, D.; Bourdolle, A.; Ouari, O.; Rosay, M.; Maas, W. E.; Tordo, P.; Lelli, M.; Emsley, L.; Lesage, A.; Copéret, C., Non-aqueous solvents for DNP surface enhanced NMR spectroscopy. *Chemical Communications* **2012**, *48* (5), 654-656.
511. Zagdoun, A.; Casano, G.; Ouari, O.; Schwarzwälder, M.; Rossini, A. J.; Aussenac, F.; Yulikov, M.; Jeschke, G.; Copéret, C.; Lesage, A.; Tordo, P.; Emsley, L., Large Molecular Weight Nitroxide Biradicals Providing Efficient Dynamic Nuclear Polarization at Temperatures up to 200 K. *Journal of the American Chemical Society* **2013**, *135* (34), 12790-12797.
512. Lelli, M.; Chaudhari, S. R.; Gajan, D.; Casano, G.; Rossini, A. J.; Ouari, O.; Tordo, P.; Lesage, A.; Emsley, L., Solid-State Dynamic Nuclear Polarization at 9.4 and 18.8 T from 100 K to Room Temperature. *Journal of the American Chemical Society* **2015**, *137* (46), 14558-14561.
513. Ong, T.-C.; Mak-Jurkauskas, M. L.; Walish, J. J.; Michaelis, V. K.; Corzilius, B.; Smith, A. A.; Clausen, A. M.; Cheetham, J. C.; Swager, T. M.; Griffin, R. G., Solvent-Free Dynamic Nuclear Polarization of Amorphous and Crystalline ortho-Terphenyl. *The Journal of Physical Chemistry B* **2013**, *117* (10), 3040-3046.
514. 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.
515. States, D. J.; Haberkorn, R. A.; Ruben, D. J., A two-dimensional nuclear overhauser experiment with pure absorption phase in four quadrants. *Journal of Magnetic Resonance (1969)* **1982**, *48* (2), 286-292.
516. Rossini, A. J.; Zagdoun, A.; Lelli, M.; Lesage, A.; Copéret, C.; Emsley, L., Dynamic Nuclear Polarization Surface Enhanced NMR Spectroscopy. *Accounts of Chemical Research* **2013**, *46* (9), 1942-1951.
517. Yarava, J. R.; Chaudhari, S. R.; Rossini, A. J.; Lesage, A.; Emsley, L., Solvent suppression in DNP enhanced solid state NMR. *Journal of Magnetic Resonance* **2017**, *277*, 149-153.
518. Marion, D.; Ikura, M.; Tschudin, R.; Bax, A., Rapid recording of 2D NMR spectra without phase cycling. Application to the study of hydrogen exchange in proteins. *Journal of Magnetic Resonance (1969)* **1989**, *85* (2), 393-399.
519. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* **2010**, *132* (15), 154104.
520. Ditchfield, R.; Hehre, W. J.; Pople, J. A., Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *The Journal of Chemical Physics* **1971**, *54* (2), 724-728.
521. Hehre, W. J.; Ditchfield, R.; Pople, J. A., Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *The Journal of Chemical Physics* **1972**, *56* (5), 2257-2261.
522. Frandl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A., Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *The Journal of Chemical Physics* **1982**, *77* (7), 3654-3665.
523. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 09*, Revision D.01; Gaussian Inc.: Wallingford CT, 2009.
524. Tomasi, J.; Mennucci, B.; Cammi, R., Quantum Mechanical Continuum Solvation Models. *Chemical Reviews* **2005**, *105* (8), 2999-3094.
525. Breneman, C. M.; Wiberg, K. B., Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *Journal of Computational Chemistry* **1990**, *11* (3), 361-373.
526. Broo, A.; Nilsson Lill, S. O., Transferable force field for crystal structure predictions, investigation of performance and exploration of different rescoring strategies using DFT-D methods. *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials* **2016**, *72* (4), 460-476.
527. *MacroModel*, Release 2017-3; Schrodinger, LLC: New York, NY, 2016.
528. Neumann, M. A.; Perrin, M.-A., Energy Ranking of Molecular Crystals Using Density Functional Theory Calculations and an Empirical van der Waals Correction. *The Journal of Physical Chemistry B* **2005**, *109* (32), 15531-15541.
529. Neumann, M. A.; Leusen, F. J. J.; Kendrick, J., A Major Advance in Crystal Structure Prediction. *Angewandte Chemie International Edition* **2008**, *47* (13), 2427-2430.
530. Kresse, G.; Hafner, J., Ab initio molecular dynamics for liquid metals. *Physical Review B* **1993**, *47* (1), 558-561.
531. Kresse, G.; Hafner, J., Ab initio molecular-dynamics simulation of the liquid-metal-amorphous-semiconductor transition in germanium. *Physical Review B* **1994**, *49* (20), 14251-14269.
532. Kresse, G.; Furthmüller, J., Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science* **1996**, *6* (1), 15-50.
533. Kresse, G.; Furthmüller, J., Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B* **1996**, *54* (16), 11169-11186.

