UNIVERSITY OF CALIFORNIA
RIVERSIDE

Low-Cost Strategies for Predicting Accurate Density Functional Theory-Based Nuclear Magnetic Resonance Chemical Shifts

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Chemistry

by

Pablo Andres Unzueta

June 2022

Dissertation Committee:

Dr. Gregory J. O Beran, Chairperson
Dr. Chia-en A. Chang
Dr. Leonard J. Mueller

The Dissertation of Pablo Andres Unzueta is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

First and foremost, I would to thank Greg Beran for his guidance throughout the PhD process. Without his tutelage, I would not have had the creative "guardrails" to embark on the machine learning journey. Furthermore, our lab outings in nature were always rejuvenating. I would like to give a special thanks to Cameron Cook, Brandon Lui, and Dr. Chandler Greenwell for their friendship and talks outside of lab. Thank you to Dr. Victor Fung at Oak Ridge National Lab for collaborating with me through the DOE SCGSR and the hands-on training with graph machine learning. I would also like to thank Dr. Joshua Hartman, Dr. Dominique Nocito, Dr. Jessica McKinley, Dr. Watit Sontising, Cody Perry, and Joshua Thompson for their insightful discussions. Lastly, this PhD would not have been possible without the overwhelming support of my family. Please know that you were with me through every step of the way. To Jack, Valentina, Victoria, and Luna, I hope you read this someday and realize that scientists are merely grown up kids with bigger toys. Mom, your mijo es un doctorado de química!

To Yolanda Unzueta, Maggie Unzueta, Jose Unzueta, and Agustin Ponce

ABSTRACT OF THE DISSERTATION

Low-Cost Strategies for Predicting Accurate Density Functional Theory-Based Nuclear Magnetic Resonance Chemical Shifts

by

Pablo Andres Unzueta

Doctor of Philosophy, Graduate Program in Chemistry
University of California, Riverside, June 2022
Dr. Gregory J. O Beran, Chairperson

Nuclear magnetic resonance (NMR) chemical shifts play a large role in the structural characterization of amorphous or disordered solids. When combined with X-ray diffraction, NMR-assisted crystallography can routinely generate Å resolution of crystals structures. However, the solid-state NMR spectrum can be complicated and often requires chemical shift prediction models to refine and/or validate candidate structures. While density functional theory (DFT) methods provide a reasonable computational "cost-to-accuracy" ratio for chemical shift prediction, current methods are (1) limited in accuracy by the usage of generalized gradient approximation functionals in planewave basis sets, or (2) become a computational bottleneck when applied to numerous structures. In this work, we describe our efforts to improve chemical shift prediction accuracy and computational cost.

First, we examine a simple monomer correction to the *de facto* planewave DFT NMR method, GIPAW. We show that one can improve accuracy by refining the intramolecular contribution and incorporating a more accurate description, such as that obtained by hybrid functionals like PBE0 that include a fraction of exact Hartree-Fock exchange.

However, not all systems are neatly described by periodic unit cells, such as biomolecular systems. In those cases, one can use a cluster approximation which models the atoms of interest and their neighbors with local atomic basis sets. To further reduce the computational cost, fragment methods, which decompose the system into many smaller/manageable calculations, can be used. We explore how polarizable continuum models (PCM) can be used on highly-charged fragments to better mimic the accuracy of cluster models with lower cost.

Lastly, we developed machine learning (ML) methods to reduce the computational cost of *ab intio* calculations. We demonstrate the use of ML methods to reproduce PBE0/6-311+G(2d,p) predictions of solution-phase organic molecules through Δ-ML. We show that Δ-ML "corrects" an inexpensive calculation and are 2–3 orders of magnitude faster than legacy calculations without sacrificing accuracy. Finally, we investigate a new class of ML models called graph neural networks for solid-state NMR predictions. We use convolutional and attentional graph operators for chemical shift prediction, and show the best accuracy for ^{15}N and ^{17}O compared to literature precedents.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The use of magnetism to improve our understanding of the natural world can be dated back to approximately 200 BC from the Han dynasty in China. [54] A simple compass made of magnetized iron called the *south pointing fish*, was used to assess which land was suitable for building houses, farming, and in the search of rare gems. Later adapted for sea navigation in the 11^{th} century, our ancestors relied on magnetism from an early age. In a more modern context, the phenomenon known as nuclear magnetic resonance (NMR) involves the detailed manipulation of nuclear spins. Initially reported by Rabi in the 1930s in vacuum, [189] and later refined by Bloch and Purcell (independently) for techniques that could be applied towards liquids and solids, [188, 24] NMR spectroscopy has grown to a high degree of sophistication.

From a single NMR experiment, one can extract the chemical shift (local structural information), spin-spin coupling constants (interactions with adjacent atoms), relaxation times (dynamics), and signal intensities (quantitative information). [141] All of these

observables are valuable experimental information for the characterization and/or quantification of molecular species and solid-state materials.

1.1 Pushing Beyond the Limits of Diffraction Alone

X-ray diffraction (XRD) is potent in resolving atomic positions in systems with long-range periodicity. [196] However, not all systems have well-defined repeating unit cells. The field of NMR crystallography bridges the gap in characterizing difficult or amorphous systems. By combining NMR spectroscopy for local information, XRD for the positions of heavy nuclei to generate plausible candidate structures, and computational chemistry to predict chemical shifts, NMR crystallography can routinely provide Ångstrom resolution.

For example, the anthracene photodimer from Bardeen, Mueller, and Beran undergoes a photoinduced solid-state cycloaddition (shown in figure 1.1) which yields appreciable expansion in the crystal. [37] If too much of the crystal is solid-state reacted for characterization, the crystal disintegrates. Thus, diffraction methods alone could not characterize the system to elucidate the mechanism of expansion. On the other hand, NMR probes the local interactions and its combination with XRD was successful in understanding the anisotropic rearrangement of the molecular contents in the unit cell that governs the expansion.

Another example in the literature is the characterization of disordered pharmaceuticals from the Emsley group. [174] Here, a locally-disordered drug called Tenapanor (used to treat irritable bowl syndrome) had two experimentally realizable forms. A side

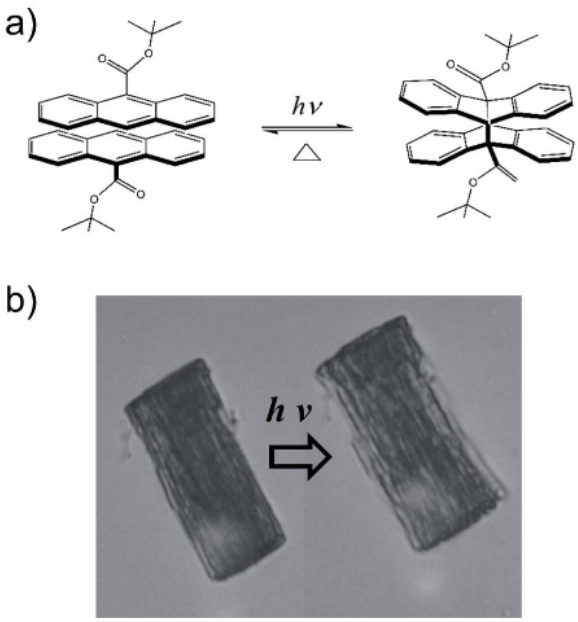


Figure 1.1: Anthracene photodimer reaction. Photomechanical crystals represent a new class of materials that generate large forces on fast timescales. After undergoing a cycloaddition reaction, the crystal appreciably expands. Using NMR crystallography, one could determine the mechanism of expansion that could not be solved using XRD alone. Figure adapted from ref. [37].

w

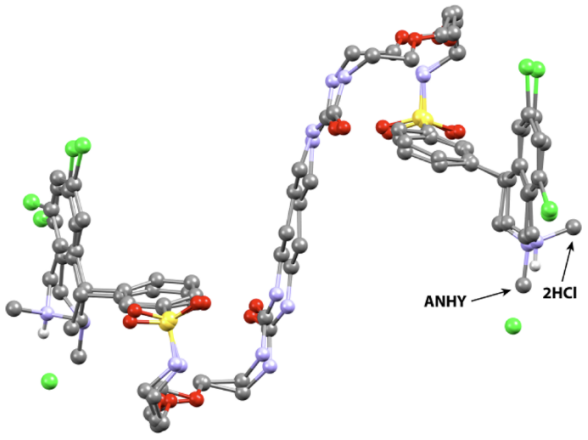


Figure 1.2: Overlay of amorphous tenapanor crystal structures. The key differences highlight the methyl group in an axial (2HCl) vs equatorial position (ANHY). By using NMR crystallography, one could determine a crystal packing strategy for the more bioavailable form, a minor product resulting from 2HCl directly related to controlling the position of the highlighted methyl group. Fig adapted from ref. [174].

product of one of the major forms was more readily bioavailable. Thus, fully characterizing the amorphous drug lead to a crystal packing strategy for production of the side product.

1.2 Improvements in Chemical Shift Prediction Expedites Structural Characterization

In both of the previous examples, chemical shift prediction plays a large role by disentangling convoluted NMR spectra and confirming candidate structures. First-principles methods such as density functional theory (DFT) are the most popular for these types of application, but they have a large computational cost relative to empirical methods. While the DFT calculations are often feasible, they can still represent a computational bottleneck, especially when high accuracy is needed or when carrying out a study on numerous structures.

For example, the most popular method for chemical shift prediction in the solid–state, GIPAW, yields a root-mean-squared-error (RMSE) of around 2.2 ppm for ^{13}C on benchmark sets. [108] However, ^{13}C shift differences between crystal polymorphs can be well below that threshold. Ideally, the statistical errors of a given method would be smaller than the shift differences between polymorphs. Furthermore, computational NMR predictions yield absolute shielding values, rather than the experimentally observed NMR chemical shift. Thus, there is a need for a high–quality mapping between shielding and shift on well-defined systems for organic and biomolecular systems.

In another example, the mechanistic determination of tryptophan synthase from the Mueller group required multiple different geometry optimizations and NMR chemical

shift calculations of various combinations of ionizable sites. [34] If one were to rapidly calculate *accurate* chemical shifts at different geometries, rather than energy-based optimization, one could arrive to the target structure faster. In fact, NMR experimental shifts can be used to monitor dynamic processes. [128] However, the steep computational scaling with respect to system size prevents the application to more complex/larger crystals, or even non-equilibrium structures.

The present work addresses the two main challenges associated with DFT-based NMR chemical shift prediction: 1) A straightforward route to **improve the accuracy of DFT-based chemical shifts**; 2) Methods to **rapidly predict chemical shifts through machine learning** (ML) algorithms. Improving the accuracy of NMR chemical shift predictions allows one to confidently generate candidate structures. While there are previously reported methods to improve accuracy, they are often computationally prohibitive. More importantly, there needs to be robust set of conversion factors. We develop these methods on curated crystal polymorph sets, and develop transferable regression parameters. Fast-moving advances in ML algorithms have developed inexpensive NMR chemical shift prediction models which bypass solving the Schrödinger equation, the main computational bottleneck. However, they are often limited in accuracy or elemental diversity. We show methods that truly reproduce DFT-based NMR predictions and explore new ML models to go beyond HCNO atom types. Improvements in both of the aforementioned areas ultimately lead to expediting characterization via NMR crystallography for accurate crystal and biomolecular structure prediction.

1.3 Overview of NMR Spectroscopy

Before one can predict chemical shifts, it is important to understand what is *physically* occurring in an NMR experiment. Electrons in molecules cause the local magnetic field to vary on a submolecular distance scale. The magnetic field of two nuclei at different sites in a molecule are different, if the electronic environments are different. For example, the protons in a $-CH_3$ group in ethanol experience a different magnetic field than protons in the $-CH_2$ group. The chemical shift is predominately an intramolecular interaction, but it does have significant intermolecular contributions. A simple example of this can be seen in how the solution-phase chemical shifts of the same molecule vary in different solvents. While the molecular structure has not changed, the intermolecular solvent/solute interactions causes changes that are observed in the chemical shift. More interestingly are the chemical shifts of formally identically molecules in the solid-state, where different packing motifs can subtly change the intermolecular contribution.

The chemical shift can be explained as a two-step process: 1) The external magnetic field $\mathbf{B}^0$ induces *currents* in the molecular electron clouds. 2) The molecular currents in turn generate a magnetic field, called the *induced field* $\mathbf{B}_j^{\text{induced}}$ of atom j. The nuclear spins thus experience an effective magnetic field $\mathbf{B}_j^{\text{eff}}$:

$$\mathbf{B}_j^{\text{eff}} = \mathbf{B}^0 - \mathbf{B}_j^{\text{induced}} \tag{1.1}$$

While the induced field, $\mathbf{B}_j^{\text{induced}}$ is typically smaller relative to the external field, $\mathbf{B^0}$, it is large enough to yield measurable chemical shifts. The induced field is linearly dependent on the externally applied field:

$$\mathbf{B}_j^{\text{induced}} = \sigma_j \cdot \mathbf{B}^0 \tag{1.2}$$

Where the symbol σ_j represents a 3×3 matrix called the *chemical shielding tensor* of site j. The chemical shielding tensor describes to what extent nucleus j is influence by the external magnetic field. We can also rewrite equation 1.1 to show how the effective magnetic field is influenced by the chemical shielding of site j:

$$\mathbf{B}_j^{\text{eff}} = \mathbf{B}^0(1 - \sigma_j) \tag{1.3}$$

In addition, chemical shielding is an anisotropic property, which means that it depends on the orientation of the molecule in the magnetic field:

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{pmatrix} \tag{1.4}$$

For each nuclear site, there are three special directions of the external magnetic field where the induced field is parallel. These special directions are always perpendicular to one another, and are called the *principal axes* of the chemical shielding tensor. Diagonalizing the shielding tensor, we obtain the following:

$$\boldsymbol{\sigma}^{PAS} = \begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{pmatrix} \tag{1.5}$$

Finally, isotropic NMR values are most often reported from experiment. We define the *isotropic* chemical shielding as:

$$\sigma_{iso} = \frac{1}{3}(\sigma_{11} + \sigma_{22} + \sigma_{33}) \tag{1.6}$$

There are many NMR conventions used in the literature to describe the line shape from experiments. The work described in chapter 2 uses the Haberlen convention [97] where the

principal components are ordered as follows:

$$|\sigma_{zz} - \sigma_{iso}| \geq |\sigma_{xx} - \sigma_{iso}| \geq |\sigma_{yy} - \sigma_{iso}| \tag{1.7}$$

This has the added advantage of describing the anisotropy and asymmetry, which yield insight into the molecular structure.

1.4 Chemical Shielding From First-Principles

Calculating chemical shielding values depends on the choice of gauge, which makes shielding values dependent on the orientation or frame of reference of the molecule. In an infinitely large basis, the gauge error disappears, [234] but one must use methods that are practical. Ditchfield's seminal work in 1974 exploits London's approach of gauge-invariant orbitals for the calculation of circular dichromism which proves a tractable route to calculate chemical shielding values that bypasses the gauge origin problem. [62, 149]

To compute the NMR chemical shielding values of a molecule, we solve the Schrödinger equation in the presence of an external magnetic field, $\mathbf{B}$.

$$\mathscr{H}(\mathbf{B}, \boldsymbol{\mu}_j)\Psi(\mathbf{B}, \boldsymbol{\mu}_j) = E(\mathbf{B}, \boldsymbol{\mu}_j)\Psi(\mathbf{B}, \boldsymbol{\mu}_j) \tag{1.8}$$

Where $\boldsymbol{\mu}_j$ is the nuclear magnetic moment of atom j. For small values of $\mathbf{B}$ and $\boldsymbol{\mu}_j$, Ψ can be expanded as a Taylor series about their zero-field values:

$$\begin{aligned}\Psi(\mathbf{B}, \boldsymbol{\mu}_j) &= \Psi^{(0)} + \sum_\alpha \left(\frac{\partial\Psi(\mathbf{B}, \boldsymbol{\mu}_j)}{B_\alpha}\right)_{\mu_{j\alpha}} B_\alpha + \sum_\alpha \left(\frac{\partial\Psi(\mathbf{B}, \boldsymbol{\mu}_j)}{\mu_{j\alpha}}\right)_{B_\alpha} \mu_{j\alpha} + ... \\ &= \Psi^{(0)} + \sum_\alpha \Psi_\alpha^{(1,0)} B_\alpha + \sum_\alpha \Psi_\alpha^{(0,1)} \mu_{j\alpha} + ...\end{aligned} \tag{1.9}$$

where we have introduced the shorthand notation $\Psi^{(1,0)}$ to indicate the partial first-derivative of the wavefunction with respect to $\mathbf{B}$ while keeping $\boldsymbol{\mu}_j$ constant and vice-versa for $\Psi^{(0,1)}$.

Similarly for the energy

$$E(\mathbf{B},\boldsymbol{\mu}_j) = E^0 + \sum_\alpha E_\alpha^{(1,0)} B_\alpha + \sum_\alpha E_{j\alpha}^{(0,1)} \mu_{j\alpha} + \frac{1}{2}\sum_\alpha \sum_\beta B_\alpha E_{\alpha,\beta}^{(2,0)} B_\beta + \sum_\alpha \sum_\beta B_\alpha E_{j\alpha,\beta}^{(1,1)} \mu_{j\beta} + \frac{1}{2}\sum_\alpha \sum_\beta \mu_{j\alpha} E_{\alpha,\beta}^{(0,2)} \mu_{j\beta} + ... \quad (1.10)$$

Alternatively, one can write equation (1.10) as the following:

$$E(\mathbf{B},\boldsymbol{\mu}_j) = E^0 - \sum_\alpha \gamma_\alpha B_\alpha - \sum_\alpha \mu_{j\alpha} B_\alpha - \frac{1}{2}\sum_\alpha \sum_\beta B_\alpha \chi_{\alpha\beta} B_\beta + \sum_\alpha \sum_\beta B_\alpha \sigma_{j\alpha\beta} \mu_{j\beta} + ... \quad (1.11)$$

In equation 1.11, γ_α is a component of the permanent magnetic moment of the molecule. The third term represents the direct interaction of the external magnetic field, and the nuclear magnetic moment. The fourth term represents the diamagnetic polarization of the molecule. And finally, the fifth term shows the induced magnetic field ($B_\alpha \sigma_{j\alpha\beta}$) in direction α. From equation 1.10 and 1.11, we see that the chemical shielding is thus defined as

$$\sigma_{\alpha,\beta}^j = \frac{\partial^2 E}{\partial B\alpha \partial \mu_{j,\beta}} \quad (1.12)$$

which describes how the electronic environment "shields" the nuclei from the magnetic field.

Recall that molecular orbitals Ψ_j are constructed by a linear combination of atomic orbitals (LCAO) ϕ_ν.

$$\Psi_j = \sum_\nu c_{\nu j} \phi_\nu \quad (1.13)$$

It becomes clear from examining 1.9 and 1.13 that one cannot use the ground state solutions (i.e. Ψ) for the perturbed wavefunctions (e.g. $\Psi^{(1,0)}$). More importantly, if one were to attempt to use 1.13 to solve for $\Psi(\mathbf{B},\boldsymbol{\mu}_j)$, it would lead to the gauge problems discussed

earlier. To include the external field in our determination of SCF solutions, we use the vector potential $\mathbf{A}(\mathbf{r})$:

$$\mathbf{A}(\mathbf{r}) = \frac{1}{2}\mathbf{B} \times \mathbf{r} \tag{1.14}$$

$$\Psi = \Psi(\mathbf{A}) \tag{1.15}$$

Formally, a gauge transformation is defined as:

$$\mathbf{A}(\mathbf{r}) \rightarrow \mathbf{A}'(\mathbf{r}) = \mathbf{A}(\mathbf{r}) + \nabla\lambda(\mathbf{r}) \tag{1.16}$$

then the wavefunction, now dependent on $\mathbf{A}$, undergoes the transformation:

$$\Psi(\mathbf{A}') = \exp(-i\Lambda)\Psi(\mathbf{A}) \tag{1.17}$$

where

$$\Lambda = \frac{1}{2c}\sum_{i}^{N}(\mathbf{B} \times \mathbf{d}) \cdot \mathbf{r}_i) \tag{1.18}$$

with c as the speed of light, $\mathbf{d}$ is the displacement vector, and $\mathbf{r}_i$ is the reference point. Equation 1.17 is exact if and only if Ψ is an exact solution to the Schrödinger equation (i.e. complete basis set limit) as seen from equation 1.18. [74]

Gauge-invariant atomic orbitals (GIAOs), χ_ν, where the real atomic basis functions ϕ_ν are multiplied by a complex factor that depends on the gauge of the vector potential $\mathbf{A}(\mathbf{r})$, bypass this issue:

$$\chi_\nu = \exp(-(i/c)\mathbf{A}_\nu \cdot \mathbf{r})\phi_\nu \tag{1.19}$$

where $\mathbf{A}_\nu$ is the value of the vector potential $\mathbf{A}$ at nuclear position $\mathbf{R}_\nu$. Since the gauge origin is now at each nuclear site, we do not have to expand equation 1.18 to the complete basis set limit to obtain gauge invariance. A more apt name for this technique would be gauge-*including* atomic orbitals, since we are including a local gauge for each atom.

1.5 DFT-Based NMR Chemical Shift Prediction

While Hartree-Fock can be used for chemical shielding calculations, it is not as accurate as one would hope. More expensive wavefunction methods like Møller-Plesset perturbation theory (MP2) or coupled-cluster methods can be used, but at a much higher computational cost. DFT provides a practical middle ground of accuracy and efficiency. Luckily, the equations defined previously for NMR apply to DFT with no additional modifications. The formalism uses the energy as a functional of the electron density (ρ):

$$E[\rho] = T_s[\{\phi_i\}] + J_{ee}[\rho] + J_{eN}[\rho] + E_{XC}[\rho] \tag{1.20}$$

Where T_s is the kinetic energy operator for a set of non-interacting electrons (in the form of Kohn-Sham molecular orbitals $\{\phi_i\}$), the electron-electron repulsion J_{ee}, electron-nuclear Coulomb attraction J_{eN}, and the exchange-correlation functional E_{XC}. For every density functional, T_s, J_{ee}, and J_{en} are identical in functional form. The only difference is in the last term E_{XC} since the exact mathematical formula is not known. [32] Researchers have thus developed a hierarchy (aka Jacob's Ladder) for density functionals incorporating various degrees of information to improve accuracy. This work focuses on two main flavors: Generalized gradient approximation (GGA) and hybrid density functionals. GGAs (like the popular PBE) incorporate contributions to the electron density from the neighboring area through the gradient of the density $\nabla\rho(r)$. Since the E_{XC} is approximate, if one where to improve the starting point description, it would result in a more accurate answer. Hybrid density functionals use this approach by using an existing density functional (like GGAs) as a starting point, and then incorporate "exact" exchange (from HF) into the GGA functional (e.g. PBE0).

For planewave DFT codes, PBE (and its variants) are the *de facto* density functionals of choice, mostly for its accuracy and computational efficiency. [181] If one were to use more accurate hybrid density functionals, the computational cost skyrockets for planewave DFT due to the increased number of integrals to evaluate, and is thus avoided by most practitioners. In addition, the core electrons are treated with a psuedopotential, and only the valence electrons receive a full quantum mechanical treatment. For solid-state NMR chemical shielding predictions, the gauge-including projector augmented wave (GIPAW) method developed by Pickard and Mauri reconstructs the true all-electron wave function with respect to the gauge problem. [185] The reconstructed wave-function is thus used for chemical shielding predictions through methods defined previously.

While planewave DFT methods work well, there are situations where increased accuracy is needed. An alternative to the planewave approach is the cluster approach, where the key atoms of interest and its neighboring atoms/unit cells are extracted to make a large cluster. Cluster methods do not suffer from the dramatic increase in computational cost associated with planewave DFT as mentioned since we have now switched to a local atomic basis. To further reduce the computational cost, one can partition the cluster into smaller subsystems (figure 1.3) via a many-body expansion approach shown in equation 1.21:

$$E_{total} = \sum_{i} E_i + \sum_{ij} \Delta E_{ij} + ... \tag{1.21}$$

where

$$\Delta E_{ij} = E_{ij} - E_i - E_j \tag{1.22}$$

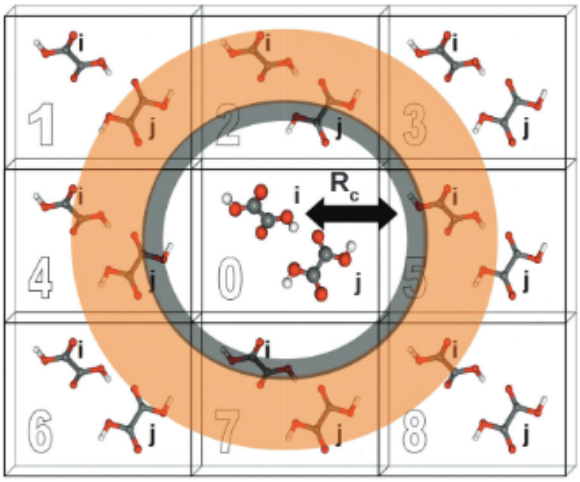


Figure 1.3: Example 2-body fragment decomposition with mixed-basis scheme from the constructed cluster. The central unit cell is treated with a large basis as well as molecular fragments within R_C. Fragments within the orange circle are treated with a smaller basis, and all atoms further out a treated with the smallest basis. Fragment pairs are constructed within some cutoff (usually 6 Å) and atoms further out are usually not included are corrected using electrostatic embedding approaches. Calculations are then run in a pairwise fashion. Since each dimer calculation is not dependent on any other fragment, the method is embarrassingly parallel. Figure adapted from ref. [113]

.

Since chemical shielding is a linear operator, we can simply differentiate equation 1.21 and obtain the following:

$$\sigma_{total} = \sum_{i} \sigma_i + \sum_{ij} \Delta\sigma_{ij} + ... \tag{1.23}$$

Equation 1.23 shows that we can achieve the total shielding of a specific atom by decomposing the system into 1-body → n-body components. Some analysis has shown that 3-body and greater contributions are not necessary. [108] In addition, one can further reduce the cost by using a mixed-basis scheme, using less dense basis sets further away from the atoms of interest. [110]

1.6 Benchmarking Solid-State Chemical Shift Accuracy

To convert from predicted chemical shieldings to experimentally observed chemical shifts, one simply calculates the chemical shielding of a standard for the atom of interest and subtract the predicted shieldings at the same model chemistry to obtain the chemical shift:

$$\delta_i = \sigma_{ref} - \sigma_i \tag{1.24}$$

For example, one could use neat tetramethylsilane (TMS) for ^{1}H and ^{13}C. [19] Alternatively, one can use a linear regression approach from a set of carefully curated experimental chemical shifts.

$$\delta_i = A + B\sigma_i \tag{1.25}$$

Comparing equation 1.24 and 1.25, it is clear that ideally A would be σ_{ref} and B is -1. By allowing these values to fluctuate, one can correct systematic errors for choice of modeling techniques (e.g. self interaction error in DFT or incomplete basis set effects). In figure 1.4, one can see the linear regression parameters for 80 measured experimental ^{1}H experimental shifts compared to a 2-body fragment predicted shieldings.

1.7 NMR Shift Prediction in the Real-World

Given a method to calculate chemical shieldings, and convert to accurate chemical shifts, how does one employ this method in the real-world? For example, in a crystal-structure prediction workflow, one generates candidate structures from XRD data and simultaneously carries out the NMR experiments to obtain the chemical shifts. [11] One then

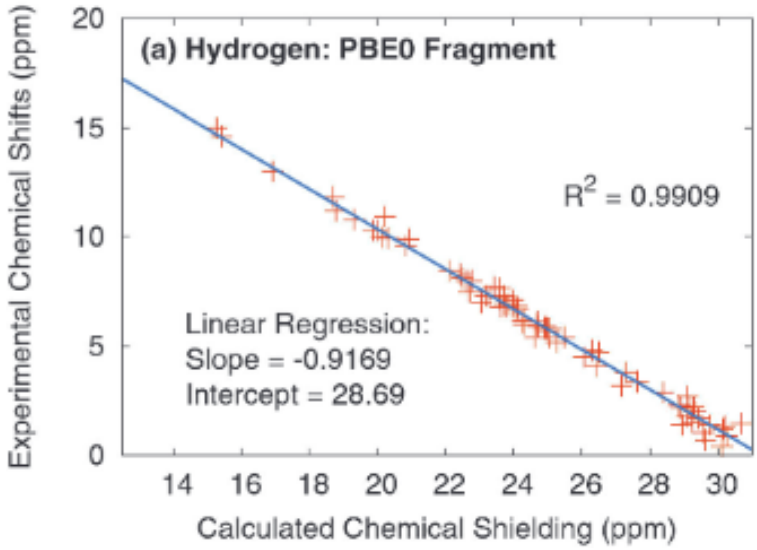


Figure 1.4: Example linear regression mapping predicted chemical shielding to measured experimental shift. Here, the slope and y-intercept deviate from the ideal values to account for the systematic errors in choice of density functional and incomplete basis set errors. Figure adapted from [113].

carries out a large-scale crystal structure landscape search to obtain plausible candidates and further refines using an accurate DFT-based geometry optimization method. Finally, one can then predict the chemical shifts on the lowest energy structures for direct comparison against experiment. Figure 1.5 demonstrates the final step in this workflow for acetaminophen using fragment PBE0 calculations. The ^{13}C predictions achieve approximately 1 ppm root-mean-squared-errors (RMSE) relative to experiment, and demonstrates the high-accuracy needed in these routines.

1.8 Research Directions

The next two chapters discuss improvements to state-of-the-art DFT-based NMR chemical shift prediction. In Chapter 2, we explore a simple monomer correction to GIPAW

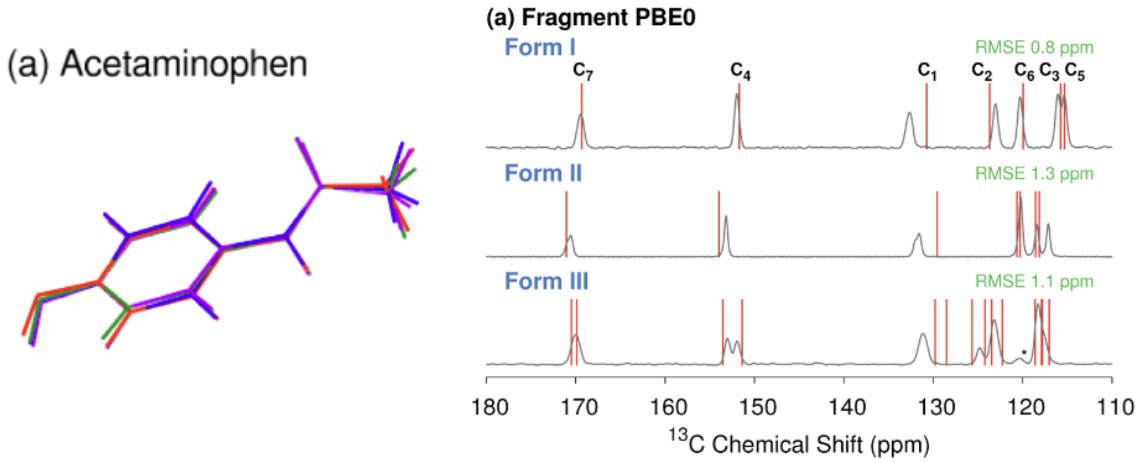


Figure 1.5: Example chemical shift predictions of various forms of acetaminophen. (*left*): Monomer overlays of the crystallographically unique monomers: forms I (red), II (blue), IIIa (green), and IIIb (purple). (*right*): Overlay of experimental chemical shift spectra of acetaminophen with their predicted shifts from fragment PBE0 calculations. Figure adapted from [112].

NMR shifts which appreciably improve prediction accuracy relative to curated molecular crystal benchmarks with experimental shifts. The monomer correction improves known deficiencies of DFT GGA functionals and incorporates a more accurate density functional or correlated wavefunction method. We then use this method on real-world test cases where GIPAW performs relatively poorly.

However, GIPAW DFT methods are not easily transferable to biomolecluar systems, which lack a small, periodic replicating unit cell. Alternatively, one can use a cluster method of the key areas of interest in biomolecules to calculate NMR chemical shifts. To further reduce the computational cost, fragment methods decompose the cluster into many smaller more manageable pieces. Isolated fragments or fragment pairs from biomolecules would have large errors in the SCF solution due to the missing "charge" stabilization. Chapter 3 explores a method to incorporate the embedding environment through the polarizable

continuum model (PCM). We show that PCM embedding fragments faithfully reproduce cluster results on our molecular crystal benchmarks. We also successfully apply the PCM fragment method to a small α-helix protein and a much larger tryptophan synthase system. Overall, the PCM fragment method allows one to obtain accurate NMR shift predictions without the computational burden of cluster methods.

In chapter 4, we briefly depart from our discussion on DFT-based NMR to outline the theoretical background of machine learning models to accelerate predictions. Machine learning (ML) is a subfield of artificial intelligence which creates a surrogate model based on a representative set of training data. ML models have a rich history of success to various problems in image recognition, natural language processing, and training intelligent agents. In the physical sciences, ML models can be used to gain insight into high dimensional data or bypass the computational cost of demanding calculations. In the later, half of this dissertation we demonstrate methods to reduce the computational burden of DFT-based chemical shift prediction with an emphasis on uncertainty estimation for transferability to real-world cases.

In chapter 5, we use ML to predict NMR chemical shieldings for solution-phase NMR. We show that a pure descriptor-based ML model is not sufficiently accurate. We thus explore Δ-ML models, where instead of learning the absolute shielding, one creates a ML model to predict a correction for an inexpensive shielding calculation. We develop our own training/testing sets from the generated databases of molecules, and show that our Δ-ML model can faithfully reproduce DFT-based NMR chemical shift predictions.

Lastly, in Chapter 6 we show that new ML models, specifically graph neural networks (GNNs), are a promising technology which out competes descriptor-based ML models. We explore GNNs for molecular crystal NMR benchmarks through the use of convolutional, attentional, and directional message-passing schemes. We show best in class accuracy for O and N in our developments in hopes of using these models for predicting experimental crystal structures via a NMR-guided geometry optimization.

Chapter 2

Improving the Accuracy of Solid-State Nuclear Magnetic Resonance Chemical Shift Prediction With a Simple Molecular Correction

2.1 Introduction

Solid-state NMR spectroscopy (ssNMR) plays an indispensable role in the characterization of solids. In past two decades, the progress of ssNMR methods has led to the development of NMR crystallography, which combines experimental ssNMR data with

theoretical simulations to obtain otherwise inaccessible insights into the structure and dynamics of solid materials. The recent rapid development of NMR crystallography has been greatly facilitated by the availability of fast and reliable computational methods that enable direct linking between structure and NMR observables. Two main approaches are used to predict NMR parameters (and other properties) of crystalline solids. First, solids can be modelled as infinite crystals using periodic boundary conditions (PBC) that ensure that a basic structural element (typically a crystallographic unit cell) is periodically repeated in all three dimensions. Second, a small part of an infinite crystal can be modelled as a molecular cluster or using small fragments. Both computational approaches have certain advantages and limitations.

The PBC approach exploits the translational repetition in crystals. Inherently periodic plane waves are used to form a basis set, instead of the local atomic orbital basis sets typically employed in molecular calculations. Because rapid variations in electron density are difficult to describe with plane waves, effective-core pseudopotentials are used to describe interactions close to the nuclei. Almost two decades ago, the gauge-including projector-augmented wave (GIPAW) procedure was developed for the prediction of the magnetic-resonance parameters in crystalline materials. [185] The method has been implemented in several density functional theory (DFT) software packages and it has been used successfully in many applications. [26, 7, 39] Unfortunately, hybrid density functionals are prohibitively demanding computationally for plane-wave calculations, and therefore, the GIPAW method has been used with the general-gradient-approximation (GGA) family of density functionals. However, many studies have demonstrated that going beyond the GGA level improves the

accuracy of the predicted NMR parameters. [108, 105, 111, 228] On the other hand, in the cluster approach, neighboring molecules or fragments are considered explicitly during the NMR calculations and traditional molecule-based software packages may be used for the calculations. [108, 228, 68, 150, 165, 10, 28, 119, 121] Although there is no fundamental limitation on the level of theory that can be used to compute the chemical shieldings in the fragments or cluster, the choice of the cluster size may be limiting, as the calculations must be maintained at a manageable size. [102] NMR parameters are generally mostly sensitive to the local environment. However, there are effects, such as electrostatic effects and ring currents, where long-range interactions are significant. It has been demonstrated that relatively large clusters have to be used for accurate predictions of NMR parameters.

Fragment methods reduce the computational costs of cluster calculations by replacing a large, many-molecule cluster with a series of electrostatically embeded monomer and dimer calculations. [21] Drawing inspiration from the earlier embedded-ion model and related approaches, the self-consistent reproduction of the Madelung potential (SCRMP) model [105] embeds these fragments electrostatically in a field of point charges designed to mimic the crystalline environment. Benchmark calculations on both isotropic shifts [105] and the principal components of the chemical shielding anisotropy (CSA) tensor [111] demonstrate very good performance of these fragment methods when hybrid density functionals are used, especially for ^{13}C and ^{15}N NMR parameters. For ^{17}O, these fragment methods exhibit a moderate degradation due to the high sensitivity of that nucleus to the electrostatic environment. Here, we propose a fast, straightforward method for computing NMR chemical shieldings that combines the advantages of both planewave and molecu-

lar computational approaches, capturing the fully periodic nature of the crystal while also obtaining the higher accuracy associated with computational models beyond GGA DFT functionals. This simple method performs a standard periodic GIPAW GGA calculation and then corrects it based on single, non-embedded gas-phase molecule calculations at any higher level of theory. This approach has roots in the incremental methods pioneered by Stoll and others decades ago. [235, 180] Recently, Boese and co-workers have presented a similar strategy for molecular crystal energies based on periodic DFT or density functional tight binding corrected with higher-level monomer and dimer corrections. [25, 63] We demonstrate that the new method significantly improves the correlations between experimental and calculated chemical shifts while adding almost no additional computational cost.

2.2 Theory and Methods

The greatest advantage of GIPAW calculations is that they inherently contain long-range interactions in crystals. On the other hand, the advantage of cluster calculations is that any computational level, such as hybrid DFT functionals or post Hartree–Fock methods can be used. The idea behind the newly proposed method is that the inaccuracy of GGA functionals for NMR shielding calculations is mostly limited to close (intramolecular) neighborhood of the nucleus of interest and long-range effects are well-approximated by the GGA-based GIPAW method. Therefore, we add a correction to the GIPAW calculated shieldings that is calculated as the difference between the shielding calculated at a higher computational level and at the GGA-level employed in the GIPAW calculation. These

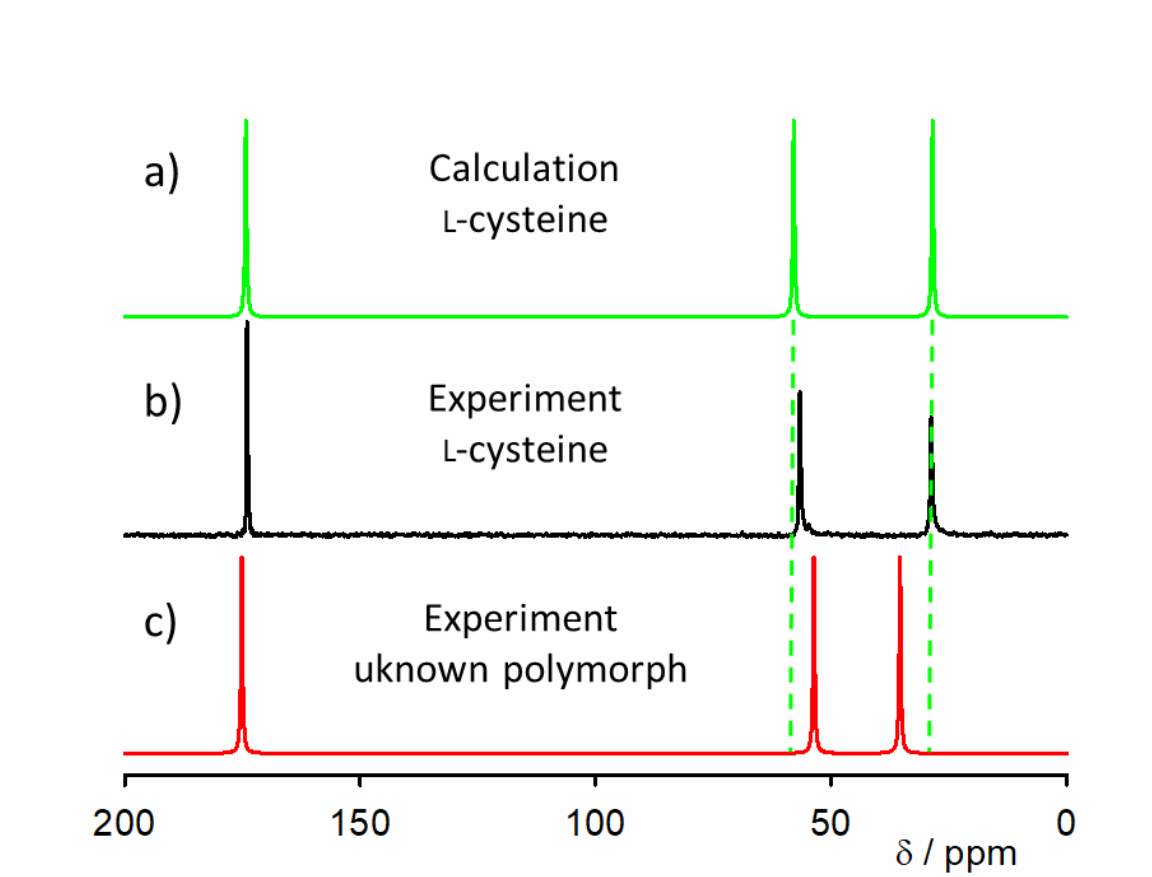


Figure 2.1: a) Calculated ^{13}C spectrum of solid L-cysteine (corrected GIPAW-PBE0); spectrum simulated using line broadening of 50 Hz. (b) Experimental CP-MAS spectrum of crystalline L -cysteine. (c) Simulated CP-MAS spectrum of cysteine using experimental chemical shifts from ref. [271] and line broadening of 50 Hz.

corrections are calculated for a single isolated molecule in the geometry taken from the crystal structure. The corrected shielding for a given atom (σ_{corr}) is calculated, for example, according to equation 2.1, where the hybrid PBE0 functional is applied to correct PBE-GIPAW shieldings.

$$\sigma_{\mathrm{corr}} = \sigma_{\mathrm{GIPAW,\ cryst.}} - \sigma_{\mathrm{PBE,\ mol.}} + \sigma_{\mathrm{PBE0,\ mol.}} \tag{2.1}$$

The proposed method consists of three basic steps: (1) geometry optimization of the crystal structure obtained by X-ray or neutron diffraction experiment and calculation of NMR chemical shieldings using the GIPAW method. (2) Calculation of NMR shieldings for a single molecule taken from the geometry-optimized structure obtained in step (1). The calculations are performed at the same level as the GIPAW calculation (typically the PBE functional) and at a higher computational level (typically a hybrid functional, such as PBE0). (3) Evaluation of the corrected shieldings according to eqn (1). Separate benchmark sets of molecular crystal structures were used to evaluate the effect of the proposed method on the agreement with experimental data of carbon, nitrogen and oxygen nuclei. All benchmark sets are based on benchmark sets used in previous studies of fragment-based chemical shift predictions in molecular crystals. [108] The benchmarks here consist of 21 structures with 132 chemical shifts in the carbon set, 16 structures and 37 shifts in the nitrogen set and 15 structures and 28 shieldings in the oxygen set. The chemical structures of all systems studied are shown in the appendix (A.1, A.2, and A.3). The NMR shieldings of the studied structures were calculated by the CASTEP program, [49] version 17.2, which is a DFT-based code that uses pseudopotentials to model the effects of core electrons, and plane waves to describe the valence electrons. Positions of all atoms were optimized prior to the NMR calculation; the unit cell parameters were fixed. Electron-correlation effects were modeled using the generalized gradient approximation of Perdew, Burke, and Ernzerhof. [181] A plane wave basis set energy cutoff of 600 eV, default 'on-the-fly generation' pseudopotentials, and a k-point spacing of 0.05 Å over the Brillouin zone via a Monkhorst–Pack grid [166] was used. The NMR calculations were performed using the GIPAW approach.

[185, 270] For comparison, the structures in the carbon set were also optimized using empirical dispersion correction, but the resulting calculated chemical shifts and corrected chemical shifts were almost identical to those obtained without the correction. [161, 243] The use of the fixed experimental unit cell parameters compensates for the artificially repulsive nature of the uncorrected density functionals. Finite temperature effects [58, 64] were not included in the calculations. However, constraining the lattice parameters to their experimental room-temperature values effectively captures the thermal expansion that occurs upon heating the crystal to room temperature. [160] DFT NMR shieldings for the isolated molecules (in vacuum) were calculated by the Gaussian16 program. [83] For co-crystals, solvates, or salts, the molecular correction was performed only on the molecule whose shielding was of interest, without the other coformer species. The gas-phase molecule input geometries were taken from the periodic DFT geometry-optimized structures and were not further optimized. To explore how the results depend on the choice of the Gaussian basis set employed, the 6-31G(d), 6-311+G(2d,p), and pcSseg-n (n = 1–3) were selected as representative basis sets for NMR shielding calculations. The pcSseg-n basis sets were obtained from basis set exchange website (`https://bse.pnl.gov/bse/portal`). [206] NMR shieldings at the coupled cluster singles and doubles (CCSD) level and 6-311+g(2d,p) basis set were calculated with CFOUR program package, which is suitable for performing high-level quantum chemical calculations on atoms and molecules. [9, 229] Corrected shieldings were obtained using eqn (1). The correlation between the corrected shieldings and experimental chemical shifts was fitted to a straight line, $\delta_i = A + B\sigma_i$ where σ_i is the computed chemical shielding and d corresponds to the experimentally observable chemical shift. The A and B parameters

of this linear correlation were used for the calculations of chemical shifts, which were then compared with experimental data. The slope B of the shielding-shift correlation 15 should equal -1 in an ideal case, but it has been shown previously that nuclear quantum effects, [66] incomplete basis sets, and other systematic errors in the DFT calculations can lead to deviations from this ideal value. Experimental chemical shifts were re-measured for a few 20 crystals in the test sets to correct issues with the earlier experiments. Solution-state NMR spectra of adenosine in DMSO-d_6 were recorded on Bruker Avance 500 (^{1}H at 500 MHz, ^{13}C at 125.8 MHz) spectrometer. The spectra were referenced to the residual solvent signal (2.50 ppm for ^{1}H and 39.7 25 ppm for ^{13}C). A combination of 1D and 2D experiments (H,H- COSY, H,C-HSQC and H,C-HMBC) was used to assign all proton and carbon signals. High-resolution ^{13}C solid-state NMR spectrum of adenosine, L-cysteine, L-glutamine, L-threonine and L-tyrosine were 30 obtained using a JEOL ECZ600R spectrometer operating at 150.9 MHz for ^{13}C and 600.2 MHz for ^{1}H and samples were packed into 3.2 mm magic angle spinning rotors (MAS) and measurements taken at MAS rate of 18 kHz using cross polarization (CP). The chemical shifts were referenced to crystalline α-glycine as a secondary reference ($\delta_{\mathrm{st}} = 176$ ppm for the carbonyl carbon). The ramped amplitude shape pulse was used during the cross polarization. The contact time for CP was 5 ms and the relaxation delays were estimated from ^{1}H saturation recovery experiments and ranged from 3s for L-threonine to 200s for adenosine. The assignment of the signals was done with the help of a CPMAS experiment with a short contact time (50 ms), where the signals of quaternary carbons are suppressed. Furthermore, a C,H-HETCOR experiment was done with the

Nucleus	Method	MAE	Max. Error
^{13}C	GIPAW	1.6	6.8
	GIPAW-corrected	0.8	3.9
	SCRMP 1 + 2-body	0.8	3.9
^{15}N	GIPAW	4.1	10.6
	GIPAW-corrected	2.8	8.3
	SCRMP 1 + 2-body	2.8	7.7
^{17}O	GIPAW	5.2	11.6
	GIPAW-corrected	4.3	10.4
	SCRMP 1 + 2-body	5.9	14.1

Table 2.1: Mean absolute errors (MAE) and maximal errors (ppm) of the predicted chemical shifts (in comparison with experiment) obtained for the conventional GIPAW method (PBE functional), corrected GIPAW (PBE0 correction, 6-311+g(2d,p) basis set) and for previously proposed SCRMP fragment method [105]

L-glutamine sample to assign unequivocally the two $C = O$ carbon signals. Experimental chemical shifts of other systems were taken from ref. [108, 68, 69] and references therein.

2.3 Results and Discussion

2.3.1 Carbon Isotropic Shifts

At first sight, the chemical shifts obtained from uncorrected GIPAW shieldings correlate well with the experimental data and the mean absolute error (MAE, 2.1) of 1.6 ppm looks also reasonable. However, a closer inspection of the data shows that 26% of the differences between experimental and calculated chemical shifts are larger than 2 ppm, 14% are larger than 3 ppm and the maximal error of 6.8 ppm is quite large.

Correcting the chemical shieldings with molecular PBE0/6-311+G(2d,p) calculations according to the newly proposed method improves agreement with experiment con-

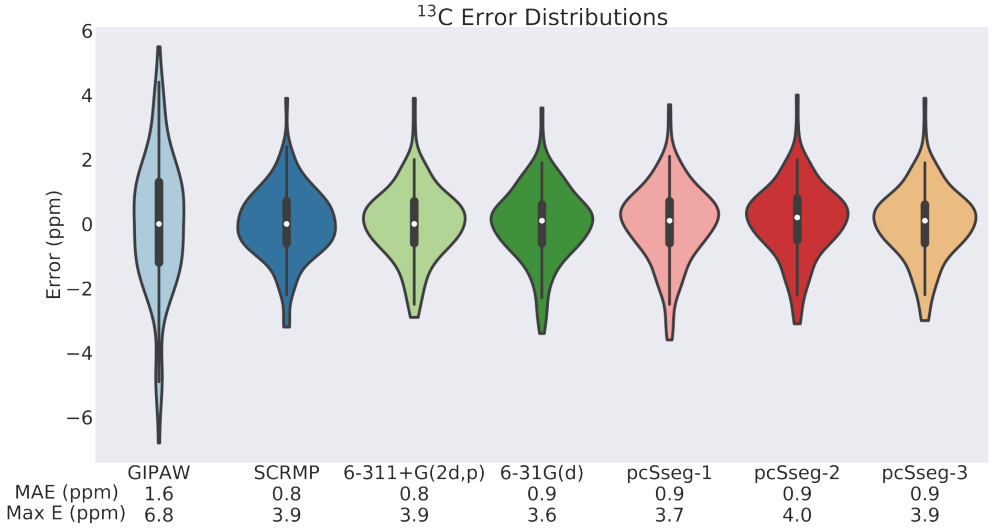


Figure 2.2: Errors in the ^{13}C chemical shift predictions from GIPAW PBE against those with the gas-phase monomer PBE0 corrections computed in various basis Gaussian sets. Violin plots indicate the kernel density estimate of the error distributions. Boxplots in the interior of each violin indicate the median (white dot), middle two quartiles (black box), and outer quartile data (within a factor of 1.5 times the inner quartile range).

siderably; the MAE drops to 0.8 ppm and the maximal error is 3.9 ppm. Only one out of the 132 (0.8%) calculated carbon chemical shifts differs by more than 3 ppm from the experimental shift and eight (6.1%) shifts differ by 2–3 ppm. All the remaining shifts (93%) are predicted with accuracy better than 2 ppm. The violin plots in Fig. 2.2 visualize how adding the PBE0 molecular correction tightens the error distribution about zero error. The corrected GIPAW results have the same mean absolute and maximum errors as the PBE0 results obtained using the self- consistent charge embedded fragment approach SCRMP, [105] as seen in Fig. 2.2. One might wonder if the combination of plane-wave GIPAW and Gaussian basis set molecular calculations here could conceivably prove problematic due to differing degrees of basis set completeness in the two calculations. To investigate this pos-

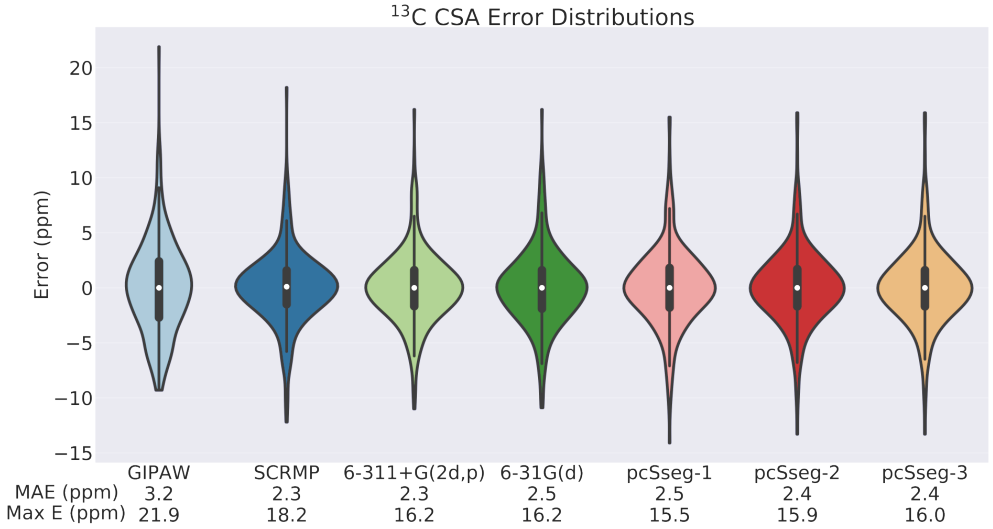


Figure 2.3: Errors in the principal components of the C chemical shift anisotropy tensor predictions from GIPAW PBE against those with the gas-phase monomer PBE0 corrections computed in various basis Gaussian sets.

sibility, the monomer correction to the chemical shielding was also evaluated with several additional Gaussian basis sets.

For each possible basis set, a new linear regression was fitted on the data to convert the shieldings to chemical shifts. As shown in Fig. 2.2, the quality of the molecular correction is quite insensitive to basis set. Even the small and computationally inexpensive 6-31G* basis gives results of nearly equal quality, with a MAE of 0.9 ppm and a maximum error of 3.6 ppm. The systematically growing pcSseg-n basis sets were also tested for n = 1–3, and all three gave similar mean absolute errors of 0.9 ppm and maximum errors ranging 3.7–4.0 ppm. Recently, Hartman and Beran used the SCRMP method to predict the three principal components (σ_{11}, σ_{22}, σ_{33}) of the chemical shielding anisotropy (CSA) tensor. [111] Using the experimental data collected there for the crystals used in the present

study, Fig. 2.3 compares the errors of each method for reproducing each experimental principal component. Employing the monomer hybrid density functional correction to GIPAW PBE CSA tensors significantly improves their accuracy, with mean absolute errors reducing from 3.2 ppm to 2.3 ppm, and giving accuracy very similar to that obtained with PBE0 using the SCRMP fragment model. Using the same computed and experimental data, the error distributions were also evaluated for the chemical shielding anisotropy and asymmetry (Haeberlen convention), as shown in Fig. A.4 and A.5 of the appendix. The behavior observed for the anisotropy, mimics that seen for the principal components in Fig. 2.3: GIPAW PBE performs well (MAE 4.5 ppm), but the SCRMP and corrected GIPAW results perform appreciably better (MAE 3.0–3.3 ppm). On the other hand, no significant difference is observed among GIPAW PBE, SCRMP PBE0, and the corrected GIPAW models for the asymmetry. All methods tested give MAE of 0.08–0.09, and maximum errors of about 0.4 ppm. The high accuracy of the corrected GIPAW approach actually helped us identify errors in the experimental data for several of the systems in the test set. When comparing the experimental and calculated carbon chemical shifts, we noticed particularly large errors for adenosine, L-cysteine, L-glutamine, L-threonineand L-tyrosine systems. Therefore, we reexamined the experimental data of these systems. The experimental ^{13}C ssNMR chemical shifts of adenosine were taken from ref. [237], where the assignment of the signals was based on a comparison of the ssNMR spectrum with solution-state spectrum. However, in the correlation of these experimental data with NMR shieldings calculated with the newly proposed method, one can notice that the assignment of carbon atoms C2' and C3' seems to be interchanged (Fig. A.6 in the appendix). We re-measured adenosine in solution and

using a combination of 1D and 2D NMR experiments, we unambiguously assigned all carbon signals. These experiments revealed that, indeed, that chemical shifts of C2' and C3' were wrongly assigned in the original report. The experimental ^{13}C ssNMR chemical shifts of L-cysteine used in previous studies for comparison with calculated data were taken from ref. [271], where chemical shifts and CSAs of 20 amino acids were reported. However, the authors of the paper admit that they measured ssNMR spectra of purchased amino acids without any recrystallization or crystal-structure determi- nation, and that some of the amino acids were racemates. The calculated chemical shifts of L -cysteine were far from these experimental values (Fig. 2.1). Therefore, we measured ^{13}C ssNMR spectrum of enantiomerically pure crystalline L-cysteine and the obtained spectrum is very close to the predicted one. L-Glutamine spectrum contains two signals of carbonyl carbons (COO and CON) at 173.0 and 176.5 ppm. Our calculations predicted the opposite assignment of these signals than that proposed in ref. [271]. Therefore, we performed a C,H-HETCOR experiment, which confirmed our prediction. A cross-peak between the signal of the hydrogen atom in position a has a strong correlation with one of the carbonyl signals (173.0 ppm), which confirms that this signal is the COO carbon adjacent and Cα (see Fig. X in the appendix).

The new experiments with L-tyrosine and L-threonine did not change the previously published assignment of the signals, but they provided slightly different carbon chemical shifts after careful referencing of the spectra. The newly determined and assigned carbon chemical shifts are used in the experiment- prediction correlations (Table 2.1). These examples demonstrate that the proposed method improves the reliability of GI-

PAW chemical shift predictions, which allows finding previously unnoticed signal or structure mis-assignment. The largest error in the GIPAW calculations of 6.8 ppm is found for the anomeric carbon atom C2 of β-D-fructopyranose; with the molecular PBE0 correction, this error drops to 2.2 ppm. Interestingly, all other saccharide anomeric carbons in this set of structures have also very large deviations of GIPAW- calculated carbon chemical shifts (4.4–5.2 ppm), the only exception is glucopyranose carbon C1 in sacharose with the error of 2.7 ppm. Apparently, the PBE functional is not reliable in the chemical shift calculations of anomeric carbon atoms, which are attached to two electronegative oxygen atoms. The errors of the corrected GIPAW chemical shifts are substantially smaller for all anomeric carbons (0.2–1.7 ppm). Similarly, analyzing the predicted CSA tensor components for the largest errors indicated that the experimental reference for L -glutamine was incorrect. The glutamine carbonyls were swapped and the HETCOR experimental spectrum in the ESI† shows the correct assignment. Furthermore, the C6 carbon of adenosine yields the largest error across all methods ranging in deviations of about 15 ppm. Although the experimental chemical shifts from the reference have been validated, the consistent errors indicate that adenosine CSA principal values should be remeasured.

Nitrogen and Oxygen Isotropic Shifts

Molecular PBE0 corrections to GIPAW chemical shifts of nitrogen ^{15}N lead also to significant improvement of the agreement with experimental data (Table 2.1, MAE decreases from 4.1 to 2.8 ppm). This MAE is identical to that of the SCRMP PBE0 model, albeit with a slightly larger maximum error (7.7 ppm for SCRMP vs. 8.3 ppm for the corrected-GIPAW

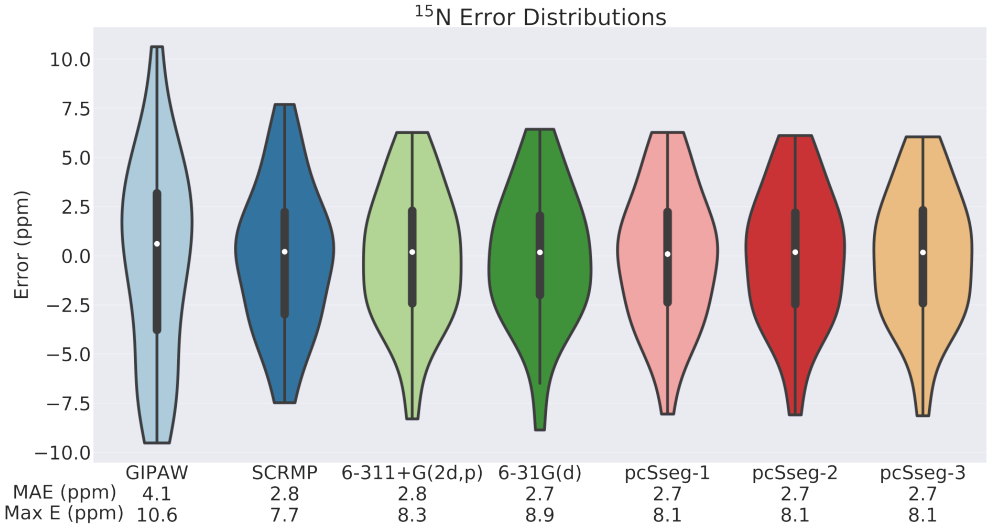


Figure 2.4: Errors in the 15 N chemical shift predictions from GIPAW PBE against those with the gas-phase monomer PBE0 corrections computed in various basis Gaussian sets.

result). The improvement of the chemical-shift prediction of oxygen nuclei is also considerable (MAE decreases from 5.2 to 4.3 ppm). Oxygen chemical shifts are highly sensitive to their electronic environment of the nucleus, making them the most difficult to predict correctly with the fragment-based SCRMP approach. Here, the monomer-corrected GIPAW approach significantly outperforms the 5.9 ppm MAE obtained with SCRMP. Somewhat smaller SCRMP errors would be obtained if a cluster-based approach were used instead of just 1-body and 2-body (mono- mer and dimer) contributions, [105] but that requires appreciably higher computational cost. These oxygen results truly highlight the advantage of combining the complete treatment of the crystalline lattice with the local higher-level correction. For co-crystals, salts, and solvates, one might conceivably perform the gas-phase correction on the entire asymmetric unit instead of just the molecule of interest. For the two such species in the carbon test set, L-asparagine monohydrate (ASPARM03) and L-serine

monohydrate (LSERMH10), the mean absolute difference in the ^{13}C monomer shielding correction obtained on the asymmetric unit versus the amino acid mole- cule only is a mere 0.02 ppm, with a max error of 0.08 ppm. Even for the CSA tensors, the mean and maximum differences to the shielding correction are only 0.04 and 0.16 ppm, respec- tively. In other words, the choice of the "monomer" used for the correction is rather unimportant. On the other hand, the effect of the monomer definition is much more significant for nitrogen and oxygen chemical shieldings. For the five multi-component crystals in the 15 N set (GEHHEH, TEJWAG, FUSVAQ, LTYRHC10, and CYSCLM; four of them are salts, one is a trihydrate), the mean absolute change in the shielding correction between using the full asymmetric unit instead of just the molecule of interest is 1.8 ppm, with a maximum change of 6.9 ppm. For these five crystals, computing the correction using only the single molecule of interest gives a slightly better MAE relative to experiment compared to using the full asymmetric unit (3.1 vs. 3.4 ppm). The impact of the monomer choice on the gas-phase correction is similar for the 17 O chemical shifts. Nine of the fifteen crystals contained in the oxygen set are amino acid hydrochloride salts. DFT suffers from delocalization error, which causes problems with charge transfer [50] and can artificially stabilize salt forms of co-crystals. [139] The MAE versus experiment for the oxygen atoms for the nine HCl salts is 3.3 ppm (max 6.3 ppm) when the correction is obtained for just the protonated amino acid, versus 4.0 ppm (max 11.4 ppm) when the full asymmetric unit is employed. Taken together, this evidence indicates that the gas-phase correction should be evaluated using only the molecule of interest. Once again, the dependence of the results on the basis set used to compute the correction is found to be fairly small (Fig. 2.4 and 2.5). For nitrogen, the

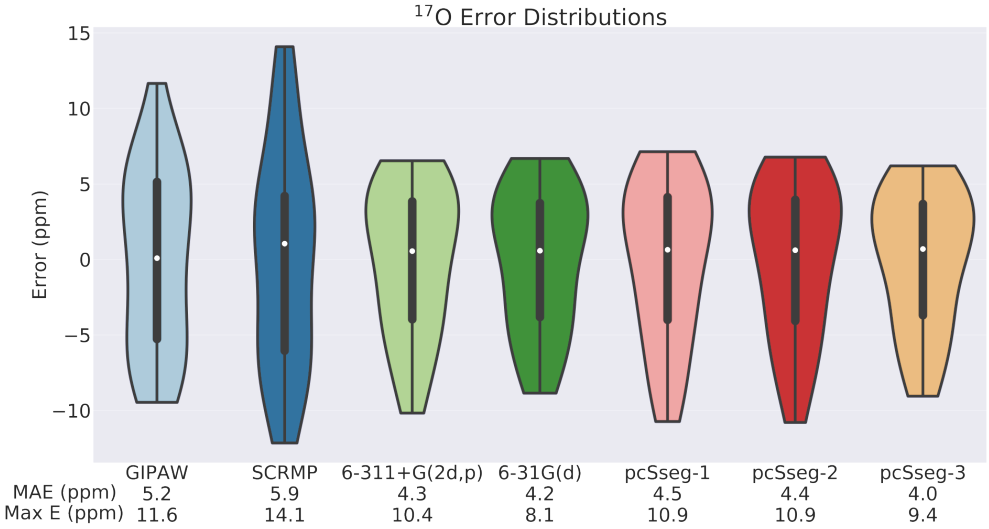


Figure 2.5: Errors in the 17 O chemical shift predictions from GIPAW PBE against those with the gas-phase monomer PBE0 corrections computed in various basis Gaussian sets.

MAE values range 2.7–2.8 ppm across the different basis sets. Larger basis set dependence is observed for the ^{17}O set, where the MAE ranges from 4.0 to 4.5 ppm, and the maximal error from 8.1 to 10.9 ppm. As before, the small-basis 6-31G* results are similar to those from larger basis sets.

Interestingly, however, all basis sets except pcSseg-3 predict a large 10 ppm error for the oxygen in cytosine (CSD refcode CYTSIN). In pcSseg-3, this error drops to less than 1 ppm. So while one generally can use small basis sets to evaluate the monomer correction, the computational cost is low enough that it is probably worthwhile to use relatively large ones in most cases. Finally, it should be noted that the nitrogen and oxygen test sets are substantially smaller and exhibit less chemical variety than the carbon test set. Further validation of the proposed method on a wider variety of systems would be appropriate. Indeed, the small test set size is probably also what causes the skewed and/or

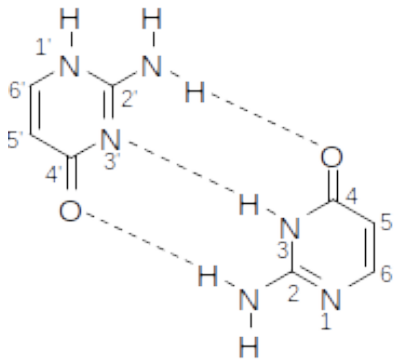


Figure 2.6: The hydrogen bonded pair of two isocytosine tautomers in solid isocytosine.

bimodal error distributions observed for most models in the ^{17}O results. Note also that experimental determi- nation of isotropic shifts of ^{17}O, which is a spin $\frac{5}{2}$ nucleus with large electric quadrupole moment, is substantially more difficult than the measurement of ^{13}C and ^{15}N shifts.

2.4 Applications

In this section, the new method is applied to three specific examples beyond the basic benchmarks described above. To test the limits of the proposed method, "difficult" examples were selected purposefully. All three systems have previously been studied by ss-NMR and DFT calculations and the limited accuracy of the GIPAW approach was stressed.

2.4.1 Isocytosine

Isocytosine is a constitutional isomer of cytosine with interest- ing biological activities. Isocytosine crystallizes as a 1:1 mixture of two tautomers, which form hydrogen

	Experiment	GIPAW	GIPAW-corrected
C2-C2'	0.0	-2.77	-2.29
C4-C4'	6.1	6.11	6.11
C5-C5'	4.0	3.24	4.50
C6-C6'	-19.3	-18.54	-19.44
MAE		1.1	0.7
N1-N1'	-73.4	-74.91	-74.34
N2-N2'	49.9	54.42	52.78
N3-N3'	-3.4	-5.17	-4.56
MAE		2.6	1.7

Table 2.2: Experimental 10 and calculated chemical shift differences (ppm) in solid isocytosine. Mean absolute errors (MAE) obtained for the conventional GIPAW method (PBE functional) and for corrected GIPAW (PBE0 correction, 6-311+G(2d,p) basis set). Atom numbering is depicted in Fig. 2.6

bonded pairs similar to pairs of guanine and cytosine in nucleic acids (Fig. 2.6). It has been shown recently that a combination of experimental and simulated chemical shifts of isocytosine may serve as a probe of proton transfer reactions and hence, rare tautomer formation. [186, 187] The presence of two non-equivalent isocytosine molecules in the crystal structure enables direct comparison of their experimental chemical shift differences against the predicted values. Table 2.2 summarizes the experimental and calculated ^{13}C and ^{15}N chemical shift differences between the two non-equivalent isocytosine molecules. Once again, applying the PBE0 correction to GIPAW predictions improves the agreement with experiment significantly.

2.4.2 Methacrylamide

In the pharmaceutical industry, solid-state NMR is commonly used for the identification of polymorphic crystal structures. ssNMR can detect polymorphic impurities and

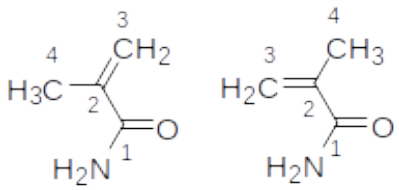


Figure 2.7: s-cis (left) and s-trans (right) conformers of methacrylamide found in monoclinic and orthorhombic polymorphs, respectively

characterize polymorphic forms of active pharmaceutical ingredients (APIs) in formulated drug products and drug carriers. [217, 268] The industrially important compound methacrylamide has two known polymorphs; the monoclinic form contains only the s-cis molecules (Fig. 2.7), whereas the orthorhombic polymorph is exclusively formed by the s-trans conformer. [96] Carbon chemical shift differences between the two forms of methacrylamide are very small (Table 2.3) and, therefore, may be used as a stringent test of chemical shift predictions. The methacrylamide molecule is small enough to allow high-level ab initio calculations of its NMR shieldings. Table 2.3 summarizes the predicted chemical shift differences between the two methacrylamide forms calculated at the GIPAW level and at the GIPAW level corrected with PBE0, MP2 or CCSD monomer calculations. Surprisingly, applying the PBE0 correc- tion slightly deteriorates the agreement with experiment, and the MAE calculated for MP2-corrected GIPAW result is almost identical to the uncorrected GIPAW one. The CCSD correction improves the MAE value only slightly. All four models repro- duce the experimental shifts to within a ppm or better. These particularly subtle differences in the chemical shifts between the two polymorphs probably represent the limit of what can be achieved by corrections computed for a single, isolated molecule. Chemical shift differences between poly- morphs are mostly governed by molecular packing and intermolecular interactions in the crystals; these intermolecular interactions are modelled with

Atom	Experiment	GIPAW	GIPAW PBE0-corrected	GIPAW MP2-corrected	GIPAW CCSD-corrected
C1	0.15	1.50	1.23	0.89	0.87
C2	0.27	0.68	1.46	0.14	0.85
C3	-0.73	-1.37	-1.95	0.97	0.28
C4	0.06	0.38	0.49	0.17	0.10
MAE		0.68	0.98	0.67	0.59

Table 2.3: Experimental 37 and calculated chemical shift differences (ppm) between the monoclinic and orthorhombic polymorphs of methacrylamide. Mean absolute errors (MAE) obtained for the conventional GIPAW method (PBE functional) and for corrected GIPAW (PBE0, MP2 and CCSD correction, 6-311+G(2d,p) basis set). Atom numbering is depicted in Fig. 2.7

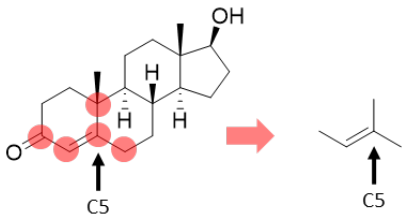


Figure 2.8: The structure of testosterone and its fragment used for the calculation of CCSD corrections.

the GGA level of theory only and are not included in the molecular correction proposed here.

2.4.3 Testosterone

Two crystal forms of testosterone have been studied by ssNMR and most of the carbon signals have been assigned using INADEQUATE carbon–carbon experiment. [103] The a form contains two crystallographically non-equivalent molecules in the asymmetric unit, while the β-form is a monohydrate. The conformation is almost identical in all three crystallographically unique molecules. [107] Carbon chemical shifts of solid testosterone

have also been calculated using the GIPAW and cluster/fragment approach. [107, 271] Most individual chemical shifts were reproduced to within a few ppm, with the notable exception of C5, which was significantly overestimated (Table 2.4) by both methods. We calculated carbon C5 chemical shift at the GIPAW level and, indeed, the agreement with experiment is surprisingly poor. The molecular PBE0 corrections improve the agreement by about 3 ppm, but the shifts are still overestimated by 10 ppm. Therefore, we calculated a CCSD correction for a partial fragment of the testosterone molecule (because CCSD chemical shielding calculations on the full testosterone would be very expensive). Starting from the GIPAW-optimized structure of the β-form of testosterone, this partial fragment consists of the C4–C5 double bond and three carbon atoms directly attached to the double bond (C3, C6 and C10, see Fig. 2.8); missing hydrogen atoms were added to saturate the dangling bonds on the terminal carbon atoms. NMR shieldings of this fragment were then calculated at the PBE and CCSD levels of theory. To allow comparison of the calculated shieldings with the chemical shifts of testosterone, we calculated NMR shieldings of a- glycine, a commonly used reference compound, at the same levels of theory. The chemical shift of C5 in the molecular fragment calculated at the PBE level is by 22.3 ppm lower than the shift of glycine carbonyl. On the other hand, the CCSD calculation predicts the C5 chemical shift lower by 34.2 than that of glycine, i.e. CCSD level of theory predicts that the shift of C5 in the molecular fragment is by 11.9 ppm lower than the value predicted by PBE. If we transfer this correction to the whole β-testosterone molecule, the GIPAW-predicted chemical shift (186.9) drops to 175 ppm, which is reasonably close to the experimental value (173.8 ppm). It is not clear, why DFT with both the PBE and PBE0 functionals fail to predict

	α form		
	Molecule u	Molecule v	β form
Experiment	170.6	172.1	173.8
GIPAW	182.6	184.1	186.9
GIPAW PBE0-corrected	179.9	181.4	183.9
Cluster/fragment	176.2	177.0	182.1
GIPAW CCSD-corrected			175.0

Table 2.4: Experimental and predicted chemical shifts of carbon C5 in solid testosterone

this particular carbon atom chemical shift correctly. However, this example demonstrates that high- level ab initio corrections may be calculated for molecular fragments and these corrections may be used to improve the agreement of predicted shifts with experiment.

2.5 Conclusion

In conclusion, this study has demonstrated a very simple strategy for improving the quality of GGA-based GIPAW NMR chemical shielding calculations in molecular crystals by evaluating a correction to the shielding computed at a higher level of theory on an isolated molecule. The new approach achieves accuracy rivaling or beating that of fragment-based methods. The correction is quite insensitive to the basis set used for the monomer calculation, ensuring that the cost of evaluating the correction is minimal. Typically one would employ a hybrid density functional for the higher level of theory. However, as some of the applications demonstrate, it is also possible to consider the use of higher-level chemical shielding calculations, such as CCSD. The CCSD correction proved essential to predicting the chemical shift of carbon C5 in β-testosterone, for example. Finally, while the work here

focused on molecular organic systems, it would be interesting to explore the application of the technique to inorganic systems with localized electronic structure as well. The testosterone example demonstrates how even a calculation on a small local "cluster" of atoms cut out from a larger covalent network may be enough to achieve a meaningful correction to the GGA-level chemical shifts.

In the next chapter, we explore methods that are more applicable to amorphous or systems without long-range periodicity. Specifically, since we are using local cluster or fragment methods, one can routinely apply hybrid DFT methods that do not plague atomic basis sets. Specifically, we show how these methods can be applied to biomolecular systems which would require unreasonably large unit cells (i.e. computationally expensive) using GIPAW DFT. We also show how to use these methods for highly charged fragment systems.

Chapter 3

Polarizable Continuum Models Provide an Effective Electrostatic Embedding Model for Fragment-Based Chemical Shift Prediction in Challenging Systems

3.1 Introduction

Determining the biochemical mechanisms occurring in the active sites of proteins often requires atomic-scale structural resolution for both hydrogen and non-hydrogen atoms. Nuclear magnetic resonance (NMR) crystallography[104, 156, 8, 169] achieves such res-

olution by combining X-ray crystallography, solid-state magic-angle-spinning NMR spectroscopy, and chemical shift prediction to reveal chemical structure and dynamics. X-ray crystallography resolves the heavy atom positions and is often used to generate candidate structures. NMR experiments probe the finer local structure and dynamical details, while computational chemical shift prediction helps map those experimentally observed chemical shifts into a three-dimensional crystal structure.

High computational cost constitutes the most significant barrier to performing NMR chemical shift prediction in larger, more complex systems. In small-molecule and inorganic crystals, periodic boundary conditions and the gauge-including projector augmented wave (GIPAW)[184] approach for planewave density functional theory (DFT) enable accurate chemical shift prediction in the solid state. The GIPAW approach has proved highly successful and is now widely used.[27, 8].

Unfortunately, periodicity is less useful in the context of biomolecules, where the unit cell frequently consists of tens of thousands of atoms or more. Cluster models provide an alternative approach for predicting chemical shifts in biomolecules and other systems which cannot be represented by small periodic unit cells. This strategy models a large, finite cluster that includes the atoms of interest and some portion of the surrounding chemical environment. However, large clusters are needed to obtain well-converged NMR chemical shieldings. For example, [126] showed that a cluster extending at least 6 Å from the key atoms of interest was necessary to converge the chemical shieldings.[126] Flaig et al similarly found that converging ^{1}H and ^{13}C chemical shieldings to within 0.1 and 0.5 ppm, respectively, required clusters extending 6–10 Å and containing hundreds or even one thousand

atoms.[79] Employing electrostatic embedding reduces the cluster size required to achieve such convergence, especially if the embedding environment is polarizable. For example, Kongsted et al showed that for acrolein in water, the radius of environment required to converge the ^{17}O chemical shieldings was 2 Å smaller compared to a non-polarizable embedding environment.[233] Further computational savings can be obtained with multi-layer (ONIOM) approaches[153, 42, 274, 167] and the use of locally dense basis sets.[43, 45]

Even with such strategies, *ab initio* chemical shift calculations in large systems remain computationally demanding. This becomes particularly problematic in the context of NMR crystallography where one might screen dozens of different candidate structures. Such candidate structures might correspond to the combinatorial number of different possible protonation states at several ionizable sites and/or the possibility of including non-crystallographic waters in the active site, for example. An NMR crystallography study of the indoline carbanionic intermediate in the β-active site of tryptophan synthase using a 7 Å cluster involved six potentially ionizable protons, from which 28 viable candidate protonation states were constructed.[136] Techniques which could reduce the computational cost of evaluating the predicted chemical shieldings for such candidates would be very useful. Even imperfect models that can effectively screen out poor candidates with relatively low cost would allow researchers to concentrate their computational resources on the most promising structures.

Fragment methods reduce the computational costs of quantum chemical calculations in large systems by replacing a single calculation on the entire system with a series of calculations on smaller subsystems whose contributions can be combined to approximate the

system as a whole. Many fragment methods have been developed, and they have been successful in predicting energies, structures, and spectroscopic properties.[93, 51, 190, 20, 117] In the context of NMR chemical shift prediction for large molecules, early research on fragment methods includes the 2007 demonstration that chemical shieldings could be obtained via calculations on fragments and their pairwise interactions[140, 241] and the 2010 adaptation of the fragment molecular orbital method to computing NMR properties.[87] Since then, many other fragment approaches for NMR chemical shift calculations have been developed, such as the automated fragmentation QM/MM,[115, 242, 276, 116, 240] the adjustable density matrix assembler,[82, 81, 251] the systematic molecular fragmentation,[195, 194, 131] and the molecules-in-molecules approaches.[157, 127] Fragment methods have also proved highly effective in molecular crystals.[236, 106, 109, 108, 107, 105, 111, 160]

Many of the fragment approaches can be unified under a common framework of a generalized many-body expansion,[158, 197] with key distinctions between methods involving whether the fragments overlap or not, the level of interactions between fragments considered explicitly (e.g. are properties computed only on individual fragments or also with pairwise or higher-order interactions?), whether all interactions are treated at the same level theory, and how electrostatic embedding or other long-range effects are included. For example, in the automated fragmentation QM/MM approach, each non-overlapping local fragment is treated via a calculation of the fragment atoms surrounded by a buffer region of atoms (i.e. atoms which are part of adjacent fragments), and each of these systems is electrostatically embedded to incorporate longer-range interactions.[240] The size of the each fragment calculations is an important question. Units as small as tripeptides have

been found useful,[252, 4] though often considerably larger clusters of atoms are used (e.g. ref [240]).

We have recently developed the self-consistent reproduction of the Madelung potential (SCRMP) model for fragment-based chemical shift prediction in organic molecular crystals.[105] This model combines chemical shielding calculations on the single molecule of interest with pairwise contributions to the shielding due to nearby molecules, all embedded in a self-consistent field of point charges that are fitted to mimic the crystalline Madelung potential. For a given density functional, the SCRMP model rivals the accuracy of GIPAW planewave DFT calculations and has lower computational cost. However, it offers two important advantages over GIPAW. First, computing the fragment contributions in a gauge-including atomic orbital (GIAO) basis instead of planewaves means that hybrid density functionals can be used at much lower computational cost, which in turn provides considerably improved accuracy.[109, 108, 120, 122] Second, GIAO-based calculations can be applied both to periodic systems like molecular crystals as well as non-periodic systems such as proteins. This broad applicability is particularly useful for referencing chemical shifts because it allows one to develop reliable linear regression models for referencing the predicted chemical shieldings on well-defined molecular crystal systems which then can be applied to biomolecular systems where solvent, flexibility, dynamics, and other factors create additional modeling complications.[23]

The SCRMP model does have some disadvantages when applied to biomolecular-type systems. Computing the self-consistent embedding environment requires computing each monomer fragment's charges repeatedly. This is relatively inexpensive in a periodic

crystal where the number of symmetrically unique monomers is small (often just a single molecule). However, it becomes more computationally demanding when each fragment is unique, as in a biomolecule. Furthermore, fragmentation and point-charge embedding are easier when the molecules are small and no covalent bonds need to be cut. Partitioning fragments across covalent bonds requires terminating the dangling bonds and addressing the unphysical interactions that can arise from placing embedding point charges very close to the terminating atoms.[72, 52]

The present study circumvents these difficulties with a simple, easy-to-implement method for performing high-quality fragment NMR calculations that is viable for both molecular crystals and proteins. Instead of developing an elaborate point-charge or multipolar polarizable embedding environment, the proposed model computes chemical shieldings through a series of calculations on individual (non-overlapping) fragments and pairs of fragments, each embedded in a polarizable continuum model (PCM). PCM embedding is widely available in quantum chemistry software packages, is computationally inexpensive, and is typically easier to use for non-experts than more elaborate self-consistent charge embedding schemes. In molecular crystal benchmarks, the PCM-embedded fragment model achieves root-mean-square (rms) errors on par with previous fragment methods for ^{13}C, provides modest improvements to predicted ^{15}N shifts, and results in slightly larger errors for ^{17}O shifts. While the errors are found to be larger for ^{17}O, they are still reasonably competitive with other approaches. The same fragment model is also successfully demonstrated in the protein piscidin-1 and the challenging example of the indoline carbanionic intermediate bound in the active site of tryptophan synthase. The combination of relatively

simple implementation, low computational cost, diverse applicability, and good accuracy make the PCM-embedded fragment approach a worthwhile approach for NMR chemical shift prediction in large systems.

3.2 Theory

One approach to reducing computational costs in large systems involves expressing the energy in terms of a many-body expansion, where the total energy is represented as a sum of 1-body, 2-body, 3-body, etc terms.

$$E_{total} = \sum_i E_i + \sum_{j>i} \Delta^2 E_{ij} + \sum_{k>j>i} \Delta^3 E_{ijk} + \cdots \tag{3.1}$$

A fragment might correspond to a single molecule in a set of interacting molecules or to a group of atoms in a larger covalent system.[117, 190, 51, 93] The 1-body terms in Eq 3.1 correspond to the energy of each isolated fragment E_i, the 2-body terms are interactions between pairs of fragments i and j,

$$\Delta^2 E_{ij} = E_{ij} - E_i - E_j \tag{3.2}$$

and the 3-body and higher terms represent higher-order non-additive contributions due to polarization and other many-body effects.

The 3×3 NMR chemical shielding tensor $\boldsymbol{\sigma}$ on atom A is defined as the second derivative of the energy with respect to the magnetic field B and nuclear magnetic moment μ:

$$\boldsymbol{\sigma}^A = \frac{\partial E}{\partial \boldsymbol{B} \partial \boldsymbol{\mu}^A} \tag{3.3}$$

Accordingly, differentiating the many-body expansion of the energy in Eq 3.1 expresses the chemical shielding tensor of atom A in fragment i in terms of a series of many-body contributions,[106, 117]

$$\boldsymbol{\sigma}^A_{total} = \boldsymbol{\sigma}^A_i + \sum_j \Delta^2\boldsymbol{\sigma}^A_{ij} + \sum_{j,k} \Delta^3\boldsymbol{\sigma}^A_{ijk} + ... \tag{3.4}$$

where $\boldsymbol{\sigma}^A_{total}$ is the chemical shielding of atom A in the complete system, $\boldsymbol{\sigma}^A_i$ is the chemical shielding of atom A, on the isolated fragment i. The term $\Delta^2\boldsymbol{\sigma}^A_{ij}$ describes two-body contributions to the chemical shielding that reflect how the presence of a second fragment j alters the shielding of atom A,

$$\Delta^2\boldsymbol{\sigma}^A_{ij} = \boldsymbol{\sigma}^A_{ij} - \boldsymbol{\sigma}^A_i \tag{3.5}$$

The non-additive three-body contribution $\Delta^3\boldsymbol{\sigma}^A_{ijk}$ is similarly defined as:

$$\Delta^3\boldsymbol{\sigma}^A_{ijk} = \boldsymbol{\sigma}^A_{ijk} - \Delta^2\boldsymbol{\sigma}^A_{ij} - \Delta^2\boldsymbol{\sigma}^A_{ik} - \boldsymbol{\sigma}^A_i \tag{3.6}$$

The fragments used here are defined to be non-overlapping—each atom of the original system is present only in a single fragment. The chemical nature of the individual fragments in these expansions will depend on the system, though. In a molecular crystal, each fragment would typically correspond to a single molecule, and the higher-order terms in the many-body expansion involve pairwise and higher non-additive interactions within small clusters of molecules. In a biomolecule, each fragment would consist of a local group of atoms, such as one or more amino acids. Care must to be taken with regards to where cuts across covalent bonds are made and how the resulting dangling bonds are terminated.[145] The specific strategies used here will be discussed later.

Although the many body expansion for the chemical shielding tensor is formally exact, the number/size of the fragment calculations combined with numerical precision issues[199, 198, 117] make evaluation of the 3-body and higher terms impractical in large systems. Instead, truncating the many-body expansion to include only one-body and two-body terms and using electrostatic embedding to capture the many-body polarization effects on those monomers and dimers has proven to be effective for calculating NMR properties:[106, 109, 108]

$$\boldsymbol{\sigma}_{total}^{A} \approx \sum_{i} \boldsymbol{\sigma}_{i}^{A,embed} + \sum_{i<j} \Delta^{2}\boldsymbol{\sigma}_{ij}^{A,embed} \tag{3.7}$$

Embedding proves particularly important for nuclei such a ^{15}N and ^{17}O, whose chemical shifts are sensitive to the electrostatic environment.[105]

In molecular crystals, self-consistent point-charge embedding schemes[230, 238, 236, 76, 105] that mimic the self-consistent Madelung potential perform well. However, the intermolecular separations between monomers in a molecular crystal are typically larger than those between covalently bonded fragments in a biomolecule. Furthermore, biomolecules frequently include highly charged residues that increase the importance of the electrostatic environment. Directly applying the fragment approach with the SCRMP polarizable point-charge embedding scheme to a charged substrate bound in an enzyme active site leads to poor-quality chemical shifts, for example. One solution is to surround the atoms of interest with a sizable quantum mechanical buffer region to ensure any point charges are further away from the key atoms. However, the use of large buffer regions increases the computational cost significantly. Alternatively, a more elaborate polarizable embedding environment[233, 231] that employs quantum-mechanically derived higher-order multipole

expansions and dipole-dipole polarizabilities at each atomic site in the fragment environment can work well, though the implementation of such models is considerably more involved.

The present study overcomes these difficulties by replacing the point-charge embedding scheme with a standard polarizable continuum model environment.[164] A PCM defines a cavity surrounding the molecule of interest, using the union of van der Waals spheres around each atom or other criteria. The electrostatic potential associated with the solute-environment interaction is then mapped onto an integral representation of the reaction potential involving an apparent charge distribution on the surface of the cavity. The solution to the integral is discretized into a set of finite elements with local point charges that interact with and respond to the molecule in the cavity based on the dielectric of the surrounding medium. These equations are then solved simultaneously and self-consistently with the Kohn-Sham equations of DFT. In contrast to explicit embedding environments, the PCM assumes a homogeneous polarization environment with constant dielectric.

In the PCM-embedded fragment approach examined here, each monomer and dimer shielding contribution is computed by embedding it in a PCM,

$$\boldsymbol{\sigma}^{A}_{total} \approx \sum_{i} \boldsymbol{\sigma}^{A,PCM}_{i} + \sum_{i<j} \Delta^{2}\boldsymbol{\sigma}^{A,PCM}_{ij} \tag{3.8}$$

In the integral equation formalism PCM approach used here, for example, the PCM embedding cavity is defined as a union of spheres derived from the van der Waals radii of each atom.[33] Note that the PCM cavity is defined separately for each fragment or pair of fragments based solely on the atoms present in that particular monomer/dimer, rather than using a single, large cavity associated with the complete unfragmented system.

Given the inhomogeneity of the protein environment surrounding any particular fragment, it is not obvious that the environment can suitably be replaced by a simple dielectric model in chemical shielding calculations. Polarizable continuum models are traditionally used to estimate solvent effects in homogeneous[152] environments implicitly via their dielectric values, without need to sample the explicit molecular details. However, we find that a fragment-based model which replaces a point-charge embedding environment with a PCM handles even highly charged residues effectively and performs well even for nuclei such as ^{15}N and ^{17}O that are sensitive to the electrostatic environment.

Employing PCM embedding in the fragment approach offers several advantages. First, the individual fragment calculations are easy to run with standard electronic structure packages once the fragments have been defined. There is no need to implement new self-consistent charge polarization schemes, for instance. Second, it simplifies the treatment when fragmentation involves covalent bond cleavages. When partitioning the system across a C-C single bond, for example, one typically terminates the dangling bond with a hydrogen or link atom. This new hydrogen atom will be very close to the position of the former carbon atom that was deleted. In a SCRMP-like point-charge embedding model, the close proximity of the new hydrogen atom to the charge on the original adjacent carbon atom introduces spurious artifacts into the chemical shifts. Various charge translation/redistribution schemes can be considered to mitigate this effect,[254, 144] but the PCM scheme avoids this problem entirely. Third, PCM embedding is computationally inexpensive, and the PCM-embedded fragmentation approach provides clear computational benefits compared to large cluster calculations. For a cluster consisting of a few dozen glycine molecules, for

example, the PCM-embedded fragment approach with single-molecule fragments requires an order of magnitude less CPU time compared to a calculation on the full cluster. The wall time savings can be even more advantageous, since the dozens of independent fragment contributions can be distributed over different groups of processors.

Introduction of the PCM embedding does raise the question of what the appropriate dielectric constant should be. For example, the appropriate dielectric constant for a protein has been much debated in the literature, with suggested values spanning a broad range from ∼2–30, and "typical" values of ∼6–9.[209, 142, 133, 5] However, these values can vary with temperature[5] and the nature of the model used to describe the system.[209] Fortunately, in practice we find that the choice of the dielectric used for the PCM embedding here has only a small impact on the accuracy of the final chemical shifts. The particular choice of dielectric constant is largely compensated for via the chemical shift referencing that maps the predicted chemical shieldings onto experimentally observable chemical shifts.

3.3 Computational Methods

3.3.1 Structures

The impact of fragmentation on the predicted chemical shieldings is first investigated for the piscidin-1 protein (PDB ID: 2MCU, Figure 3.2)[183]. The structure of this 22-amino-acid α-helix with C-terminus amidation was solved with atomic resolution by NMR crystallography. Fragmentation of this system was performed using the FragIt software tool.[232] Covalent bonds were cleaved at the Cα-C$'$ single bond between amino acids as described in the original FragIt paper. Hydrogen capping atoms were placed at the

cleaved covalent bonds according to the scheme from Deev and Collins,[59] which places the hydrogen atoms along the original C-C bond vector at a distance determined by the original C-C distance scaled by the ratio of typical C-C and C-H bond lengths. Figure 3.1 presents a sample of how the peptide fragments might be constructed and where the capping atoms are placed.

Specific cut locations are described in Section 3.4.1. Net charges for each biomolecular fragment were then assigned using an in-house code that recognizes the charges of common functional groups in SMILES string notation.