534. Beran, G. J. O., Approximating quantum many-body intermolecular interactions in molecular clusters using classical polarizable force fields. *The Journal of Chemical Physics* **2009**, *130* (16).
535. Beran, G. J. O.; Nanda, K., Predicting Organic Crystal Lattice Energies with Chemical Accuracy. *The Journal of Physical Chemistry Letters* **2010**, *1* (24), 3480-3487.
536. Frisch, M. J. T.; G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16*, Revision A.03; Gaussian, Inc.: Wallingford CT, 2016.
537. *BIOVIA Materials Studio*, Release 2017; BIOVIA, Dassault Systèmes: San Diego, 2017.
538. Sun, H.; Jin, Z.; Yang, C.; Akkermans, R. L. C.; Robertson, S. H.; Spenley, N. A.; Miller, S.; Todd, S. M., COMPASS II: extended coverage for polymer and drug-like molecule databases. *Journal of Molecular Modeling* **2016**, *22* (2), 47.
539. 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* (45), 11225-11236.
540. Banks, J. L.; Beard, H. S.; Cao, Y.; Cho, A. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; Maple, J. R.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Zhang, L. Y.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, R. M., Integrated Modeling Program, Applied Chemical Theory (IMPACT). *Journal of Computational Chemistry* **2005**, *26* (16), 1752-1780.
541. *ffld_server*, Release 2017-3; Schrödinger, LLC: New York, NY, 2017.
542. Frolov, A. I.; Kiselev, M. G., Prediction of Cosolvent Effect on Solvation Free Energies and Solubilities of Organic Compounds in Supercritical Carbon Dioxide Based on Fully Atomistic Molecular Simulations. *The Journal of Physical Chemistry B* **2014**, *118* (40), 11769-11780.
543. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79* (2), 926-935.
544. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C., GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry* **2005**, *26* (16), 1701-1718.
545. Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E., GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19-25.
546. Bussi, G.; Donadio, D.; Parrinello, M., Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **2007**, *126* (1), 014101.
547. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **1984**, *81* (8), 3684-3690.
548. Parrinello, M.; Rahman, A., Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **1981**, *52* (12), 7182-7190.
549. Nosé, S.; Klein, M. L., Constant pressure molecular dynamics for molecular systems. *Molecular Physics* **1983**, *50* (5), 1055-1076.
550. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: AnN-log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993**, *98* (12), 10089-10092.
551. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G., A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **1995**, *103* (19), 8577-8593.
552. Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M., LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **1997**, *18* (12), 1463-1472.
553. Chakraborty, A.; Hanson, L.; Robinson, D.; Lewis, H.; Bickerton, S.; Davies, M.; Polanski, R.; Whiteley, R.; Koers, A.; Atkinson, J.; Baker, T.; del Barco Barrantes, I.; Ciotta, G.; Kettle, J. G.; Magiera, L.; Martins, C. P.; Peter, A.; Wigmore, E.; Underwood, Z.; Cosulich, S.; Niedbala, M.; Ross, S., AZD4625 is a Potent and Selective Inhibitor of KRASG12C. *Molecular Cancer Therapeutics* **2022**, *21* (10), 1535-1546.
554. Kettle, J. G.; Bagal, S. K.; Bickerton, S.; Bodnarchuk, M. S.; Boyd, S.; Breed, J.; Carbajo, R. J.; Cassar, D. J.; Chakraborty, A.; Cosulich, S.; Cumming, I.; Davies, M.; Davies, N. L.; Eatherton, A.; Evans, L.; Feron, L.; Fillery, S.; Gleave, E. S.; Goldberg, F. W.; Hanson, L.; Harlfinger, S.; Howard, M.; Howells, R.; Jackson, A.; Kemmitt, P.; Lamont, G.; Lamont, S.; Lewis, H. J.; Liu, L.; Niedbala, M. J.; Phillips, C.; Polanski, R.; Raubo, P.; Robb, G.; Robinson, D. M.; Ross, S.; Sanders, M. G.; Tonge, M.; Whiteley, R.; Wilkinson, S.; Yang, J.; Zhang, W., Discovery of AZD4625, a Covalent Allosteric Inhibitor of the Mutant GTPase KRASG12C. *Journal of Medicinal Chemistry* **2022**, *65* (9), 6940-6952.
555. Filik, J.; Ashton, A. W.; Chang, P. C. Y.; Chater, P. A.; Day, S. J.; Drakopoulos, M.; Gerring, M. W.; Hart, M. L.; Magdysyuk, O. V.; Michalik, S.; Smith, A.; Tang, C. C.; Terrill, N. J.; Wharmby, M. T.; Wilhelm, H., Processing two-dimensional X-ray diffraction and small-angle scattering data in DAWN 2. *Journal of Applied Crystallography* **2017**, *50* (3), 959-966.
556. Soper, A. K.; Barney, E. R., Extracting the pair distribution function from white-beam X-ray total scattering data. *Journal of Applied Crystallography* **2011**, *44* (4), 714-726.