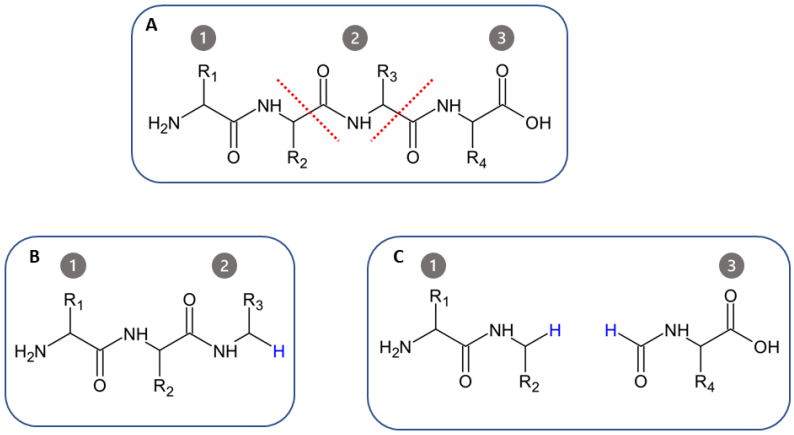


Figure 3.1: Schematic showing fragmentation in a covalent peptide. (a) This sample peptide is divided into three fragments by cleaving C-C bonds. To compute the chemical shieldings for fragment 1, two-body fragment contributions will be computed for fragment 1 alone plus two-body corrections involving (b) fragment dimer (1,2) and (c) fragment dimer (1,3). Any dangling bonds in the monomer or dimer fragments are capped with hydrogen atoms (shown in blue).

To assess the quality of chemical shieldings predicted from the PCM-embedded fragment model against experiment, chemical shifts were predicted for a benchmark set of 47 molecular crystals from ref [70] and references therein. This set includes 132 ^{13}C, 37 ^{15}N, and 28 ^{17}O experimental isotropic shifts. A complete list of Cambridge Structure Database

(CSD) Reference Codes and experimental chemical shifts can be found in the appendix. The DFT-optimized crystal structures were taken from ref [70]. The set of chemical shieldings predicted for each nucleus type were referenced to experiment via linear regression, as described in Section 3.3.3. In these molecular crystals, each individual fragment correspond to a single molecule.

Finally, the PCM-embedded fragment model is applied to a more complex example of the indoline carbanionic intermediate bound within the β-subunit of tryptophan synthase. The structure of this system, including key protonation states in the active site, was solved via NMR crystallography in an earlier study.[136] That study measured isotropic chemical shifts for 13 key isotopically carbon, nitrogen, and oxygen atoms of the substrate via solid state NMR. It then modeled the system with a 612-atom cluster model consisting of the indoline carbanionic substrate and surrounding atoms from the protein. The cluster includes the substrate and all surrounding protein residues within 7 Å of the substrate, and the structure relaxed via a mixed DFT (substrate) and semi-empirical (protein side chains) approach. See Ref [136] for details surrounding the determination of the crystal structure and the development and computational relaxation of the cluster model. The cluster model was further refined in Ref [34] using DFT and a locally dense mixed basis scheme for the full cluster. Fragmentation of this system was again performed using FragIt, placing capping hydrogen atoms to terminate and cleaved C-C bonds. Specific fragmentation patterns are described in Section 3.4.3.

3.3.2 Chemical Shielding Calculations

NMR chemical shielding tensors were calculated using our hybrid-many body interaction (HMBI) code to manage the fragment contributions.[22] A locally dense basis set scheme[43, 45] was employed that reduces computational costs associated with computing the shielding tensors without significantly impacting the chemical shift accuracy.[110, 109] In molecular crystals, the locally dense basis scheme employs a 6-311+G(2d,p) basis for the molecules in the asymmetric unit and any other atoms lying within 2 Å, 6-311G(d,p) for atoms up to 4 Å away, and 6-31G for all atoms beyond. Two-body interactions were computed between the asymmetric unit and any other fragment which has an atom lying within 4 Å of any atom within the asymmetric unit. This cutoff has proved sufficiently converged in earlier fragment model studies that employed different electrostatic embedding environments.[109, 108, 105] Test calculations here find that the same cutoff works well with PCM embedding too. For the ^{15}N molecular crystal benchmarks described in Section 3.4.2, for example, the root-mean-square chemical shift error versus experiment varies by only 0.02 ppm when cutoffs of 4, 6, or 8 Å are used. See for details.

The calculations on biological systems employ the same distance criteria for the basis sets, albeit with a particular central fragment or enzyme substrate defining the atoms of interest instead of the asymmetric unit. All NMR shielding calculations were calculated using the hybrid PBE0 density functional in Gaussian 09[84], with an integration grid of 150 radial and 974 Lebedev angular points. The PBE0 functional was chosen based on its strong performance in previous molecular crystal benchmarks that compared six different

functionals.[108] It is possible that some other functional might predict the chemical shifts even more accurately.

The polarizable continuum environment was represented using the integral equation formalism (IEFPCM).[33, 163] Several different solvent environments were selected with dielectrics ranging from 1.4–181: argon (ϵ=1.4), acetic acid (ϵ=6.3), dichloromethane (ϵ=8.9), ethanol (ϵ=24.9), water (ϵ=78.4), and an n-methylformamide mixture (ϵ=181.6). The cavities surrounding the molecules were generated using the default Gaussian 09 PCM model settings which define the cavity as the union of universal force field (UFF) atomic radii scaled by 1.1.

For the indoline substrate of tryptophan synthase studied in Section 3.4.3, the proton is believed to exchange dynamically between the Schiff base nitrogen and the phenolic oxygen.[35] In accord with that earlier work, the chemical shifts for this system were computed according to a two-site fast proton exchange scheme. In brief, chemical shifts are predicted separately for the phenolic and protonated Schiff base structures. The final shifts for each atom i are modeled as a linear combination of the two sets of predicted shifts,

$$\delta_i = c_1 \delta_i^{phenolic} + c_2 \delta_i^{Schiff} \tag{3.9}$$

with the coefficients c_1 and c_2 fitted to minimize the reduced χ_r^2 statistic representing the level of agreement between the predicted and experimental shifts. Our testing suggests that the optimal mixture of the phenolic oxygen and Schiff base nitrogen contributions varies slightly between B3LYP functional (77% phenolic oxygen) used in Ref [35] and the PBE0 one used here (75% phenolic oxygen). These two mixtures are effectively the same within

the modeling errors and indicate that the system is dominated by the phenolic form and has smaller contributions from the protonated Schiff base form.[35]

3.3.3 Chemical Shift Referencing

In this work, isotropic chemical shielding values σ_{iso}^{A} are converted to experimentally observable chemical shifts δ_{iso}^{A} via a linear regression approach.[148, 23]

$$\delta_{iso}^{A} = m\sigma_{iso}^{A} + b \tag{3.10}$$

Ideally the slope m would equal -1 and the intercept would correspond to the chemical shielding of the reference atom. In practice, least-squares fitting the chemical shifts versus chemical shieldings results in a slope that deviates from unity by a few percent. This linear regression approach helps compensate for systematic deficiencies in the chosen level of theory and finite basis set errors. The regression parameters are specific to the nuclide and the chosen model chemistry: density functional, basis set, embedding model, etc. Assuming the training set contains well-defined structures and diverse chemical shielding environments, the resulting regression parameters should be transferable across different chemical systems.[148, 23]

To maximize the predictive power of the regression approach, the regression parameters should be fitted against training data that is distinct from the system(s) to which the regression model will be applied.[23] Here, the regression parameters are fitted based on molecular crystal benchmarks, which have well-defined structures and exhibit far less dynamical motion than typical biomolecules. The applicability of these particular molec-

ular crystal chemical shift regressions to other molecular crystals[108, 107, 267] and to biomolecules[272, 35] has been demonstrated previously.

3.4 Results and Discussion

The following sections first investigate how well PCM-embedded fragment calculations reproduce the chemical shieldings that would be obtained from a calculation on the full system for a small α-helix protein. Next, to establish the performance of the model relative to experiment and to develop linear regression models for referencing the chemical shifts, the PCM fragment approach is applied to predict chemical shifts in a set of organic molecular crystals. Finally, chemical shifts for an intermediate bound in the active site of tryptophan synthase are predicted with and without fragmentation and compared against experiment.

3.4.1 Systematic Analysis of Piscidin-1

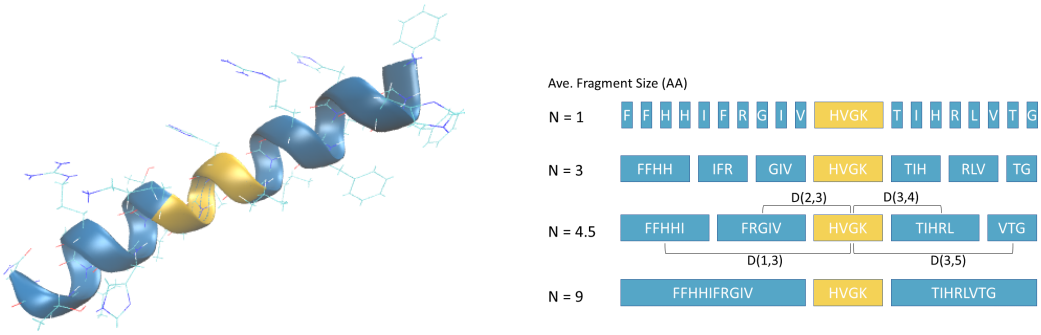


Figure 3.2: ***(left)*** Piscidin 1 protein is a 22-residue, cationic protein which adopts an α helix and exhibits antimicrobial activities. The coloring highlights the central fragment used for all calculations in yellow and the surrounding protein environment in blue. ***(right)*** The amino acid sequence and four different fragmentation schemes considered. Colored boxes indicate the contents of each fragment, and the schemes are labeled according to the average fragment size N (excluding the central HVGK fragment). The shieldings are always computed for the central HVGK fragment. For the N=4.5 case, the fragment dimers $D(i,j)$ involving the central fragment are shown (with fragments numbered 1–5 from left to right).

To begin, the performance of the PCM-embedded fragment approach is analyzed on the piscidin-1 protein (Figure 3.2). This 22-amino acid protein is small enough to allow fully quantum mechanical prediction of the chemical shieldings without any fragmentation, which can serve as a benchmark for the more approximate fragment models. The charged lysine, arginine, and N$'$ terminus give it a net +4 charge. Because our interests particularly lie in the ability to predict chemical shifts of key, isotopically labeled atoms in a central region (e.g. for a substrate bound in an enzyme active site), an HVGK peptide fragment near the center of the peptide was defined. After addition of two capping hydrogens to saturate the carbon-carbon bond cuts at either end, it has the chemical formula $C_{19}H_{32}N_7O_4^+$. This specific HVGK fragment was chosen to include the cationic lysine residue since charged

sites are particularly sensitive to the electrostatic environment and can disproportionately impact the chemical shifts of neighboring atoms.

Several different possible fragmentation schemes for the remaining 18 amino acids surrounding this central HVGK fragment were constructed, averaging anywhere from 1–9 amino acids per fragment. These schemes are shown in Figure 3.2 and are denoted by average number of residues per fragment (excluding the central HVGK fragment). The chemical shieldings of the central HVGK fragment were computed via inclusion of the 1-body (i.e. the isolated HVGK fragment) and 2-body (pairwise combinations of the central HVGK fragment with each of the other side fragments). Errors in the chemical shieldings on the central HVGK fragment atoms were computed relative to the shieldings obtained for those same atoms in the full peptide with no fragmentation. Reference shieldings for the full system were obtained with and without a PCM ($\epsilon = 8.9$) for appropriate comparison with the corresponding fragment models.

First, simply computing the chemical shieldings on the central fragment alone (1-body approximation) leads to errors that are far too large to be useful. For example, the rms error for the ^{13}C chemical shieldings computed for the isolated fragment relative to those for the same atoms in the full protein is $\sim$6 ppm, regardless of whether PCM embedding is employed. The errors for the ^{17}O shieldings exceed 10 ppm. The 18 surrounding amino acid residues clearly impact the chemical shieldings on the central fragment amino acids appreciably.

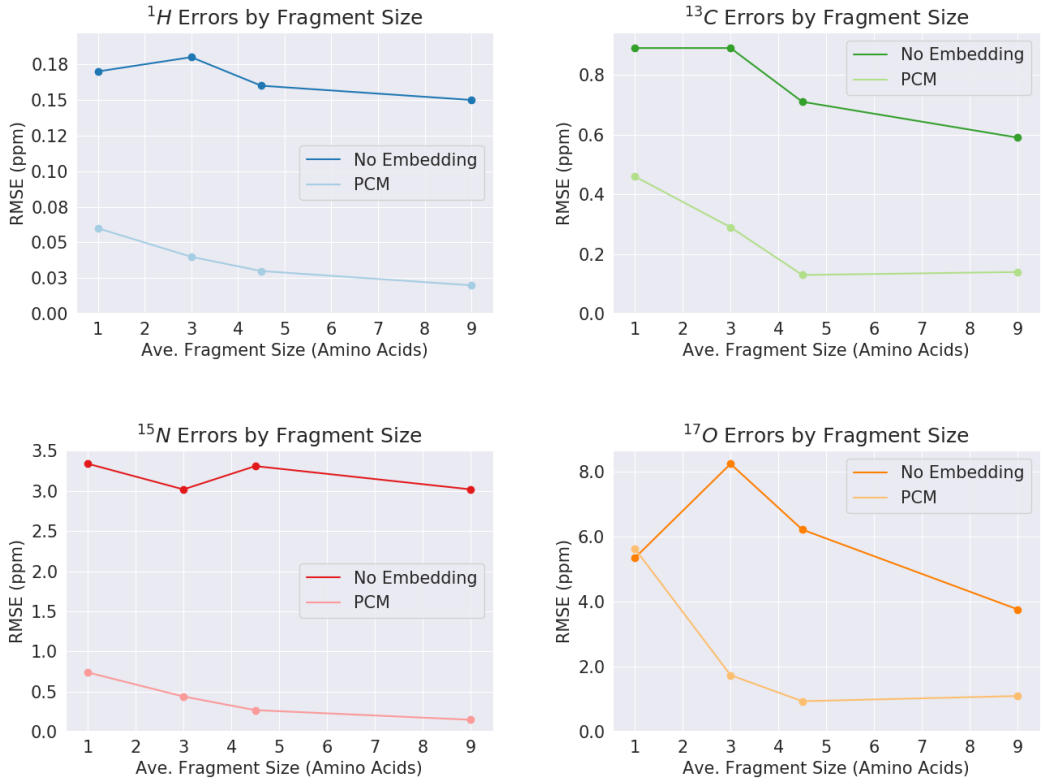


Figure 3.3: Root-mean-square errors in reproducing the 17 ^{1}H, 19 ^{13}C, 7 ^{15}N, and 4 ^{17}O chemical shieldings on the central HVGK fragment of piscidin-1 when using a 1-body and 2-body many-body expansion. The reference chemical shieldings were computed using the entire unfragmented protein, with or without the PCM ($\epsilon = 8.9$) as appropriate.

Introducing the contributions of those other 18 amino acids in a pairwise fashion substantially improves the quality of the shielding. Figure 3.3 plots the rms errors for the ^{1}H, ^{13}C, ^{15}N, and ^{17}O chemical shieldings computed with the 1+2-body fragmentation model as a function of fragment size. Even with fragments composed of just one amino acid each, the rms errors for the 1+2-body models are considerably smaller than the 1-body only models mentioned above. For example, the ^{13}C rms error falls from around 6 ppm to below 1 ppm upon inclusion of pairwise fragment contributions involving single amino acids.

Without PCM embedding, the rms errors generally decrease with increasing fragment size, though the convergence is slow and sometimes non-monotonic. A full analysis of PCM vs. No PCM can be seen in section in the appendix.

Employing PCM embedding typically decreases the rms errors by a factor of 2–3 compared to the non-embedded case, even for the smallest fragment sizes. The only notable exception occurs for ^{17}O with fragment size of one, where the no embedding case exhibits anomalously small errors. Even when the fragments contain just a single amino acid, the shielding error introduced by PCM-embedded fragmentation is already considerably smaller than the typical chemical shift errors exhibited between DFT and experiment. For example, molecular crystal benchmarks discussed below and published previously[105, 70] that employ the same PBE0 functional and basis sets found rms errors of $\sim$0.3 ppm for ^{1}H, $\sim$1.3 ppm for ^{13}C, $\sim$4 ppm for ^{14}N, and $\sim$7.5 ppm for ^{17}O. As the fragment size increases, the errors introduced by fragmentation decrease systematically. Including 4.5–9 residues per fragment and employing PCM embedded leads to chemical shieldings that are essentially converged. The associated errors introduced by fragmentation are insignificant relative to typical errors versus experiment.

These piscidin-1 results highlight how the use of PCM embedding can accelerate the convergence of the many-body expansion and enable even low-order one- and two-body expansions to capture the full system result. However, use of the PCM does change the absolute chemical shieldings (hence the use of different benchmark shieldings for the full system in the embedded and non-embedded cases). On the other hand, the absolute shielding value is far less important than the experimentally observable chemical shift. The

next section investigates the performance of the embedded and non-embedded models for reproducing experimental chemical shifts in a benchmark set of molecular crystals.

3.4.2 Molecular Crystal Benchmarks Against Experiment

The performance of the PCM-embedded fragment approach is examined for benchmark molecular crystal test sets of ^{13}C, ^{15}N, and ^{17}O chemical shifts. The relatively static and well-characterized nature of molecular crystals, along with the ability to partition them into fragments without cleaving covalent bonds makes them particularly useful for benchmarking purposes and for fitting the chemical shift referencing regression lines. ^{1}H chemical shifts are not considered here, due to the bigger role dynamics and nuclear quantum effects have for those shifts.[65]

We first examine the set of 132 ^{13}C chemical shifts taken from 21 crystals, including amino acids, nucleosides, sugars, and other small molecules. Figure 3.4 plots the error distributions for several different models. For each model, the absolute shieldings were computed and a linear regression was performed between the computed shieldings and experimental chemical shifts to establish the chemical shift referencing for that model. The distribution of chemical shift errors obtained from each model is represented via a violin plot. A box plot inscribed within each violin shows the median error (white dot), middle 50th percentile (black box), and the range of errors (black lines). The GIPAW PBE and SCRMP-embedded 2-body fragment PBE0 results shown here were taken from Ref [105].

To begin, consider the GIPAW PBE, SCRMP PBE0, and no embedding PBE0 results. As has been discussed previously,[105, 108, 109] the SCRMP and GIPAW models perform similarly for ^{13}C chemical shifts when the same density functional is used. However,

the fragment approaches make it less expensive to employ hybrid density functionals, and hybrid functionals like PBE0 reduce the rms error versus experiment by about a third compared to a generalized gradient approximation (GGA) density functional like PBE (i.e. from 2.1 ppm with PBE to 1.3 ppm with PBE0). The SCRMP PBE0 error distribution is also peaked much more sharply about zero than the GIPAW PBE one, and the maximum error is reduced from 6.8 ppm (GIPAW) to 3.7 ppm (SCRMP PBE0). Unlike nitrogen and oxygen chemical shifts, carbon chemical shifts are relatively insensitive to the electrostatic environment. As a result, the non-embedded model performs only moderately worse than the embedded ones, with an rms error of 1.4 ppm instead of 1.3 ppm.

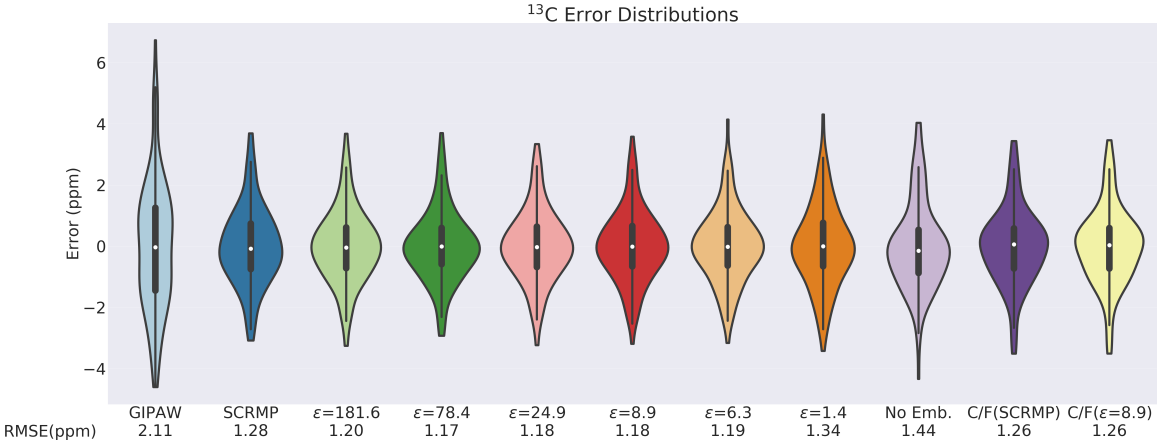


Figure 3.4: Errors in predicting 132 experimental ^{13}C chemical shifts for 21 molecular crystals. The GIPAW model employs the PBE functional; all other models use the hybrid PBE0 functional. The SCRMP and PCM embedding models with various dielectrics all employ 1- and 2-body fragment approximations, while the clusterfragment (CF) model employs a large central cluster plus longer-range 2-body shielding contributions. The No Embedding model control omits any electrostatic embedding when computing the 1- and 2-body shielding contributions.

Next, fragment calculations are performed with PCM embedding instead of the SCRMP point-charge embedding. A series of implicit PCM solvent environments are considered, with dielectric constants ranging ϵ=1.4–181. Intermediate dielectric constants in

the $\epsilon \sim 6$–25 range are consistent with typical organic environments that one might associate with a neutral organic crystal, while the higher dielectrics reflect more highly polarizable environments. As shown in Figure 3.4, the ^{13}C chemical shift error distribution is insensitive to the specific choice of the embedding dielectric, with rms errors of 1.2 ppm for nearly all dielectrics. Only as the dielectric constant decreases toward the no-embedding limit ($\epsilon \to 1$) does the error begin to increase noticeably. The rms errors for the PCM-embedded models are about 0.1 ppm smaller than those from the SCRMP embedding, though that may not be significant relative to the precision of the experimental chemical shifts (typically ± 0.1 ppm for ^{13}C). Altering the PCM dielectric environment changes the absolute shielding values obtained, but these differences are compensated for by the chemical shift referencing regression.

Finally, earlier work also investigated the performance of a combined cluster/fragment model.[108] The cluster/fragment model calculates the chemical shielding on a larger cluster consisting of all molecules lying within 4 Å of the central monomer of interest, ensuring that the local many-body effects are explicitly captured in the quantum mechanical treatment. Longer-range contributions are approximated in a pairwise fashion between the central molecule and any other molecule out to 6 Å away (the same cutoff used in the 1+2-body fragment model) and via the electrostatic embedding. Previous work with point-charge embedding has found that the 2-body fragment and cluster/fragment models perform similarly for ^{13}C chemical shifts.[109, 108, 105] That finding is echoed here with the PCM embedding. As shown in Figure 3.4, one obtains the same 1.3 ppm rms error for the cluster/fragment model with either embedding approach. Taken together, these results indicate that when

computing ^{13}C chemical shifts via a fragment approach, one can replace self-consistent point-charge embedding with a PCM model and that the specific choice of the environment dielectric is largely unimportant, as long as it is somewhat more polarizable than a vacuum. Because the PCM embedding effectively captures the long-range and many-body contributions in the system, the accuracy benefits of including non-additive terms beyond 1+2-body is modest.

Nitrogen chemical shifts are typically more sensitive to electrostatic environment than carbon, which makes them a more rigorous test for the PCM embedding model. Figure 3.5 plots error distributions for 37 isotropic ^{15}N chemical shifts from 16 crystals. Indeed, predicting the ^{15}N chemical shifts via the 2-body fragment PBE0 approach without any electrostatic embedding leads to a poor chemical shift referencing regression line and a 39 ppm rms error relative to experiment, emphasizing the importance of treating long-range and many-body effects when predicting chemical shifts in the condensed phase. Due to those extremely large errors, the data for the non-embedded model is omitted from Figure 3.5. Examining the fragment models with SCRMP and PCM embedding, we see once again that the fragment models employing the hybrid PBE0 functional predict the shifts with considerably smaller rms errors of 4.0-4.2 ppm, compared to 5.6 ppm for GIPAW PBE. The dependence of the errors on the dielectric environment is once again small, spanning 0.2 ppm for dielectrics ranging from 6.3 to 181. In most cases, the errors for the PCM-embedded model are marginally smaller than those from the SCRMP approach, though these variations are not particularly significant relative to experimental precision. Once

again, the errors begin to increase as the dielectric constant approaches 1 (vacuum). For example, reducing ϵ from 6.3 to 1.4 increases the rms error from 4.0 ppm to 15.5 ppm.

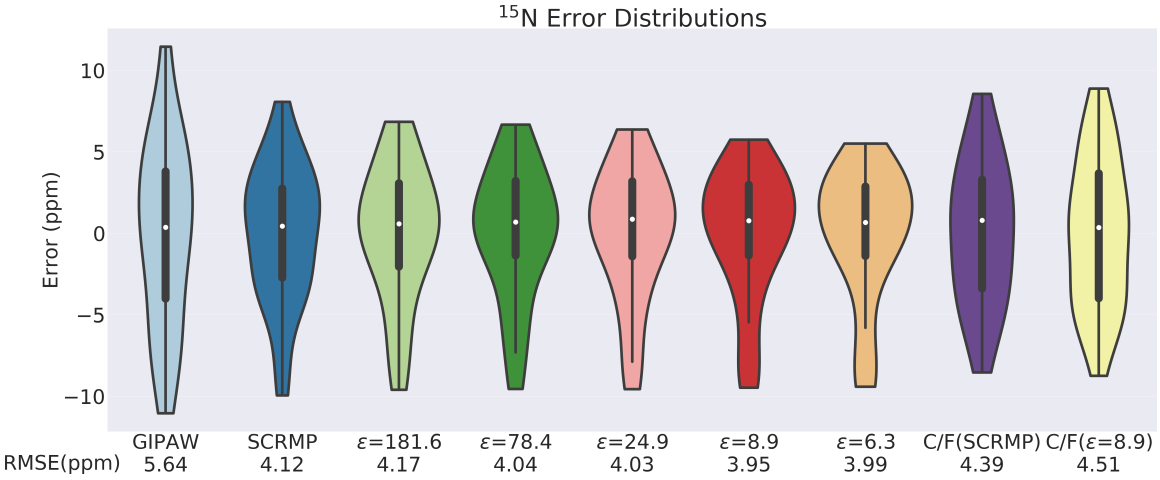


Figure 3.5: Errors in predicting 37 experimental ^{15}N chemical shifts for 16 molecular crystals. See Figure 3.4 for a more detailed description of the plot. Data for the $\epsilon = 1.4$ and No Embedding cases are omitted here because their errors are very large.

Further validation for the PCM-embedding comes from the ^{15}N cluster/fragment results. Once again, the SCRMP- and PCM-embedded variants of the cluster/fragment model perform similarly to each other and to the 1+2-body fragment approach. Consistent with what has been found previously,[105] the cluster/fragment errors relative to experiment are a few tenths of a ppm larger than the 1+2-body ones. This is probably fortuitous, since the cluster/fragment approach captures more of the many-body polarization effects with explicit quantum mechanics and therefore ought to be more reliable. Use of a larger benchmark data set to fit the chemical shift referencing regression would hopefully produce the expected behavior in which the cluster/fragment model is more accurate than the 1+2-body one. Regardless, the accuracy differences between the 1+2-body and cluster/fragment

models are modest, demonstrating once again that the embedding environment captures the important many-body effects.

Finally, ^{17}O chemical shifts exhibit even greater sensitivity to electrostatic embedding. Without any electrostatic embedding, the 2-body fragment model exhibits rms errors of over 80 ppm. For models which do include many-body/embedding effects, prior work with fixed point-charge embedding found the 2-body model performs considerably worse than the cluster/fragment one for ^{17}O, with rms errors of 9.8 and 7.6 ppm, respectively.[108] Switching to the self-consistent SCRMP embedding to polarize the point charges reduced the errors and narrowed the difference between the two models dramatically, to 7.5 ppm for the 1+2-body model and 7.2 ppm for the cluster/fragment model.[105] Given that experimental uncertainties in oxygen chemical shifts are typically half a ppm or more, a 0.3 ppm difference in accuracy between the two models is small. Nevertheless, the limitations of the SCRMP embedding are highlighted by the fact that GIPAW PBE performs comparably to SCRMP PBE0 (see Figure 3.6). As another recent study showed, correcting GIPAW PBE calculations with gas-phase monomer PBE0 improves the rms errors for ^{17}O chemical shifts by another $\sim$2 ppm.[70]

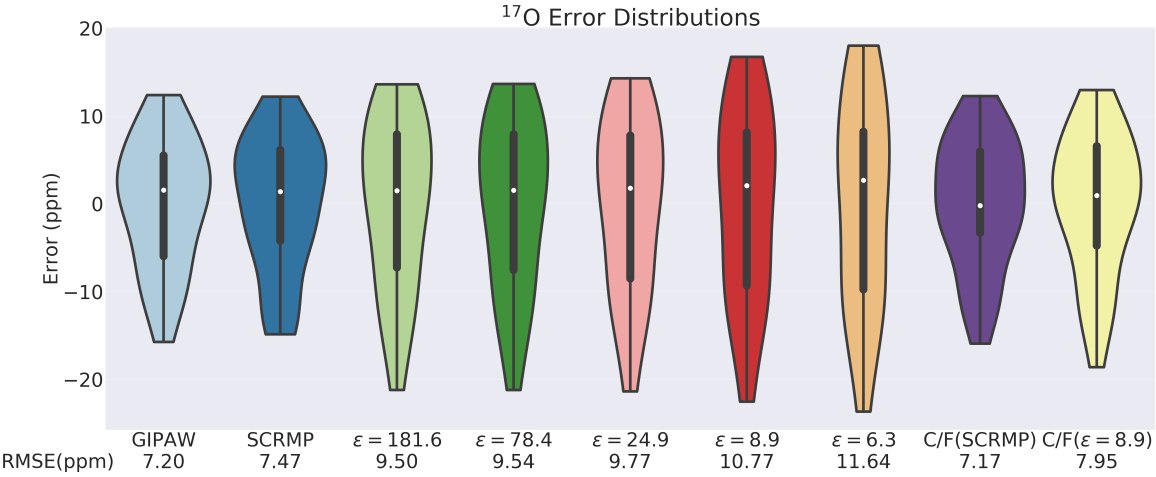


Figure 3.6: Errors in predicting 28 experimental ^{17}O chemical shifts for 15 molecular crystals. See Figure 3.4 for a more detailed description of the plot. Data for the $\epsilon = 1.4$ and No Embedding cases are omitted because their errors are very large.

For the present study, Figure 3.6 shows that the PCM model performs somewhat worse than the self-consistent SCRMP point charge embedding for ^{17}O in a set of 28 chemical shifts for 15 crystals. Nevertheless, the errors with larger dielectrics (e.g. ϵ=25, 78, or 181) are still slightly smaller than what was obtained from the earlier fixed point charge embedding schemes.[108] In other words, the performance of the PCM-embedded models for ^{17}O is reasonable, even if it is less accurate than the self-consistently polarized SCRMP-embedded[105] or monomer-corrected GIPAW [70] models. Unsurprisingly given the important role of the electrostatic embedding, the ^{17}O chemical shift errors are more sensitive to the dielectric constant than either carbon or nitrogen chemical shifts are, with the larger dielectric constants here performing best. In the limit of low dielectric constants, the errors rise steeply. For example, reducing ϵ from 6.3 to 1.4 doubles the rms error to 23 ppm, and with no embedding ($\epsilon = 1$) it exceeds 80 ppm. These results reiterate how important embedding environment is for a 1+2-body fragment model to perform reasonably.

Finally, the PCM-embedded cluster/fragment model improves considerably upon the PCM-embedded 1+2-body fragment results and gives results that are closer to those obtained with SCRMP embedding. Explicitly treating the local many-body contributions with quantum mechanics reduces the sensitivity of the predicted chemical shifts to the environment. Of the three nuclides considered here, ^{17}O benefits the most from explicit inclusion of terms beyond 1+2-body.

Overall, these molecular crystal benchmarks show that the electrostatic embedding treatment can effectively be replaced by a polarizable continuum model. Although the optimal dielectric for a given system might be difficult to define, the impact of the specific dielectric constant ranges from small (carbon and nitrogen nuclei) to modest (oxygen). Moderate or high dielectric constants appear to perform best. Based on the fact that ^{13}C and ^{15}N chemical shifts are measured far more frequently and accurately than ^{17}O ones, a dielectric of ϵ=8.9 (dicholoromethane) is adopted for the remainder of the paper. That value is reasonably consistent with the $\sim$6–9 range of dielectric constants often cited as "typical" for proteins.[142, 133, 5]

3.4.3 Indoline Carbanionic Intermediate Bound Within the β-Subunit of Tryptophan Synthase

Having seen that the PCM-embedded fragment approach can effectively reproduce both the chemical shieldings from a full-system calculation in piscidin-1 and experimental chemical shifts in molecular crystals, we now turn to the fragment-based prediction of experimental chemical shifts of the indoline carbanionic intermediate bound within the active site of the β-subunit of tryptophan synthase. Typtophan synthase catalyzes the last two steps in

tryptophan biosynthesis and belongs to a class of pyridoxial-5'-phosphate (PLP) enzymes. PLP dependent enzymes carry out vital biochemistry such as racemization, transamination, decarboxylation, elimination, and substitution reactions.[245, 114, 38] Detailed mechanisms of PLP enzymes like tryptophan synthase are incompletely understood at the atomic level and have been the subject of considerable study.[246, 169]

The NMR chemical shifts in the tryptophan synthase active site are considerably more challenging to model with the fragment approach than those examined in the sections above. The enzyme has far more tertiary structure than piscidin-1, meaning that the substrate is deeply embedded in a three-dimensional protein environment. It also is highly charged, including a -4 net charge on the substrate. Intermolecular interactions stabilize those substrate charges in the full system, but their presence can be problematic in individual fragment calculations that omit most of the environment unless appropriate electrostatic embedding is employed.

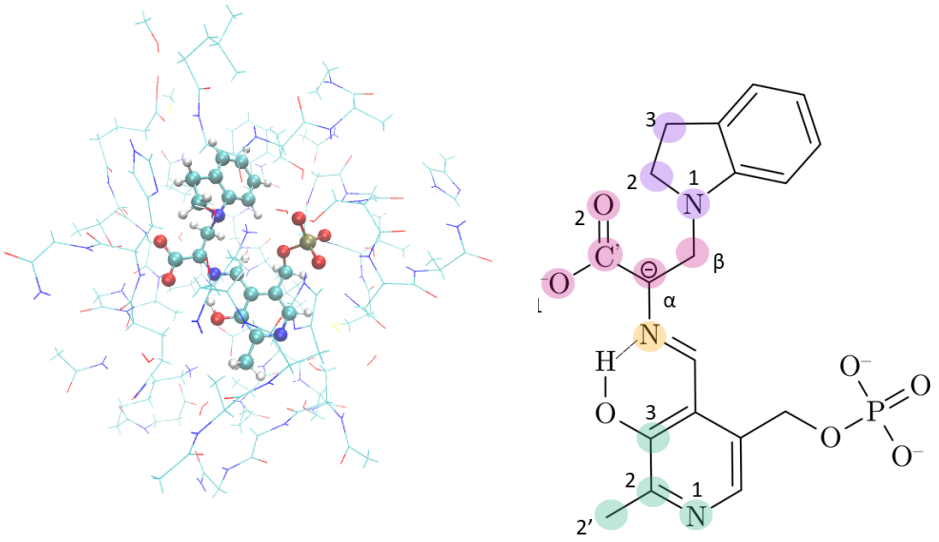


Figure 3.7: Indoline carbanionic System from PDB-ID 3PR2, showing (a) indoline-protein cluster model used to study this system and (b) the isotopically labeled sites in the substrate for which experimental chemical shifts have been measured (colored circles).

Following prior work,[136, 35] this system is modeled here via a finite cluster model consisting of the indoline quininoid intermediate substrate and ∼7 Å of surrounding protein (Figure 3.7a). The cluster model also includes seven explicit crystallographic water molecules and a sodium cation. The specific cluster geometry was taken directly from Ref [35]. That study reports 13 isotopically labeled 13-C, 15-N, and 17-O chemical shifts in the indoline carbanionic substrate (Figure 3.7b).

The present study compares the chemical shifts predicted for the substrate in the entire 612-atom cluster against those obtained from the PCM-embedded fragment approach. Two different protein fragmentation schemes are compared. The first scheme performs highly aggressive fragmentation, cutting the protein environment into fragments consisting of just one amino acid each (cleaved at the Cα-C$'$ single bond). The substrate has 48 atoms, and the median and maximum dimer sizes in this scheme are 61 and 87 atoms, respectively.

The second scheme employs larger, more chemically appealing fragments obtained largely from the peptide fragments that were naturally created when extracting the cluster model from the full protein. New covalent cuts were made only for the largest peptide fragments. In this second scheme, the median dimer fragment has 79 atoms, and largest dimer fragment contains 133 atoms (substrate plus an 8-amino acid peptide). Figure S1 shows the fragments used in this second scheme. In both fragmentation schemes, the explicit water molecules and Na^+ ion were grouped into a single fragment for convenience. It should be noted that the use of PCM embedding is compatible with the explicit inclusion of chemically important solvent molecules.

Table 3.1 reports the key experimental and predicted chemical shifts for the indoline carbanionic intermediate. Chemical shieldings for the PCM-embedded fragment methods were converted to chemical shifts using the corresponding 1+2-body fragment linear regression parameters obtained from the molecular crystal systems above (Section 3.4.2). For comparison purposes, predicted chemical shifts are also reported for the full cluster without fragmentation or embedding, matching how this system was modeled previously.[136, 35] Those values were referenced using the cluster/fragment scaling parameters reported in ref [108].

The two-site proton exchange model described in Section 3.3.2 was employed to represent the final chemical shifts of this system as a weighted average of the protonated Schiff base nitrogen and protonated phenolic oxygen structures. See ref [35] for details. In each case, the optimal mixing ratio of the two structures that minimizes the value of the reduced χ_r^2 statistic was found. The optimal mixing ratio varies little across the different

models, ranging from 70-77% phenolic oxygen (and 23–30% Schiff base nitrogen). These variations do not alter the picture of the system being dominated by the phenolic oxygen form. The χ_r^2 statistic was computed as,

$$\chi_r^2 = \frac{1}{N} \sum_i \frac{(\delta_i^{pred} - \delta_i^{expt})^2}{s_i^2} \tag{3.11}$$

where N is the degrees of freedom (12 for the two-site models, due to 13 measured shifts and 1 fitted parameter). The denominator s_i^2 corresponds to the expected root-mean-square error for the given model for each nucleus, as taken from the molecular crystal benchmarks above. For example, for the 1+2-body fragment approach with $\epsilon = 8.9$ dielectric, $s_i = 1.18$ ppm for ^{13}C, 3.95 ppm for ^{15}N, and 10.77 ppm for ^{17}O, as reported in Figures 3.4–3.6. The χ^2 statistic effectively normalizes for the fact that errors relative to experiment are typically largest for oxygen, intermediate for nitrogen, and smallest for ^{13}C chemical shifts.

	Chemical Shift Errors vs. Experiment						
	Single Amino Acid Fragments				Larger Fragments	Full	Expt.
	ϵ=181.6	ϵ=78.4	ϵ=24.9	ϵ=8.9	ϵ=8.9	Cluster	(ref [35])
PLP C2	-3.5	-3.4	-3.5	-3.4	-1.5	-0.1	145.4
PLP C2′	2.1	2.1	2.1	2.2	2.4	2.8	17.0
PLP C3	-2.0	-1.8	-1.7	-1.3	-1.1	-0.6	154.1
Serine Cα	0.4	0.3	0.0	-0.4	-0.3	-0.5	103.5
Serine C′	-2.3	-2.1	-2.2	-2.1	-1.7	-2.4	173.0
Serine Cβ	-2.9	-2.9	-2.9	-2.8	-3.0	-3.1	54.1
Indoline C2	-1.1	-1.1	-1.2	-1.1	-1.0	-0.1	50.5
Indoline C3	0.8	0.9	0.9	1.1	1.1	-1.9	28.5
PLP N1	-1.7	-1.9	-1.3	0.0	2.4	0.1	265.0
Schiff Base N	-5.4	-5.6	-5.0	-3.2	2.4	0.6	296.0
Indoline N1	-6.1	-5.8	-4.9	-4.0	-5.3	-1.1	83.5
Serine O1	19.1	21.8	18.5	23.7	10.8	8.1	243.0
Serine O2	20.8	19.9	17.0	12.3	-0.6	-8.6	233.0
^{13}C rms error	2.1	2.1	2.1	2.0	1.8	1.7	
^{15}N rms error	4.8	4.8	4.1	3.0	1.0	3.6	
^{17}O rms error	20.0	20.9	17.7	18.9	8.4	7.6	
χ^2_r (two-site)	3.15	3.26	2.92	2.63	1.66	1.26	
% Phenolic Oxygen	70%	70%	70%	71%	73%	77%	

Table 3.1: Errors in the PBE0 chemical shifts (ppm) computed for the full cluster and those from the PCM-embedded 2-body fragment approach relative to experiment. Fragment calculations were performed using either single amino acid fragments in various dielectrics or larger fragments and $\epsilon = 8.9$. The reduced χ^2_r statistic for the chemical shift errors is reported for each case for the optimal mixture of phenolic oxygen and Schiff base nitrogen.

Several features can be observed in the predicted shifts reported in Table 3.1. For the aggressive partitioning with single amino acid fragments, the predicted chemical shifts are again only modestly sensitive to the dielectric constant. The best agreement with experiment and smallest χ^2_r of 2.63 is obtained with $\epsilon = 8.9$. However, the mean-absolute change in chemical shifts from $\epsilon = 181$ to $\epsilon = 8.9$ is only 0.3 ppm for carbon, a moderate 2.0 ppm for nitrogen, and a somewhat larger 6.6 ppm for oxygen. This sensitivity to the dielectric constant is larger than what was seen for the molecular crystals, but it is still probably reasonable given the smaller numbers of shifts behind these statistics (especially for nitrogen and oxygen) and the highly charged nature of this system. On the other hand,

the chemical shifts from this aggressive fragmentation model differ appreciably from the full cluster results, with mean absolute differences of 1.1 ppm for carbon, 2.3 ppm for nitrogen, and 18.3 ppm for oxygen for the $\epsilon = 8.9$ dielectric. The errors relative to experiment are also somewhat worse than those obtained for the full cluster, especially for oxygen. The χ_r^2 statistics are correspondingly larger, with values ranging 2.6–3.3. For the twelve degrees of freedom here, models with a $\chi_r^2 \geq 1.75$ can be ruled out with 95% confidence. In other words, the errors introduced by fragmenting the protein down to single amino acids are large enough to impact the agreement with experiment meaningfully.

On the other hand, using larger fragments and the same $\epsilon = 8.9$ that worked well for the molecular crystals and the aggressive fragmentation of this current system reduces several of the largest errors versus experiment (PLP C2 and the two serine oxygens) that occur when the smaller fragments are used. In particular, the improvements for the oxygen chemical shifts reduce the χ_r^2 statistic considerably to 1.66. This χ_r^2 value indicates that the model chemical shifts are consistent with experiment at the 95% confidence level. The errors for the fragmented model are still larger than those obtained with the full cluster model, particularly for nitrogen, but this is a trade off for the considerably lower computational cost of the fragment approach.

Overall, the results suggest that the PCM-embedded fragment approach provides sufficient accuracy to be useful in NMR crystallography applications for challenging protein systems like this as long as reasonably large fragments are used (several amino acids each). Previous studies of intermediates bound in the active site of tryptophan synthase found that many structure candidates have large $\chi_r^2 \gg 10$.[136, 35]. On that scale, the errors introduced

by fragmentation would be small enough to allow elimination of a significant number of candidate structures. One could, for example, use the fragment approach as an inexpensive initial screening tool for considering large numbers of structures before performing more accurate and expensive chemical shift predictions on the best candidates using larger cluster models. Furthermore, the fragment approach might also enable including more or all of the entire protein environment, instead of just the several Ångstrom cluster around the active site used here.

3.5 Conclusions

Chemical shift prediction in condensed-phase systems can be accelerated via electrostatically embedded fragmentation schemes. The present work demonstrates that a PCM can provide a simple embedding strategy that is readily available in many electronic structure software packages. The performance of the PCM embedding approach was demonstrated here for systems ranging from molecular crystals to biomolecules. It was found to perform on par with self-consistent charge embedding schemes like SCRMP in molecular crystals, while also being effective in the charged protein environments. The accuracy of the predicted chemical shifts is rather insensitive to the specific dielectric constant chosen, thereby avoiding concerns about picking the appropriate dielectric for the system at hand. A dielectric environment consistent with dichloromethane ($\epsilon = 8.9$) generally appears to perform well, though only very small dielectrics become problematic.

Even if one has concerns about the errors introduced by fragmentation in challenging systems such as the tryptophan synthase one considered here, the fragment approach

could still be used as a very effective screening model for ruling out candidate structures whose chemical shifts disagree markedly with experiment. Finally, the fact that the same model can be used on both molecular crystals and biomolecules is advantageous. Chemical shift referencing regressions can be fitted based on the well-characterized and (mostly) static molecular crystal structures which are completely distinct from the biomolecular systems. Those regressions can then be applied to reference the chemical shifts in more complicated and dynamic biomolecules.

We have now shown methods that can routinely predict chemical shifts with very good agreement compared to experiment, via fragment and planewave methods. We are now in a position to explore techniques which can expedite these calculations. In the following chapter, we discuss some of the mathematical foundations of machine learning, a powerful technique which can act as a surrogate model to quantum chemistry calculations, with an emphasis on NMR chemical shift prediction. We discuss the some of the general principles of machine learning as well give a prospective on the field.

Chapter 4

Machine Learning

Machine learning (ML) is a subfield of artificial intelligence that creates a surrogate model to make predictions or decisions based on a representative set of training data. The simplest machine learning model one can construct is a linear regression: given known x and y values that seem to follow linearity, we can create a model with 2 parameters – the slope and y-intercept. However, we wish to model complicated phenomena with highly non-linear behavior. Here, the goal is bypass the computational burden of solving the Schrödinger equation by using ML. The overall ML workflow is depicted in figure 4.1.

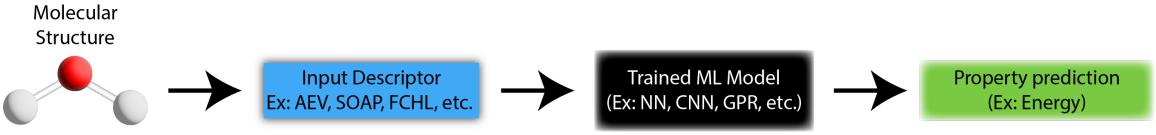


Figure 4.1: Cartoon of ML workflow. The structure is converted to an input descriptor, from there it is "fed" into the ML model and based on the training will output the predicted values.

4.1 A Pedagogical Workflow for Machine Learning

This section outline some key pragmatic questions for the chemist who wishes to embark on developing a machine learning solution. They are designed for the non-specialist and intend to be a rough template for designing a ML project.

4.1.1 Are There Other *Simpler* Solutions Than ML? Is the Current ML Model Good Enough?

It is quite vogue for one to create a ML solution when the current state-of-the-art works well or the task at hand is not computationally expensive. This first question highlights a simple step that many practitioners often ignore due to oversight or profits. For example, a societally important task is the recidivism risk prediction tool to allocate bail amounts. While there are complex ML models that allocate bail based on a huge number of input factors, [29] a simple model with 4 `if/then` statements has been shown to work equally as well. [244] While not directly related to physical sciences, this simple example highlights how one can over-engineer a solution (e.g. fit a million parameters to the equation of a line). Thus, one must carry out their due diligence to ensure that ML is a worthy endeavor to even start.

The second question is more subjective dependent on the level of accuracy needed. If one is trying to quickly gauge the energy differences of small molecules, existing ML potentials like ANI or SchNet will work just fine. [220, 208] But if one is trying to calculate kJ/mol accuracy on a molecular crystal dataset, then more care will be needed.

4.1.2 What Data is Currently Available for the Prediction Task? Is the Dataset Representative?

Broadly speaking, when ML research in chemistry began to take off (circa 2014), sharing data was not popular. Nowadays, sharing large datasets between research groups is the norm due to changes in best practices and a healthy amount of skepticism. There are plenty of resources for quantum chemistry data through MolSSI (`https://qcarchive.molssi.org/`), community maintained websites (`http://quantum-machine.org/datasets/`), larger ML competitions (`https://paperswithcode.com/dataset/ogb-lsc`), and national labs (`https://materialsproject.org/`). [219, 125, 244] These resources represent just the surface of large dataset repositories and are good starting points for properly labelled structures with properties (e.g. energies, HOMO-LUMO gaps, dipole, etc.) one would study with quantum chemistry. In addition, the next latest and greatest ML papers are often posted on `https://arxiv.org/` with code and/or data year(s) before publication due to the peer review process. An exhaustive search for representative ML data will save one much frustration, computer time, and human time for data processing. One can only imagine how large datasets will reach in the future at the time of this writing. For best practices in sharing data, see this excellent publication by Walsh, Isayev, and friends. [6]

Representative, high-quality data is the most important ingredient in your ML recipe. For example, ref. [269] created a ML model with two freely available datasets for ^{1}H chemical shift prediction (see figure 3 in the paper). The more sanitized dataset, SHIFTX, is only one-tenth the size of the second dataset, RefDB, but training both (independently) yielded the same out-of-sample prediction accuracy. Thus, double-checking if the data is

clean is a worthwhile step especially if one hasn't curated the data themselves. This step can be painful, but section 4.1.4 outlines some anecdotal tips for data sanitization.

4.1.3 How Does One Construct a Dataset?

Unfortunately, there will be instances where the data is not readily available and one has to construct a dataset for this new ML task. One consideration is the computational "allowance" available to construct a dataset. For example, ref. [226] wanted to create a ML model to predict CCSD quality energies, but the calculations to create a reliable database would be too computationally expensive to curate. Some simple benchmarking to calculate the number of structures at a desired level of theory will help at this stage. Properly creating a dataset also depends on the type of chemistries one wishes to model later down the line. One cannot train a ML model solely on aliphatic hydrocarbons and expect great performance on aromatic rings. Thus properly data sampling can be broken down to two steps: 1) Use an existing input descriptor to quantify the relationship between structures. 2) A sampling algorithm for each data point.

Input Descriptors

An input descriptor (sometimes called features) is a mathematical description of the molecular structure (local or periodic) to compare points across chemical space. Here, we define chemical space to be some abstract hyper-dimensional surface that contains all possible structures as well as their changes in geometries with respect to some coordinate, such as the energy. Similar to the E_{XC} in DFT, an exact input descriptor is not known and all descriptors make some approximations. The basic premise of each descriptor is as follows:

each molecular geometry should have a unique mathematical representation in which the ML model should be able to tease apart. Creating an input descriptor is an active field of research in ML for quantum chemistry. There are several excellent literature reviews which benchmark various precedents and describes the design principles a descriptor should contain. [137, 56, 193] In short, an ideal input descriptor should be translationally and rotationally invariant so as not to introduce degeneracies into the training procedure. In our work, we use the AEV, a modified version of atom-centered symmetry functions, due to its popularity in the ANI neural network potential papers. [220, 226, 61, 275] The mathematical details are worked out in section 5.2.2, but the graphical description in figure 4.2 highlights how changes in geometry for the *same* molecular structure are reflected in the input descriptor. Do not forget that the input descriptor **is not** the molecule, but rather some lower dimensional projection which is not exact. Thus, some input descriptors perform better for certain tasks than others and benchmarking may be needed. Now that we have some way to characterize whether two structures are mathematically similar, we can now properly sample.

Sampling Chemical Space

This section will focus on static data sampling, rather than dynamic (also known as active learning) data sampling. The premise between the two is quite simple, we construct the dataset based on sampling the dataset once (static) vs. continuously sampling and updating the dataset/ML model (active learning). For an excellent example on active learning, see the trial-by-committee paper [224] or this primer in nature computational materials. [30]

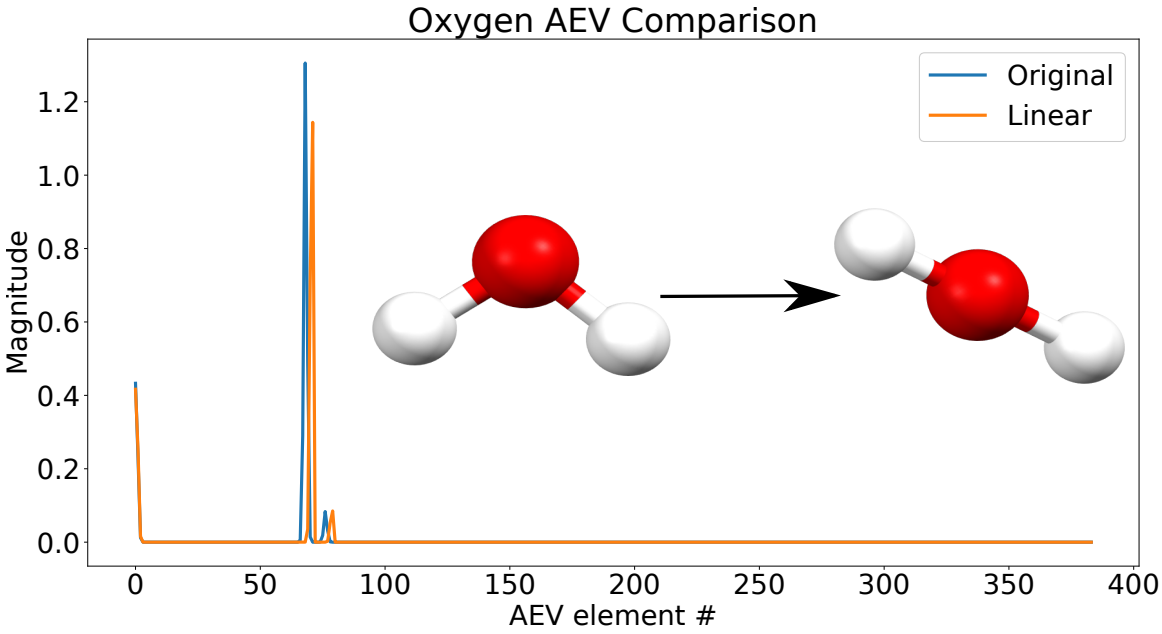


Figure 4.2: Sample AEV (input descriptor) for oxygen in two different geometries of water. While they have the same stochiometry, the change in the bond angle is reflected in changes in the input descriptor.

Ultimately, sampling will depend on how one will **test** the data. A straightforward sampling scheme is to access a dataset repository (such as one of the ones listed above) and search for structures that match the desired criteria. From there, one can use an input descriptor to label each structure and compare similarities. For example, in the ShiftML paper, the authors accessed the Cambridge Structural Database and found 61 thousand structures that met their critera. [177] They then carried out furthest point sampling to select 2000 structures for their training dataset and the test dataset was then a random sample of the remaining structures. In another example, the authors of ref. [250] trained on the GDB(N=1-8) set and tested on a smaller sample of the GDB17, with the idea that a properly trained ML model should transfer to larger structures. While in principle this could work, it would have been even better if the authors created one large dataset with

a good mix of GDB(N=1-8) through GDB17 and then finally tested on a "holdout" set. Ultimately, there are many ways to sample, but using an input descriptor to help narrow down your search is recommended.

After careful curation, it is time to create the **train/validation/test** split. Anecdotally, this is usually set to 0.80/0.05/0.15 in the ML literature. The **train** set is used solely to train the model, the best errors on the **validation** set will be used indentify the best performing model, and finally the **test** set (or "holdout" data) evaluates the final performance. Ideally, the validation and test errors should be similar. However, do not be surprised if the test errors are much larger. The train/validation/test step is usually automated through another external package, but do not forget to fix the random seed if one wishes to directly compare the effect of training size.

4.1.4 What is Wrong With my own Dataset?

ML datasets become large very fast! It is unreasonable to spot check every input and output. Thus, here are some key tips to make sure that your ML model is performing poorly due to the model and not some silly mistake in processing.

- If your ML model fails spectacularly, start with this step. The order of files and values must be kept consistent. Some processing tools and scripting may reorder and cause a huge mess.

- Plot the range of *target* values. This may reveal an outlier or a specific region that is problematic.

- Plot the range of *predicted* values from the ML model. A constant line indicates that you do not have sufficient amounts of data or the ML model is not powerful enough to discriminate between data points. Make sure you try other ML models or hyperparamaters to make sure it's not the former.

- Double check files from repositories. They are usually well maintained, but in some instances there may be errors.

- Use packages that maintain database files like pandas (`https://pandas.pydata.org/`). [176] Pandas has integration with other popular ML software and store data in a compressed format.

- Lastly, backup data to an external drive and an external computer. This is for a disaster scenario, but also allows for ease of use when working on other computers.

4.1.5 How Does One Improve a ML Model?

Improving a ML model could be as simple as changing some hyperparamaters (ex: learning rate and batch size in neural networks) or trying different input descriptors. However, it is typically not that simple if one isn't approaching the desired accuracy with routine methods. More often than not, one will need to redeploy the ML model of choice through improvements in model architecture motivated by physics-inspired modifications. We cannot give specific recommendations to the best performing models since they will be obsolete by the time this document is published. However, there are plenty of literature references cited throughout this chapter which outline key precedents.

A final note which is more philosophical in nature. ML models are a black box, meaning there is no clear path for interpretability and/or model improvement. If one were to easily deconstruct the model into its part and it were interpretable by humans from established rules, did one need the ML model to begin with? Another important question is how one interprets the ML model. If one were to have a successful model, and then trains another ML model to decipher the first model, is one truly capturing the correct details? Or is one simply learning the details of the second ML model? Or even worse, associating correlation as cause.

4.1.6 Code Resources for ML Packages

Lastly, when this author started ML, there weren't as many freely available codes. It is highly recommended for a first ML chemistry project to augment an existing code, rather than build one from scratch. Now, it is quite easy to find a code that carries out something similar to what one is trying to achieve. Doing an exhaustive search for data will usually uncover code for that project. If not, a thorough search on `https://github.com/` will typically be fruitful. Github is the *de facto* repository for code, and many people will upload their works in progress. Following code repo's is like following your favorite artist for their next album.

For other resources on ML projects, at the time of writing, pytorch (`https://pytorch.org/`) is the most popular ML package for neural networks (see section 4.2). There are many free tutorials on the website and includes a helpful forum. Hands-on training books are also quite popular, and a worthwhile investment if one wishes to learn even more details. Overall, there will be no shortage of resources to construct models, but it may seem

perplexing to even begin. Trying to construct a ML model, and iteratively augmenting it is much better than reading endlessly and understanding all the details. One will learn much more along the way than trying to digest all the material all at once.

4.2 Neural Networks

This section outlines the most popular ML solution, neural networks (NNs). NNs will be used throughout the rest of this work, and have been quite successful in predicting energies at the DFT-level, but with the computational cost of force field methods. While the mathematical foundations are also presented, one can skip the gory details if they are looking for a review of success stories.

4.2.1 Mathematical Formulation of NNs

This section heavily draws on inspiration from ref. [135]. We use artificial neural networks (NNs) which are based on a collection of connected nodes called neurons that loosely mimic the neurons in a biological brain. NNs have seen explosive growth in the past few decades with advancements in natural language processing and image recognition. A NN has a surprisingly straightforward functional form:

$$f(x_1, ..., x_n) = \rho\left(\sum_{i=1}^{n} x_i w_i - b\right) \tag{4.1}$$

where an *artificial neuron*, x, with *weights* w, *bias* b and activation function ρ, similar to how a neuron "fires" in a brain. While there are numerous activation functions in the literature, by far the most successful and used throughout this work is the rectifier linear

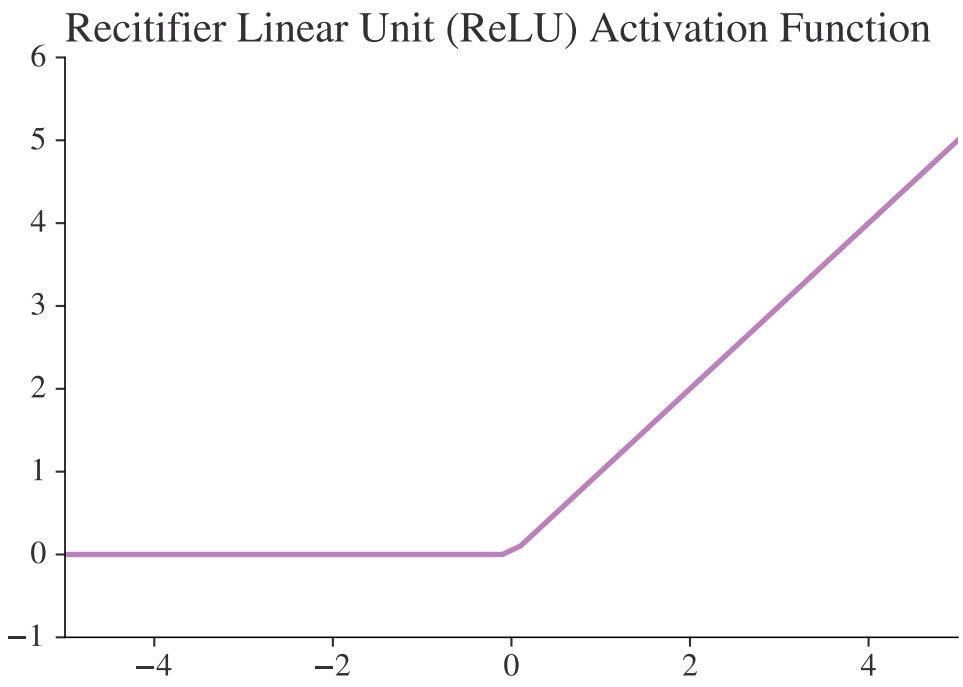


Figure 4.3: Plot of rectifier linear unit activation function $\rho(x) = \max(0, x)$

unit (ReLU) activation function (shown in figure 4.3):

$$\rho(x) = \max(0, x) \tag{4.2}$$

The success of ReLU comes from its simple piecewise linear structure, which has been shown to be advantageous and performs superior to other more sophisticated activation functions. The neurons in a neural network have a particular structure composed of affine linear maps (linear function with a translation). We define the linear transformation functions T_l of layer l to be:

$$T_l = W^{(l)}x + b^{(l)} \tag{4.3}$$

where W^l and b^l are the weight and bias matrices of layer l. We can now formulate the deep neural network Φ as a collection of these transformations of total depth L:

$$\Phi(x) = T_{L\rho}(T_{L-1\rho}(...\rho(T_1(x)))) \tag{4.4}$$

We are now ready to *train* our NN using the following optimization scheme:

$$\min_{(W^{(l)},b^{(l)})} \sum_{i=1}^{m} \mathscr{L}(\Phi_{(W^{(l)},b^{(l)})_l}(x_i), y^i) + \lambda P((W^{(l)}, b^{(l)})_l) \tag{4.5}$$

where the loss function $\mathscr{L}$ determines how close the NN is to the known values $y^{(i)}$ from the training data. An additional term called a regularizer P controlled by another biasing parameter λ to avoid overfitting (i.e. inject small amounts of noise to the loss function). Overall, equation 4.5 tells us we are trying to minimize the loss of the neural network through the weights of each neuron and the bias of layer l. We use the mean-squared-error loss function throughout NN trainings to evaluate performance:

$$\mathscr{L} = \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)} \right)^2 \tag{4.6}$$

where $\hat{y}^{(i)}$ are the predicted and $y^{(i)}$ are the ground truth values, respectively. The most common method to optimize neural network parameters is through *stochastic gradient descent* of small batch sizes. After training, our NN has trained weights and biases with a similar functional form to equation 4.4:

$$\Phi_{(W^{(l)},b^{(l)})}(x) = T_{L\rho}(T_{L-1\rho}(...\rho(T_1(x)))) \tag{4.7}$$

We are finally ready to *test* our NN through analysis of the following:

$$\Phi_{(W^{(l)},b^{(l)})}(x^{(i)}) \approx y^{(i)} \tag{4.8}$$

where the input descriptors $x^{(i)}$ are then fed into a trained NN $\Phi_{(W^{(l)},b^{(l)})}$, and evaluated to see if they correspond to the target value $y^{(i)}$. This is figure 4.1 but in mathematical form.

4.2.2 NNs in Computational Chemistry

The most successful ML model in computational chemistry are NNs. Typically trained to predict the potential energy, we now routinely see accurate NN potentials rival force field methods. The previously mentioned ANI papers are by far the largest breakthroughs in predicting energies with chemical accuracy. Other NN potentials come from Unke and Meuwly called PhysNet. [247] Another success story is rather than learning the eigenvalues (i.e. energy) one can learn the eigenvectors (ψ). PauliNet achieves just that by using a deep neural network to obtain the converged wavefunction. [118] Simply googling NN and computational chemistry will reveal a plethora of examples. Overall, NNs have shown to be accurate methods for energies. If they work well for predicting energies, they should also work well in predicting properties of atoms. We will use NNs as a starting point in the next chapter to predict chemical shifts.

4.3 Δ-Machine Learning

Ideally, a ML model will be able to reasonably predict the target property (e.g. energy) without too much tinkering. However, in some instances given the size of the data or sensitivity of the target property, the ML model isn't performing as desired. Δ-ML makes this prediction task easier by computing the target value using a less expensive method. Overall, the learning problem then becomes the following:

$$y_{target} = y_{cheap} + \Delta_{ML} \tag{4.9}$$

where Δ_{ML} is the correction learned by the ML model. Δ-ML has been used successfully before by von Lilenfeld and coworkers to correct DFT or semi-empirical methods up to G4MP2 methods on 5k small molecule structures. Here, they demonstrate that one needs significantly fewer data points to reach chemical accuracy, and is more transferable. [192]. Another more recent example is from Margraf, where a semi-empirical method (DFTB) is corrected up to DFT for predicting the crystal structure landscape of molecule XXIII from the latest blind test. [259] These two papers highlight how Δ-ML is a possible route to improve prediction accuracy without necessarily having to redesign a ML model or when data is scarce.

4.4 Gaussian Process Regression

It is quickly worth highlighting that there are other ML models than NNs. Gaussian process regression (GPRs) are another popular technique in computational chemistry and have been very successfully in creating potentials. GPRs are formally exact (if the input descriptor is also exact), but suffer from long inference times with large datasets since the trained model depends on the size of training. See examples of excellent work from von Lilienfeld, Csányi, and Ceriotti. [47, 173] For a full review, see ref. [60].

4.5 Graph Neural Networks

In addition to traditional ML approaches, we began to explore state-of-the-art methods such as graph neural networks (GNNs). Unlike previously defined NN models, GNNs do not have predefined feature descriptors that are fed into the model. The GNN

model learns an appropriate descriptor 'on-the-fly' based on the dataset. In some ways, this is more powerful since the ML model is free from specialist input, and one simply needs to define the type of graph layers to use. However, a thorough discussion in chapter 6 has recommendations for using GNNs for molecular crystal structures. In addition, they excel on large datasets. See ref. [123] on how large these dataset for GNNs can get. Overall, we define the inner workings, and highlight some success as we embark on trying to use this models for solid-state NMR chemical shift prediction.

4.5.1 Mathematical Background of GNNs

This section draws heavily from ref. [55]. A graph is a compact data structure which describes the relations (*edges*) between entities (*nodes*). The information stored in each edge and node are called *embeddings*. Given a graph, G, and fixing an arbitrary order for nodes n, we can visualize the connectivity of the graph through the adjacency matrix ($\boldsymbol{A}$ shown in figure 4.4):

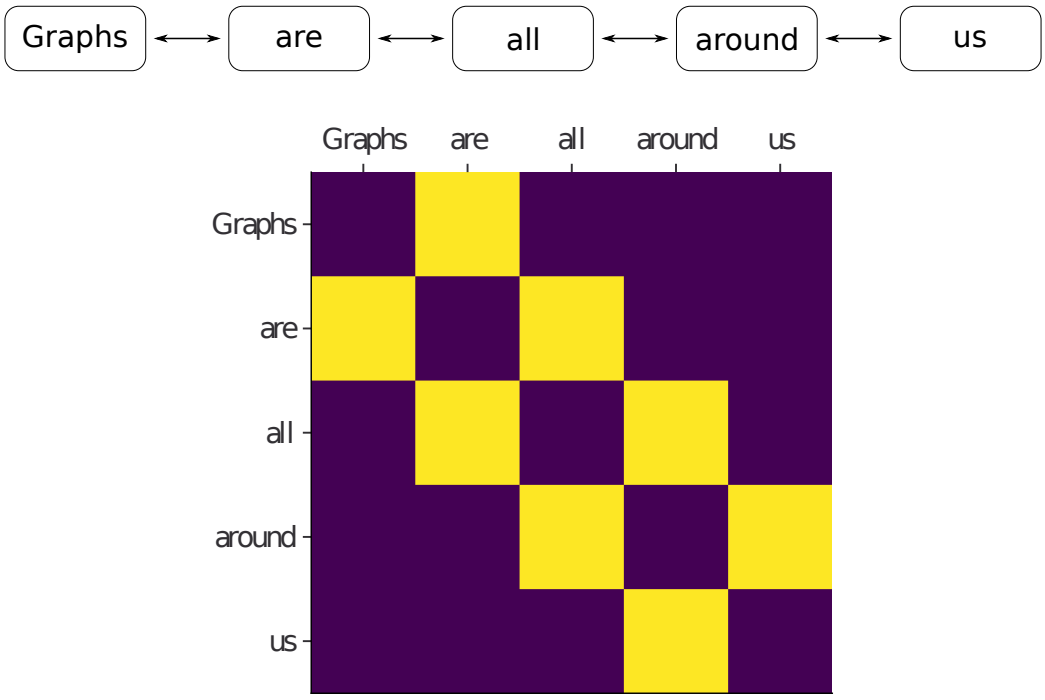


Figure 4.4: Example 0 - 1 adjacency matrix using the sentence "Graphs are all around us" as an example graph. Elements shaded in yellow have values of 1 meaning the row and column pair are indeed neighbors, while values in dark purple are 0 indicating they are not adjacent.

We then define the diagonal degree matrix D of G as:

$$D_\nu = \sum_u A_{\nu u} \tag{4.10}$$

where $A_{\nu u}$ denotes the entry in the row corresponding to ν and column u. The degree of each node represents the number of edges at node ν. The graph Laplacian L is then defined as the square $n \times n$ matrix: $L = D - A$. The graph Laplacian gets its name from the Laplacian operator from calculus, and contains similar information to the adjacency matrix, A. However, L has many interesting properties that can be exploited. We can build polynomials of L of the form:

$$\begin{aligned} p_w(L) &= w_0 I_n + w_1 L + w_2 L^2 + ... + w_d L^d \\ &= \sum_{i=0}^{d} w_i L^i \end{aligned} \tag{4.11}$$

where we can just store the vector of coefficients $w = [w_0, ..., w_d]$.

For simplicity, consider that all the node features of our graph are 1-dimensional. We can stack all the node features x_ν to a get a vector $\mathbf{x}$. Once we have the feature vector, we define the convolution of $\mathbf{x}$ as :

$$\mathbf{x}' = p_w(L)\mathbf{x} \tag{4.12}$$

In the simplest case, consider $w_0 = 1$ and all other coefficents are 0, the convolution $\mathbf{x}'$ is $\mathbf{x}$:

$$\mathbf{x}' = p_w(L)\mathbf{x} = \sum_{i=0}^{d} w_i L^i \mathbf{x} = w_0 I_n \mathbf{x} = \mathbf{x} \tag{4.13}$$

Consider the case were $w_1 = 1$ for node ν:

$$\begin{aligned} x'_\nu &= (L\mathbf{x})_\nu = (L_\nu \mathbf{x}) \\ &= \sum_{u \in G} L_{\nu u} x_u \\ &= \sum_{u \in G} (D_{\nu u} - A_{\nu u}) x_u \\ &= D_\nu x_\nu - \sum_{u \in G} x_u \end{aligned} \tag{4.14}$$

We see that convolution operation of node ν is combined with the node features of its immediate neighbors in the graph in the last line of equation 4.14. These convolutions are the key in GNNs and are seen as the **message-passing** step. Thus, how does the degree d of the Laplacian influence the convolution? This is shown in ref. [99]:

$$\text{dist}_G(\nu, u) > i \Longrightarrow L^i_{\nu u} = 0 \tag{4.15}$$

which states that if the distance of node v and u on the graph G are more than i hops away, then the laplacian of that degree is equal to zero. Let's see apply the convolutions to a general example

$$\begin{aligned} x'_\nu &= (p_w(L)\mathbf{x})_\nu = ((p_w L)_\nu \mathbf{x}) \\ &= \sum_{i=0}^{d} w_i L^i_\nu \mathbf{x} \\ &= \sum_{i=0}^{d} w_i \sum_{u \in G} L^i_{\nu u} x_u \\ &= \sum_{i=0}^{d} w_i \sum_{\substack{u \in G \\ \text{dist}_G(\nu,u) \leq i}} L^i_{\nu u} x_u \end{aligned} \tag{4.16}$$

which shows that convolution of node ν occurs only with nodes u if they are not more than d hops away. The polynomial filters are localized. The degree of localization is completely governed by d.

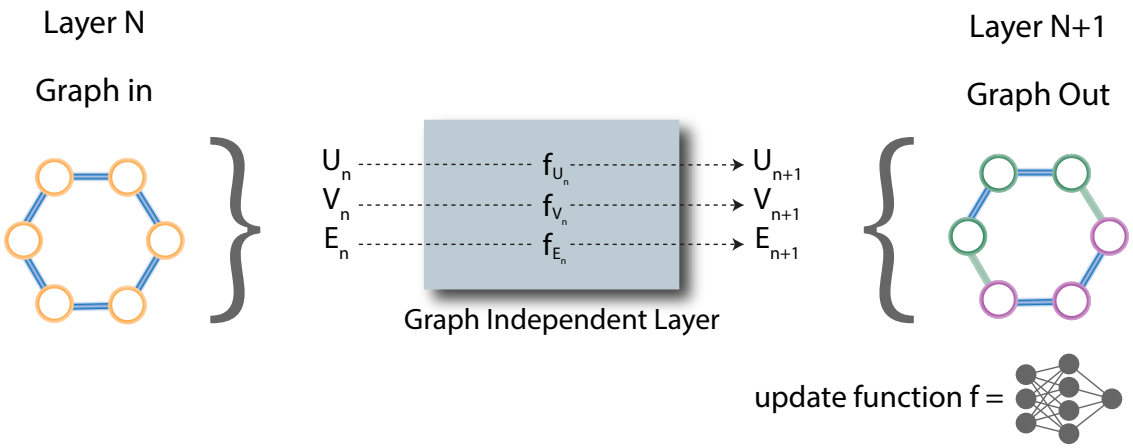


Figure 4.5: Schematic of GNN layers. The input graph embeddings undergo a convolution, and then are fed through a graph independnt layer which updates the embeddings. The update function is a NN and is passed through many graph independent layers to fully "learn" the descriptor.

We now can carry out convolutions on a generic graph G, and copy the embeddings through a NN. This is schematically shown in figure 4.5. In this work, we consider two main classes of graph layers. We initialize our embeddings through the following:

$$h_\nu^{(0)} = x_\nu \tag{4.17}$$

And then update embeddings via the following update schemes. The convolutional form:

$$h_\nu^{(k)} = f^{(k)}\left(C^{(k)} \cdot \frac{\sum_u h_u^{(k-1)}}{\mathcal{N}} + B^{(k)} \cdot h_\nu^{(k-1)}\right) \tag{4.18}$$

The convolutional layer simply takes the average of neighbor inputs (first term in the parenthesis of eq. 4.18) and add the embeddings of the previous step (second term in paranthesis) fed into a NN, $f^{(k)}$. The attentional form has a slightly different formulation:

$$h_\nu^{(k)} = f^{(k)}\left(W^{(k)} \cdot \left[\sum_u \alpha_{\nu u}^{(k-1)} h_u^{(k-1)} + \alpha_{\nu\nu}^{(k-1)} h_\nu^{(k-1)}\right]\right) \tag{4.19}$$

Similar to equation 4.18, equation 4.19 also uses the embeddings of the previous step, but the $\alpha_{vu}^{(k-1)}$ are now *weighted* averages. Ideally, these should work better since there is more control and a learnable parameter. In general, we will use these graph layers in combination with dense neural networks for property prediction in chapter 6 for prediction of NMR chemical shifts in molecular crystals.

4.5.2 GNNs in Computational Chemistry

GNN models are exploding in popularity at the time of this writing. They have found excellent success in molecules, inorganic solids, catalysts, etc. Here, we highlight some notable examples for the interested reader to obtain some sense of what is possible with GNNs.

Molecules

The "pandoras box" of GNNs for quantum chemistry is the landmark paper from Google brain in 2017. [91] Here, they showed best in class accuracy across the board for properties like electronic energy, dipole, enthalpy, etc. for the QM9 dataset (contains 134k molecules consisting of various combinations of CNOF). [191] Improvements to the MPNN scheme came from Tkatchenko and Müller which used continuous convolutions, rather than just one. [208] Lastly, these previous models only used 2-body information (pairwise distances). In principle this should work, but depending on the accuracy and system, higher n-body information is needed. The directional message-passing papers from Klicpera are an excellent starting point for understanding how one can incorporate 3-body information. [130]

Crystals

By far, the largest test have been on inorganic solids rather than molecular crystals. The crystal graph convolutional neural network paper is an excellent starting point for understanding convolutions in large systems. [264]. Other examples include using the previously defined attention graph layer scheme. [205]. There are also attempts to make these models even better through regularization and modifications to stack more layers. [92, 175] One can also use the methods for molecules on crystalline systems if one properly treats the periodic boundary conditions. See the Open-Catalys-Project github repo as an example of how one can transform molecular GNN's for pbc (`https://github.com/Open-Catalyst-Project/ocp`). In addition, there are now software solutions to benchmark the most popular GNN models against each other. [86]

4.6 Conclusion

In the next chapter, we focus on using NNs to rapidly predict chemical shieldings for small molecules. We then augment the approach to include an inexpensive quantum mechanical description in addition to the NN, similar to Δ-ML described previously. Δ-ML shows significant transferability to systems never seen before in its training, and is 2-3 orders of magnitude faster than running the target (i.e. expensive) chemical shielding calculation. Overall, we show that NN ensembles are the key to reducing errors, and validating uncertainties.

Chapter 5

Predicting Density Functional Theory-Quality Nuclear Magnetic Resonance Chemical Shifts via Δ-Machine Learning

5.1 Introduction

Nuclear magnetic resonance (NMR) chemical shifts are among the most useful spectroscopic observables in chemistry. They can be used to characterize molecular species, perform quantitative analysis, and monitor molecular dynamics. Given the widespread impact of NMR spectroscopy, there has been a heavy emphasis on NMR chemical shift prediction through first-principles and density functional theory (DFT)[23, 8, 148, 40, 262,

261, 13, 151, 134]. Chemical shift predictions can help assign peaks in an experimental NMR spectrum, refine structures, or even discriminate among multiple plausible structures. [27, 20] NMR crystallography, the combination of solid-state NMR spectroscopy, x-ray diffraction, and chemical shift prediction, has proven to be a potent combination in resolving Ångstrom-resolution crystal structures with applications towards molecular crystals, materials, and biomolecules.[75, 101, 169, 156, 67, 8]

Unfortunately, the computational cost associated with first-principles DFT chemical shift predictions can be significant. This in turn has spawned considerable interest in data-driven empirical and machine learning (ML) models that can be evaluated several orders of magnitude faster. Such models have been used in biological systems extensively[265, 3, 170, 162, 100, 212, 214, 215, 132, 46, 154, 48, 80] and in other more specialized systems such as acrylonitrile copolymers[129] or amorphous silicon oxides.[53] In all of these examples, the machine learning problem is facilitated by needing to learn only a relatively narrow subset of chemical space (e.g. proteins are composed of only 20 unique amino acids).

There have also been recent efforts to build more general ML models that can predict DFT-quality chemical shieldings for any organic molecule. In 2015, Rupp et al [201] built ML models based on the Coulomb matrix descriptor and kernel ridge regression that predict DFT chemical shieldings in organic molcules with root-mean-square (rms) accuracy of 0.42 ppm for ^{1}H and 5.8 ppm for ^{13}C. More recently, the IMPRESSION model based on kernel ridge regression demonstrated improved rms errors of 0.35 ppm for ^{1}H and 3.9 ppm for ^{13}C.[88] In 2018, Paruzzo et al [177] developed an ML model for solid-state organic

molecule chemical shieldings based on Gaussian process regression and the smooth overlap of atomic positions (SOAP) [15] kernel representation of the local atomic environment. Training on gauge-including projector augmented wave (GIPAW) DFT chemical shieldings computed for large numbers of organic molecular crystal structures with the PBE functional, they developed an ML model capable of predicting those PBE shieldings in organic crystals with rms errors of 0.49 ppm for ^{1}H, 4.3 ppm for ^{13}C, 13.3 ppm for ^{15}N, and 17.7 ppm for ^{17}O. With the exception of hydrogen, these errors in the ML shielding predictions relative to DFT are 2–3 times larger than what one would expect for the target GIPAW PBE calculations relative to experiment: 0.33–0.43 ppm for ^{1}H, 1.9–2.2 ppm for ^{13}C, 5.4 ppm for ^{15}N, and 7.2 ppm for ^{17}O.[203, 122, 108] Nevertheless, they showed that the ML chemical shift predictions could aid discrimination between candidate structures in the context of NMR crystallography.[177] Liu et al [147] subsequently developed their multi-resolution 3D-DenseNet convolutional neural network architecture which predicts chemical shifts based on representations of the electron density around each atom in the system. This approach led to rms errors in the chemical shieldings that were up to 24% smaller compared to those in ref [177]: 0.37 for ^{1}H, 3.3 ppm for ^{13}C, 10.2 ppm for ^{15}N, and 15.3 ppm for ^{17}O.

These recent successes emphasize how the highly local nature of the chemical shielding tensor makes it amenable to machine-learning based on local geometric descriptors that capture the chemical environment within several Ångstroms from the atom of interest. At the same time, ample evidence demonstrates that chemical shieldings can be influenced by surrounding atoms lying 5–8 Å away,[126, 277, 110] outside the range of local atomic environment descriptors typically used in present-day ML models. Despite the excellent

progress in ML chemical shielding prediction discussed above, the errors in current state-of-the-art ML models relative to first-principles DFT remain substantially larger than the errors between DFT and experiment. The errors introduced by the ML model mimicking DFT would ideally be considerably smaller than the errors inherent in DFT itself.

The present study improves the performance of the ML chemical shielding prediction and incorporates longer-range interactions via Δ-ML.[192, 71, 14, 211, 143, 256, 239, 273] Specifically, we perform an inexpensive, low-accuracy calculation to obtain an initial approximate isotropic chemical shielding σ_{cheap} and utilize a trained neural network (NN) to correct it (Δ_{ML}) up to the accuracy of a much more demanding, higher-accuracy "target" chemical shielding prediction, σ_{target}:

$$\sigma_{target} = \sigma_{cheap} + \Delta_{ML} \tag{5.1}$$

The Δ-ML approach improves the accuracy of the chemical shielding prediction in two ways. First, by capturing some of the details of how a given atom's chemical shielding depends on its specific chemical environment, the Δ-ML approach simplifies the learning problem to that of learning only the residual correction Δ_{ML}, which hopefully has a smoother functional form. Second, the inexpensive baseline shielding calculation directly incorporates long-range quantum mechanical interactions into the final shielding. This contrasts other models[88] which include large molecules in the ML training sets to capture those effects. The results presented below will demonstrate how a model trained on small-molecule chemical shieldings and with a local atomic environment descriptor exhibits improved transferability to larger molecules when Δ-ML is employed.

The target chemical shieldings here are obtained at the PBE0/6-311+G(2d,p) level of theory. In molecular crystal benchmarks against experiment, the hybrid PBE0 functional and this basis set perform as well as or better than the PBE GIPAW results cited above, with rms errors of 0.33 ppm for ^{1}H, 1.44 ppm for ^{13}C, 3.86 ppm for ^{15}N, and 7.47 ppm for ^{17}O.[105, 70] We then investigate several potential models for the "cheap" chemical shieldings, including the local density approximation functional SVWN,[218, 253] the generalized gradient approximation (GGA) functional PBE,[182] or the hybrid functional PBE0.[2] These baseline shieldings will be computed in the minimal STO-3G basis set or the small double-zeta 6-31G basis set (without polarization functions), neither of which would typically be considered viable for standalone chemical shielding predictions. Nevertheless, this work will demonstrate how training a NN to correct such low-cost shieldings can lead to predictions that mimic the target level of theory with precision that is superior to the experimental accuracy of the target functional. While it may seem surprising that such small basis sets would be useful in this context, previous work[110, 109] using locally dense basis sets[43, 45, 44] has demonstrated that even a simple basis like 6-31G can effectively capture longer-range contributions to the chemical shielding.

In the end, we demonstrate that while reasonable chemical shielding predictions can be obtained with Δ-ML corrections to any of these inexpensive functional and basis set combinations, the best results are obtained for the Δ-ML model based on PBE0/6-31G. For the gas-phase molecule testing set here containing thousands of molecules with up to 17 heavy atoms, this Δ-ML model predicts the target shieldings with rms errors of 0.11 ppm for ^{1}H, 0.70 ppm for ^{13}C, 1.69 ppm for ^{15}N, and 2.47 ppm ^{17}O. Though the test

systems here differ from those in earlier studies,[201, 88, 177, 147] these errors are several times smaller those obtained with the previously reported pure ML models described above. More importantly, these errors are only a fraction of the aforementioned errors typically found for DFT versus experiment. We demonstrate this point further by investigating the performance of this Δ-ML model for predicting solution-phase experimental chemical shifts for a set of nine pharmaceutical molecules in either DMSO or $CDCl_3$. Because it involves a first-principles DFT calculation, the computational cost of the Δ-ML approach is considerably higher than that of pure ML approaches. On the other hand, performing the DFT calculations at the inexpensive level of theory is still 1–2 orders of magnitude cheaper than doing so at the target level of theory. Finally, we show how the standard deviation among the predictions obtained from an ensemble of NN models can be related to the uncertainty in the predicted chemical shieldings. Overall, the excellent performance of the models here highlights how Δ-ML-based chemical shielding models can potentially seamlessly replace much more expensive DFT calculations without sacrificing quantum mechanical accuracy.

5.2 Computational Details

5.2.1 ML Training and Testing Data

The training, validation, and testing data were aggregated from various sources. For the training and validation data, a set of all possible small molecules with up to eight heavy (non-hydrogen) atoms and containing only the elements C, N, and O was obtained from the ANI-1 data set[222], which uses the GDB11 database[78] as a starting point. The

minimum-energy geometry for each molecule at the ωB97X/6-31G(d) level of theory was extracted from the set. After removing six molecules with improper numbers of hydrogen atoms or unlikely nuclear contacts, 57,456 molecules remained in the trainingvalidation set.

For the testing set, 3780 molecules with 12–17 heavy atoms and containing only the elements H, C, N, and O were drawn randomly from the GDB17 database.[200] The specific molecules are listed in the appendix Section C.3. These molecules were initially obtained in SMILES notation. To convert them to three-dimensional coordinates, RDKit (`www.rdkit.org`) was used to saturate the molecules with hydrogen atoms and perform preliminary MMFF94[98] force field geometry optimizations. Finally, the molecular geometries were optimized in Gaussian 09[83] at the same ωB97X/6-31G(d) level of theory as the training molecule set.

NMR chemical shieldings were then computed for every atom in every molecule in the training and testing sets. In total, the ∼60,000 molecules contain over one million chemical shieldings (Table 5.1), with a little more than half being ^{1}H, about a third being ^{13}C, 9% being ^{15}N, and 6% being ^{17}O. The target shieldings were computed with the hybrid PBE0 density functional and the 6-311+G(2d,p) basis set, which has performed well in previous NMR chemical shift benchmarking studies.[108] The inexpensive chemical shielding models used in the Δ-ML approach will be discussed in Section 5.2.2. The NMR chemical shielding calculations used in the machine learning training and testing were performed in Gaussian 09 using the default "FineGrid," a pruned 75 radial and 302 Lebedev angular point integration grid. Sample input files are provided in the appendix Section C.2. Computational timings for the ML workflow are reported using ORCA v4.2.1[171, 172] instead

of Gaussian due to software licensing restrictions. The ORCA calculations employed density fitting with the chain-of-spheres approximation (RIJCOSX) for the ωB97X/6-31G(d) geometry optimization. The density-fitted NMR calculations employed Coulomb and/or exchange density fitting (RIJ or RIJK) and the appropriate def2/J or def2/JK auxiliary basis sets[257, 258] for the pure and hybrid functionals. NMR calculations without density fitting utilized analytic integrals.

Table 5.1: Summary of the numbers of species and atoms of each type in the training/validation and testing data sets. N refers to the number of heavy (non-hydrogen) atoms.

	Training/Validation GDB11 (N=1–8)	Testing GDB17 (N=12–17)
Molecules	57,456	3,780
C atoms	298,081	44,146
H atoms	496,275	65,524
N atoms	89,010	9,961
O atoms	60,403	7,549

5.2.2 Feature Representation and Neural Network Architecture

Geometric information is encoded in the NN input descriptor via the atomic environment vector (AEV)[221]. The AEV builds on the Behler and Parrinello atomic symmetry functions[17], and it is one of several atomic descriptors[16, 202, 15] that effectively describes the local chemical environment of an atom in an orientationally invariant manner. The AEV was chosen as the descriptor based on its success in predicting energies[221] and charges [216] of small molecules. Other studies using the AEV as the descriptor have predicted the 2-body energy term of the many-body expansion for molecular crystals[159].

The AEV has been described in detail previously,[221] so we only review the main features briefly. The 384-element AEV for a given atom i used here consists of 64 radial and 320 angular elements. The radial AEV elements $G^R_{a,s}$ are given as,

$$G^R_{a,s} = \sum_{j\neq i}^{\text{all atoms}} e^{\eta(R_{ij}-R_s)^2} f_C(R_{ij}) \tag{5.2}$$

where η equals 16 (see ref [221]) and j runs over all other atoms in the system. The species used here only include H, C, N, and O atoms. The 64 radial AEV elements are indexed by a and s. The first index a corresponds to the four possible atom types that atom i might interact with, while the second index s denotes the 16 "bins" corresponding to different fixed distances R_s from atom i. The set of distances R_s are given in Å by $R_s = 0.9 + n\frac{a_0}{2}$, where a_0 is the Bohr radius (0.529177 Å) and n ranges 0–15.[221] The use of Gaussian functions in Eq 5.2 means that an atom j lying distance R_{ij} from the central atom i will contribute significantly when R_{ij} is similar to R_s. The local cutoff function f_C effectively decreases the weights for more distant atoms, and it is given by,

$$f_C(R_{ij}) = \begin{cases} 0.5 \times \left(1 + cos\left(\frac{\pi R_{ij}}{R_C}\right)\right) & \text{for } R_{ij} \leq R_C \\ 0.0 & \text{for } R_{ij} > R_C \end{cases} \tag{5.3}$$

Atoms j lying further away than $R_C = 5.2$ Å from central atom i do not contribute to the AEV.

The 320 angular AEV elements are similarly defined for atoms j and k surrounding central atom i as,

$$G^{A_{mod}}_{a,b,m,n} = 2^{1-\zeta} \sum_{j,k\neq i}^{\text{all atoms}} (1 + cos(\theta_{ijk} - \theta_m))^\zeta \exp\left[-\eta\left(\frac{R_{ij} + R_{ik}}{2} - R_n\right)^2\right] f_C(R_{ij}) f_C(R_{ik}) \tag{5.4}$$

In this case, there are ten possible pairs of atom types a and b, the central atom may form an angle with (CC, CH, CN, CO, etc.). The "bins" are now defined in terms of radial (R_n) and angular (θ_m) values to probe specific regions of the angular environment. The following values are used: R_n = (0.90, 1.55, 2.20, 2.85) Å and $\theta_m = (\frac{\pi}{16}, \frac{3\pi}{16}, \frac{5\pi}{16}, \frac{7\pi}{16}, \frac{9\pi}{16}, \frac{11\pi}{16}, \frac{13\pi}{16}, \frac{15\pi}{16})$. The combination of ten atom combinations a and b, four radial bins n, and eight angular bins m leads to 320 total elements in the angular portion of the AEV. As in the original AEV work, a radial cutoff $R_C = 3.5$ Å is used for equation 5.4. [227, 221] Given this set of bins, the normalization constant $\zeta = 32$.

The AEV is computed for each atom in a molecule, summing over all radial atom pairs (Eq 5.2) and angular triplets (Eq 5.4). It provides a fingerprint for chemical environment that is fed into the NN for the purpose of predicting the isotropic chemical shielding or Δ-ML shielding correction. After generating the AEV and isotropic shieldings for each atom, a `pandas`[176, 260] dataframe file, separated by atom type, was created for the training and testing sets.

Separate neural network (NN) models were then trained to predict chemical shieldings for each of the four nuclei considered here: ^{1}H, ^{13}C, ^{15}N, and ^{17}O. The NNs were constructed using Tensorflow 2.0 [1] and the `keras` backend version 2.2.4 (`www.keras.io`). The NN architecture is depicted in Figure 5.1. The model used for training consists of 1 input layer containing 384 neurons (equivalent to the size of the AEV descriptor), 3 hidden layers of 128 neurons each, and 1 output layer consisting of 1 neuron. Each hidden layer neuron used the rectified linear unit (ReLU) activation function. The mean-squared-error loss function was used for all trainings. Initial testing found similar performance between stan-

dardized and non-standardized isotropic shielding data for NN training RMSEs. Therefore, the shielding data was not standardized in the final models for simplicity.