557. Lu, C.; Wu, C.; Ghoreishi, D.; Chen, W.; Wang, L.; Damm, W.; Ross, G. A.; Dahlgren, M. K.; Russell, E.; Von Bargen, C. D.; Abel, R.; Friesner, R. A.; Harder, E. D., OPLS4: Improving Force Field Accuracy on Challenging Regimes of Chemical Space. *Journal of Chemical Theory and Computation* **2021**, *17* (7), 4291-4300.
558. *Desmond Molecular Dynamics System*, Schrödinger Release 2021-4; D. E. Shaw Research: New York, NY, 2021.
559. Bowers, K. J.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A., Scalable algorithms for molecular dynamics simulations on commodity clusters. In *Proceedings of the 2006 ACM/IEEE conference on Supercomputing - SC '06*, 2006.
560. *BIOVIA Materials Studio*, Release 2020; BIOVIA, Dassault Systèmes: San Diego, 2020.
561. Lin, J., Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* **1991**, *37* (1), 145-151.
562. Yang, Y.; Yu, H.; York, D.; Cui, Q.; Elstner, M., Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method: Third-Order Expansion of the Density Functional Theory Total Energy and Introduction of a Modified Effective Coulomb Interaction. *The Journal of Physical Chemistry A* **2007**, *111* (42), 10861-10873.
563. Choy, W.-Y.; Forman-Kay, J. D., Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *Journal of Molecular Biology* **2001**, *308* (5), 1011-1032.
564. Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M., Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR Residual Dipolar Couplings. *Journal of the American Chemical Society* **2009**, *131* (49), 17908-17918.
565. Kragelj, J.; Ozenne, V.; Blackledge, M.; Jensen, M. R., Conformational Propensities of Intrinsically Disordered Proteins from NMR Chemical Shifts. *ChemPhysChem* **2013**, *14* (13), 3034-3045.
566. Becke, A. D., Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **1993**, *98* (7), 5648-5652.
567. 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. *The Journal of Physical Chemistry* **2002**, *98* (45), 11623-11627.
568. Akkermans, R. L. C.; Spenley, N. A.; Robertson, S. H., COMPASS III: automated fitting workflows and extension to ionic liquids. *Molecular Simulation* **2020**, *47* (7), 540-551.
569. Nosé, S., A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics* **1984**, *52* (2), 255-268.
570. Hoover, W. G., Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A* **1985**, *31* (3), 1695-1697.
571. Martyna, G. J.; Tobias, D. J.; Klein, M. L., Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics* **1994**, *101* (5), 4177-4189.
572. Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Židek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D., Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577* (7792), 706-710.
573. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D., Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583-589.

Curriculum Vitae

First name	Manuel
Last name	Cordova
Date of birth	May 13 th , 1996
Place of birth	La Chaux-de-Fonds, Switzerland
Nationality	Swiss
Address	Rue de l'Aurore 2, 2345 Les Breuleux, Switzerland
Email	manucordova@bluewin.ch
Phone number	+41 79 602 82 87

Professional experience

2019-2023	Doctoral Assistant, Laboratory of Magnetic Resonance École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland Development of computational and data-driven methods for solid-state nuclear magnetic resonance (NMR) with main emphasis on NMR crystallography. 20% of the time was spent as a teaching assistant at EPFL and UNIL (supervision of master students and teaching of undergraduate chemistry courses).
2019	Undergraduate Research Assistant, Laboratory for Computational Molecular Design École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland Construction of a synthetic database of homogeneous Ni-based catalysts and development of a machine learning model to predict their efficiency for aryl ether cleavage.