The NNs for each atom type were trained independently using the N=1-8 small-molecule data set for training and validation. Specifically, for each atom type, a 10-fold cross-validation scheme was employed in which the training data was divided into ten bins with approximately equal numbers of data points each. For each of the ten cross-fold fits, data from one bin was excluded from the fitting process. 10% of the remaining training data was randomly held as validation data, and NN fitting was performed against the rest. The validation data was employed to monitor for early stopping to reduce the risk of over-training. Specifically, the fits were stopped once errors on the validation data set started increasing and did not drop below their previous best value over 10 subsequent epochs. See the appendix Section C.8 for training and validation errors. The NN weights from each cross-fold fit were saved to become a member of the final NN ensemble. As shown in Figure 5.1, the final ML model prediction is computed as the mean value of the predictions from each of the 10 cross-fold fits. The standard deviation of those ensemble member predictions is used to estimate the uncertainties. Ensemble models have been shown to have better predictive performance than any individual NN model.[225, 221, 227] Once the cross-validation training was complete, the final ensemble model was tested on molecules sampled randomly from GDB17. Table 5.1 summarizes the distribution of atom types in the training/validation and testing sets.

The present work focuses primarily on Δ-ML models, though NN models that employ the AEV alone, without any Δ-ML contribution, are also trained as a control. The

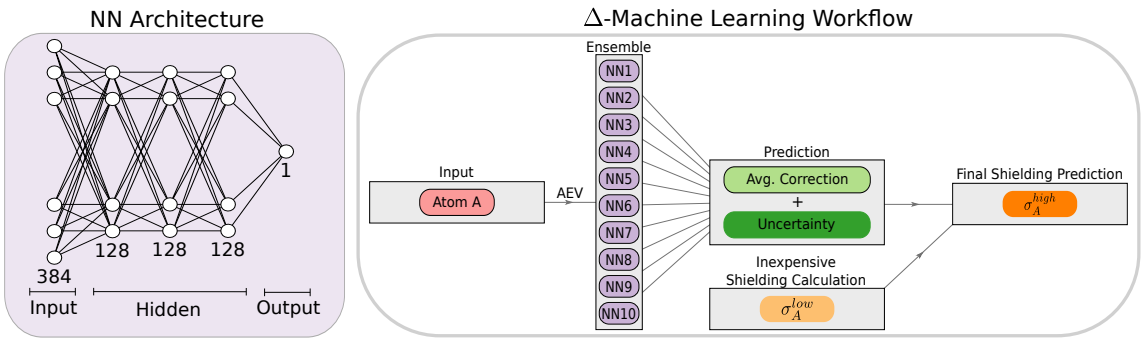


Figure 5.1: The basic NN architecture here for a given atom type employs a 384-element AEV input descriptor for the atom of interest, three hidden layers with 128 neurons each, and a final output layer consisting of a single neuron. The NN output for the Δ-ML models represents the correction to the inexpensive shielding value. The final prediction is computed as the mean value from an ensemble of 10 cross-fold NN fits, and the uncertainty in the prediction is estimated from the standard deviation among the ensemble member predictions.

Δ-ML NNs were fitted to reproduce the difference between the low-level chemical shielding and the target PBE0/6-311+G(2d,p) shielding, as shown in equation 5.1. Six different possible inexpensive chemical shielding model chemistries are considered: the SVWN,[218, 253] PBE,[182] or PBE0[2] density functionals in either the STO-3G or 6-31G basis sets. These models were chosen to explore the interplay between cost and accuracy in the Δ-ML approach. Generally speaking, more accurate baseline shielding models will be easier to correct with the Δ-ML approach, but the greater computational expense will also reduce the efficiency advantages of the Δ-ML calculation relative to conventional larger basis DFT calculations.

A hybrid functional like PBE0 generally predicts experimental chemical shifts with root-mean square errors that are up to 30% smaller than for a GGA like PBE,[105] albeit at additional computational expense. The SVWN local density approximation requires even less computational effort than a GGA, but it will also likely provide worse accuracy. The

minimal STO-3G basis set is too small to make useful chemical shielding predictions on its own. The more flexible double-zeta 6-31G basis set will improve representation of the electron density somewhat, but it still lacks polarization functions. Without polarization functions, 6-31G might perform tolerably for simple hydrocarbons, but that performance is expected to degrade as more polar functional groups are added or when nuclei such as ^{15}N and ^{17}O are considered. On the other hand, omitting polarization functions from the 6-31G basis set ensures the low-level shielding calculations remain fast. The 6-31G basis set places only nine basis functions on a carbon atom, compared to five in STO-3G, 15 in 6-31G(d), and 27 in the target 6-311+G(2d,p) basis.

Although the small-basis DFT models are not expected to be accurate on their own, they should be useful in the Δ-ML context for capturing the long-range contributions missing from the AEV. For example, the widely used and successful locally dense basis set approach[43, 45, 44] in chemical shift prediction employs large basis sets on the atoms of interest, while smaller basis sets are used on more distant atoms. In fact, previous work has shown that the 6-31G basis can describe long-range contributions to chemical shieldings well, despite the lack of polarization functions.[109, 110]

Finally, a hyperparameter search was conducted to validate the hyperparameter choices described above. This search was performed using a Bayesian search algorithm with Gaussian processes, as implemented in the `scikit-optimize` package (`scikit-optimize.github.io`). Bayesian optimization provides an alternative to the popular grid search method of hyperparameter optimization when the time to train the model prohibits the use of an extensive grid search. Hyperparameter searches were performed for the best-

performing PBE0/6-31G + Δ-ML model. For each optimization, 1, 2, 3, 4, or 5 layers were used, while the number of neurons per layer varied between either (32, 128) or (128, 500). These two sets represent networks with relatively few neurons per layer or a larger number of neurons per layer, respectively. The performance of the neural networks on the training and testing sets varied by only a few hundredths of a ppm across all the hyperparameter searches. Thus, the model architecture used here (Figure 5.1) appears to be well-converged with respect to the hyperparameter choices. See the appendix Section C.6 for more details.

5.2.3 Experimental Structures and Referencing

The machine-learning model is trained to predict either the PBE0/6-311+G(2d,p) chemical shieldings directly (pure AEV model) or the Δ-ML shielding correction to the inexpensive shielding values. To compare against experimentally measured chemical shifts, predicted shieldings σ_i must be referenced appropriately. Multiple referencing strategies exist;[148, 23] here we adopt the linear regression approach in which the final chemical shift δ_i is given as,

$$\delta_i = a\sigma_i + b \tag{5.5}$$

where a and b are empirical parameters fitted via linear regression between a set of predicted chemical shieldings and known experimental chemical shifts. Ideally, the slope a would equal -1 and the intercept b would correspond to the shielding of the reference compound (e.g. tetramethylsilane for ^{13}C). In practice, the parameters deviate from these values due to solvent effects and other inherent approximations present in the shielding prediction models. The fitted parameters for Eq 5.5 are unique to a specific computational model used

to generate the chemical shieldings, and new linear regression parameters are generated for each different nuclide, level of theory, basis set, ML model, and solvent.

Here, linear regression parameters were fitted for two common solvents, DMSO and $CDCl_3$, using separate data sets of experimental chemical shifts for each. The $CDCl_3$ experimental chemical shift regression parameters were generated using the data set of molecules and experimental shifts provided by the CHESIRE NMR chemical shift repository[148]. Molecules including atom types other than H, C, N, or O were removed from the regression data set, which left the 57 structures with 163 experimental ^{13}C chemical shifts listed in Section S4. For DMSO, 23 species with 45 experimental ^{13}C shifts were curated from refs [94] and [85]. Details of these experimental data sets are provided in the appendix Section C.4.

After fitting a chemical shift referencing line for each chemical shielding prediction model, the regression parameters were used to predict experimental chemical shifts for nine relatively rigid drug molecules. Rigidity should reduce the chemical shift errors introduced by neglecting conformational sampling. Initial geometries for the drug molecules were taken from their crystal structures, as extracted from the Cambridge Structure Database (reference codes given in parentheses): acetaminophen (HXACAN14), aspirin (ACSALA14), benzoic acid (BENZAC02), cortisone acetate (ACPRET), estrone (ESTRON11), mefenamic acid (XYANAC07), nalidixic acid (NALIDX01), nitrofurantoin (LABJON), and trimethoprim (AMXBPM12). Experimental ^{13}C chemical shifts measured in either DMSO or $CDCl_3$ were obtained from the National Institute of Advanced Industrial Science and Technology website (`https://sdbs.db.aist.go.jp`) and are listed in Section S5. Geometry optimiza-

tion and chemical shielding calculations for all species was performed at the same levels of theory as described for the data sets in Section 5.2.1.

5.3 Results and Discussion

Table 5.2: Summary of RMSE (in ppm) per Δ-ML model separated by atom type for the small-molecule GDB11 set used to train the models and for the set of larger molecules from GDB17 used to test the final models. For brevity, only selected density functional/basis set combinations are shown here for ^{1}H, ^{15}N, and ^{17}O shieldings. [a] Control mapping the low-level shieldings onto the target ones via simple linear regression. [b] Control using only the AEV descriptor to predict the target shieldings. See the appendix Section C.7 for a full comparison.

Model	Training: N=1–8		Testing: GDB17	
	No Δ-ML[a]	w/ Δ-ML	No Δ-ML[a]	w/ Δ-ML
^{13}C Chemical Shieldings				
AEV only (No Δ-ML)[b]		2.15		4.74
SVWN/STO-3G	9.97	1.34	8.54	2.44
PBE/STO-3G	9.62	1.33	8.20	2.51
PBE0/STO-3G	9.00	1.39	7.18	2.49
SVWN/6-31G	2.99	0.52	3.31	0.93
PBE/6-31G	2.75	0.45	3.01	0.82
PBE0/6-31G	1.77	0.38	1.54	0.70
^{1}H Chemical Shieldings				
AEV only (No Δ-ML)[b]		0.225		0.360
PBE0/STO-3G	0.651	0.110	0.675	0.214
PBE0/6-31G	0.247	0.060	0.23	0.110
^{15}N Chemical Shieldings				
AEV only (No Δ-ML)[b]		4.85		13.86
PBE0/STO-3G	21.65	3.14	20.78	5.79
PBE0/6-31G	5.63	0.84	5.40	1.69
^{17}O Chemical Shieldings				
AEV only (No Δ-ML)[b]		8.09		18.22
PBE0/STO-3G	31.95	4.50	30.01	9.06
PBE0/6-31G	7.1	1.39	6.68	2.47

5.3.1 Δ-ML Performance for ^{13}C Shielding

We begin by examining how the choice of the inexpensive chemical shielding calculation model impacts the performance of the Δ-ML model for predicting ^{13}C chemical shieldings. The insights gained for ^{13}C chemical shieldings prove transferable to the other

three nuclides discussed later. All results presented here represent the mean value predictions obtained from the 10-fold cross-validated NN trainings (Figure 5.1). Table 5.2 summarizes the resulting root-mean-square error (RMSE) for different model combinations on the small-molecule (N=1–8 heavy atoms) training data and for the larger molecules from GDB17 that were exclusively used for testing the final models.

First, we examine how well a NN based solely on the AEV performs for predicting the full ^{13}C chemical shieldings, without any lower-level shielding calculation to correct via Δ-ML. The AEV-only model performs fairly well for the training set, with RMSE of 2.2 ppm, though this performance deteriorates to 4.7 ppm on the GDB17 testing set of larger molecules. For comparison, previous ML studies reported ^{13}C chemical shielding RMSE of 3.3–4.9 ppm in small molecules or molecular crystals.[201, 88, 177, 147]. So while the AEV itself provides a reasonable starting point, achieving quantitative chemical shielding prediction beyond what has been shown previously clearly requires a better ML model. The large generalization gap between training and testing data is indicative of over-fitting to the training data and bodes poorly for how the AEV-only model will perform on unseen data.

The relatively poor transferability of the AEV NN model to larger molecules likely reflects the local nature of the AEV. The local chemical environment described by the AEV dominates the physics governing the chemical shielding, but electrostatics/polarization contributions from the longer-range environment that are ignored by the AEV also impact the shieldings. Moreover, training the NNs on all possible molecules with up to eight heavy atoms should provide a representative set of chemical environments within the 5.2 Å AEV radial distance cutoff, but only a relatively small fraction of the atoms in the training set

will exhibit significant shielding contributions from longer-range interactions beyond that cutoff.

The Δ-ML approach approximates those missing longer-range contributions via the inexpensive chemical shielding calculation on the entire molecule. As shown in Table 5.2, the SVWN/STO-3G Δ-ML model training set RMSE of 1.3 ppm already represents substantial improvement over using the AEV descriptor alone. More importantly, the Δ-ML model proves considerably more transferable to the larger molecules in the testing data set, with an RMSE of 2.4 ppm—a generalization gap of only 1.1 ppm. Both the RMSE values and the generalization gap are about half what was obtained from the AEV alone. Similar Δ-ML performance is found for PBE/STO-3G and PBE0/STO-3G, with RMSEs of 1.3–1.4 ppm and ∼2.5 ppm for the training and testing sets, respectively. In other words, increasing the quality of the the density functional has little impact when a minimal basis set is used.

Switching to the larger 6-31G basis set for the Δ-ML models improves performance further. SVWN/6-31G already achieves sub-ppm accuracy in reproducing the target shieldings in the training set (0.52 ppm), and only a modestly worse error of 0.93 ppm on the testing set. Moving up Jacob's ladder of density functionals to GGA and hybrid functionals further reduces the error by about 0.1 ppm per rung. Not only is the PBE0/6-31G the best-performing approach here with training and testing errors of 0.38 ppm and 0.70 ppm respectively, it also exhibits the smallest generalization gap of only 0.32 ppm.

The 0.70 ppm testing set RMSE for the Δ-ML model based on PBE0/6-31G is particularly noteworthy, since it represents only a fraction of the ∼1.2–1.5 ppm RMS

errors expected for ^{13}C chemical shift predictions relative to experiment in the best case scenarios.[105, 70, 248] To our knowledge, the Δ-ML approach here is the first one to predict DFT chemical shieldings with precision that is considerably better than the accuracy of the target DFT approach relative to experiment. Earlier ML models exhibit RMSEs that are up to 2–3 times larger than the accuracy of DFT itself.[147, 177, 201, 88] The trade-off, of course, is that the Δ-ML models require a small-basis DFT chemical shielding calculation, which is considerably more expensive than simply evaluating a neural network (though it is still at least an order of magnitude faster than a first-principles PBE0/6-311+G(2d,p) calculation).

Deeper insight into the performance of the Δ-ML models can be gained by investigating the difficulty of learning the chemical shieldings. Figure 5.2a plots the kernel density estimate (KDE) for the distribution of errors in the PBE0/6-31G chemical shieldings before and after adding the Δ-ML correction as a function of chemical shielding. The highest density of points lies in the ∼100–175 ppm chemical shielding (not chemical shift) range, which correlates to aliphatic carbon environments which are described relatively well with even the simple 6-31G basis set. In contrast, the ∼0–50 ppm shielding region corresponds to functional groups such as carbonyls and aromatics, for which the omission of polarization functions is problematic. Indeed, the PBE0/6-31G shielding errors roughly vary linearly with the target PBE0/6-311+G(2d,p) chemical shieldings. Examining Figure 5.2a, it almost seems as if the learning problem could largely be solved using a simple linear regression scheme instead of a NN.

As a control experiment to assess how much value the Δ-ML correction provides, we mapped the small-basis shieldings onto the larger-basis ones via a linear regression. The resulting RMSEs for all Δ-ML model combinations are listed in the "No Δ-ML" columns of Table 5.2. For PBE0/6-31G, this simple linear regression gives errors of 1.8 and 1.5 ppm for training and testing sets. While those errors are surprisingly small, they are 2–3 times larger than the PBE0/6-31G Δ-ML model gives. Moreover, the performance of the PBE0/6-31G-based Δ-ML model is essentially independent of the chemical shielding—the performance in the aliphatic and carbonyl regions is similar. For the STO-3G model, the difference between the simple linear regression (7.2-10.0 ppm RMSE) and the ML models (∼1.8–2.8 ppm) is even more dramatic. The simpler linear mapping performs even worse than the AEV-only model with no Δ-ML contribution. These results highlight how the NN is learning the nuanced relationship between atomic environment and the isotropic chemical shielding. This effect will be even more dramatic when we consider the performance for ^{15}N and ^{17}O in Section 5.3.2.

In summary, increasing the size of the basis set used in the baseline ^{13}C chemical shielding model from STO-3G to 6-31G has a large impact on the performance of the Δ-ML model. Improving the quality of the density functional has a smaller effect (∼0.1–0.2 ppm) on the ^{13}C shieldings, but it may still be worthwhile given the generally low cost of even a PBE0/6-31G chemical shielding calculation. As will be discussed in Section 5.4.3, the small basis shielding calculation requires only a small fraction of the computatinal time required to optimize the geometry, making the cost differences between functionals a relatively minor factor. Accordingly, the remainder of the paper focuses on the performance

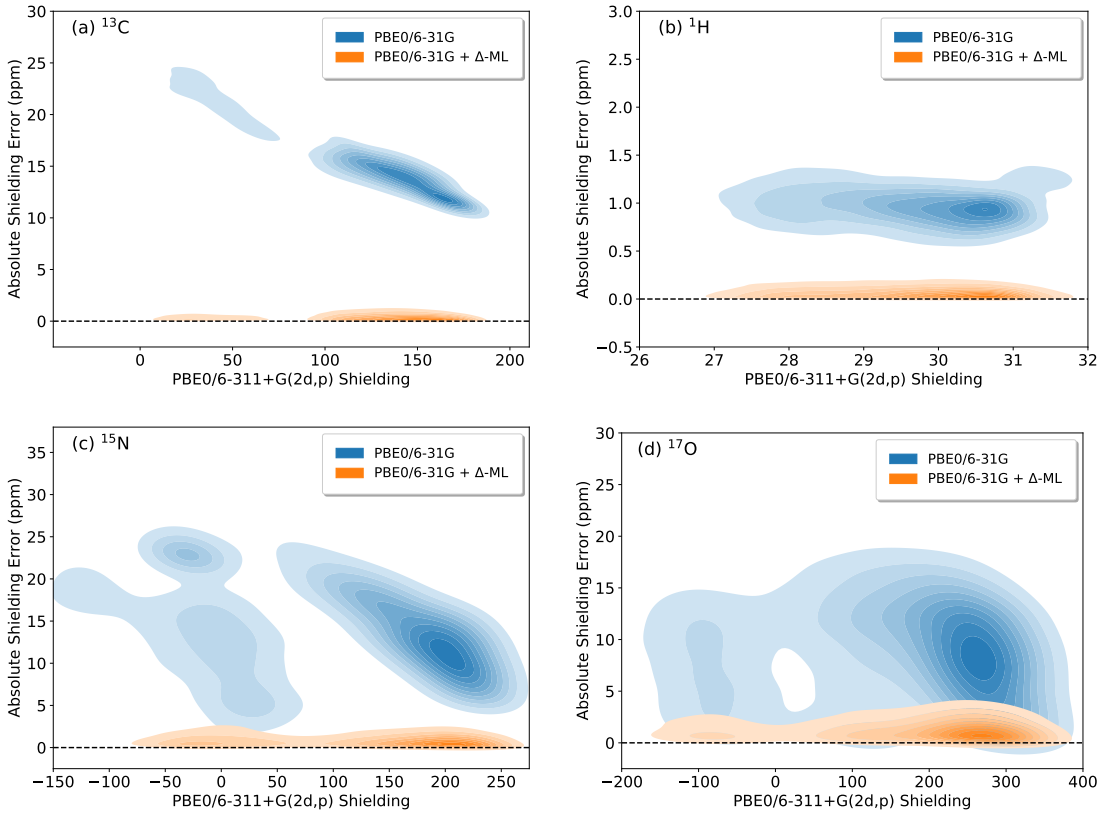


Figure 5.2: Kernel density estimate plots showing the errors of the inexpensive PBE0/6-31G and PBE0/6-31G + Δ-ML model shieldings relative to the target PBE0/6-311+G(2d,p) shieldings versus the target PBE0/6-311+G(2d,p) shieldings for (a) ^{13}C, (b) ^{1}H, (c) ^{15}N and (d)^{17}O. Darker regions indicate a higher density of data points.

of the best-performing PBE0/6-31G Δ-ML model for other atom types and for predicting experimental chemical shifts. However, if the small-basis calculation were to become a significant bottleneck in a particular application, one could opt for a less expensive density functional like PBE with little loss in accuracy.

5.3.2 Δ-ML Performance of ^{1}H, ^{15}N, and ^{17}O

The excellent performance of the PBE0/6-31G Δ-ML model in reproducing the ^{13}C target shieldings is now demonstrated for ^{1}H, ^{15}N, and ^{17}O. Table 5.2 summarizes the key results for these atom types; results for all possible Δ-ML functional and basis set combinations are provided in the appendix Section C.7. Because the experimental chemical shift ranges differ considerably for the different nuclei, the error magnitudes will also vary. Nevertheless, the general trends and relative fidelity of the ML models to the target PBE0/6-311+G(2d,p) chemical shieldings are similar across all four nuclei.

For ^{1}H, the AEV alone performs reasonably once again, with an RMSE of 0.23 ppm for the training set and 0.36 ppm for the testing set. These errors from the pure AEV model are similar to the 0.35–0.49 ppm accuracy obtained from previously published ML models.[147, 177, 88] This error range is also comparable to the expected ∼0.3–0.4 ppm accuracy for large-basis DFT relative to experiment.[203, 105]

The ^{1}H Δ-ML models perform far better than the AEV alone, especially when the 6-31G basis is used. For example, Δ-ML based on PBE0/6-31G reproduces the target shieldings with RMSE of 0.06 ppm and 0.11 ppm for the training and testing sets, respectively. The generalization gap of ∼0.04–0.05 ppm for the 6-31G Δ-ML models is also considerably smaller than what is observed for the STO-3G or AEV-only models. Similar

to ^{13}C, the KDE plot for ^{1}H in Figure 5.2b shows a fairly linear relationship between the target PBE0/6-311+G(2d,p) chemical shieldings and the PBE0/6-31G errors relative to those shieldings. Nevertheless, the Δ-ML model once again performs ∼2–3 times better than a simple linear regression model that attempts to map the small-basis shieldings onto the target ones, emphasizing the value of the NN. Overall, these PBE0/6-31G Δ-ML model errors are small compared to the typical DFT error versus experiment.

Nitrogen and oxygen are more interesting test cases. Several factors potentially make machine learning of the chemical shieldings for these two nuclei more challenging. First, ^{15}N and ^{17}O chemical shieldings are more sensitive to their electrostatic environment compared to ^{1}H and ^{13}C, and their chemical shifts also exhibit broader absolute chemical shift ranges. Typical errors for PBE0 chemical shifts versus experiment in solid state systems are ∼4 ppm for ^{15}N and ∼7–8 ppm for ^{17}O, versus ∼1.2–1.5 ppm for ^{13}C.[105, 70] Second, the molecular data sets used here contain far fewer data samples for ^{15}N and ^{17}O (Table 5.1). As shown in Figures 5.2c–d, these data samples are also less-uniformly distributed across the chemical shielding range. For example, most of the training samples for ^{15}N occur within the 200-215 ppm chemical shielding range (e.g. 1amine functional groups), while the ones for oxygen are concentrated in the 200-350 ppm shielding range (hydroxyl and ether groups). Third, Figures 5.2c–d also emphasize the highly non-linear relationships between the the target shieldings and the errors in the small-basis shieldings, suggesting that the ML correction to the small-basis shieldings will be particularly important. Simple linear fits between the small and large-basis PBE0 shieldings perform poorly, with errors of ∼20–30 ppm (Table 5.2).

Despite these challenges, the ^{15}N and ^{17}O ML model performance follows the same trends as were seen for ^{13}C and ^{1}H. The AEV alone does not perform especially well, but considerable improvements are obtained by the Δ-ML models (Table 5.2). The STO-3G Δ-ML models perform fairly well on the training set, but they generalize poorly to the testing set, especially for oxygen (RMSE 4.50 ppm for training, but 9.06 ppm for testing). Finally, the PBE0/6-31G Δ-ML model performs very well, with small RMSEs overall (e.g. testing set errors of 1.7 ppm for ^{15}N and 2.5 ppm for ^{17}O) and generalization gaps of about 1 ppm between the training and testing sets. Comparison of the performance of the PBE0/6-31G with (orange) and without (blue) the Δ-ML correction in Figures 5.2c–d highlights how effectively the NN learns the correction to the small basis shieldings. These RMSEs for the PBE0/6-31G Δ-ML models are once again only a small fraction of the typical DFT chemical shift errors relative to experiment. Overall, comparing to the training set chemical shielding ranges that span roughly 12 ppm for ^{1}H, 240 ppm for ^{13}C, 560 ppm for ^{15}N, and 850 ppm for ^{17}O, the testing set RMSEs from Table 5.2 amount to fractional errors of only 0.3% for ^{13}C, ^{15}N, and ^{17}O, and 1% for ^{1}H. In other words, despite the variations in RMSE for different nuclei, the ML models are performing similarly well across the nuclei in relative terms.

In summary, Δ-ML NN corrections to inexpensive PBE0/6-31G chemical shieldings can reproduce larger-basis PBE0 shieldings for ^{13}C, ^{1}H, ^{15}N, and ^{17}O with fidelity that is superior to the expected accuracy of DFT versus experiment. The relationship between the small- and larger-basis shieldings varies in complexity depending on the nuclide, but in all cases, the NN learns the correction well. The small generalization gaps between the

small-molecule training set and larger-molecule testing set suggest that the Δ-ML approach is effectively capturing the long-range contributions to the chemical shielding that are absent in the AEV.

5.3.3 Uncertainty Quantification

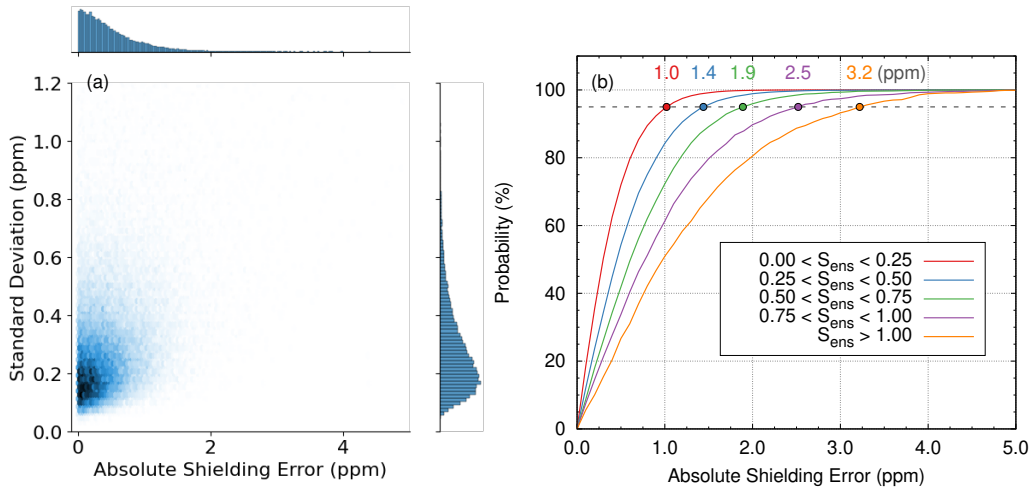


Figure 5.3: (a) 2-dimensional kernel density plot showing the distribution of ^{13}C chemical shielding errors vs the standard deviation S_{ens} in the ensemble prediction for the GDB17 testing data set (44,146 data points) using PBE0/6-31G Δ-ML. Darker shading indicates a higher density of data points. The histograms on the sides of each axis show the distribution of data relative to that axis. (b) Curves showing the probability of having an absolute error less a given amount for different ranges of S_{ens}. For each S_{ens} window, the numbers at the top of the figure indicate the absolute shielding error for which 95% of predictions will fall below.

Estimating the uncertainty associated with a given prediction represents one of the major challenges of machine learning models, but ensemble modeling can help with uncertainty quantification. Here, the final Δ-ML correction is computed as the mean value of the predictions from ten NNs that were trained as part of a 10-fold cross validation. A neural

network is unlikely to perform well if the input descriptors fall too far outside the space spanned by the training data.[71] Disagreement among the members of the ensemble, as measured by the standard deviation of the individual model predictions, can indicate that the prediction lies in a region of space that was ill-constrained by the training data. Therefore, the standard deviation S_{ens} of the ensemble mean can inform about the uncertainty inherent in the prediction.

Here, we examine how the standard deviation S_{ens} relates to the fidelity of the prediction to the target DFT chemical shielding in the GDB17 testing data set. Figure 5.3a plots the distribution of errors in the machine-learning predicted ^{13}C chemical shieldings (relative to the target PBE0/6-311+G(2d,p) values) versus the standard deviation among the ensemble members for the testing set. This figure reveals that the shielding errors generally increase as the standard deviation among the ensemble grows.

For further insight, the data was partitioned into several different windows of S_{ens}. About 50% of the 44,146 shielding predictions have $S_{ens} < 0.25$ ppm, 37% have $0.25 < S_{ens} < 0.5$ ppm, and 9% have $0.5 < S_{ens} < 0.75$ ppm. Only 3% have ensemble standard deviations $0.75 < S_{ens} < 1.00$ ppm, and 2% have $S_{ens} > 1$. Within each window of S_{ens}, the distribution of chemical shielding error data points was integrated to determine the fraction of data points lying within various chosen maximum chemical shielding error thresholds. Figure 5.3b plots the resulting probability curves. These probability curves highlight that smaller standard deviations among the predictions within the ensemble are associated with increased probability of predicting the chemical shielding accurately. For example, 95% of the predictions with $S_{ens} < 0.25$ ppm have a shielding error of 1.0 ppm

or less relative to the target DFT shieldings. If $0.50 < S_{ens} < 0.75$ ppm, the probability of having a larger error in the predicted shielding increases moderately, and 95% of the values fall within 1.9 ppm of the target shielding. Analogous data for the other three nuclei is presented in the appendix Section C.7.5.

Given that the subset of GDB17 molecules were randomly chosen and are chemically distinct from the training molecules, these values should provide reasonable general estimates for the 95% confidence intervals for chemical shielding predictions from the ensemble model, especially for the smaller values of S_{ens} for which many data points are present in this set. As noted previously, RMSEs in DFT-predicted ^{13}C chemical shifts relative to experiment are often found to lie in the ∼1.5–2.5 ppm range, depending on the context (solid state vs solution phase, etc). The present estimates suggest that, with 95% confidence, the uncertainty in the machine learning model prediction will be comparable to or less than the inherent DFT errors when $S_{ens} < 1.0$ ppm. Caution may be warranted in interpreting the confidence intervals if S_{ens} is substantially larger than 1.0 due to the relative sparsity of data in that regime. For example, only 903 of the 44,146 shieldings in the set have $S_{ens} \geq 1$ ppm, and S_{ens} exceeds 2 ppm for only 63 of those.

Finally, it should be emphasized that these uncertainty estimates reflect the uncertainty in the ML prediction of the DFT chemical shielding, rather than the uncertainty in the chemical shifts relative to experiment. Nevertheless, these uncertainty estimates can still be valuable. In a scenario where a predicted chemical shift differs markedly between theory and experiment, for example, a large S_{ens} might indicate limitations of the ML model

training data, while a smaller S_{ens} might point to errors stemming from other factors such as having an incorrect molecular structure or conformation.

5.4 Predicting Experimental Chemical Shifts

Perhaps the most important feature of an ML model for chemical shielding prediction is how well it predicts experimental chemical shifts. In this section, we use the ML models developed above to predict experimental ^{13}C chemical shifts for small molecules in two different solvents, from which chemical shift linear regression referencing models are obtained. After assessing the performance of the ML models for predicting the experimental chemical shifts on these small training sets, we then predict experimental shifts for several fairly rigid pharmaceutical species which were not present in either the ML or chemical shielding regression training sets to give insight to the"real-world" performance of the ML model. Rigidity in these molecules reduces the need for conformational sampling.

Before proceeding, note that DFT chemical shift errors relative to experiment are typically larger than those found for the solid state. For example, B3LYP/6-311+G(2d,p) chemical shift errors in a molecular crystal test set obtained an RMSE of 1.5 ppm,[108] while the same functional and basis set give an RMSE of 3.3 ppm for molecules in solution (when the solvent environment is neglected).[148, **?**] These larger solution-phase errors arise due to factors such as the neglect of solute-solvent interactions and the greater conformational dynamics that can occur in solvent compared to the crystalline state. Accordingly, the errors obtained relative to experiment below for the solution-phase NMR will be larger than the best-case scenario errors that have been discussed earlier in this study.

5.4.1 Chemical Shielding Regression Parameters

To reference the predicted ^{13}C chemical shieldings so that they can be compared against experimental chemical shifts, we perform the commonly-used linear regression referencing approach described in Section 5.2.3. Specifically, a given DFT or ML model is used to make chemical shielding predictions for a set of small molecules with known ^{13}C experimental chemical shifts. Those predicted chemical shieldings are fitted onto the experimental shifts via Eq. 5.5. This referencing process is performed here separately for two common solvents: DMSO and $CDCl_3$. No implicit or explicit solvent environment was including in any of the chemical shielding calculations. Rather, the present work assumes for simplicity that the solvent effects on the gas-phase shieldings can be captured in the shift referencing model. While imperfect, this approach allows a Δ-ML model trained in gas-phase to be transferable to any desired solvent. All of the DMSO and most of the $CDCl_3$ species were present in the ML training set. That is not a significant limitation, because the purpose here is to fit a model for mapping predicted shieldings onto experimentally observable shifts. The performance of this shielding regression model for molecules outside of the ML training will be assessed in Section 5.4.2

The $CDCl_3$ training data set (163 experimental ^{13}C shifts) for the chemical shift referencing is much larger than the DMSO set (44 ^{13}C shifts), which means that the DMSO fitting is therefore probably somewhat less robust. The full set of species and the regression parameters used are listed in the appendix Section C.4.1 and **??**. The neglect of solvent effects in the shielding calculation may also have a larger impact for chemical shifts in DMSO, since that solvent is considerably more polar.

Table 5.3: Root-mean-square errors (ppm) from the linear regressions of predicted shieldings against experimental chemical shifts for the $CDCl_3$ and DMSO training sets.

	$CDCl_3$ (N=163)		DMSO (N=44)	
	Raw	+Δ-ML	Raw	+Δ-ML
AEV		4.82		2.27
PBE0/STO-3G	11.0	3.01	12.0	2.34
PBE0/6-31G	2.00	1.80	2.12	2.35
PBE0/6-311+G(2d,p)	1.82		2.42	

Figure 5.4 plots example regression lines and the error residuals for experimental chemical shifts using the target PBE0/6-311+G(2d,p) DFT calculations or the PBE0/6-31G + Δ-ML ones. The shieldings obtained with these two models are so similar that their corresponding data points and regression lines in the upper panels of Figure 5.4 cannot be clearly distinguished. The fitted slope and intercept parameters differ by no more than 0.1% for $CDCl_3$ and 0.3% for DMSO. In fact, using the regression parameters from the PBE0/6-311+G(2d,p) model for the PBE0/6-31G + Δ-ML shieldings changes the RMSE by less than 0.01 ppm. Comparison of the residuals in the lower panels and the RMS errors for each also show the excellent agreement between the two models. Finally, the fitted slopes deviate from unity by 2.5% or less, which is consistent with the modest level of systematic error one expects from DFT calculations.[148, 108]

Table 5.3 summarizes the experimental errors for generating the regression parameters with and without Δ-ML to evaluate the typical errors relative to experiment. The AEV alone is insufficient to predict experimental shifts reliably, with an RMSE of 4.8 ppm in $CDCl_3$. Surprisingly, it performs much better for the DMSO set, with an RMSE of 2.3

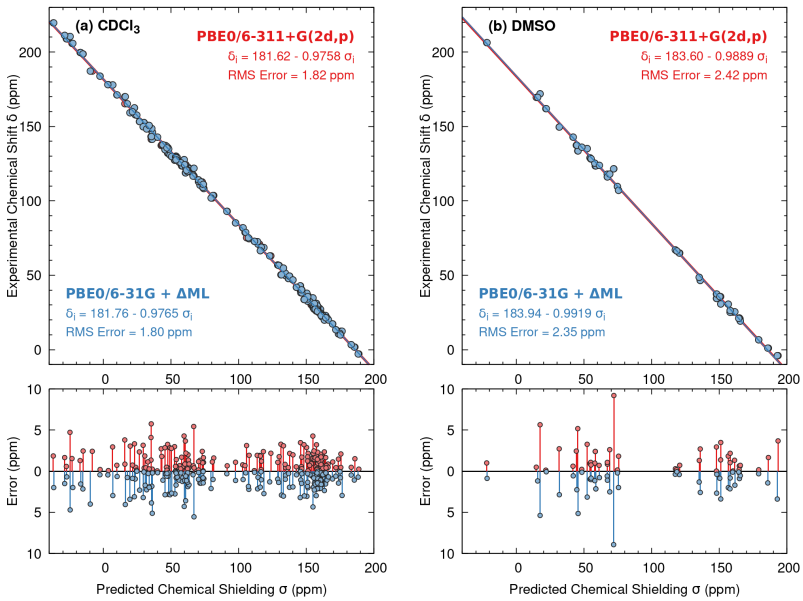


Figure 5.4: Sample linear regressions and absolute values of the residuals for the predicted chemical shieldings versus experimental chemical shifts in the (a) $CDCl_3$ and (b) DMSO small molecule sets using either pure PBE0/6-311+G(2d,p) (red) or PBE0/6-31G + Δ-ML (blue). The data and regression lines for the two models in the upper panels are nearly indistinguishable.

Table 5.4: Root-mean-square errors (ppm) from the linear regressions of the "cheap" predicted shifts versus the PBE0/6-311+G(2d,p) ones after the regression against experiment has been performed.

	$CDCl_3$ (N=163)		DMSO (N=44)	
	Raw	+Δ-ML	Raw	+Δ-ML
PBE0/STO-3G	10.4	2.10	11.4	0.90
PBE0/6-31G	1.63	0.55	1.25	0.29

ppm that is marginally smaller than the error from larger-basis PBE0/6-311+G(2d,p) or either of the Δ-ML models.

Next, note that even without any Δ-ML contribution, the PBE0/6-31G model performs very well. The PBE0/6-31G RMS errors of around 2 ppm relative to experiment are competitive with the target PBE0/6-311+G(2d,p) calculations alone. This highlights an important point when discussing the accuracy of the ML models relative to experiment: the Δ-ML correction makes the small-basis shieldings more faithful to the target level of theory, but that does not necessarily translate to improved agreement with experiment. The static structure, gas-phase PBE0/6-311+G(2d,p) chemical shielding calculations have their own deficiencies which will not be addressed by the Δ-ML correction. To see this, compare the errors versus experiment in Table 5.3 to those against the target shieldings in Table 5.4. The latter table shows that the Δ-ML correction reduces the shielding errors relative to the target PBE0/6-311+G(2d,p) shielding by 3–4 fold, even if this improvement does not reduce the RMSE relative to experiment.

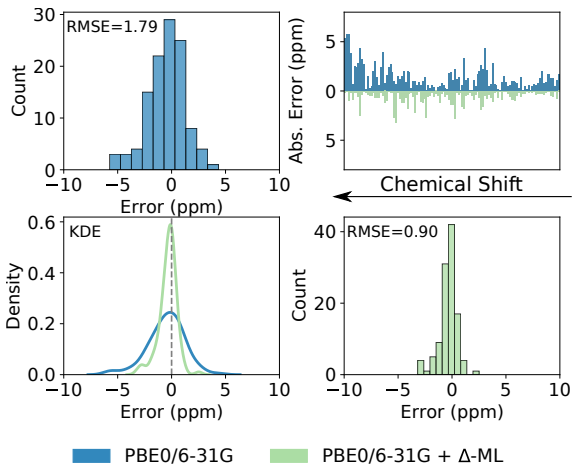


Figure 5.5: Comparison of the PBE0/6-31G chemical shifts with and without Δ-ML correction against the target PBE0/6-311+G(2d,p) ones for the set of drug molecules.

5.4.2 Predicting Experimental Chemical Shifts For Pharmaceutical Molecules

After establishing regression parameters for DMSO and $CDCl_3$, we then predicted 114 experimental ^{13}C chemical shifts for the nine pharmaceutical molecules which were not present in any of the earlier training or testing sets. 78 experimental chemical shifts for acetaminophen, aspirin, estrone, mefanamic acid, nalidixic acid, nitrofurantoin, and trimethoprim were obtained in DMSO, while 36 shifts in $CDCl_3$ come from aspirin (again), benzoic acid, and cortisone acetate. The analysis here combines data from both solvents to establish broad trends. With 10–28 heavy atoms (and up to 59 atoms with hydrogens included), some of these drug molecules are considerably larger than the ones used in the earlier training and testing sets, making them a nice test of the "real-world" applicability of the ML model to solution-phase NMR.

Figure 5.5 compares the performance of PBE0/6-31G and PBE0/6-31G + Δ-ML ^{13}C chemical shifts against the target PBE0/6-311+G(2d,p) shifts. The Δ-ML correction reduces the shielding errors considerably, decreasing the RMSE from 1.79 ppm to 0.90 ppm. The top-right panel of Figure 5.5 highlights how the PBE0/6-31G model exhibits the largest errors for the more polar functional groups which are characterized by larger chemical shifts, as expected due to the omission of polarization functions in the basis set. The Δ-ML correction reduces these largest errors, tightening the error distribution appreciably (bottom-left panel). In other words, the Δ-ML correction is behaving as expected for these drugs, bringing the small-basis PBE0 shifts into better agreement with the large-basis target ones.

Using the uncertainty estimates from the GDB17 set in Figure 5.3b, the difference between the predicted Δ-ML and target shieldings lies within the estimated 95% confidence intervals for 94.7% of the atoms. For the remaining 5.3% (six atoms), the target shielding lies only 0.3 ppm or less outside the predicted confidence interval. In other words, the performance of the Δ-ML model on these drug molecules is consistent with the GDB17 uncertainty estimates described in Section 5.3.3. See the appendix Section C.5.11 for more details.

Next, Figure 5.6 summarizes the performance of PBE0/6-31G, PBE0/6-31G + Δ-ML, and PBE0/6-311+G(2d,p) relative to experiment for these drugs. Examining the diagonal panels, we see the RMSEs of these three models vary from 2.3 to 2.8 ppm. With an RMSE of 2.3 ppm, PBE0/6-311+G(2d,p) exhibits a relatively tight error distribution around zero with only five errors larger than 5 ppm and a maximum error of 7.9 ppm.

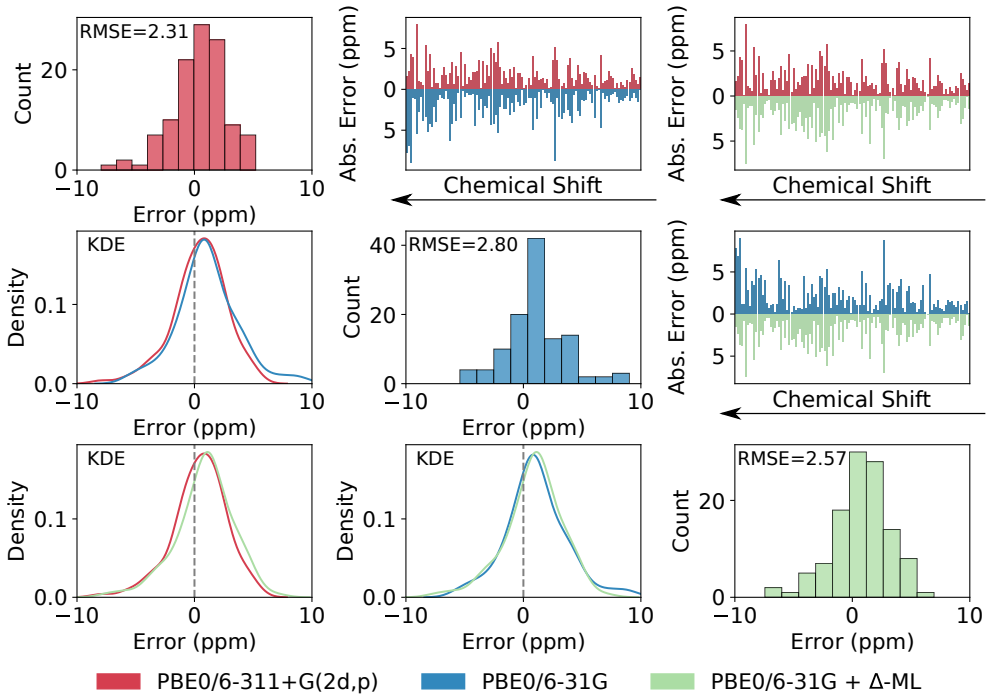


Figure 5.6: Comparison of the drug molecule experimental shift errors among various models. Along the diagonal of this plot shows the ^{13}C error histograms for the target PBE0/6-311+G(2d,p), the baseline PBE0/6-31G, and the Δ-ML -corrected PBE0/6-31G models. The bottom-left 3 panels compare the kernel density representations (KDE) for each model. The upper-right panels compare the error residuals for each model sorted by descending experimental chemical shifts (left to right).

PBE0/6-31G has an RMSE of 2.8 and shows a somewhat similar error distribution, albeit with nine errors exceeding 5 ppm and a maximum error of 9.0 ppm. The Δ-ML correction modestly reduces the PBE0/6-31G RMSE to 2.6 ppm, exhibits seven errors greater than 5 ppm, and decreases the maximum error to 7.4 ppm. In other words, the Δ-ML model results are more similar to the PBE0/6-311+G(2d,p) results.

Sorting the errors by chemical shift (three top-right panels of Figure 5.6) shows some of the same trends as were observed in Figure 5.5. PBE0/6-31G generally exhibits larger errors relative to experiment for larger chemical shifts, which again reflects the in-

adequacies of that basis set for describing carbonyl functional groups and aromatic carbon environments. The errors exhibited by the target PBE0/6-311+G(2d,p) model are somewhat more uniform across the chemical shift range, and the PBE0/6-31G + Δ-ML model mimics this better behavior (top-right panel). For experimental chemical shifts greater than 150 ppm, for example, the PBE0/6-31G model gives an RMSE of 4.0 ppm, compared to 2.8–2.9 ppm for PBE0/6-311+G(2d,p) and the PBE0/6-31G + Δ-ML model.

Overall, the Δ-ML model predicts experimental ^{13}C chemical shifts with accuracy approaching that of the target PBE0/6-311+G(2d,p) model. The 0.9 ppm RMSE errors introduced to the shieldings by the ML model are relatively small and are not strongly correlated with the DFT errors versus experiment, such that the ML model increases the overall RMSE versus experiment by a mere 0.3 ppm. It is surprising how well the baseline PBE0/6-31G chemical shifts perform relative to experiment, even without any ML contribution. The evidence presented in Table 5.4 and Figure 5.5 highlight how much the Δ-ML correction improves the low-cost shieldings relative to the target one. Accordingly, the good performance of PBE0/6-31G likely reflects some fortuitous error cancellation for the PBE0/6-31G model due to inadequacies of the target model relative to experiment (such as the neglect of solvent and dynamics) and the nearly linear variation of its errors with respect to the carbon chemical shielding environment (Figure 5.2). Given the highly non-linear relationships between the PBE0/6-31G and target shieldings for ^{15}N or ^{17}O in Figure 5.2, one would expect much greater differences in the experimental accuracy of PBE0/6-31G with and without the Δ-ML correction for those nuclei.

5.4.3 Computational Timings

Finally, to give some perspective on the computational costs of the Δ-ML models, Table 5.5 summarizes single-core wall timings in Orca for the geometry optimization and subsequent NMR shielding calculation on five of the drug molecules studied above. Density-fitting algorithms were used throughout, except for the values listed in parentheses. Timings with the SVWN functional are not shown in the table, but they are about 10% faster than the PBE ones on average. Timings for evaluating the AEV and NN ensemble are also not shown explicitly, since they require only hundredths of a second per molecule once the software libraries have been loaded into memory, or less than two seconds each if library loading is included.

In the traditional scheme of geometry optimization with ωB97X/6-31G(d) followed by a PBE0/6-311+G(2d,p) NMR chemical shielding calculation, the NMR calculation constitutes about a third of the computational time, while the geometry optimization occupies the other two-thirds. Using the small-basis Δ-ML models reduces the cost of the NMR calculation by 1–2 orders of magnitude, such that the NMR calculation constitutes no more than a few percent of the geometry optimization time.

The discussion above focused primarily on Δ-ML using PBE0/6-31G shieldings. Without density fitting algorithms (as was done in the Gaussian calculations used to generate the results above), the PBE0 functional costs only about 30% more than PBE, potentially making the minor accuracy gains of PBE0/6-31G Δ-ML worthwhile. With density fitting, PBE/6-31G shielding calculations become 4–5 times faster than PBE0/6-31G, in which case the minor loss in accuracy of the Δ-ML model (Table 5.2) is arguably out-

Species	Geom Opt ωB97X 6-31G(d)	NMR Shielding calculation PBE0 6-311+G(2d,p)	PBE0 6-31G	PBE 6-31G	PBE0 STO-3G	PBE STO-3G
Acetaminophen ($C_8H_9NO_2$)	19.8	6.0 (61.1)	0.7 (2.1)	0.2 (1.8)	0.3	0.2
Aspirin ($C_9H_8O_4$)	20.0	11.2 (106)	1.0 (4.0)	0.3 (3.2)	0.4	0.2
Nitrofurantoin ($C_8H_6N_4O_5$)	32.2	18.7 (165)	1.8 (4.8)	0.4 (3.6)	0.6	0.3
Nalidixic Acid ($C_{12}H_{12}N_2O_3$)	38.8	28.4 (257)	2.3 (7.6)	0.6 (5.8)	0.9	0.4
Cortisone Acetate ($C_{23}H_{30}O_6$)	461	205 (1540)	14.2 (39.0)	2.1 (28.4)	4.8	1.3
Mean Δ-ML NMR Savings Factor			11 (33)	50 (42)	31	80

Table 5.5: Timings (in minutes) for the ωB97X/6-31G(d) geometry optimization and subsequent NMR chemical shielding calculations with several different model chemistries in Orca. Most timings utilized density fitting algorithms, though select timings without density fitting are given in parentheses. All timings utilized a single AMD EPYC 7282 core with 4 GB RAM and a solid-state hard disk.

weighed by the computational savings. On the other hand, the cost of either Δ-ML shielding calculation is trivial compared to the geometry optimization. In the end, the most appropriate low-cost Δ-ML shielding calculation will depend on the application: When obtaining the geometry represents the computational bottleneck, the more accurate Δ-ML models are probably worthwhile given the small marginal cost. Alternatively, if one were sampling many structures along a molecular dynamics trajectory or looking at very large systems, the less-expensive Δ-ML models become more attractive.

5.5 Conclusion

We have developed Δ-ML models using various combinations of density functionals and basis sets to predict isotropic chemical shieldings quickly. We first assessed the performance Δ-ML models for ^{13}C data sets to establish trends regarding the roles of the functional and basis set. The Δ-ML-corrected PBE0/6-31G model proved to be the best-performing model of the six combinations tested to predict PBE0/6-311+G(2d,p) shieldings,

though several other Δ-ML models tested performed only slightly worse. The PBE0/6-31G + Δ-ML model performs well across all four nuclei tested, including the highly non-linear learning cases of ^{15}N and ^{17}O. These errors are several-fold smaller than both what has been obtained previously in the literature using pure ML models and and are a fraction of the errors expected for the target DFT model relative to experiment.

Using the PBE0/6-31G Δ-ML models, we showed that uncertainty quantification is possible from the ensemble of predicted chemical shieldings. Specifically, larger standard deviations among the ensemble members are associated with greater uncertainty in the shielding predictions. Such uncertainty quantification could be useful for interpreting the level of agreement or disagreement between the Δ-ML-predicted shieldings and experimental shifts.

As a final test, we evaluated the accuracy of predicted ^{13}C chemical shifts to known experimental shifts. We first employed experimental chemical shifts in a set of small molecules to develop regression parameters that convert our predicted chemical shieldings to experimental chemical shifts in DMSO and $CDCl_3$. We then used these parameters to predict chemical shifts for a set of rigid pharmaceutical molecules with RMSE that are almost as good as those of the target DFT predictions. Despite the potential inadequacies associated with our target gas-phase PBE0/6-311+G(2d,p) model for predicting experimental chemical shifts, we demonstrate that the lower-cost Δ-ML approach predicts shifts with accuracy that is only marginally worse. In other words, the Δ-ML model does exactly what it is trained for, which is to improve the "cheap" shielding calculation relative to the target level of theory.

The low-cost and particularly high fidelity of the Δ-ML chemical shieldings to the DFT ones open a number of potentially interesting opportunities for the future. For example, dynamical averaging of chemical shifts and explicit treatment of local solvent effects are known to be important in many NMR problems, and the Δ-ML approach could potentially be used for inexpensively averaging over snapshots from a molecular dynamics trajectory without sacrificing DFT-accuracy. The accurate Δ-ML models here also potentially expand the role for NMR-aided geometry optimizations that combine energy and chemical shift data to solve structures directly, circumventing the traditional trial-and-error process of generating candidate structures, computing shifts, and assessing agreement with experiment. Effective approaches will require cheap chemical shielding predictions that don't sacrifice quantum mechanical accuracy, such as the Δ-ML approach here. It remains to be seen how much additional training data would be required for the NNs to learn how to predict chemical shieldings for non-equilibrium structures, though the use of Δ-ML could potentially simplify the process by capturing a substantial fraction of the geometry-dependent variations in the low-cost shielding. In the longer term, it will also be important to extend the models to molecules containing atoms other than hydrogen, carbon, nitrogen, and oxygen.

In the next chapter, we begin to explore ML chemical shift prediction in the solid state. While we have shown Δ-ML is the most accurate ML chemical shielding prediction model, it would be even better if there was no additional computational overhead with the initial "cheap" calculation. To overcome these bottlenecks, we discuss our implementation of graph neural networks in various flavors (convolutional, attentional, and message-passing)

for existing ssNMR datasets, as well as show that they are easily transferable to more diverse stochiometeries.

Chapter 6

Predicting Solid-State Nuclear Magnetic Resonance Chemical Shifts Using Graph Neural Networks

6.1 Introduction

Nuclear magnetic resonance (NMR) chemical shifts play an important role in the structural determination of materials, pharmaceuticals, and biologically relevant molecules.[196] In combination with X-ray diffraction and chemical shift prediction models, one can routinely resolve atomic positions beyond the limits of diffraction alone. The goal of chemical shift prediction is to then refine or validate candidate structures that closely align

with experimental shifts. While popular quantum chemistry methods such as density functional theory (DFT) work well for structure optimization and chemical shift prediction, the rapidly increasing computational demands become a significant bottleneck in large systems or when many candidate structures must be tested. Furthermore, extending DFT predictions to disordered solids becomes non-trivial, since the deviation from periodicity often requires calculations on many different local environments.[168] Less-expensive empirical shift models have been developed, but they work best for specific classes of systems, such as proteins.[266, 213, 155]

More recently, low-cost machine-learning (ML) chemical shift prediction models have started transforming the NMR crystallography of organic and inorganic materials.[178, 146, 36] In crystal structure prediction, fast ML shift prediction can be used to rapidly screen large numbers candidate structures against experimental NMR data to identify the correct structure.[178, 73] They can dramatically reduce the cost of computing ensemble–averaged chemical shifts from molecular dynamics (MD) trajectories. They can be used to assess the impacts of different local environments in disordered minerals and amorphous pharmaceuticals, sometimes using simulation cells that are far too large for conventional DFT.[57] In the future, sufficiently accurate NMR ML models could help interpret recent NMR experiments which monitored crystallization processes,[128] characterize complex and/or larger organic semiconductor materials,[210] and perhaps even improve our ability to invert from NMR spectrum to 3-D structure by enabling NMR–driven geometry optimizations.[204, 12]

Unfortunately, existing solid–state NMR ML models are limited in their elemental diversity (H, C, N, and O) and often perform noticeably worse for NMR parameters of

atom types other than ^{1}H. To remedy the accuracy limitations, we have developed a Δ-ML model which predicts DFT-quality NMR chemical shifts in small molecules by "learning" how to correct an inexpensive chemical shift prediction up to a more accurate one.[250] This Δ-ML model is the first to achieve DFT-accuracy chemical shifts in small molecules, but the neural network based on atom-centered symmetry function (ACSF) descriptors [18] requires extensive training data that will hinder its generalization to a wider variety of chemical environments in solid–state NMR.

Here, we present our efforts to develop an improved and more data-efficient ML model based on graph neural networks (GNNs) that achieve DFT–quality solid state NMR chemical shift prediction for organic materials. GNNs learn compact feature representations that can be easily transferred to other atom types. Improvements in chemical shift prediction accuracy will facilitate the use of NMR to characterize the structures of molecular crystals, polymers, pharmaceuticals, and other complex organic materials. We begin by discussing GNNs in the context of materials prediction, and current methodologies. We then begin to apply these methods to existing solid–state chemical shift datasets. After training, we scrutinize the GNN models to show that exisiting models are not sufficiently powerful to discern between H and C environments, but can be used for improvements in N and O.

6.2 Theory and Methods

A graph convolutional neural network learns a representation from the graph nodes, edges, and subgraphs in a low-dimensional vector. [263] While there is a rich his-

tory in the literature for GNNs in cheminformatics, GNNs for quantum chemistry have become increasingly more popular for energy, forces, and other property predictions like charges and dipoles. [91] GNNs for quantum chemistry are popular because one does not rely on predefined features such as ACSFs or similarity-based descriptions of the atomic environment like the smooth overlap of atomic positions (SOAP). [15] GNNs rather learn a robust feature description on-the-fly. More importantly, the GNN learned input description (also known as latent space embedding) does not grow exponentially when incorporating additional atom types allowing one to seamlessly incorporate more diverse data. GNNs have been used previously for chemical shift prediction, however only for the narrow case of solution-phase proteins. [269]

6.2.1 ML Training and Testing Data

For the training and testing dataset, we used the Cambridge Structural Database sets created from the original ShiftML paper. [177] In summary, the training dataset contains 2000 structures (CSD-2k) and 500 for testing (CSD-500). The CSD-2k set was constructed via furthest point sampling using the SOAP kernel to assess similarity from the subset containing HCNO and other size cutoffs called CSD-61k. The remaining structures were randomly sampled to create the CSD-500. Other works have used this train/test split, [146] and we adopt it here as well for direct comparison. The breakdown of atomic species in the set is shown in table 6.1 as well as the best prediction accuracy per atom type.

We also began to explore other atom types using the CSD-Drug subset as a starting point. [31] Here, we targeted crystal structures containing S, F, Cl, Br, I, and P atom types

Atoms	Train CSD-2k	Test CSD-500	Best RMSE (ppm)
C	76,174	29,913	0.37
H	58,148	26,607	3.3
N	27,814	2,713	10.2
O	25,924	5,404	15.3

Table 6.1: Breakdown of CSD-2k and CSD-500 training and testing datasets, respectively. The datasets are lacking a N and O atom types in each data partition which explains previously reported large errors upon testing. The best RMSE comes from the performance of an ensemble neural network model reported in ref. [146] on the CSD-500.

in addition to HCNO in hopes to augment the original CSD-2k and CSD-500. Crystal structures from the CSD-2k and CSD-500 sets have previously been geometry optimized and their GIPAW chemical shieldings tabulated. We use the same methods from ref. [177] on our augmented data. Briefly, we use the Quantum ESPRESSO v6.5 [90, 89] for all calculations to optimize an additional 1512 crystal structures. GIPAW reconstruction was carried out with ultrasoft pseudopotentials: H.pbe-kjpaw_psl.0.1.UPF, C.pbe-n-kjpaw_psl.0.1.UPF, N.pbe-n-kjpaw_psl.0.1.UPF, Br.pbe-n-kjpaw_psl.1.0.0.UPF, Cl.pbe-n-kjpaw_psl.0.1.UPF, F.pbe-n-kjpaw_psl.0.1.UPF, I.pbe-n-kjpaw_psl.1.0.0.UPF, P.pbe-n-kjpaw_psl.1.0.0.UPF, S.pbe-n-kjpaw_psl.0.1.UPF, and O.pbe-n-kjpaw_psl.0.1.UPF. Optimizations were carried out using the GGA density functional PBE, with the Grimme dispersion correction. [95] All lattice constants were kept fixed while atomic positions were allowed to relax. For the optimizations a 60 Ry energy cut-off and a 240 Ry charge density cut-off were used. For the GIPAW calculations, the parameters were tightened to 100 Ry and 400 Ry, respectively. We also used a stringent SCF cutoff of 10^{-12} Ry to avoid any residual error as noted in the SI of ref. [177].

6.2.2 Graph Neural Network Details

The GNNs described in this project are implemented in the graph neural network package for materials, MatDeepLearn. [86] There are very few solutions for crystal property prediction using GNNs, and MatDeepLearn fills that role with easy to implement routines since it is built through pytorch and pytorch-geometric. [179, 77] Individual structure graphs were constructed with respect to periodic boundary conditions through atomic simulation environment. [138] Node neighbors were generated with local cutoffs of 4, 6, or 8 Å to limit the size of the neighbor list. During the processing step, the edge distances are expanded in a Gaussian basis set to provide a continuous description of the interatomic distances (eq. 6.1).

$$e_{i,j} = \exp(-\gamma(d_{i,j} - \mu)^2) \tag{6.1}$$

In practice, we use $\gamma = 0.5$ as the normalization factor, $d_{i,j}$ the interatomic distance in Å, and μ, as the shifting parameter. We sample 50 equally spaced points between each interatomic distance. Each node, edge, and graph is then fed through a graph-independent layer to update the embedding environment, where the NNs try to "learn" the features of each graph, edge, or node through each pass.

The overall workflow of MatDeepLearn is depicted in figure 6.1. Each node is decomposed into its local graph via some cutoff distance. The neighbor list is then constructed and each edge attribute is then created. Then, each node is passed through the graph-independent layer constantly updating the embeddings via a NN. Between each graph layer, we use batch normalization to ensure the distribution of each layer's inputs remains constant during training, with respect to the parameters of the previous layers. [124] Af-

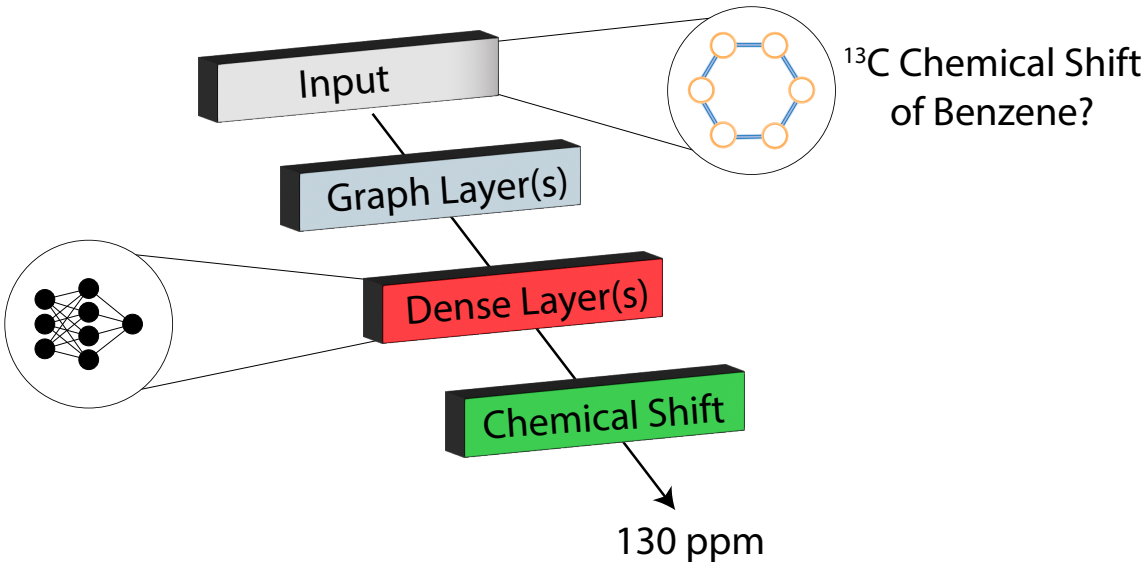


Figure 6.1: **GNN workflow** Each structure (graph) is decomposed into its elementary nodes. The nodes then generate a local neighbor list based on some cutoff distance. The nodes, edges, and graph features are then fed into the graph layer(s) as depicted in figure 4.5. After each node and edge are embedded, the structure is then passed through a fully-connected simply NN to predict the chemical shielding of each atom (node). d

ter the graph layers, the graphs are then passed through a traditional dense NN, where a chemical shielding is predicted for each node. As described in section 4.5.1, we use the convolutional and attentional operators to "pass messages" between each node and its respective neighbor. In addition, we benchmark against the SOAP kernel, a predefined feature descriptor, and other GNN flavors such as SchNet and MEGNet. [207, **?**, 41] SchNet and MEGNet are also convolutational flavors, except MEGNet also tries to learn the embeddings of each edge. Lastly, we began to explore heterographs, where each node and edge type are labeled, which are different than homogeneous graphs. [255] The motivation behind discretized node and edge types is similar to how chemical structures have various atom

and bond types. We further modified MatDeepLearn to process crystal structures and label each bond type based on atom connectivity for this functionality. In the message passing step, the embeddings are then the aggregation of each node type with respect to the other node types interacting with each other.

6.3 Results and Discussion

6.3.1 Benchmarking GNNs for Chemical Shift Prediction

We begin our discussion by analyzing the correlation plots from predicted shielding of a single training instance to ground truth DFT GIPAW values. At first glance in figure 6.2, the overall correlation seems to be reasonable and follows the $y = x$ line closely. However, upon calculating the RMSE, the errors are larger than one would hope. In particular, the 0.55 ppm RMSE for ^{1}H is notably large considered that GIPAW errors relative to experiment are usually smaller than 0.3 ppm RMSE. The ^{13}C errors are also large, with a RMSE of 4.83 ppm which is 1.5 ppm larger than the 3.3 ppm reported in table 6.1. The ^{1}H and ^{13}C correlation plots also reveal a handful of predictions in the most deshielded regions (10-15 ppm for H and -30 ti -50 for C). There are a cluster of points with a constant value indicating that the GNN model is not learning a representation of these types of chemical environments. The RMSE is not dominated by these points, and thus are not the root cause of the large RMSEs.

While it may seem like a bad training instance, table 6.2 summarizes the best performing models for for each atom type when testing on the CSD-500 set. Overall, the

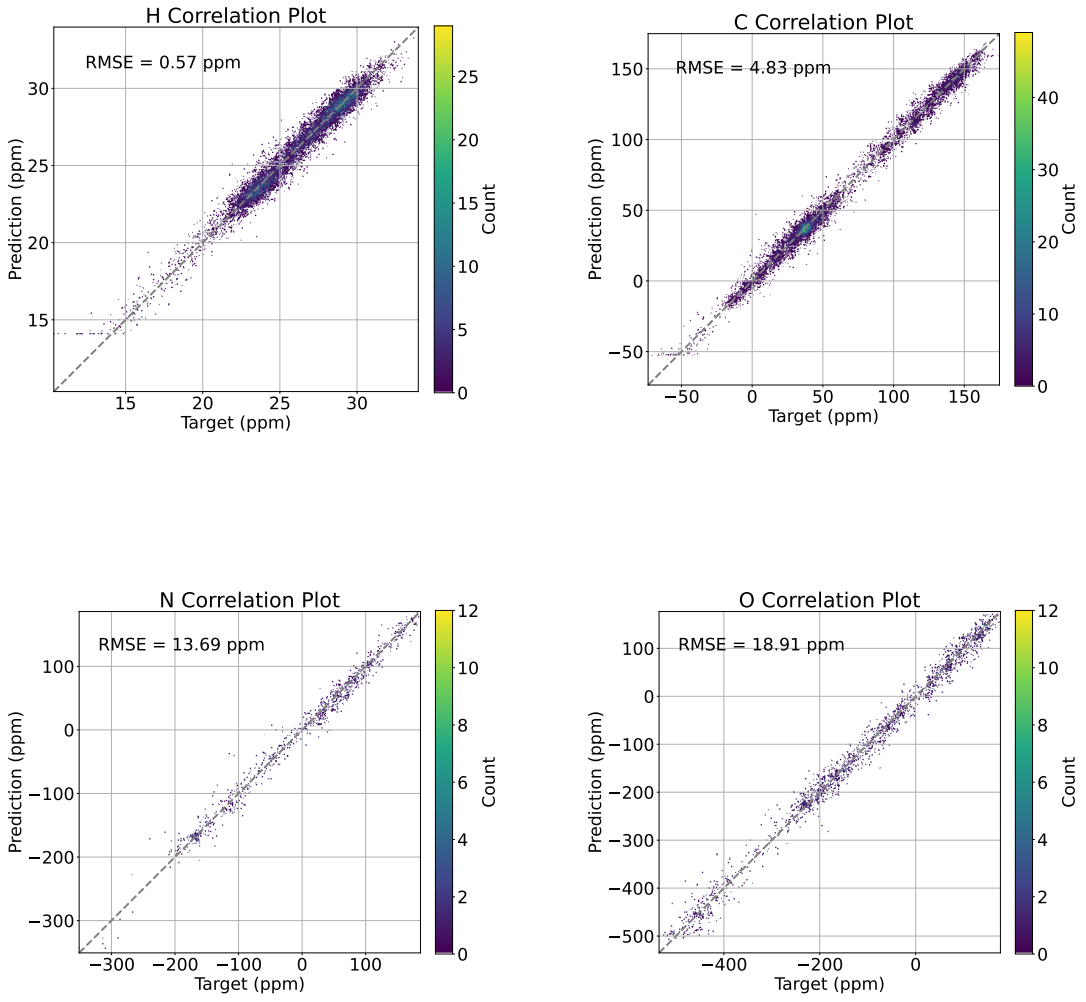


Figure 6.2: Target vs. predicted chemical shieldings from a **single** trained GATGNN model on the testing dataset (CSD-500). See table 6.1 for atom count in testing data and literature precedent accuracy.

accuracy for each atom type seems to plateau regardless of model or hyperparameters (0.5 ppm for H, 4.8 ppm for C, 12.3 ppm for N, and 17-18 for O). Some other interesting trends that are noticeable are the effect of cutoff. For ^{1}H, longer cutoffs appear to reduce errors. However, this trend is not seen in other atom types. Another interesting trend is the repeating occurrence of GAT as one of the more accurate models. Attentional models like GAT are used over traditional convolutional models since the attention coefficients are weighted averages of the interactions. However, while the errors for ^{15}N and ^{17}O (13.69 and 18.91 ppm, respectively) are larger relative to the best reported model in table 6.1, these errors come closer to the previously reported ShiftML model. [177]. As show previously in other publications, [146, 177, 249, 223] ensemble models outperform any single model due to error cancellation. Thus, we then used an ensemble model to examine the performance boost (if any) relative to a single model. In figure 6.3, the average prediction of the top 10 performers for ^{1}H are shown. Surprisingly, the ensemble models reduce the errors of ^{1}H and ^{13}C prediction errors significantly (0.57 to 0.49, and 4.8 to 4.1, respectively). Furthermore, the ^{15}N and ^{17}O and prediction errors are appreciably better than literature precedents. The distribution of points also tightens around the $y = x$ line, indicating better correlation to the target values.

Lastly, we compare the newly constructed GNN models to predefined feature descriptors, such as the SOAP kernel. In figure 6.4, we see that SOAP models without hyperparameter tuning are significantly worse than the constructued GNN models. In ref. [177],

	Model	GC Layers	FC Layers	Dim1	Dim2	Cutoff (Å)	Act. Function	RMSE (ppm)
^{1}H models	GAT	5	2	64	64	8	softplus	0.55
	GAT	5	2	100	100	8	softplus	0.56
	GAT	4	2	64	64	6	softplus	0.56
	GAT	5	3	64	64	6	softplus	0.57
	GAT	4	2	100	100	8	softplus	0.57
	GAT	5	1	100	100	6	softplus	0.57
	GAT	5	3	100	100	6	softplus	0.57
	CGCNN	4	1	32	32	4	relu	0.57
	GAT	5	3	32	32	8	softplus	0.57
	GAT	5	3	64	64	8	softplus	0.57
^{13}C models	GAT	5	2	100	100	8	softplus	4.75
	GAT	5	3	100	100	8	softplus	4.79
	CGCNN	5	2	32	32	4	relu	4.79
	MEGNet	2	3	64	64	4	relu	4.79
	MEGNet	5	1	32	32	4	relu	4.79
	GAT	4	2	100	100	8	softplus	4.83
	GAT	5	3	32	32	8	softplus	4.84
	MEGNet	4	2	64	64	4	relu	4.85
	GAT	5	2	64	64	8	softplus	4.86
	MEGNet	3	1	100	100	4	relu	4.90
^{15}N models	MEGNet	3	3	64	64	4	relu	11.68
	GAT	5	3	100	100	6	softplus	12.32
	MEGNet	3	1	64	64	4	relu	12.33
	MEGNet	4	2	100	100	4	relu	12.33
	GAT	3	3	32	32	6	softplus	12.35
	MEGNet	5	3	32	32	4	relu	12.36
	GAT	3	2	64	64	4	softplus	12.45
	CGCNN	4	1	32	32	4	relu	12.55
	MEGNet	3	1	100	100	4	relu	12.58
	GAT	5	3	64	64	6	softplus	12.60
^{17}O models	GAT	5	2	64	64	8	softplus	17.62
	SchNet	3	1	100	100	4	relu	17.82
	MEGNet	4	1	32	32	4	relu	18.32
	MEGNet	4	2	64	64	4	relu	18.36
	GAT	5	3	100	100	8	softplus	18.36
	GAT	5	3	100	100	6	softplus	18.39
	MEGNet	3	3	64	64	4	relu	18.40
	GAT	5	2	100	100	8	softplus	18.44
	MEGNet	2	3	64	64	4	relu	18.46
	MEGNet	3	2	64	64	4	relu	18.58

Table 6.2: Comparison of top ten most accurate GNN models per atom type. Here, the number of graph layers, dense layers, dimensionality of the graph layers, dimensionality of the dense layers, the local cutoff used, and the type of activation function is listed.

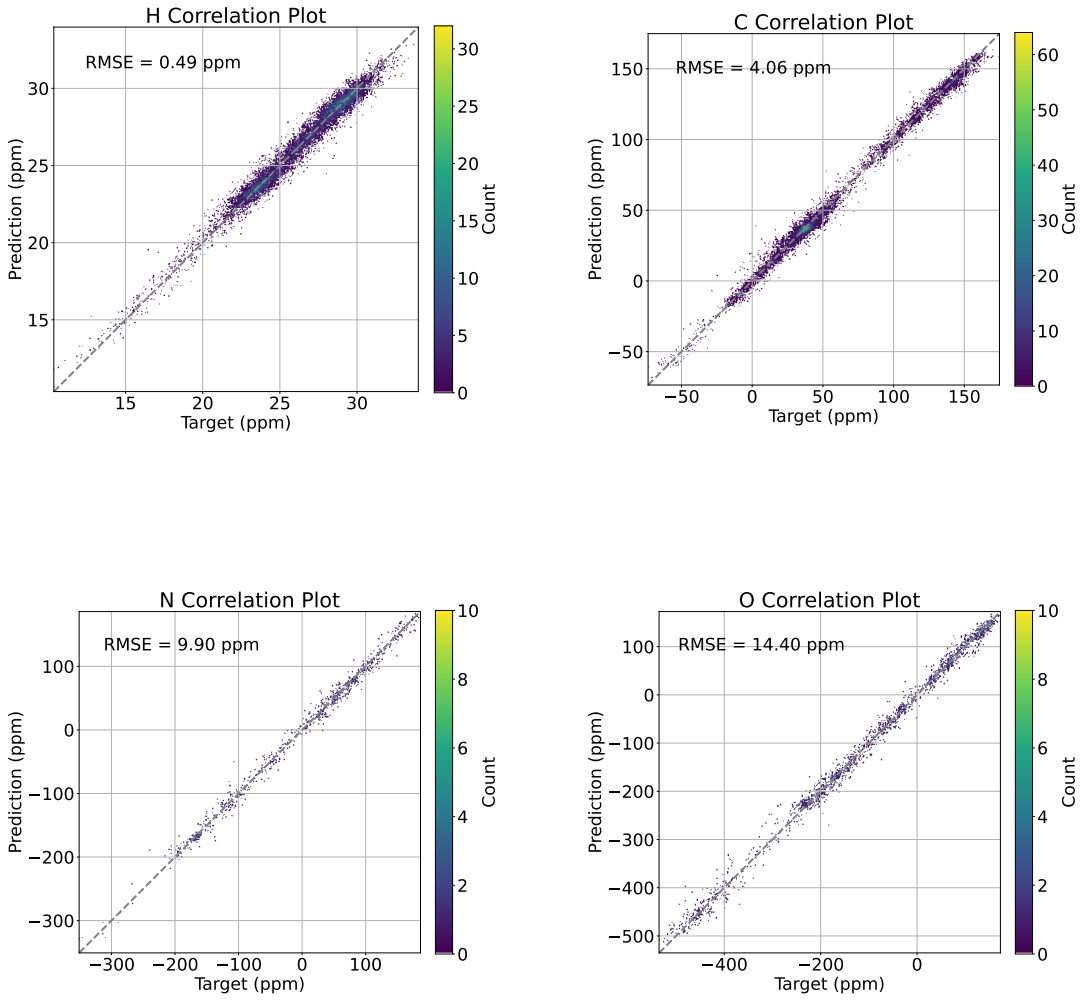


Figure 6.3: Target vs. predicted chemical shieldings from an **ensemble** of GNN models on the testing dataset (CSD-500). See table 6.1 for atom count in testing data and literature precedent accuracy.

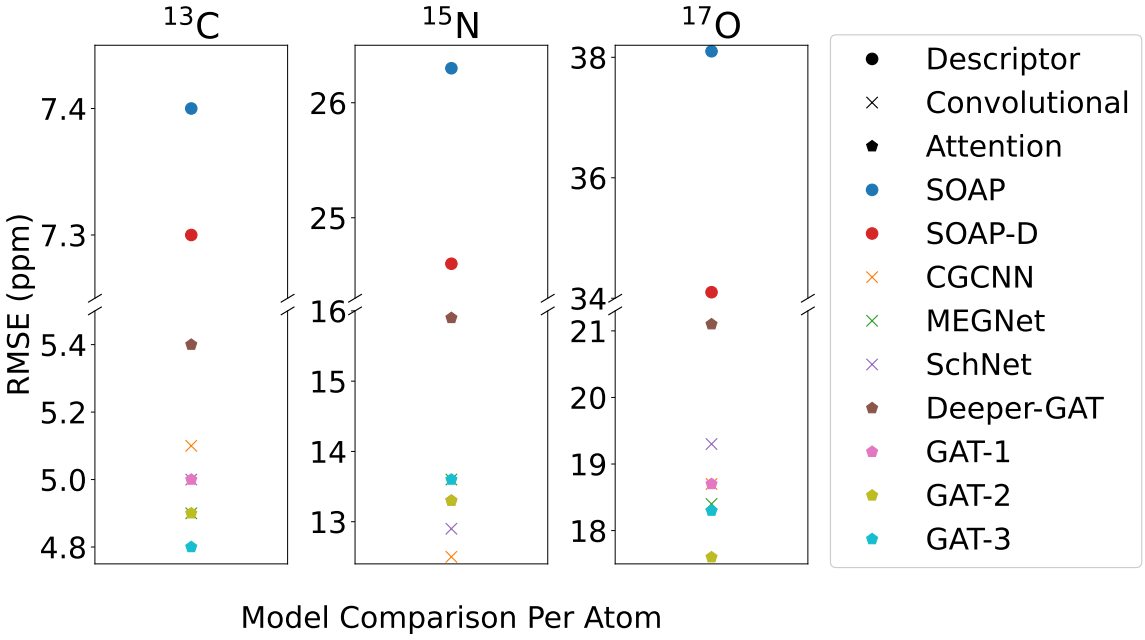


Figure 6.4: GNN models compared against two different variations of SOAP for CNO atom types. The SOAP descriptor consistently underperforms relative to the GNN models. Some SOAP kernel tuning may be necessary to improve performance, but the boon of GNN models are the simplicity in which one can use these models. Here, we see that attentional type work slightly better for ^{13}C and ^{17}O atoms, while convolutional work best for ^{15}N. GAT-1 through GAT-3 represent different hyperparameters of the same model. Interestingly, there are no clear "winners" from these methods.

the SOAP kernel is used to predict chemical shieldings, but the method is a highly parameterized SOAP multi-scale kernel for predicting chemical shifts.

6.4 Conclusion

In summary, solid–state chemical shift prediction models based on GNN architectures have been constructed. Our models perform slightly worse than literature precedents for ^{1}H and ^{13}C, but perform better for ^{15}N and ^{17}O using ensemble models. We show that 2-body graph attentional models are the most accurate for all atom types. Crucially, the cutoff distances for attentional models improves accuracy relative to convolutional type models. The large errors for ^{1}H and ^{13}C may be due to lack of 3-body information which has been shown to improve predictions in molecular datasets. [130] A future study would directly compare the 3-body vs 2-body GNN models. Furthermore, Δ-ML methods have been shown to reduce errors albeit at a larger computational cost. [250] One could imagine performing a Δ-ML model which learns GIPAW chemical shieldings using a smaller planewave cutoff. Or to further increase computational savings, one could use the crystal monomers or use cluster approximations. Either of these methods will the be the subject of future studies to remedy the accuracy problem with DFT-based NMR chemical shifts. Lastly, since a generalizable method for improving accuracy for all atom types was not found, we did not further explore the augmented structures. Once a robust GNN architecture is discovered, one could easily expand the diversity of chemical shift prediction.

Chapter 7

Conclusions

In conclusion, we report the success of various methods to improve the accuracy of DFT-based NMR chemical shift predictions for the solution–phase and solid–state. We benchmark each method, and show their use on real-world systems.

1. The monomer-corrected GIPAW chemical shieldings yield excellent agreement with experimental shifts for ^{13}C, ^{15}N, ^{17}O on crystal benchmarks. In some cases, they even outperform the more accurate SCRMP method. We then demonstrate this correction on real world test cases such as testosterone, methacrylamide, and isocytosine in the solid-state.

2. Our PCM model allows us to easily extend a 2-body fragment method to biomolecular system to "shield" highly charged regions. We develop high–quality regression parameters on the crystal benchmark set. Then, from the regression parameters, we calculate solid–state NMR chemical shifts of an intermediate of tryptophan synthase and faithfully reproduce the two-site proton exchange.

3. We developed a training and testing NMR dataset of single component molecules containing approximately 58k structures. We then use this datset to develop the Δ-ML NMR method which shows true reproduction of target NMR shielding values 2–3 orders of magnitude cheaper than legacy calculations. We show that ensemble methods generate the best performance, as well as gauge the uncertainty of predictions. Using pre-trained Δ-ML models, we reproduce solution-phase chemical shifts of nine pharmaceutical molecules, such as acetaminophen and aspirin.

4. GNN models perform slightly worse for ^{1}H and ^{13}C predictions, but improve upon ^{15}N and ^{17}O predictions on the CSD-500 test set. This seems to be a limitation of 2-body GNN models, but ensemble models "clean" up a significant amount of errors.

There are a handful of research directions that warrant further exploration. First, current experimental benchmarks are quite limited and are solely focused on HCNO. If one is to predict reliable DFT-based chemical shifts for more diverse atom types, more experimental mappings would be needed. Next, extending the Δ-ML chemical shift models beyond single-component systems. Preliminary unreported data showed large errors and uncertainty when predicting shieldings from the monomer to the crystalline environment. Since the N=1–8 dataset is limited in intermolecular interactions, it is not hard to imagine that the inclusion of more intermolecular environments would decrease errors. Next, the GNN models used throughout the study were limited to only include 2-body information. Newer GNN models which modify the message-passing scheme to incorporate 3-body information have been developed. However, they were not implemented in our code to be extensively studied due to time constraints. Just like force fields, one can imagine that

the inclusion of 3-body or higher terms in the expression could improve accuracy. Another avenue is the Δ-ML scheme in GNN. Could Δ-ML improve performance? Lastly, extracting the "chemistry" from the ML model to help understand what is actually being learned in the ML algorithm could be impactful in designing new techniques for chemical shift prediction. One could use ML models to tease out the necessary information the ML black box. It is clear that more accurate ML models will be used for rapid chemical shift predictions in the future, reserving actual calculations when more information is truly necessary.

Bibliography

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, Xiaoqiang Zheng, and Google Brain. TensorFlow: A system for Large-Scale machine learning.

[2] C. Adamo and V. Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.*, 110:6158, 1999.

[3] João Aires-de Sousa, Markus C Hemmer, and Johann Gasteiger. Prediction of ^{1}H NMR Chemical Shifts Using Neural Networks. *Anal. Chem.*, 74(1):80–90, jan 2002.

[4] Roger Amos and Rika Kobayashi. Ab Initio NMR Chemical Shift Calculations Using Fragment Molecular Orbitals and Locally Dense Basis Sets. *J. Phys. Chem. A*, 120(44):8907–8915, nov 2016.

[5] Liaoyuan An, Yefei Wang, Ning Zhang, Shihai Yan, Ad Bax, and Lishan Yao. Protein apparent dielectric constant and its temperature dependence from remote chemical shift effects. *J. Am. Chem. Soc.*, 136(37):12816–9, sep 2014.

[6] Nongnuch Artrith, Keith T Butler, François-Xavier Coudert, Seungwu Han, Olexandr Isayev, Anubhav Jain, and Aron Walsh. Best practices in machine learning for chemistry. *Nat. Chem.*, 13(6):505–508, June 2021.

[7] Sharon E Ashbrook and David McKay. Combining solid-state NMR spectroscopy with first-principles calculations - a guide to NMR crystallography. *Chem. Commun.*, 52(45):7186–7204, June 2016.

[8] Sharon Elizabeth Ashbrook and David McKay. Combining Solid-State NMR Spectroscopy with First-Principles Calculations – A Guide to NMR Crystallography. *Chem. Commun.*, 52:7186–7204, 2016.

[9] Alexander A Auer and Jürgen Gauss. Triple excitation effects in coupled-cluster calculations of indirect spin–spin coupling constants. *J. Chem. Phys.*, 115(4):1619–1622, July 2001.

[10] Martin Babinský, Kateřina Bouzková, Matej Pipíška, Lucie Novosadová, and Radek Marek. Interpretation of crystal effects on NMR chemical shift tensors: electron and shielding deformation densities. *J. Phys. Chem. A*, 117(2):497–503, January 2013.

[11] Maria Baias, Jean-Nicolas Dumez, Per H Svensson, Staffan Schantz, Graeme M Day, and Lyndon Emsley. De novo determination of the crystal structure of a large drug molecule by crystal structure prediction-based powder NMR crystallography. *J. Am. Chem. Soc.*, 135(46):17501–17507, November 2013.

[12] Martins Balodis, Manuel Cordova, Albert Hofstetter, Graeme M Day, and Lyndon Emsley. De novo crystal structure determination from machine learned chemical shifts. *J. Am. Chem. Soc.*, 144(16):7215–7223, April 2022.

[13] Giampaolo Barone, Luigi Gomez-Paloma, Dario Duca, Arturo Silvestri, Raffaele Riccio, and Giuseppe Bifulco. Structure validation of natural products by quantum-mechanical GIAO calculations of 13C NMR chemical shifts. *Chemistry*, 8(14):3233–3239, July 2002.

[14] Albert P. Bartok, Michael J. Gillan, Frederick R. Manby, and Gabor Csanyi. Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water. *Phys. Rev. B*, 88(5):054104, aug 2013.

[15] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B Condens. Matter*, 87(18):184115, May 2013.

[16] Albert P Bartók, Mike C Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, 104(13):136403, April 2010.

[17] Jörg Behler and Michele Parrinello. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.*, 98(14):146401, apr 2007.

[18] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98(14):146401, April 2007.

[19] G J O Beran. Calculating nuclear magnetic resonance chemical shifts from density functional theory: A primer. *eMagRes*, 2007.

[20] G. J. O. Beran. Modeling polymorphic molecular crystals with electronic structure theory. *Chem. Rev.*, 116:5567–5613, 2016.

[21] G. J. O. Beran, J. .D. Hartman, and Y. N. Heit. Predicting molecular crystal properties from first principles: Finite-temperature thermochemistry to NMR crystallography. *Acc. Chem. Res.*, 49:2501–2508, 2016.

[22] G. J. O. Beran and K. Nanda. Predicting organic crystal lattice energies with chemical accuracy. *J. Phys. Chem. Lett.*, 1:3480–3487, 2010.

[23] Gregory J. O. Beran. Calculating nuclear magnetic resonance chemical shifts from density functional theory: A primer. *eMagRes*, 8:215–226, 2019.

[24] F Bloch. Nuclear induction. *Phys. Rev.*, 70(7-8):460–474, October 1946.

[25] A Daniel Boese and Joachim Sauer. Embedded and DFT calculations on the crystal structures of small alkanes, notably propane. *Cryst. Growth Des.*, 17(4):1636–1646, April 2017.

[26] Christian Bonhomme, Christel Gervais, Florence Babonneau, Cristina Coelho, Frédérique Pourpoint, Thierry Azaïs, Sharon E Ashbrook, John M Griffin, Jonathan R Yates, Francesco Mauri, and Chris J Pickard. First-principles calculation of NMR parameters using the gauge including projector augmented wave method: a chemist's point of view. *Chem. Rev.*, 112(11):5733–5779, November 2012.

[27] Christian Bonhomme, Christel Gervais, Florence Babonneau, Cristina Coelho, Frédérique Pourpoint, Thierry Azaïs, Sharon E Ashbrook, John M Griffin, Jonathan R Yates, Francesco Mauri, and Chris J Pickard. First-Principles Calculation of NMR Parameters Using the Gauge Including Projector Augmented Wave Method: A Chemist's Point of View. *Chem. Rev.*, 112(11):5733–5779, nov 2012.

[28] Kateřina Bouzková, Martin Babinský, Lucie Novosadová, and Radek Marek. Intermolecular interactions in crystalline theobromine as reflected in electron deformation density and (13)c NMR chemical shift tensors. *J. Chem. Theory Comput.*, 9(6):2629–2638, June 2013.

[29] Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Crim. Justice Behav.*, 36(1):21–40, January 2009.

[30] Keith A Brown. MODEL, GUESS, CHECK: Wordle as a primer on active learning for materials research. *npj Computational Materials*, 8(1):1–3, May 2022.

[31] Mathew J. Bryant, Simon N. Black, Helen Blade, Robert Docherty, Andrew G.P. Maloney, and Stefan C. Taylor. The CSD Drug Subset: The Changing Chemistry and Crystallography of Small Molecule Pharmaceuticals. *J. Pharm. Sci.*, 108(5):1655–1662, may 2019.

[32] Kieron Burke and Lucas O Wagner. DFT in a nutshell. *Int. J. Quantum Chem.*, 113(2):96–101, January 2013.

[33] E. Cancès, B. Mennucci, and J. Tomasi. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *J. Chem. Phys.*, 107(8):3032–3041, aug 1997.

[34] B. G. Caulkins, R. P. Young, R. A. Kudla, C. Yang, T. Bittbauer, B. Bastin, M. J. Marsella, M. F. Dunn, and L. J. Mueller. NMR Crystallography of a Carbanionic Intermediate in Tryptophan Synthase: Chemical Structure, Tautomerization, and Reaction Specificity. *submitted*, 2016.

[35] Bethany G. Caulkins, Robert P. Young, Ryan A. Kudla, Chen Yang, Thomas J. Bittbauer, Baback Bastin, Eduardo Hilario, Li Fan, Michael J. Marsella, Michael F. Dunn, and Leonard J. Mueller. NMR Crystallography of a Carbanionic Intermediate in Tryptophan Synthase: Chemical Structure, Tautomerization, and Reaction Specificity. *J. Am. Chem. Soc.*, 138(46):15214–15226, nov 2016.

[36] Ziyad Chaker, Mathieu Salanne, Jean-Marc Delaye, and Thibault Charpentier. NMR shifts in aluminosilicate glasses via machine learning. *Phys. Chem. Chem. Phys.*, 21(39):21709–21725, October 2019.

[37] Kevin R Chalek, Xinning Dong, Fei Tong, Ryan A Kudla, Lingyan Zhu, Adam D Gill, Wenwen Xu, Chen Yang, Joshua D Hartman, Alviclér Magalhães, Rabih O Al-Kaysi, Ryan C Hayward, Richard J Hooley, Gregory J O Beran, Christopher J Bardeen, and Leonard J Mueller. Bridging photochemistry and photomechanics with NMR crystallography: the molecular basis for the macroscopic expansion of an anthracene ester nanorod. *Chem. Sci.*, 12(1):453–463, January 2021.

[38] Monique Chan-Huot, Alexandra Dos, Reinhard Zander, Shasad Sharif, Peter M Tolstoy, Shara Compton, Emily Fogle, Michael D Toney, Ilya Shenderovich, Gleb S Denisov, and Hans-Heinrich Limbach. NMR studies of protonation and hydrogen bond states of internal aldimines of pyridoxal 5'-phosphate acid-base in alanine racemase, aspartate aminotransferase, and poly-l-lysine. *J. Am. Chem. Soc.*, 135(48):18160–18175, December 2013.

[39] Thibault Charpentier. The PAW/GIPAW approach for computing NMR parameters: a new dimension added to NMR study of solids. *Solid State Nucl. Magn. Reson.*, 40(1):1–20, July 2011.

[40] Thibault Charpentier. The PAW/GIPAW approach for computing NMR parameters: a new dimension added to NMR study of solids. *Solid State Nuc. Magn. Reson.*, 40(1):1–20, jul 2011.

[41] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.*, 31(9):3564–3572, May 2019.

[42] Xi Chen and Chang-Guo Zhan. First-principles studies of C-13 NMR chemical shift tensors of amino acids in crystal state. *J. Mol. Struct. (THEOCHEM)*, 682(1-3):73–82, aug 2004.

[43] D. B. Chesnut and K. D. Moore. Locally dense basis sets for chemical shift calculations. *J. Comp. Chem.*, 10(5):648–659, jul 1989.

[44] D B Chesnut, B E Rusiloski, K D Moore, and D A Egolfs. Use of Locally Dense Basis Sets for Nuclear Magnetic Resonance Shielding Calculations. *J. Comp. Chem.*, 14(11):1364–1375, 1993.

[45] D.B. Chesnut and C.G. Phung. Ab initio determination of chemical shielding in a model dipeptide. *Chem. Phys. Lett.*, 183(6):505–509, sep 1991.