Education

2019-2023	<p>Ph.D. in Computational and Physical Chemistry</p> <p>École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland</p> <p>Thesis supervisor: Prof. Lyndon Emsley</p> <p>Thesis title: “NMR Crystallography in the Big Data Era: New Methods and Applications Powered by Machine Learning”</p>
2017-2019	<p>M.Sc. in Molecular and Biological Chemistry</p> <p>Minor in Computational Science and Engineering</p> <p>École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland</p> <p>Final grade: 5.61 / 6</p> <p>Thesis supervisor: Prof. Clémence Corminboeuf</p> <p>Thesis title: “Large-Scale Screening of Nickel Catalysts for Aryl Ether Cleavage Using Machine Learning”</p>
2014-2017	<p>B.Sc. in Chemistry and Chemical Engineering</p> <p>École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland</p> <p>Final grade: 5.45 / 6</p>

Industrial collaborations

2019-2023	AstraZeneca
-----------	--------------------

Languages

- French (native)
- English (proficient)

Awards and distinctions

- BASF Monthey SA Prize (2019). Awarded to the best master thesis in molecular and biological chemistry.
- SCGC Excellency award. Awarded to the students in the top 8% final bachelor grades in chemistry and chemical engineering.

Invited oral presentations

1. "A General Method for the Structure Determination of Amorphous Drugs by NMR", 73rd annual meeting of the American Crystallographic Association. Baltimore, Maryland (USA), July 2023.
2. "NMR Crystallography in the Big Data Era: New Methods and Applications Powered by Machine Learning", IBM Research Zürich seminar. Online, July 2023.
3. "Machine Learning and Data-Driven Methods in Solid-State NMR: Unlocking New Tools for NMR Crystallography", IQ computational chemistry working group seminar. Online, August 2022.
4. "Machine-Learned Chemical Shifts for Next-Generation NMR Crystallography", Bruker Machine Learning seminar. Online, January 2022.

Conference presentations

1. "Chemical shift-dependent interaction maps in molecular solids", Swiss NMR symposium. Bern, Switzerland, September 2022 (poster presentation).
2. "Chemical shift-dependent interaction maps in molecular solids", SCS Fall Meeting. Zurich, Switzerland, September 2022 (poster presentation).
3. "Machine Learning and Data-Driven Methods in Solid-State NMR: Unlocking New Tools for NMR Crystallography", Experimental Nuclear Magnetic Resonance Conference (ENC). Orlando, Florida (USA), April 2022 (oral presentation).
4. "Bayesian Probabilistic Assignment of Organic Solids", Euromar. Online, July 2021, (oral presentation).
5. "Data-Mining Large-Scale Chemical Shift Calculations for Probabilistic Assignment of Organic Solids", Experimental Nuclear Magnetic Resonance Conference (ENC). Online, March 2021 (poster presentation).
6. "Structure Determination of an Amorphous Drug through Large-Scale NMR Predictions", AstraZeneca PT&D Pharmaceutical Research Day. Online, October 2020 (poster presentation).
7. "Data-Mining Chemical Shifts in the Solid State for Automated Assignment of Organic Crystals", Swiss Chemical Society (SCS) Fall Meeting. Online, August 2020 (poster presentation).
8. "Large-Scale Screening of Ni Catalysts for Aryl Ether Cleavage Using Machine Learning", ICAT - ETH workshop on Catalysis. Zurich, Switzerland, September 2019 (poster presentation).