[46] Ming-Sin Cheung, Mahon L. Maguire, Tim J. Stevens, and R. William Broadhurst. DANGLE: A Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure. *J. Magn. Res.*, 202(2):223–233, feb 2010.

[47] Anders S Christensen, Lars A Bratholm, Felix A Faber, David R Glowacki, and O Anatole von Lilienfeld. FCHL revisited: faster and more accurate quantum machine learning. September 2019.

[48] Anders S. Christensen, Troels E. Linnet, Mikael Borg, Wouter Boomsma, Kresten Lindorff-Larsen, Thomas Hamelryck, and Jan H. Jensen. Protein Structure Validation and Refinement Using Amide Proton Chemical Shifts Derived from Quantum Mechanics. *PLoS ONE*, 8(12):e84123, dec 2013.

[49] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I.J. Probert, K. Refson, and M. C. Payne. First principles methods using CASTEP. *Z. Kristallogr.*, 220:567–570, 2005.

[50] Aron J Cohen, Paula Mori-Sánchez, and Weitao Yang. Challenges for density functional theory. *Chem. Rev.*, 112(1):289–320, January 2012.

[51] Michael A. Collins and Ryan P. A. Bettens. Energy-Based Molecular Fragmentation Methods. *Chem. Rev.*, 115:5067–5642, 2015.

[52] Qiang Cui and Martin Karplus. Molecular Properties from Combined QM/MM Methods. 2. Chemical Shifts in Large Molecules. *J. Phys. Chem. B*, 104(15):3721–3743, apr 2000.

[53] Jérôme Cuny, Yu Xie, Chris J. Pickard, and Ali A. Hassanali. Ab Initio Quality NMR Parameters in Solid-State Materials Using a High-Dimensional Neural-Network Representation. *J. Chem. Theory Comput.*, 12(2):765–773, feb 2016.

[54] Nianzu Dai. Compass. In Jueming Hua and Lisheng Feng, editors, *Thirty Great Inventions of China: From Millet Agriculture to Artemisinin*, pages 663–683. Springer Singapore, Singapore, 2020.

[55] Ameya Daigavane, Balaraman Ravindran, and Gaurav Aggarwal. Understanding convolutions on graphs. *Distill*, 6(8), August 2021.

[56] Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.*, 18(20):13754–13769, May 2016.

[57] Stefan T. Norberg Anna Svensk Ankarberg Staffan Schantz Lyndon Emsley Manuel Cordova Martins Balodis Albert Hofstetter Federico Paruzzo Sten O Nilsson Lill Emma S E Eriksson Pierrick Berruyer Bruno Simões de Almeida, Michael J. Quayle. Structure determination of an amorphous drug through large-scale NMR predictions.

[58] Itzam De Gortari, Guillem Portella, Xavier Salvatella, Vikram S Bajaj, Patrick C A van der Wel, Jonathan R Yates, Matthew D Segall, Chris J Pickard, Mike C Payne, and Michele Vendruscolo. Time averaging of NMR chemical shifts in the MLF peptide in the solid state. *J. Am. Chem. Soc.*, 132(17):5993–6000, May 2010.

[59] Vitali Deev and Michael A Collins. Approximate ab initio energies by systematic molecular fragmentation. *J. Chem. Phys.*, 122(15):154102, apr 2005.

[60] Volker L Deringer, Albert P Bartók, Noam Bernstein, David M Wilkins, Michele Ceriotti, and Gábor Csányi. Gaussian process regression for materials and molecules. *Chem. Rev.*, 121(16):10073–10141, August 2021.

[61] Christian Devereux, Justin S. Smith, Kate K Davis, Kipton Barros, Roman Zubatyuk, Olexandr Isayev, and Adrian E. Roitberg. Extending the applicability of the ANI deep learning molecular potential to Sulfur and Halogens. *J. Chem. Theory Comput.*, 2020.

[62] Robert Ditchfield. Self-consistent perturbation theory of diamagnetism. *Mol. Phys.*, 27(4):789–807, April 1974.

[63] Grygoriy A Dolgonos, Oleksandr A Loboda, and A Daniel Boese. Development of embedded and performance of density functional methods for molecular crystals. *J. Phys. Chem. A*, 122(2):708–713, January 2018.

[64] Martin Dračínský, Petr Bouř, and Paul Hodgkinson. Temperature dependence of NMR parameters calculated from path integral molecular dynamics simulations. *J. Chem. Theory Comput.*, 12(3):968–973, March 2016.

[65] Martin Dračínský, Petr Bouř, and Paul Hodgkinson. Temperature Dependence of NMR Parameters Calculated from Path Integral Molecular Dynamics Simulations. *J. Chem. Theory Comput.*, 12(3):968–973, mar 2016.

[66] Martin Dračínský and Paul Hodgkinson. Effects of quantum nuclear delocalisation on NMR parameters from path integral molecular dynamics. *Chemistry*, 20(8):2201–2207, February 2014.

[67] Martin Dračínský and Paul Hodgkinson. Solid-state NMR studies of nucleic acid components. *RSC Adv.*, 5(16):12300–12310, 2015.

[68] Martin Dračínský, Petr Jansa, Kari Ahonen, and Miloš Buděšínský. Tautomerism and the protonation/deprotonation of isocytosine in liquid- and solid-states studied by NMR spectroscopy and theoretical calculations. *European J. Org. Chem.*, 2011(8):1544–1551, March 2011.

[69] Martin Dračínský, Eliška Procházková, Jiří Kessler, Jaroslav Šebestík, Pavel Matějka, and Petr Bouř. Resolution of organic polymorphic crystals by raman spectroscopy. *J. Phys. Chem. B*, 117(24):7297–7307, June 2013.

[70] Martin Dračínský, Pablo Unzueta, and Gregory J. O. Beran. Improving the accuracy of solid-state nuclear magnetic resonance chemical shift prediction with a simple molecular correction. *Phys. Chem. Chem. Phys.*, 21:14992–15000, 2019.

[71] Pavlo O. Dral. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.*, 11(6):2336–2347, mar 2020.

[72] Simon M Eckard, Andrea Frank, Ionut Onila, and Thomas E Exner. Approximations of Long-Range Interactions in Fragment-Based Quantum Chemical Approaches. In Robert Zalesny, Manthos G. Papadopoulos, Paul G. Mezey, and Jerzy Leszczynski, editors, *Linear Scaling Techniques in Computational Chemistry and Physics*, volume 13 of *Challenges and Advances in Computational Chemistry and Physics*, pages 157–173. Springer Netherlands, Dordrecht, 2011.

[73] Edgar A Engel, Andrea Anelli, Albert Hofstetter, Federico Paruzzo, Lyndon Emsley, and Michele Ceriotti. A bayesian approach to NMR crystal structure determination. *Phys. Chem. Chem. Phys.*, 21(42):23385–23400, November 2019.

[74] S T Epstein. Gauge invariance, current conservation, and GIAO's. *J. Chem. Phys.*, 58(4):1592–1595, February 1973.

[75] J. C. Facelli and D. M. Grant. Determination of molecular symmetry in crystalline naphthalene using solid-state NMR. *Nature*, 365:325–327, 1993.

[76] M B Ferraro and J C Facelli. Modeling NMR chemical shifts: surface charge representation of the electrostatic embedding potential modeling of crystalline intermolecular effects in ^{19}F solid state NMR chemical shifts. *J. Mol. Struct.*, 603:159–164, 2002.

[77] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch geometric. March 2019.

[78] Tobias Fink and Jean-Louis Reymond. Virtual exploration of the chemical universe up to 11 atoms of c, n, o, f: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.*, 47(2):342–353, March 2007.

[79] Denis Flaig, Matthias Beer, and Christian Ochsenfeld. Convergence of Electronic Structure with the Size of the QM Region: Example of QM/MM NMR Shieldings. *J. Chem. Theory Comput.*, 8(7):2260–2271, jul 2012.

[80] Aaron T Frank, Sean M Law, and Charles L Brooks. A Simple and Fast Approach for Predicting ^{1}H and ^{13}C Chemical Shifts: Toward Chemical Shift-Guided Simulations of RNA. *J. Phys. Chem. B*, 118(42):12168–12175, oct 2014.

[81] Andrea Frank, Heiko M. Moller, and Thomas E Exner. Toward the Quantum Chemical Calculation of NMR Chemical Shifts of Proteins. 2. Level of Theory, Basis Set, and Solvents Model Dependence. *J. Chem. Theory Comput.*, 8(4):1480–1492, apr 2012.

[82] Andrea Frank, Ionut Onila, Heiko M Möller, and Thomas E Exner. Toward the quantum chemical calculation of nuclear magnetic resonance chemical shifts of proteins. *Proteins*, 79(7):2189–202, jul 2011.

[83] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian 09 Revision E.01, 2009. Gaussian Inc. Wallingford CT.

[84] M J Frisch, G W Trucks, H B Schlegel, G E Scuseria, M A Robb, J R Cheeseman, G Scalmani, V Barone, B Mennucci, G A Petersson, H Nakatsuji, M Caricato, X Li, H P Hratchian, A F Izmaylov, J Bloino, G Zheng, J L Sonnenberg, M Hada, M Ehara, K Toyota, R Fukuda, J Hasegawa, M Ishida, T Nakajima, Y Honda, O Kitao, H Nakai, T Vreven, J A Montgomery, Jr., J E Peralta, F Ogliaro, M Bearpark, J J Heyd, E Brothers, K N Kudin, V N Staroverov, R Kobayashi, J Normand, K Raghavachari, A Rendell, J C Burant, S S Iyengar, J Tomasi, M Cossi, N Rega, J M Millam, M Klene, J E Knox, J B Cross, V Bakken, C Adamo, J Jaramillo, R Gomperts, R E Stratmann, O Yazyev, A J Austin, R Cammi, C Pomelli, J W Ochterski, R L Martin, K Morokuma, V G Zakrzewski, G A Voth, P Salvador, J J Dannenberg, S Dapprich, A D Daniels, Ö Farkas, J B Foresman, J V Ortiz, J Cioslowski, and D J Fox. Gaussian09 Revision E.01.

[85] Gregory R Fulmer, Alexander J M Miller, Nathaniel H Sherden, Hugo E Gottlieb, Abraham Nudelman, Brian M Stoltz, John E Bercaw, and Karen I Goldberg. NMR chemical shifts of trace impurities: Common laboratory solvents, organics, and gases in deuterated solvents relevant to the organometallic chemist. *Organometallics*, 29(9):2176–2179, May 2010.

[86] Victor Fung, Jiaxin Zhang, Eric Juarez, and Bobby G Sumpter. Benchmarking graph neural networks for materials chemistry. *npj Computational Materials*, 7(1):1–8, June 2021.

[87] Qi Gao, Satoshi Yokojima, Dmitri G Fedorov, Kazuo Kitaura, Sakurai M, and Nakamura S. Fragment-Molecular-Orbital-Method-Based ab Initio NMR Chemical-Shift

Calculations for Large Molecular Systems. *J. Chem. Theory Comput.*, 6:1428–1444, 2010.

[88] Will Gerrard, Lars A. Bratholm, Martin J. Packer, Adrian J. Mulholland, David R. Glowacki, and Craig P. Butts. IMPRESSION – prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chem. Sci.*, 11(2):508–515, 2020.

[89] P Giannozzi, O Andreussi, T Brumme, O Bunau, M Buongiorno Nardelli, M Calandra, R Car, C Cavazzoni, D Ceresoli, M Cococcioni, N Colonna, I Carnimeo, A Dal Corso, S de Gironcoli, P Delugas, R A DiStasio, A Ferretti, A Floris, G Fratesi, G Fugallo, R Gebauer, U Gerstmann, F Giustino, T Gorni, J Jia, M Kawamura, H-Y Ko, A Kokalj, E Küçükbenli, M Lazzeri, M Marsili, N Marzari, F Mauri, N L Nguyen, H-V Nguyen, A Otero-de-la Roza, L Paulatto, S Poncé, D Rocca, R Sabatini, B Santra, M Schlipf, A P Seitsonen, A Smogunov, I Timrov, T Thonhauser, P Umari, N Vast, X Wu, and S Baroni. Advanced capabilities for materials modelling with Quantum ESPRESSO. *J. Phys. Condens. Mat.*, 29(46):465901, nov 2017.

[90] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, Andrea Dal Corso, Stefano de Gironcoli, Stefano Fabris, Guido Fratesi, Ralph Gebauer, Uwe Gerstmann, Christos Gougoussis, Anton Kokalj, Michele Lazzeri, Layla Martin-Samos, Nicola Marzari, Francesco Mauri, Riccardo Mazzarello, Stefano Paolini, Alfredo Pasquarello, Lorenzo Paulatto, Carlo Sbraccia, Sandro Scandolo, Gabriele Sclauzero, Ari P Seitsonen, Alexander Smogunov, Paolo Umari, and Renata M Wentzcovitch. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Mat.*, 21(39):395502, 2009.

[91] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. April 2017.

[92] Jonathan Godwin, Michael Schaarschmidt, Alexander Gaunt, Alvaro Sanchez-Gonzalez, Yulia Rubanova, Petar Veličković, James Kirkpatrick, and Peter Battaglia. Very deep graph neural networks via noise regularisation. June 2021.

[93] Mark S Gordon, Dmitri G Fedorov, Spencer R Pruitt, and L Slipchenko. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev.*, 112:632–672, aug 2012.

[94] Hugo E Gottlieb, Vadim Kotlyar, and Abraham Nudelman. NMR chemical shifts of common laboratory solvents as trace impurities. *J. Org. Chem.*, 62(21):7512–7515, October 1997.

[95] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.*, 132(15):154104, apr 2010.

[96] Chengyun Guo, Magali B Hickey, Evan R Guggenheim, Volker Enkelmann, and Bruce M Foxman. Conformational polymorphism of methacrylamide. *Chem. Commun.*, (17):2220–2222, May 2005.

[97] U Haeberlen. High resolution NMR in solids. *undefined*, 1976.

[98] Thomas A. Halgren. . *J. Comp. Chem.*, 17:490–519, 1996.

[99] David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Appl. Comput. Harmon. Anal.*, 30(2):129–150, March 2011.

[100] Beomsoo Han, Yifeng Liu, Simon W. Ginzinger, and David S. Wishart. SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR*, 50(1):43–57, may 2011.

[101] Robin K Harris. NMR studies of organic polymorphs & solvates. *The Analyst*, 131(3):351, 2006.

[102] Robin K Harris, Paul Hodgkinson, Chris J Pickard, Jonathan R Yates, and Vadim Zorin. Chemical shift computations on a crystallographic basis: some reflections and comments. *Magn. Reson. Chem.*, 45 Suppl 1:S174–86, December 2007.

[103] Robin K Harris, Siân A Joyce, Chris J Pickard, Sylvian Cadars, and Lyndon Emsley. Assigning carbon-13 NMR spectra to crystal structures by the INADEQUATE pulse sequence and first principles computation: a case study of two forms of testosterone. *Phys. Chem. Chem. Phys.*, 8(1):137–143, January 2006.

[104] Robin K. Harris, Roderick E. Wasylishen, and Melinda J. Duer, editors. *NMR Crystallography*. John Wiley & Sons, West Sussex, UK, 2009.

[105] J. D. Hartman, A. Balaji, and G. J. O. Beran. Improved electrostatic embedding for fragment-based chemical shift calculations in molecular crystals. *J. Chem. Theory Comput.*, 13:6043–6051, 2017.

[106] J. D. Hartman and G. J. O. Beran. Fragment-based electronic structure approach for computing nuclear magnetic resonance chemical shifts in molecular crystals. *J. Chem. Theory Comput.*, 10:4862–4872, 2014.

[107] J. .D. Hartman, G. M. Day, and G. J. O. Beran. Enhanced NMR discrimination of pharmaceutically relevant molecular crystal forms through fragment-based ab initio chemical shift predictions. *Cryst. Growth Des.*, 16:6479–6493, 2016.

[108] J. D. Hartman, R. A. Kudla, G. M. Day, L. J. Mueller, and G. J. O. Beran. Benchmark fragment-based ^{1}H, ^{13}C, ^{15}N and ^{17}O chemical shift predictions in molecular crystals. *Phys. Chem. Chem. Phys.*, 18:21686–21709, 2016.

[109] J. D. Hartman, S. Monaco, B. Schatschneider, and G. J. O. Beran. Fragment-based ^{13}C nuclear magnetic resonance chemical shift predictions in molecular crystals: An alternative to plane-wave methods. *J. Chem. Phys.*, 143:102809, 2015.

[110] J. D. Hartman, T. J. Neubauer, B. G. Caulkins, L. J. Mueller, and G. J. O. Beran. Converging nuclear magnetic shielding calculations with respect to basis and system size in protein systems. *J. Biomol. NMR*, 62:327–340, 2015.

[111] Joshua D. Hartman and Gregory J. O. Beran. Accurate ^{13}C and ^{15}N molecular crystal chemical shielding tensors from fragment-based electronic structure theory. *Solid State Nucl. Magn. Reson.*, 96:10–18, 2018.

[112] Joshua D Hartman, Graeme M Day, and Gregory J O Beran. Enhanced NMR discrimination of pharmaceutically relevant molecular crystal forms through Fragment-Based ab initio chemical shift predictions. *Cryst. Growth Des.*, 16(11):6479–6493, November 2016.

[113] Joshua D Hartman, Ryan A Kudla, Graeme M Day, Leonard J Mueller, and Gregory J O Beran. Benchmark fragment-based (1)h, (13)c, (15)n and (17)o chemical shift predictions in molecular crystals. *Phys. Chem. Chem. Phys.*, 18(31):21686–21709, August 2016.

[114] H Hayashi. Pyridoxal enzymes: mechanistic diversity and uniformity. *J. Biochem.*, 118(3):463–473, September 1995.

[115] Xiao He, Bing Wang, and Kenneth M Merz. Protein NMR chemical shift calculations based on the automated fragmentation QM/MM approach. *J. Phys. Chem. B*, 113(30):10380–8, jul 2009.

[116] Xiao He, Tong Zhu, Xianwei Wang, Jinfeng Liu, and John Z H Zhang. Fragment quantum mechanical calculation of proteins and its applications. *Acc. Chem. Res.*, 47(9):2748–57, sep 2014.

[117] John M. Herbert. Fantasy versus reality in fragment-based quantum chemistry. *J. Chem. Phys.*, 151(17):170901, nov 2019.

[118] Jan Hermann, Zeno Schätzle, and Frank Noé. Deep-neural-network solution of the electronic schrödinger equation. *Nat. Chem.*, pages 1–7, September 2020.

[119] Sean T Holmes, Robbie J Iuliucci, Karl T Mueller, and Cecil Dybowski. Density functional investigation of intermolecular effects on 13C NMR chemical-shielding tensors modeled with molecular clusters. *J. Chem. Phys.*, 141(16):164121, October 2014.

[120] Sean T Holmes, Robbie J Iuliucci, Karl T Mueller, and Cecil Dybowski. Density functional investigation of intermolecular effects on ^{13}C NMR chemical-shielding tensors modeled with molecular clusters. *J. Chem. Phys.*, 141(16):164121, oct 2014.

[121] Sean T Holmes, Robbie J Iuliucci, Karl T Mueller, and Cecil Dybowski. Critical analysis of cluster models and Exchange-Correlation functionals for calculating magnetic shielding in molecular solids. *J. Chem. Theory Comput.*, 11(11):5229–5241, November 2015.

[122] Sean T. Holmes, Robbie J. Iuliucci, Karl T. Mueller, and Cecil Dybowski. Critical Analysis of Cluster Models and Exchange-Correlation Functionals for Calculating Magnetic Shielding in Molecular Solids. *J. Chem. Theory Comput.*, 11(11):5229–5241, 2015.

[123] Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. OGB-LSC: A Large-Scale challenge for machine learning on graphs. March 2021.

[124] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. February 2015.

[125] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, July 2013.

[126] Erin R. Johnson and Gino A. DiLabio. Convergence of calculated nuclear magnetic resonance chemical shifts in a protein with respect to quantum mechanical model size. *J. Mol. Struct. (THEOCHEM)*, 898(1-3):56–61, mar 2009.

[127] K. V. Jovan Jose and Krishnan Raghavachari. Fragment-Based Approach for the Evaluation of NMR Chemical Shifts for Large Biomolecules Incorporating the Effects of the Solvent Environment. *J. Chem. Theory Comput.*, 13(3):1147–1158, mar 2017.

[128] Marie Juramy, Romain Chèvre, Paolo Cerreia Vioglio, Fabio Ziarelli, Eric Besson, Stéphane Gastaldi, Stéphane Viel, Pierre Thureau, Kenneth D M Harris, and Giulia Mollica. Monitoring crystallization processes in confined porous materials by dynamic nuclear polarization Solid-State nuclear magnetic resonance. *J. Am. Chem. Soc.*, 143(16):6095–6103, April 2021.

[129] Jaspreet Kaur and Ajaib S. Brar. An approach to predict the ^{13}C NMR chemical shifts of acrylonitrile copolymers using artificial neural network. *Eur. Polymer J.*, 43(1):156–163, jan 2007.

[130] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. March 2020.

[131] Rika Kobayashi, Roger D. Amos, David M. Reid, and Michael A. Collins. Application of the Systematic Molecular Fragmentation by Annihilation Method to ab Initio NMR Chemical Shift Calculations. *J. Phys. Chem. A*, 122(46):9135–9141, nov 2018.

[132] Kai J Kohlhoff, Paul Robustelli, Andrea Cavalli, Xavier Salvatella, and Michele Vendruscolo. Fast and Accurate Predictions of Protein NMR Chemical Shifts from Interatomic Distances. *J. Am. Chem. Soc.*, 131(39):13894–13895, oct 2009.

[133] Predrag Kukic, Damien Farrell, Lawrence P McIntosh, Bertrand García-Moreno E, Kristine Steen Jensen, Zigmantas Toleikis, Kaare Teilum, and Jens Erik Nielsen. Protein dielectric constants determined from NMR chemical shift perturbations. *J. Am. Chem. Soc.*, 135(45):16968–76, nov 2013.

[134] Andrei G Kutateladze and D Sai Reddy. High-Throughput in silico structure validation and revision of halogenated natural products is enabled by parametric corrections to DFT-Computed 13C NMR chemical shifts and Spin-Spin coupling constants. *J. Org. Chem.*, 82(7):3368–3381, April 2017.

[135] Gitta Kutyniok. The mathematics of artificial intelligence. March 2022.

[136] Jinfeng Lai, Dimitri Niks, Yachong Wang, Tatiana Domratcheva, Thomas R M Barends, Friedrich Schwarz, Ryan A Olsen, Douglas W Elliott, M Qaiser Fatmi, Chia-en A Chang, Ilme Schlichting, Michael F Dunn, and Leonard J. Mueller. X-ray and NMR Crystallography in an Enzyme Active Site: The Indoline Quinonoid Intermediate in Tryptophan Synthase. *J. Am. Chem. Soc.*, 133(1):4–7, jan 2011.

[137] Marcel F Langer, Alex Goeßmann, and Matthias Rupp. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *npj Computational Materials*, 8(1):1–14, March 2022.

[138] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017.

[139] Luc M LeBlanc, Stephen G Dale, Christopher R Taylor, Axel D Becke, Graeme M Day, and Erin R Johnson. Pervasive delocalisation error causes spurious proton transfer in organic Acid-Base Co-Crystals. *Angew. Chem. Int. Ed Engl.*, 57(45):14906–14910, November 2018.

[140] Adrian M. Lee and Ryan P A Bettens. First principles NMR calculations by fragmentation. *J. Phys. Chem. A*, 111(23):5111–5115, 2007.

[141] Malcolm H Levitt. *Spin Dynamics: Basics of Nuclear Magnetic Resonance*. John Wiley & Sons, May 2013.

[142] Lin Li, Chuan Li, Zhe Zhang, and Emil Alexov. On the Dielectric "Constant" of Proteins: Smooth Dielectric Function for Macromolecular Modeling and Its Implementation in DelPhi. *J. Chem. Theory Comput.*, 9(4):2126–2136, apr 2013.

[143] Pengfei Li, Xiangyu Jia, Xiaoliang Pan, Yihan Shao, and Ye Mei. Accelerated Computation of Free Energy Profile at ab Initio Quantum Mechanical/Molecular Mechanics Accuracy via a Semi-Empirical Reference Potential. I. Weighted Thermodynamics Perturbation. *J. Chem. Theory Comput.*, 14(11):5583–5596, 2018.

[144] Hai Lin and Donald G Truhlar. Redistributed charge and dipole schemes for combined quantum mechanical and molecular mechanical calculations. *J. Phys. Chem. A*, 109(17):3991–4004, May 2005.

[145] Hai Lin and Donald G. Truhlar. QM/MM: what have we learned, where are we, and where do we go from here? *Theor. Chem. Acc.*, 117(2):185–199, jul 2006.

[146] Shuai Liu, Jie Li, Kochise Bennett, Brad Ganoe, Tim Stauch, Martin Head-Gordon, Alexander Hexemer, Daniela Ushizima, and Teresa Head-Gordon. A Multi-Resolution 3D-DenseNet for chemical shift prediction in NMR crystallography. *J. Phys. Chem. Lett.*, July 2019.

[147] Shuai Liu, Jie Li, Kochise C. Bennett, Brad Ganoe, Tim Stauch, Martin Head-Gordon, Alexander Hexemer, Daniela Ushizima, and Teresa Head-Gordon. Multiresolution 3D-DenseNet for Chemical Shift Prediction in NMR Crystallography. *J. Phys. Chem. Lett.*, 10(16):4558–4565, aug 2019.

[148] Michael W Lodewyk, Matthew R Siebert, and Dean J Tantillo. Computational Prediction of 1 H and 13 C Chemical Shifts: A Useful Tool for Natural Product, Mechanistic, and Synthetic Organic Chemistry. *Chem. Rev.*, 112(3):1839–1862, mar 2012.

[149] F London. Théorie quantique des courants interatomiques dans les combinaisons aromatiques. *J. Phys. Radium*, 8(10):397–409, October 1937.

[150] Kateřina Maliňáková, Lucie Novosadová, Matej Pipíška, and Radek Marek. Chemical shift tensors in isomers of adenine: relation to aromaticity of purine rings? *Chemphyschem*, 12(2):379–388, February 2011.

[151] Maribel O Marcarino, Marı A M Zanardi, Soledad Cicetti, and Ariel M Sarotti. NMR calculations with quantum methods: Development of new tools for structural elucidation and beyond. *Acc. Chem. Res.*, 53(9):1922–1932, September 2020.

[152] Alberto Marini, Aurora Muñoz-Losa, Alessandro Biancardi, and Benedetta Mennucci. What is solvatochromism? *J. Phys. Chem. B*, 114(51):17128–17135, December 2010.

[153] Phineus R. L. Markwick and Michael Sattler. Site-Specific Variations of Carbonyl Chemical Shift Anisotropies in Proteins. *J. Am. Chem. Soc.*, 126(37):11424–11425, sep 2004.

[154] Osvaldo A. Martin, Jorge A. Vila, and Harold A. Scheraga. CheShift-2: Graphic validation of protein structures. *Bioinformatics*, 28(11):1538–1539, 2012.

[155] Osvaldo A Martin, Jorge A Vila, and Harold A Scheraga. CheShift-2: graphic validation of protein structures. *Bioinformatics*, 28(11):1538–1539, June 2012.

[156] C. Martineau, J. Senker, and F. Taulelle. NMR Crystallography. *Ann. Rep. NMR Spectros.*, 82:1–57, 2014.

[157] Nicholas J Mayhall and Krishnan Raghavachari. Molecules-in-Molecules: An Extrapolated Fragment-Based Approach for Accurate Calculations on Large Molecules and Materials. *J. Chem. Theory Comput.*, 7(5):1336–1343, may 2011.

[158] Nicholas J. Mayhall and Krishnan Raghavachari. Many-Overlapping-Body (MOB) Expansion: A Generalized Many Body Expansion for Nondisjoint Monomers in Molecular Fragmentation Calculations of Covalent Molecules. *J. Chem. Theory Comput.*, 8(8):2669–2675, aug 2012.

[159] David McDonagh, Chris-Kriton Skylaris, and Graeme M. Day. Machine-Learned Fragment-Based Energies for Crystal Structure Prediction. *J. Chem. Theory Comput.*, 15(4):2743–2758, apr 2019.

[160] Jessica L. McKinley and Gregory J. O. Beran. Improving predicted nuclear magnetic resonance chemical shifts using the quasi-harmonic approximation. *J. Chem. Theory Comput.*, 15:5259–5274, 2019.

[161] Erik R McNellis, Jörg Meyer, and Karsten Reuter. Azobenzene at coinage metal surfaces: Role of dispersive van der waals interactions. *Phys. Rev. B Condens. Matter*, 80(20):205414, November 2009.

[162] J Meiler. PROSHIFT: protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR*, 26(1):25–37, 2003.

[163] B. Mennucci, E. Cancès, and J. Tomasi. Evaluation of Solvent Effects in Isotropic and Anisotropic Dielectrics and in Ionic Solutions with a Unified Integral Equation Method: Theoretical Bases, Computational Implementation, and Numerical Applications. *J. Phys. Chem. B*, 101(49):10506–10517, dec 1997.

[164] Benedetta Mennucci. Polarizable continuum model. *WIREs Comput Mol Sci*, 2(3):386–404, May 2012.

[165] Mahmoud Mirzaei and Nasser L Hadipour. An investigation of hydrogen-bonding effects on the nitrogen and hydrogen electric field gradient and chemical shielding tensors in the 9-methyladenine real crystalline structure: a density functional theory study. *J. Phys. Chem. A*, 110(14):4833–4838, April 2006.

[166] Hendrik J Monkhorst and James D Pack. Special points for brillouin-zone integrations. *Phys. Rev. B Condens. Matter*, 13(12):5188–5192, June 1976.

[167] Seongho Moon and David A Case. A comparison of quantum chemical models for calculating NMR shielding parameters in peptides: mixed basis set and ONIOM methods combined with a complete basis set extrapolation. *J. Comp. Chem.*, 27(7):825–36, may 2006.

[168] Robert F Moran, Daniel M Dawson, and Sharon E Ashbrook. Exploiting NMR spectroscopy for the study of disorder in solids. *Int. Rev. Phys. Chem.*, 36(1):39–115, February 2017.

[169] Leonard J. Mueller and Michael F Dunn. NMR Crystallography of Enzyme Active Sites: Probing Chemically Detailed, Three-Dimensional Structure in Tryptophan Synthase. *Acc. Chem. Res.*, 46:2008–2017, 2013.

[170] Stephen Neal, Alex M. Nip, Haiyan Zhang, and David S. Wishart. Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J. Biomol. NMR*, 26(3):215–240, 2003.

[171] Frank Neese. The orca program system. *WIREs Comput. Molec. Sci.*, 2:73–78, 2012.

[172] Frank Neese and Markéta L. Munzarová. Historical Aspects of EPR Parameter Calculations. In Martin Kaupp, Michael Bühl, and Vladimir G Malkin, editors, *Calculation of NMR and EPR Parameters*, chapter 3, pages 21–32. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG, jun 2004.

[173] Jigyasa Nigam, Sergey Pozdnyakov, and Michele Ceriotti. Recursive evaluation and iterative contraction of n-body equivariant features. *J. Chem. Phys.*, 153(12):121101, September 2020.

[174] Sten O Nilsson Lill, Cory M Widdifield, Anna Pettersen, Anna Svensk Ankarberg, Maria Lindkvist, Peter Aldred, Sandra Gracin, Norman Shankland, Kenneth Shankland, Staffan Schantz, and Lyndon Emsley. Elucidating an amorphous form stabilization mechanism for tenapanor hydrochloride: Crystal structure analysis using x-ray diffraction, NMR crystallography, and molecular modeling. *Mol. Pharm.*, 15(4):1476–1487, April 2018.

[175] Sadman Sadeed Omee, Steph-Yves Louis, Nihang Fu, Lai Wei, Sourin Dey, Rongzhi Dong, Qinyang Li, and Jianjun Hu. Scalable deeper graph neural networks for high-performance materials property prediction. *Patterns Prejudice*, page 100491, April 2022.

[176] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[177] Federico M. Paruzzo, Albert Hofstetter, Félix Musil, Sandip De, Michele Ceriotti, and Lyndon Emsley. Chemical shifts in molecular solids by machine learning. *Nature Commun.*, 9(1):4501, dec 2018.

[178] Federico M Paruzzo, Albert Hofstetter, Félix Musil, Sandip De, Michele Ceriotti, and Lyndon Emsley. Chemical shifts in molecular solids by machine learning. *Nat. Commun.*, 9(1):4501, October 2018.

[179] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani,

Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, High-Performance deep learning library. December 2019.

[180] Beate Paulus. The method of increments—a wavefunction-based ab initio correlation method for solids. *Phys. Rep.*, 428(1):1–52, May 2006.

[181] J P Perdew, K Burke, and M Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77(18):3865–3868, October 1996.

[182] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865, 1996.

[183] B Scott Perrin, Jr, Ye Tian, Riqiang Fu, Christopher V Grant, Eduard Y Chekmenev, William E Wieczorek, Alexander E Dao, Robert M Hayden, Caitlin M Burzynski, Richard M Venable, Mukesh Sharma, Stanley J Opella, Richard W Pastor, and Myriam L Cotten. High-resolution structures and orientations of antimicrobial peptides piscidin 1 and piscidin 3 in fluid bilayers reveal tilting, kinking, and bilayer immersion. *J. Am. Chem. Soc.*, 136(9):3491–3504, March 2014.

[184] Chris Pickard and Francesco Mauri. All-electron magnetic response with pseudopotentials: NMR chemical shifts. *Phys. Rev. B*, 63(24):245101, may 2001.

[185] Chris J Pickard and Francesco Mauri. All-electron magnetic response with pseudopotentials: NMR chemical shifts. *Phys. Rev. B Condens. Matter*, 63(24):245101, May 2001.

[186] Radek Pohl, Ondřej Socha, Michal Šála, Dominik Rejman, and Martin Dračínský. The control of the tautomeric equilibrium of isocytosine by intermolecular interactions. *European J. Org. Chem.*, 2018(37):5128–5135, October 2018.

[187] Radek Pohl, Ondřej Socha, Michal Šála, Dominik Rejman, and Martin Dračínský. The control of the tautomeric equilibrium of isocytosine by intermolecular interactions. *European J. Org. Chem.*, 2018(37):5128–5135, October 2018.

[188] E M Purcell, H C Torrey, and R V Pound. Resonance absorption by nuclear magnetic moments in a solid. *Phys. Rev.*, 69(1-2):37–38, January 1946.

[189] I I Rabi, J R Zacharias, S Millman, and P Kusch. A new method of measuring nuclear magnetic moment. *Phys. Rev.*, 53(4):318–318, February 1938.

[190] Krishnan Raghavachari and Arjun Saha. Accurate Composite and Fragment-Based Quantum Chemical Models for Large Molecules. *Chem. Rev.*, 115:5643–5677, 2015.

[191] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data*, 1:140022, August 2014.

[192] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.*, 11(5):2087–2096, may 2015.

[193] Raghunathan Ramakrishnan and O Anatole von Lilienfeld. Machine learning, quantum chemistry, and chemical space. In Abby L Parrill and Kenny B Lipkowitz, editors, *Reviews in Computational Chemistry*, volume 432 of *Reviews in Computational Chemistry*, pages 225–256. John Wiley & Sons, Inc., Hoboken, NJ, USA, April 2017.

[194] David M. Reid and Michael A. Collins. Approximating CCSD(T) nuclear magnetic shielding calculations using composite methods. *J. Chem. Theory Comput.*, 11(11):5177–5181, 2015.

[195] David M. Reid and Michael A. Collins. Calculating nuclear magnetic resonance shieldings using systematic molecular fragmentation by annihilation. *Phys. Chem. Chem. Phys.*, 17(7):5314–5320, 2015.

[196] Bernd Reif, Sharon E Ashbrook, Lyndon Emsley, and Mei Hong. Solid-state NMR spectroscopy. *Nature Reviews Methods Primers*, 1(1):2, January 2021.

[197] Ryan M Richard and John M Herbert. A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory. *J. Chem. Phys.*, 137(6):064113, aug 2012.

[198] Ryan M Richard, Ka Un Lao, and John M Herbert. Aiming for benchmark accuracy with the many-body expansion. *Acc. Chem. Res.*, 47(9):2828–36, sep 2014.

[199] Ryan M. Richard, Ka Un Lao, and John M. Herbert. Understanding the many-body expansion for large systems. I. Precision considerations. *J. Chem. Phys.*, 141(1):014108, jul 2014.

[200] Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.*, 52(11):2864–2875, nov 2012.

[201] Matthias Rupp, Raghunathan Ramakrishnan, and O. Anatole von Lilienfeld. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J. Phys. Chem. Lett.*, 6(16):3309–3313, aug 2015.

[202] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.*, 108(5):058301, jan 2012.

[203] Elodie Salager, Graeme M Day, Robin S Stein, Chris J Pickard, Benedicte Elena, and Lyndon Emsley. Powder Crystallography by Combined Crystal Structure Prediction and High-Resolution 1 H Solid-State NMR Spectroscopy. *J. Am. Chem. Soc.*, 132(8):2564–2566, mar 2010.

[204] Sérgio M Santos, João Rocha, and Luís Mafra. NMR crystallography: Toward chemical Shift-Driven crystal structure determination of the β-Lactam antibiotic amoxicillin trihydrate. *Cryst. Growth Des.*, 13(6):2390–2395, June 2013.

[205] Jonathan Schmidt, Love Pettersson, Claudio Verdozzi, Silvana Botti, and Miguel A L Marques. Crystal graph attention networks for the prediction of stable materials. *Sci Adv*, 7(49):eabi7948, December 2021.

[206] Karen L Schuchardt, Brett T Didier, Todd Elsethagen, Lisong Sun, Vidhya Gurumoorthi, Jared Chase, Jun Li, and Theresa L Windus. Basis set exchange: a community database for computational sciences. *J. Chem. Inf. Model.*, 47(3):1045–1052, May 2007.

[207] Kristof T Schütt, Pieter-Jan Kindermans, Huziel E Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. June 2017.

[208] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. SchNet–A deep learning architecture for molecules and materials. *J. Chem. Phys.*, 148(24):241722, 2018.

[209] C N Schutz and A Warshel. What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins*, 44(4):400–17, sep 2001.

[210] Martin Seifrid, G N Manjunatha Reddy, Bradley F Chmelka, and Guillermo C Bazan. Insight into the structures and dynamics of organic semiconductors through solid-state NMR spectroscopy. *Nature Reviews Materials*, 5(12):910–930, September 2020.

[211] Lin Shen, Jingheng Wu, and Weitao Yang. Multiscale Quantum Mechanics/Molecular Mechanics Simulations with Neural Networks. *J. Chem. Theory Comput.*, 12(10):4934–4946, oct 2016.

[212] Yang Shen and Ad Bax. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J. Biomol. NMR*, 38(4):289–302, 2007.

[213] Yang Shen and Ad Bax. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J. Biomol. NMR*, 38(4):289–302, August 2007.

[214] Yang Shen and Ad Bax. SPARTA+: A modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR*, 48(1):13–22, 2010.

[215] Yang Shen, Frank Delaglio, Gabriel Cornilescu, and Ad Bax. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR*, 44(4):213–223, aug 2009.

[216] Andrew E. Sifain, Nicholas Lubbers, Benjamin T. Nebgen, Justin S. Smith, Andrey Y. Lokhov, Olexandr Isayev, Adrian E. Roitberg, Kipton Barros, and Sergei Tretiak. Discovering a Transferable Charge Assignment Model Using Machine Learning. *J. Phys. Chem. Lett.*, 9:4495–4501, 2018.

[217] Ewa Skorupska, Sławomir Kaźmierski, and Marek J Potrzebowski. Solid state NMR characterization of Ibuprofen:Nicotinamide cocrystals and new idea for controlling release of drugs embedded into mesoporous silica particles. *Mol. Pharm.*, 14(5):1800–1810, May 2017.

[218] J. C. Slater. A simplification of the Hartree-Fock Method. *Phys. Rev.*, 81:385, 1951.

[219] Daniel G A Smith, Doaa Altarawy, Lori A Burns, Matthew Welborn, Levi N Naden, Logan Ward, Sam Ellis, Benjamin P Pritchard, and T Daniel Crawford. The MolSSI QCA rchive project: An open-source platform to compute, organize, and share quantum chemistry data. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 11(2), March 2021.

[220] J S Smith, O Isayev, and A E Roitberg. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.*, 8(4):3192–3203, April 2017.

[221] J. S. Smith, O. Isayev, and A. E. Roitberg. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.*, 8(4):3192–3203, 2017.

[222] Justin S. Smith, Olexandr Isayev, and Adrian E. Roitberg. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific Data*, 4:170193, dec 2017.

[223] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. ANI-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci Data*, 4:170193, December 2017.

[224] Justin S Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E Roitberg. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.*, 148(24):241733, June 2018.

[225] Justin S. Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E. Roitberg. Less is more: sampling chemical space with active learning. *J. Chem. Phys.*, 148(24), jan 2018.

[226] Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian E Roitberg. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.*, 10(1):2903, July 2019.

[227] Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian E Roitberg.

Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature Commun.*, 10(1):2903, dec 2019.

[228] Ondřej Socha, Paul Hodgkinson, Cory M Widdifield, Jonathan R Yates, and Martin Dračínský. Exploring systematic discrepancies in DFT calculations of chlorine nuclear quadrupole couplings. *J. Phys. Chem. A*, 121(21):4103–4113, June 2017.

[229] J. F. Stanton, J. Gauss, L. Cheng, M. E. Harding, D. A. Matthews, and P. G. Szalay. CFOUR, Coupled-Cluster techniques for Computational Chemistry, a quantum-chemical program package. With contributions from A. Asthana, A.A. Auer, R.J. Bartlett, U. Benedikt, C. Berger, D.E. Bernholdt, S. Blaschke, Y. J. Bomble, S. Burger, O. Christiansen, D. Datta, F. Engel, R. Faber, J. Greiner, M. Heckert, O. Heun, M. Hilgenberg, C. Huber, T.-C. Jagau, D. Jonsson, J. Jusélius, T. Kirsch, M.-P. Kitsaras, K. Klein, G.M. Kopper, W.J. Lauderdale, F. Lipparini, J. Liu, T. Metzroth, L.A. Mück, D.P. O'Neill, T. Nottoli, J. Oswald, D.R. Price, E. Prochnow, C. Puzzarini, K. Ruud, F. Schiffmann, W. Schwalbach, C. Simmons, S. Stopkowicz, A. Tajti, J. Vázquez, F. Wang, J.D. Watts, C. Zhang, X. Zheng, and the integral packages MOLECULE (J. Almlöf and P.R. Taylor), PROPS (P.R. Taylor), ABACUS (T. Helgaker, H.J. Aa. Jensen, P. Jørgensen, and J. Olsen), and ECP routines by A. V. Mitin and C. van Wüllen. For the current version, see http://www.cfour.de.

[230] Eugene V. Stefanovich and Thanh N. Truong. A Simple Method for Incorporating Madelung Field Effects into ab Initio Embedded Cluster Calculations of Crystals and Macromolecules. *J. Phys. Chem. B*, 102(16):3018–3022, apr 1998.

[231] Casper Steinmann, Lars Andersen Bratholm, Jógvan Magnus Haugaard Olsen, and Jacob Kongsted. Automated Fragmentation Polarizable Embedding Density Functional Theory (PE-DFT) Calculations of Nuclear Magnetic Resonance (NMR) Shielding Constants of Proteins with Application to Chemical Shift Predictions. *J. Chem. Theory Comput.*, 13(2):525–536, feb 2017.

[232] Casper Steinmann, Mikael W. Ibsen, Anne S. Hansen, and Jan H. Jensen. FragIt: A Tool to Prepare Input Files for Fragment Based Quantum Chemical Calculations. *PLoS ONE*, 7(9):e44480, sep 2012.

[233] Casper Steinmann, Jógvan Magnus Haugaard Olsen, and Jacob Kongsted. Nuclear Magnetic Shielding Constants from Quantum Mechanical/Molecular Mechanical Calculations Using Polarizable Embedding: Role of the Embedding Potential. *J. Chem. Theory Comput.*, 10(3):981–988, mar 2014.

[234] R M Stevens, R M Pitzer, and W N Lipscomb. Perturbed Hartree—Fock calculations. i. magnetic susceptibility and shielding in the LiH molecule. *J. Chem. Phys.*, 38(2):550–560, January 1963.

[235] Hermann Stoll. The correlation energy of crystalline silicon. *Chem. Phys. Lett.*, 191(6):548–552, April 1992.

[236] Dirk Stueber. The Embedded Ion Method: A New Approach to the Electrostatic Description of Crystal Lattice Effects in Chemical Shielding Calculations. *Conc. Magn. Reson. A*, 28(January 2006):347–368, 2006.

[237] Dirk Stueber and David M Grant. 13C and (15)n chemical shift tensors in adenosine, guanosine dihydrate, 2'-deoxythymidine, and cytidine. *J. Am. Chem. Soc.*, 124(35):10539–10551, September 2002.

[238] Dirk Stueber, Flavien N. Guenneau, and David M. Grant. The calculation of ^{13}C chemical shielding tensors in ionic compounds utilizing point charge arrays obtained from Ewald lattice sums. *J. Chem. Phys.*, 114(21):9236–9243, 2001.

[239] Geng Sun and Philippe Sautet. Toward Fast and Reliable Potential Energy Surfaces for Metallic Pt Clusters by Hierarchical Delta Neural Networks. *J. Chem. Theory Comput.*, 15(10):5614–5627, oct 2019.

[240] Jason Swails, Tong Zhu, Xiao He, and David A. Case. AFNMR: automated fragmentation quantum mechanical calculation of NMR chemical shifts for biomolecules. *J. Biomol. NMR*, 63(2):125–139, oct 2015.

[241] Hwee-Jia Tan and Ryan P A Bettens. Ab initio NMR chemical-shift calculations based on the combined fragmentation method. *Phys. Chem. Chem. Phys.*, 15(20):7541–7, may 2013.

[242] Sishi Tang and David A Case. Calculation of chemical shift anisotropy in proteins. *J. Biomol. NMR*, 51(3):303–12, nov 2011.

[243] Alexandre Tkatchenko and Matthias Scheffler. Accurate molecular van der waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.*, 102(7):073005, February 2009.

[244] N Tollenaar and P G M van der Heijden. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *J. R. Stat. Soc. Ser. A Stat. Soc.*, 176(2):565–584, February 2013.

[245] Michael D Toney. Reaction specificity in pyridoxal phosphate enzymes. *Arch. Biochem. Biophys.*, 433(1):279–287, January 2005.

[246] Michael D. Toney. Controlling reaction specificity in pyridoxal phosphate enzymes. *Biochim. Biophys. Acta*, 1814(11):1407–1418, nov 2011.

[247] Oliver T Unke and Markus Meuwly. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.*, 15(6):3678–3693, June 2019.

[248] Pablo A. Unzueta and Gregory J. O. Beran. Polarizable continuum models provide an effective electrostatic embedding model for fragment-based chemical shift prediction in challenging systems. *J. Comp. Chem.*, 41:2251–2265, 2020.