Scientific publications

1. **Cordova, M.**; Moutzouri, P.; Nilsson Lill, S. O.; Cousen, A.; Kearns, M.; Norberg, S. T.; Svensk Ankarberg, A.; McCabe, J.; Pinnon, A. C.; Schantz, S.; Emsley, L., Atomic-Level Structure Determination of Amorphous Molecular Solids by NMR. *In press* **2023**.
2. **Cordova, M.**; Emsley, L., Chemical Shift-Dependent Interaction Maps in Molecular Solids. *Journal of the American Chemical Society* **2023**, *145* (29), 16109-16117.
3. Datta, K.; Caiazza, A.; Hope, M. A.; Li, J.; Mishra, A.; **Cordova, M.**; Chen, Z.; Emsley, L.; Wienk, M. M.; Janssen, R. A. J., Light-Induced Halide Segregation in 2D and Quasi-2D Mixed-Halide Perovskites. *ACS Energy Letters* **2023**, *8* (4), 1662-1670.
4. Moutzouri, P.; **Cordova, M.**; Simões de Almeida, B.; Torodii, D.; Emsley, L., Two-dimensional Pure Isotropic Proton Solid State NMR. *Angewandte Chemie International Edition* **2023**, *62* (21), e202301963.
5. **Cordova, M.**; Moutzouri, P.; Simões de Almeida, B.; Torodii, D.; Emsley, L., Pure Isotropic Proton NMR Spectra in Solids using Deep Learning. *Angewandte Chemie-International Edition* **2023**, *62* (8), e202216607.
6. Morales-Melgares, A.; Casar, Z.; Moutzouri, P.; Venkatesh, A.; **Cordova, M.**; Mohamed, A. K.; Scrivener, K. L.; Bowen, P.; Emsley, L., Atomic-Level Structure of Zinc-Modified Cementitious Calcium Silicate Hydrate. *Journal of the American Chemical Society* **2022**, *144* (50), 22915-22924.
7. **Cordova, M.**; Engel, E. A.; Stefaniuk, A.; Paruzzo, F.; Hofstetter, A.; Ceriotti, M.; Emsley, L., A Machine Learning Model of Chemical Shifts for Chemically and Structurally Diverse Molecular Solids. *Journal of Physical Chemistry C* **2022**, *126* (39), 16710-16720.
8. Balodis, M.; **Cordova, M.**; Hofstetter, A.; Day, G. M.; Emsley, L., De Novo Crystal Structure Determination from Machine Learned Chemical Shifts. *Journal of the American Chemical Society* **2022**, *144* (16), 7215-7223.
9. **Cordova, M.**; Balodis, M.; Simões de Almeida, B.; Ceriotti, M.; Emsley, L., Bayesian probabilistic assignment of chemical shifts in organic solids. *Science Advances* **2021**, *7* (48), eabk2341.
10. **Cordova, M.**; Balodis, M.; Hofstetter, A.; Paruzzo, F.; Lill, S. O. N.; Eriksson, E. S. E.; Berruyer, P.; Simões de Almeida, B.; Quayle, M. J.; Norberg, S. T.; Svensk Ankarberg, A.; Schantz, S.; Emsley, L., Structure determination of an amorphous drug through large-scale NMR predictions. *Nature Communications* **2021**, *12* (1), 2964.
11. Hope, M. A.; Nakamura, T.; Ahlawat, P.; Mishra, A.; **Cordova, M.**; Jahanbakhshi, F.; Mladenovic, M.; Runjhun, R.; Merten, L.; Hinderhofer, A.; Carlsen, B. I.; Kubicki, D. J.; Gershoni-Poranne, R.; Schneeberger, T.; Carbone, L. C.; Liu, Y. H.; Zakeeruddin, S. M.; Lewinski, J.; Hagfeldt, A.; Schreiber, F.; Rothlisberger, U.; Gratzel, M.; Milic, J. V.; Emsley, L., Nanoscale Phase Segregation in Supramolecular pi-Templating for Hybrid Perovskite Photovoltaics from NMR Crystallography. *Journal of the American Chemical Society* **2021**, *143* (3), 1529-1538.
12. Stevanato, G.; Casano, G.; Kubicki, D. J.; Rao, Y.; Hofer, L. E.; Menzildjian, G.; Karoui, H.; Siri, D.; **Cordova, M.**; Yulikov, M.; Jeschke, G.; Lelli, M.; Lesage, A.; Ouari, O.; Emsley, L., Open and Closed Radicals: Local Geometry around Unpaired Electrons Governs Magic-Angle Spinning Dynamic Nuclear Polarization Performance. *Journal of the American Chemical Society* **2020**, *142* (39), 16587-16599.
13. **Cordova, M.**; Wodrich, M. D.; Meyer, B.; Sawatlon, B.; Corminboeuf, C., Data-Driven Advancement of Homogeneous Nickel Catalyst Activity for Aryl Ether Cleavage. *ACS Catalysis* **2020**, *10* (13), 7021-7031.
14. Begušić, T.; **Cordova, M.**; Vaniček, J., Single-Hessian thawed Gaussian approximation. *The Journal of Chemical Physics* **2019**, *150* (15).