[249] Pablo A. Unzueta, Chandler Greenwell, and Gregory J. O. Beran. Predicting density functional theory-quality nuclear magnetic resonance chemical shifts via Δ-machine learning. *J. Chem. Theory Comput.*

[250] Pablo A Unzueta, Chandler S Greenwell, and Gregory J O Beran. Predicting density functional Theory-Quality nuclear magnetic resonance chemical shifts via Δ-Machine learning. *J. Chem. Theory Comput.*, 17(2):826–840, February 2021.

[251] Andrea Victora, Heiko M. Möller, and Thomas E. Exner. Accurate ab initio prediction of NMR chemical shifts of nucleic acids and nucleic acids/protein complexes. *Nucl. Acids Res.*, 42(22):1–10, 2014.

[252] Jorge A Vila and Harold A Scheraga. Assessing the accuracy of protein structures by quantum mechanical computations of $^{13}C^{\alpha}$ chemical shifts. *Acc. Chem. Res.*, 42(10):1545–53, oct 2009.

[253] S. H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: A critical analysis. *Can. J. Phys.*, 58:1200–11, 1980.

[254] Bo Wang and Donald G Truhlar. Tuned and balanced redistributed charge scheme for combined quantum mechanical and molecular mechanical (QM/MM) methods and fragment methods: Tuning based on the CM5 charge model. *J. Chem. Theory Comput.*, 9(2):1036–1042, February 2013.

[255] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Peng Cui, P Yu, and Yanfang Ye. Heterogeneous graph attention network. March 2019.

[256] Logan Ward, Ben Blaiszik, Ian Foster, Rajeev S. Assary, Badri Narayanan, and Larry Curtiss. Machine Learning Prediction of Accurate Atomization Energies of Organic Molecules from Low-Fidelity Quantum Chemical Calculations. *MRS Commun.*, 9(3):891–899, jun 2019.

[257] F. Weigend. Accurate Coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.*, 8:1057–1065, 2006.

[258] F. Weigend. Hartree-Fock exchange fitting basis sets for H to Rn. *J. Comput. Chem.*, 29:167–175, 2008.

[259] Simon Wengert, Gábor Csányi, Karsten Reuter, and Johannes T Margraf. Data-efficient machine learning for molecular crystal structure prediction. *Chem. Sci.*, November 2021.

[260] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.

[261] Keith W Wiitala, Christopher J Cramer, and Thomas R Hoye. Comparison of various density functional methods for distinguishing stereoisomers based on computed (1)h or (13)c NMR chemical shifts using diastereomeric penam beta-lactams as a test set. *Magn. Reson. Chem.*, 45(10):819–829, October 2007.

[262] Patrick H Willoughby, Matthew J Jansma, and Thomas R Hoye. A guide to small-molecule structure assignment through computation of (^{1}h and ^{13}c) NMR chemical shifts. *Nat. Protoc.*, 9(3):643–660, March 2014.

[263] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE Trans Neural Netw Learn Syst*, 32(1):4–24, January 2021.

[264] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120(14):145301, April 2018.

[265] X P Xu and D A Case. Automated prediction of 15N, 13Ca, 13Cb, and 13C$'$ chemical shifts in proteins using a density functional database. *J. Biomol. NMR*, 21:321–333, 2001.

[266] X P Xu and D A Case. Automated prediction of 15n, 13calpha, 13cbeta and 13c' chemical shifts in proteins using a density functional database. *J. Biomol. NMR*, 21(4):321–333, December 2001.

[267] C. Yang, L. Zhu, R. A. Kudla, J. D. Hartman, R. O. Al-Kaysi, S. Monaco, B. Schatschneider, A. Magalhaes, G. J. O. Beran, C. J. Bardeen, and L. J. Mueller. Crystal structure of the meta-stable intermediate in the photomechanical, crystal-to-crystal reaction of 9-tertbutyl anthracene ester. *CrystEngComm*, 18:7319–7329, 2016.

[268] X Yang, T C Ong, V K Michaelis, S Heng, J Huang, R G Griffin, and A S Myerson. Formation of organic molecular nanocrystals under rigid confinement with analysis by solid state NMR. *CrystEngComm*, 16(39):9345–9352, October 2014.

[269] Ziyue Yang, Maghesree Chakraborty, and Andrew D White. Predicting chemical shifts with graph neural networks. August 2020.

[270] Jonathan R Yates, Chris J Pickard, and Francesco Mauri. Calculation of NMR chemical shifts for extended systems using ultrasoft pseudopotentials. *Phys. Rev. B Condens. Matter*, 76(2):024401, July 2007.

[271] Chaohui Ye, Riqiang Fu, Jianzhi Hu, Lei Hou, and Shangwu Ding. Carbon-13 chemical shift anisotropies of solid amino acids. *Magn. Reson. Chem.*, 31(8):699–704, August 1993.

[272] Robert P. Young, Bethany G. Caulkins, Dan Borchardt, Daryl N. Bulloch, Cynthia K. Larive, Michael F. Dunn, and Leonard J. Mueller. Solution-State ^{17}O Quadrupole

Central-Transition NMR Spectroscopy in the Active Site of Tryptophan Synthase. *Angew. Chem. Int. Ed.*, 55(4):1350–1354, 2016.

[273] Peter Zaspel, Bing Huang, Helmut Harbrecht, and O. Anatole Von Lilienfeld. Boosting Quantum Machine Learning Models with a Multilevel Combination Technique: Pople Diagrams Revisited. *J. Chem. Theory Comput.*, 15(3):1546–1559, 2019.

[274] Anmin Zheng, Minghui Yang, Yong Yue, Chaohui Ye, and Feng Deng. ^{13}C NMR shielding tensors of carboxyl carbon in amino acids calculated by ONIOM method. *Chem. Phys. Lett.*, 399(1-3):172–176, nov 2004.

[275] Peikun Zheng, Wudi Yang, Wei Wu, Olexandr Isayev, and Pavlo O Dral. Toward chemical accuracy in predicting enthalpies of formation with General-Purpose Data-Driven methods. *J. Phys. Chem. Lett.*, 13(15):3479–3491, April 2022.

[276] Tong Zhu, Xiao He, and John Z H Zhang. Fragment density functional theory calculation of NMR chemical shifts for proteins with implicit solvation. *Phys. Chem. Chem. Phys.*, 14(21):7837–45, jun 2012.

[277] J. Zienau, J. Kussmann, and C. Ochsenfeld. Quantum-chemical simulation of solid-state NMR spectra: The example of a molecular tweezer host-guest complex. *Mol. Phys.*, 108:333–342, 2010.

Appendix A

Improving the Accuracy of Solid-State Nuclear Magnetic Resonance Chemical Shift Prediction With a Simple Molecular Correction

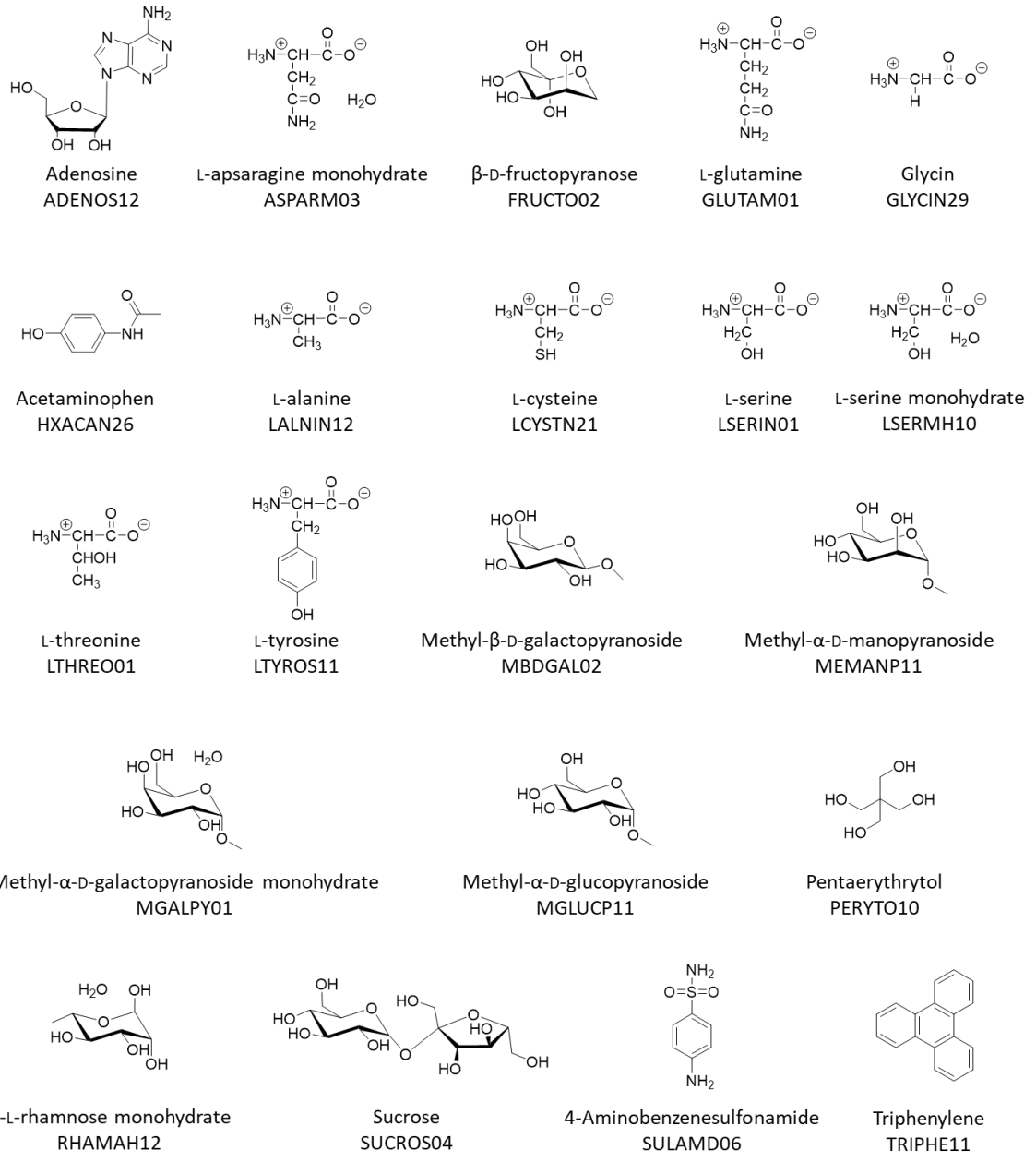


Figure A.1: Crystal structures and corresponding CSD reference codes included in the ^{13}C benchmark set.

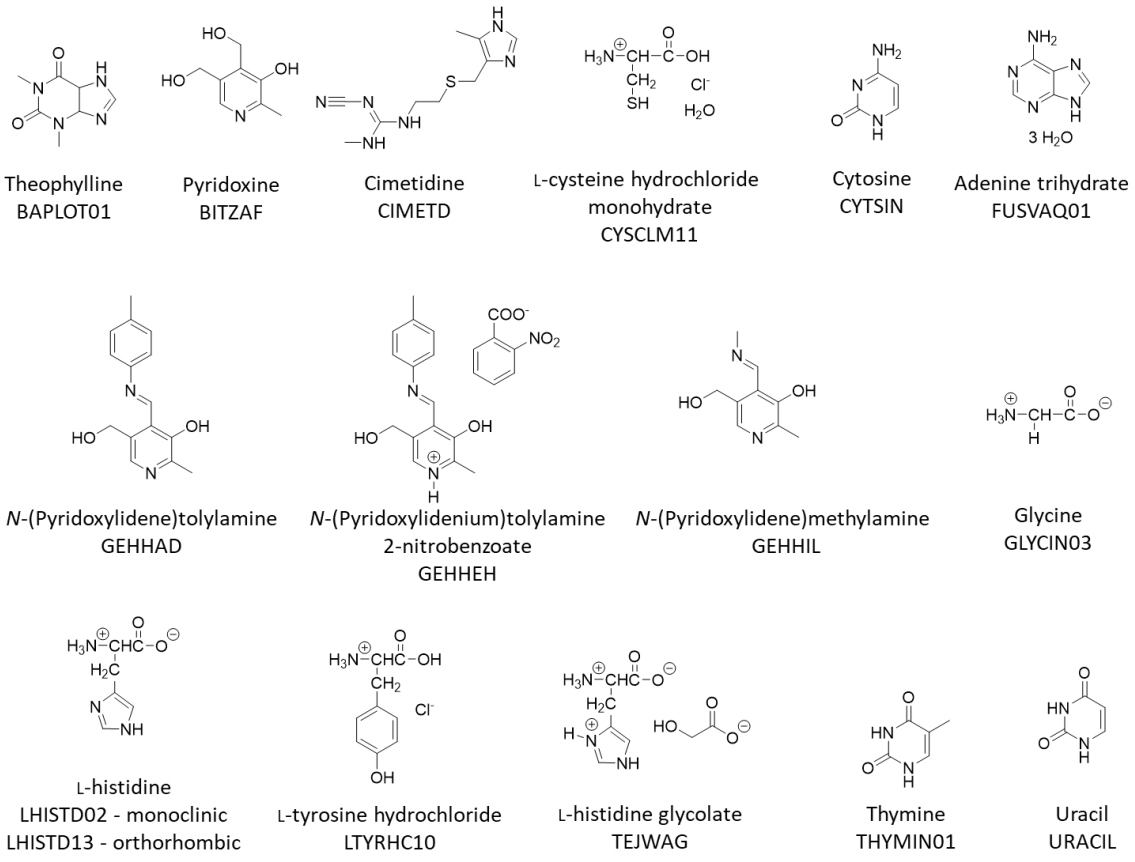


Figure A.2: Crystal structures and corresponding CSD reference codes included in the ^{15}N benchmark set.

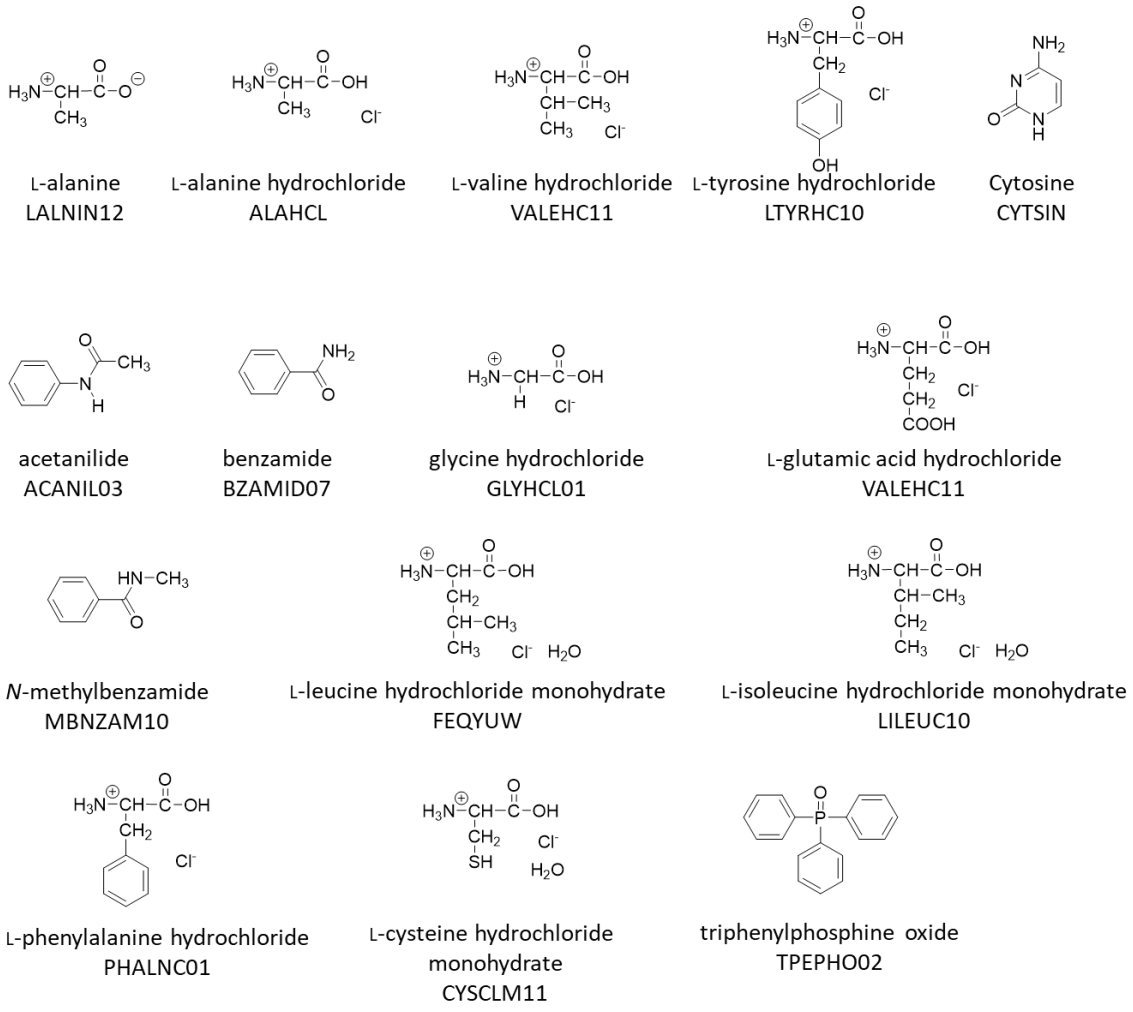


Figure A.3: Crystal structures and corresponding CSD reference codes included in the ^{17}O benchmark set.

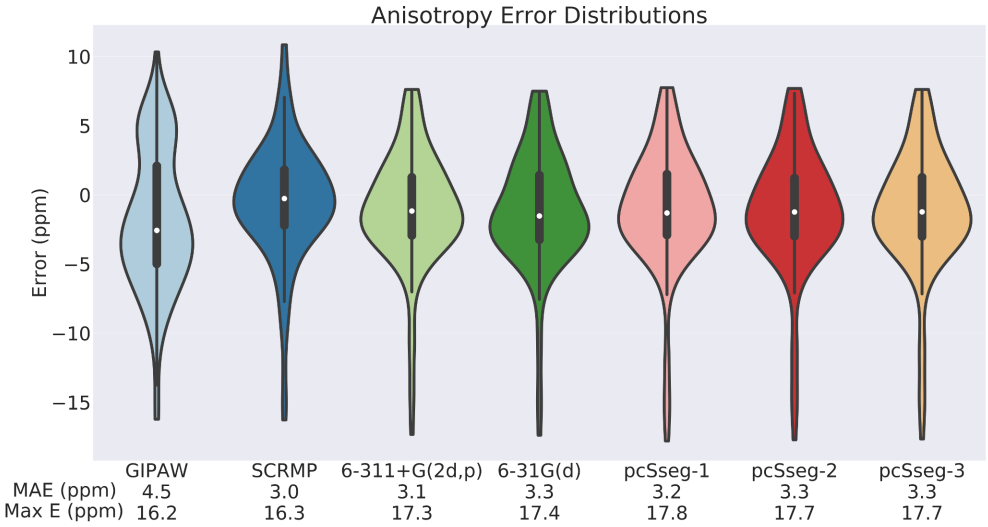


Figure A.4: Errors in reproducing the experimental ^{13}C anisotropy calculated from the principal components.

Anisotropy (note here that δ is not the chemical shift, but is the anisotropy which comes from the original paper [97]):

$$\delta = \frac{3}{2}(\sigma_{zz} - \sigma_{iso}) \tag{A.1}$$

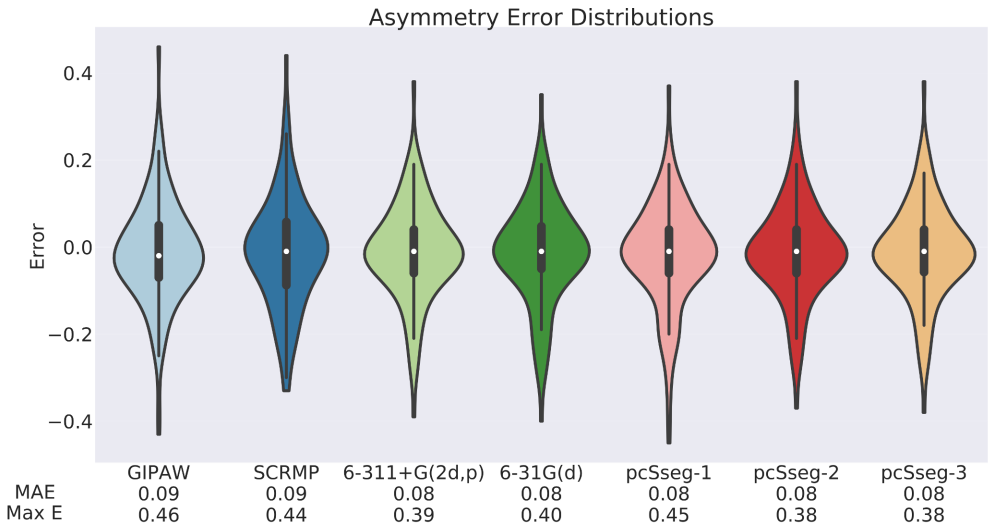


Figure A.5: Errors in reproducing the experimental ^{13}C asymmetry calculated from the principal components.

Asymmetry

$$\eta = \frac{\sigma_{yy} - \sigma_{xx}}{\delta} \tag{A.2}$$

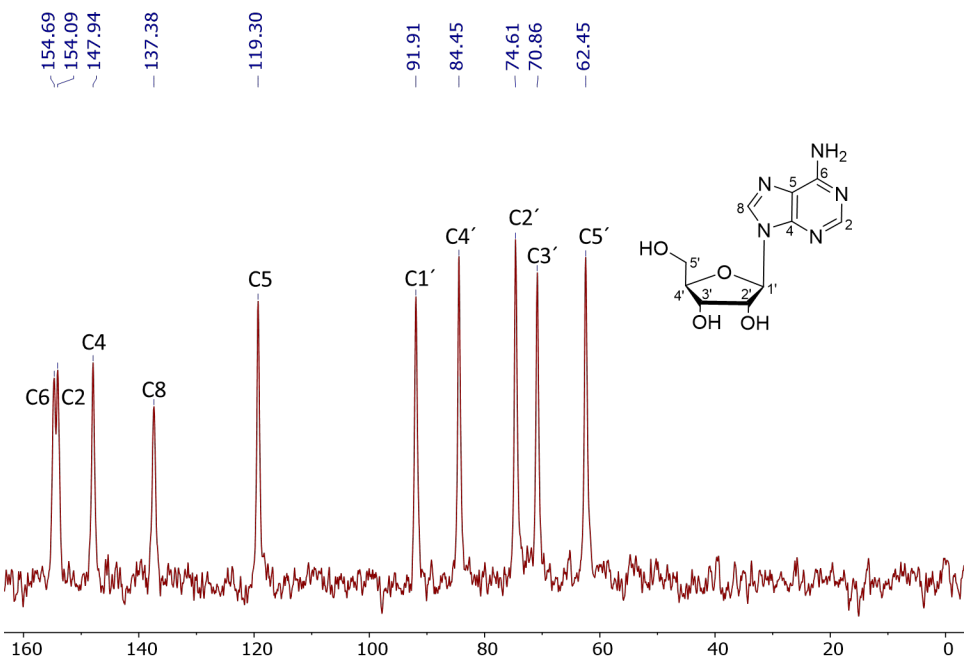


Figure A.6: ^{13}C CP-MAS spectrum of adenosine.

Appendix B

Polarizable Continuum Models Provide an Effective Electrostatic Embedding Model for Fragment-Based Chemical Shift Prediction in Challenging Systems

B.1 Piscidin-1

Tables B.1 and B.2 present the absolute shieldings for the fragment models and the errors relative to the non-fragmented calculation, with and without PCM embedding.

Table B.1: Errors in reproducing full piscidin-1 NMR isotropic shieldings **with PCM embedding**.

		Model					Errors			
XYZ order	Atom	Cluster	1AA	3AA	4.5AA	9AA	1AA	3AA	4.5AA	9AA
					^{13}C Shieldings					
1	C	5.48	7.14	6.62	5.90	5.79	-1.65	-1.14	-0.41	-0.30
4	C	126.31	126.26	126.43	126.44	126.44	0.05	-0.12	-0.14	-0.14
5	C	157.13	157.08	157.07	157.06	157.09	0.05	0.06	0.07	0.04
6	C	36.00	35.93	35.86	35.97	36.10	0.07	0.14	0.03	-0.10
8	C	66.80	66.85	66.91	66.89	66.74	-0.05	-0.11	-0.10	0.06
9	C	50.95	50.50	50.67	50.85	51.00	0.45	0.29	0.10	-0.04
18	C	9.54	9.72	9.45	9.37	9.28	-0.18	0.09	0.17	0.26
21	C	122.32	122.31	122.21	122.20	122.20	0.02	0.12	0.12	0.12
22	C	157.99	157.57	157.84	157.85	157.83	0.42	0.15	0.14	0.15
23	C	166.26	166.23	166.27	166.28	166.28	0.03	-0.01	-0.02	-0.03
24	C	166.81	166.90	166.86	166.88	166.87	-0.10	-0.05	-0.07	-0.06
34	C	10.33	11.09	10.32	10.19	10.09	-0.76	0.00	0.14	0.24
37	C	142.17	142.17	142.11	142.09	142.08	0.00	0.06	0.09	0.10
41	C	10.70	11.66	11.03	10.62	10.43	-0.96	-0.33	0.07	0.26
44	C	126.70	126.33	126.61	126.60	126.66	0.38	0.10	0.10	0.05
45	C	156.16	155.87	156.09	156.25	156.23	0.29	0.06	-0.10	-0.07
46	C	162.03	161.91	162.04	162.03	162.03	0.12	-0.01	-0.01	0.00
47	C	155.70	155.53	155.66	155.70	155.73	0.17	0.03	-0.01	-0.04
48	C	142.76	142.69	142.73	142.72	142.75	0.07	0.04	0.04	0.01
						RMSE	**0.51**	**0.29**	**0.13**	**0.14**
					^{1}H Shieldings					
11	H	24.42	24.34	24.50	24.46	24.43	0.08	-0.08	-0.03	0.00
12	H	28.05	28.00	28.04	28.04	28.05	0.05	0.01	0.01	0.00
13	H	28.16	28.14	28.14	28.12	28.13	0.01	0.02	0.04	0.03
14	H	28.87	28.86	28.88	28.88	28.90	0.01	0.00	-0.01	-0.03
15	H	23.84	23.89	23.97	23.90	23.88	-0.05	-0.13	-0.06	-0.05
16	H	23.98	23.95	23.95	23.96	23.97	0.04	0.03	0.03	0.01
17	H	23.90	23.93	23.92	23.87	23.88	-0.03	-0.02	0.02	0.02
25	H	24.94	24.98	25.01	24.96	24.95	-0.04	-0.07	-0.02	-0.01
26	H	28.53	28.57	28.58	28.56	28.56	-0.04	-0.05	-0.03	-0.03
27	H	29.15	29.23	29.15	29.13	29.13	-0.08	0.00	0.02	0.01
28	H	31.03	31.03	31.06	31.05	31.05	0.00	-0.03	-0.02	-0.02
29	H	30.87	30.88	30.89	30.89	30.89	-0.01	-0.02	-0.02	-0.02
30	H	31.15	31.14	31.12	31.12	31.12	0.01	0.03	0.03	0.03
31	H	30.06	30.06	30.05	30.05	30.04	0.00	0.01	0.01	0.02
32	H	31.65	31.61	31.64	31.65	31.65	0.04	0.01	0.00	0.00
33	H	31.09	31.10	31.08	31.10	31.11	-0.01	0.01	-0.01	-0.02
38	H	25.10	25.23	25.15	25.09	25.09	-0.13	-0.04	0.01	0.02
39	H	27.72	27.75	27.76	27.72	27.72	-0.03	-0.04	0.00	0.00
40	H	28.22	28.28	28.24	28.19	28.20	-0.06	-0.01	0.03	0.02
50	H	24.84	25.00	24.91	24.82	24.80	-0.15	-0.06	0.03	0.05
51	H	28.04	28.20	28.10	28.04	28.06	-0.15	-0.06	0.01	-0.02
52	H	29.47	29.55	29.52	29.50	29.50	-0.08	-0.05	-0.03	-0.03
53	H	30.56	30.55	30.54	30.51	30.55	0.01	0.02	0.05	0.01
54	H	31.03	31.02	31.01	30.99	31.01	0.01	0.02	0.04	0.02
55	H	29.80	29.85	29.82	29.76	29.78	-0.06	-0.02	0.03	0.02
56	H	30.96	30.97	30.96	30.94	30.95	-0.01	0.00	0.02	0.01
57	H	30.52	30.53	30.53	30.52	30.53	0.00	-0.01	0.01	0.00
58	H	28.83	28.84	28.84	28.81	28.83	-0.01	-0.01	0.01	0.00
59	H	28.79	28.86	28.82	28.78	28.79	-0.07	-0.04	0.01	0.00
60	H	27.36	27.38	27.37	27.35	27.35	-0.02	0.00	0.01	0.01
61	H	26.19	26.21	26.19	26.17	26.19	-0.01	0.00	0.02	0.00
62	H	26.38	26.39	26.39	26.37	26.38	0.00	0.00	0.02	0.01
						RMSE	**0.06**	**0.04**	**0.03**	**0.02**
					^{15}N Shieldings					
3	N	124.37	123.37	124.38	124.71	124.65	1.01	-0.01	-0.34	-0.27
7	N	-44.80	-44.87	-44.67	-45.01	-44.80	0.07	-0.13	0.21	0.00
10	N	87.84	88.58	88.21	88.07	87.86	-0.74	-0.37	-0.23	-0.02
20	N	121.96	121.98	122.23	122.01	121.94	-0.02	-0.27	-0.05	0.02
36	N	137.24	138.38	137.44	137.37	137.29	-1.13	-0.20	-0.12	-0.04
43	N	117.83	119.03	118.89	118.38	118.10	-1.20	-1.05	-0.55	-0.27
49	N	207.96	207.97	208.01	208.01	208.06	-0.01	-0.05	-0.05	-0.10
						RMSE	**0.78**	**0.44**	**0.27**	**0.15**
					^{17}O Shieldings					
2	O	-52.53	-57.02	-54.64	-53.00	-52.48	4.49	2.11	0.47	-0.05
19	O	-31.97	-35.18	-32.24	-31.03	-30.56	3.21	0.28	-0.93	-1.40
35	O	-43.38	-51.72	-42.73	-42.14	-41.61	8.34	-0.65	-1.24	-1.77
42	O	-39.37	-48.90	-42.05	-40.31	-39.18	9.53	2.68	0.95	-0.19
						RMSE	**6.91**	**1.74**	**0.94**	**1.13**

Table B.2: Errors in reproducing full piscidin-1 NMR isotropic shieldings **without PCM embedding**.

			Model				Errors			
XYZ order	Atom	Cluster	1AA	3AA	4.5AA	9AA	1AA	3AA	4.5AA	9AA
					^{13}C Shieldings					
1	C	5.45	7.68	7.46	6.67	6.38	-2.23	-2.01	-1.23	-0.93
4	C	125.77	125.88	126.08	125.99	125.90	-0.10	-0.31	-0.21	-0.13
5	C	157.07	157.17	156.99	156.82	156.86	-0.10	0.08	0.25	0.21
6	C	29.80	28.21	28.71	28.71	29.01	1.59	1.09	1.10	0.80
8	C	70.78	71.36	71.64	71.48	71.07	-0.58	-0.86	-0.70	-0.30
9	C	55.43	54.79	54.33	55.08	55.33	0.64	1.10	0.36	0.11
18	C	12.19	12.69	13.03	12.75	12.42	-0.50	-0.84	-0.56	-0.23
21	C	121.75	121.88	121.90	121.74	121.66	-0.13	-0.15	0.01	0.09
22	C	157.73	157.26	157.45	157.43	157.49	0.46	0.27	0.30	0.24
23	C	166.19	166.30	166.24	166.28	166.27	-0.11	-0.05	-0.09	-0.08
24	C	166.51	166.66	166.53	166.58	166.53	-0.15	-0.02	-0.07	-0.01
34	C	11.62	12.22	12.67	12.22	11.92	-0.60	-1.05	-0.60	-0.30
37	C	142.42	142.46	142.51	142.43	142.39	-0.04	-0.09	-0.01	0.03
41	C	11.99	12.55	12.58	12.49	12.04	-0.57	-0.60	-0.50	-0.05
44	C	128.25	128.48	128.45	128.33	128.30	-0.23	-0.20	-0.08	-0.05
45	C	156.19	155.99	155.73	156.04	156.02	0.20	0.45	0.15	0.17
46	C	162.24	162.21	162.14	162.21	162.22	0.03	0.10	0.04	0.02
47	C	154.07	153.56	153.51	153.64	153.75	0.51	0.56	0.44	0.32
48	C	143.57	141.32	141.32	141.33	141.43	2.26	2.25	2.24	2.14
						RMSE	**0.89**	**0.89**	**0.71**	**0.59**
					^{1}H Shieldings					
11	H	24.55	24.39	24.76	24.66	24.59	0.16	-0.22	-0.11	-0.04
12	H	28.43	28.44	28.44	28.42	28.45	-0.01	-0.01	0.01	-0.02
13	H	28.07	28.09	28.15	28.03	28.02	-0.03	-0.09	0.03	0.05
14	H	28.64	28.57	28.53	28.53	28.58	0.08	0.12	0.12	0.06
15	H	23.78	23.85	24.01	23.90	23.87	-0.08	-0.23	-0.12	-0.09
16	H	24.17	24.18	24.15	24.15	24.17	-0.01	0.02	0.02	0.00
17	H	24.68	24.88	24.79	24.84	24.83	-0.20	-0.11	-0.15	-0.14
25	H	25.01	25.13	25.29	25.19	25.14	-0.12	-0.28	-0.17	-0.12
26	H	28.57	28.54	28.51	28.53	28.55	0.03	0.06	0.04	0.02
27	H	29.18	29.38	29.28	29.25	29.23	-0.21	-0.10	-0.08	-0.05
28	H	30.98	30.93	30.90	30.92	30.95	0.05	0.08	0.06	0.03
29	H	30.78	30.78	30.72	30.74	30.76	0.01	0.07	0.05	0.03
30	H	31.00	31.00	31.00	30.97	30.97	0.01	0.01	0.03	0.04
31	H	30.21	30.34	30.34	30.30	30.27	-0.13	-0.13	-0.09	-0.06
32	H	31.56	31.46	31.45	31.47	31.50	0.10	0.10	0.09	0.05
33	H	31.02	31.07	31.02	31.01	31.01	-0.05	0.00	0.01	0.01
38	H	25.45	25.69	25.74	25.62	25.58	-0.24	-0.29	-0.17	-0.13
39	H	27.53	27.51	27.44	27.44	27.46	0.03	0.09	0.09	0.07
40	H	28.38	28.46	28.45	28.40	28.38	-0.07	-0.06	-0.01	0.00
50	H	25.23	25.36	25.40	25.34	25.27	-0.13	-0.16	-0.10	-0.04
51	H	28.05	28.14	28.10	27.99	28.03	-0.09	-0.05	0.06	0.02
52	H	29.38	29.41	29.37	29.39	29.39	-0.03	0.01	-0.01	-0.01
53	H	30.78	30.70	30.72	30.66	30.72	0.08	0.06	0.12	0.06
54	H	31.07	30.96	30.96	30.92	30.96	0.11	0.11	0.14	0.11
55	H	29.53	29.60	29.55	29.49	29.49	-0.07	-0.02	0.04	0.04
56	H	30.86	30.91	30.89	30.88	30.89	-0.05	-0.03	-0.02	-0.03
57	H	30.68	30.61	30.64	30.62	30.65	0.07	0.03	0.06	0.03
58	H	28.84	28.68	28.70	28.68	28.71	0.16	0.14	0.16	0.13
59	H	28.49	28.61	28.57	28.51	28.51	-0.12	-0.09	-0.03	-0.02
60	H	28.01	27.67	27.66	27.65	27.66	0.34	0.34	0.36	0.35
61	H	27.05	26.56	26.52	26.50	26.51	0.49	0.53	0.55	0.54
62	H	27.05	26.61	26.65	26.63	26.65	0.45	0.40	0.43	0.40
						RMSE	**0.17**	**0.18**	**0.16**	**0.15**
					^{15}N Shieldings					
3	N	126.48	126.24	127.72	127.56	127.33	0.24	-1.24	-1.08	-0.84
7	N	-69.35	-71.27	-70.45	-72.20	-71.52	1.93	1.10	2.85	2.17
10	N	94.74	101.77	100.25	101.02	100.36	-7.02	-5.50	-6.27	-5.61
20	N	118.79	118.54	119.33	118.86	118.67	0.26	-0.54	-0.07	0.12
36	N	139.51	141.16	141.55	140.95	140.71	-1.65	-2.04	-1.44	-1.20
43	N	124.79	125.81	127.15	126.90	126.40	-1.02	-2.36	-2.11	-1.61
49	N	201.35	205.96	205.95	206.00	206.12	-4.60	-4.59	-4.64	-4.77
						RMSE	**3.34**	**3.03**	**3.31**	**3.02**
					^{17}O Shieldings					
2	O	-63.00	-68.04	-72.47	-69.18	-66.97	5.05	9.47	6.18	3.97
19	O	-38.55	-44.13	-46.24	-43.02	-40.37	5.59	7.69	4.48	1.82
35	O	-56.18	-61.99	-66.61	-62.28	-60.08	5.81	10.42	6.10	3.90
42	O	-66.40	-71.25	-70.08	-74.06	-71.12	4.85	3.67	7.66	4.72
						RMSE	**5.34**	**8.23**	**6.21**	**3.76**

B.2 Molecular Crystal Benchmarks

Crystal Structures: The set of 47 crystal structures, optimized geometries, and experimental chemical shifts for the molecular crystal benchmarks were taken directly from ref [70]. For convenience, the list of species and Cambridge Structure Database Reference Codes is provided below in Tables B.3–B.5. Note that a few crystals occur in more than one of the sets.

Two-Body Cutoff: The chemical shifts presented here computed two-body contributions for all molecules for which any atom lies within 4 Å of an atom in the central fragment/asymmetric unit. To demonstrate convergence of the chemical shifts with respect to this cutoff, Table B.6 shows that the root-mean-square error in the ^{15}N chemical shifts for 4, 6, and 8 Å cutoffs are virtually identical, 4.04–4.06 ppm. The average individual shift change upon increasing the cutoff from 4 Å to 6 Å or 8 Å is 0.13–0.14 ppm, with the maximum change of 0.34 ppm. These results are similar to what has been found in earlier fragment studies with alternate electrostatic embedding environments[109, 108, 105] and show that the 4 Å cutoff is suitable.

Linear Regression Models: Figure B.1 shows a sample linear regression plot for converting the predicted absolute chemical shieldings to experimentally observable chemical shifts for the ^{13}C molecular crystal test set. As can be seen from this plot, the fitted slope is close to the ideal value of -1. Table B.7 lists the root-mean-square (rms) errors versus experiment and linear regression parameters associated with the predicted chemical shifts for all models and nucleus types. Aside from a few poorly-performing models (e.g. some of the "No Embedding" or PCM models with very low dielectric constants) that are not

recommended, most linear regression slopes in this table lie within ±5%,of the ideal value of -1, as recommended in Ref [148]. The slopes deviate slightly more (6–8%) from unity for the ^{17}O chemical shift models, but this is true for all models, including GIPAW. It reflects the general difficulty associated with predicting ^{17}O chemical shifts.

Table B.3: 21 molecular crystals contained in the ^{13}C test set.

RefCode	**Species**
ADENOS12	Adenosine
ASPARM03	L-Asparagine monohydrate
FRUCTO02	β-D-Fructopyranose
GLUTAM01	L-Glutamine
GLYCIN03	Glycine (α polymorph)
HXACAN26	Acetaminophen (form I)
LALNIN12	L-Alanine
LCYSTN21	L-Cysteine (form I)
LSERIN01	L-Serine (form I)
LSERMH10	L-Serine monohydrate
LTHREO01	L-Threonine
LTYROS11	L-Tyrosine
MBDGAL02	Methyl β-D-galactopyranoside
MEMANP11	Methyl-D-mannopyranoside
MGALPY01	Methyl α-D-galactopyranoside monohydrate
MGLUCP11	Methyl α-D-glucopyranoside
PERYTO10	Pentaerythritol
RHAMAH12	α-L-Rhamnose monohydrate
SUCROS04	Sucrose
SULAMD06	Sulfanilamide (β polymorph)
TRIPHE11	Triphenylene

Table B.4: 16 molecular crystals contained in the ^{15}N test set.

RefCode	**Species**
BAPLOT01	Theophylline (form II)
BITZAF	Pyridoxine
CIMETD	Cimetidine (monoclinic A polymorph)
CYSCLM11	L-cysteine hydrochloride monohydrate
CYTSIN	Cytosine ($P2_12_12_1$ polymorph)
FUSVAQ01	Adenine trihydrate
GEHHAD	N-(Pyridoxylidene)tolylamine
GEHHEH	N-(Pyridoxylidenium)tolylamine 2-nitrobenzoate
GEHHIL	N-(Pyridoxylidene)methylamine
GLYCIN03	Glycine (α polymorph)
LHISTD02	L-Histidine (monoclinic polymorph)
LHISTD13	L-Histidine (orthorhombic polymorph)
LTYRHC10	L-Tyrosine hydrochloride
TEJWAG	L-Histidine glycolate
THYMIN01	Thymine
URACIL	Uracil

Table B.5: 15 molecular crystals contained in the ^{17}O test set.

RefCode	**Species**
ACANIL03	Acetanilide
ALAHCL	L-Alanine hydrochloride
BZAMID07	Benzamide (monoclinic polymorph)
CYSCLM11	L-cysteine hydrochloride monohydrate
CYTSIN	Cytosine ($P2_12_12_1$ polymorph)
FEQYUW	L-Leucine hydrochloride monohydrate
GLYHCL01	Glycine hydrochloride
LALNIN12	L-Alanine
LGLUTA03	L-Glutamic acid hydrochloride
LILEUC10	L-Isoleucine hydrochloride monohydrate
LTYRHC10	L-Tyrosine hydrochloride
MBNZAM10	N-Methylbenzamide
PHALNC01	L-Phenylalanine hydrochloride
TPEPHO02	Triphenylphosphine oxide (monoclinic polymorph)
VALEHC11	L-Valine hydrochloride

Table B.6: Convergence of the ^{15}N chemical shifts with respect to the two-body cutoff. Both the RMSE versus experiment and the maximum absolute change in a chemical shift in the set are listed.

Cutoff	RMSE vs Experiment	Max Change vs. 4 Å cutoff
4 Å	4.04 ppm	
6 Å	4.06 ppm	0.34 ppm
8 Å	4.05 ppm	0.34 ppm

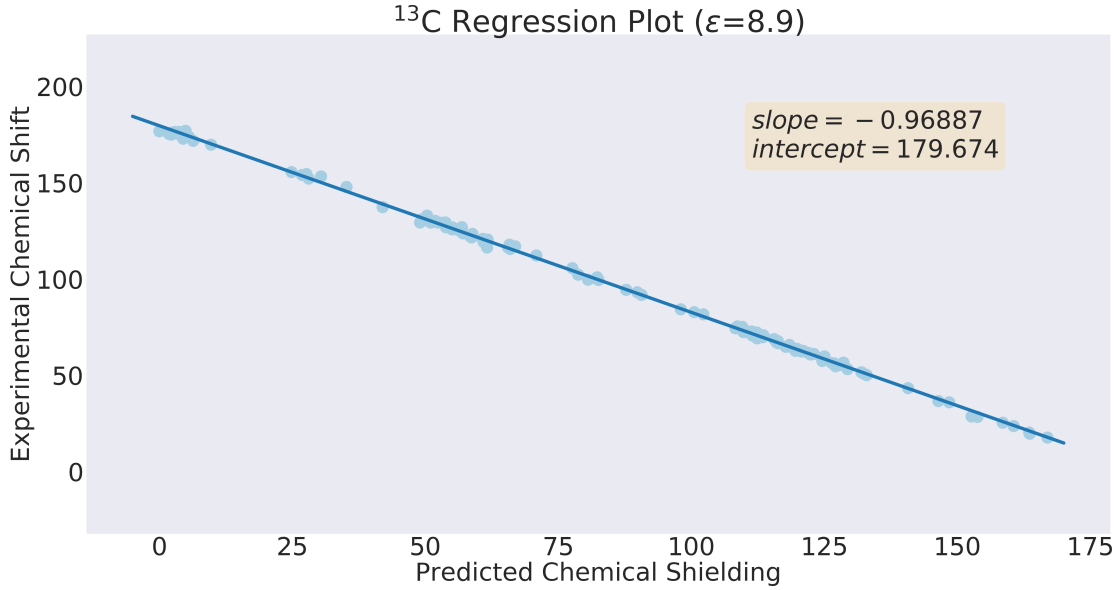


Figure B.1: A sample linear regression mapping the predicted absolute chemical shieldings σ_i to the experimentally observed chemical shifts δ_i via $\delta_i = a\sigma_i + b$. The data shown represents the 1+2-body fragment approach embedded in a PCM with dielectric $\epsilon = 8.9$ for the ^{13}C molecular crystal set.

Table B.7: Regression parameters used to convert isotropic chemical shieldings to chemical shifts by atom type. Regression parameters are generated by least-squares fitting ($\delta_{iso} = m\sigma_{iso} + b$).

Atom	Model	RMSE (ppm)	Slope	Intercept
Carbon	GIPAW	2.11	-0.99243	169.226
	SCRMP	1.28	-0.96433	179.290
	2-Body Frag ϵ=181.6	1.20	-0.96549	179.348
	2-Body Frag ϵ=78.4	1.17	-0.96647	179.478
	2-Body Frag ϵ=24.9	1.18	-0.96685	179.484
	2-Body Frag ϵ=8.9	1.18	-0.96887	179.674
	2-Body Frag ϵ=6.3	1.19	-0.96981	179.755
	2-Body Frag ϵ=1.4	1.43	-0.97320	179.896
	2-Body Frag No Embedding	1.44	-0.97375	179.838
	Cluster/Frag ϵ=8.9	1.26	-0.96295	179.176
	SCRMP Cluster/Frag	1.26	-0.96343	179.197
Nitrogen	GIPAW	5.64	-1.03030	185.120
	SCRMP	4.12	-1.02192	197.619
	2-Body Frag ϵ=181.6	4.17	-1.03380	197.830
	2-Body Frag ϵ=78.4	4.04	-1.03234	197.596
	2-Body Frag ϵ=24.9	4.03	-1.03298	197.630
	2-Body Frag ϵ=8.9	3.95	-1.03257	197.366
	2-Body Frag ϵ=6.3	3.99	-1.03293	197.268
	2-Body Frag ϵ=1.4	15.53	-1.05342	200.970
	2-Body Frag No Embedding	38.84	-0.91139	198.382
	Cluster/Frag ϵ=8.9	4.51	-1.01600	197.789
	SCRMP Cluster/Frag	4.39	-1.01519	197.652
Oxygen	GIPAW	7.20	-1.06627	248.302
	SCRMP	7.47	-1.03196	268.502
	2-Body Frag ϵ=181.6	9.50	-1.07867	271.118
	2-Body Frag ϵ=78.4	9.54	-1.07792	270.887
	2-Body Frag ϵ=24.9	9.77	-1.07523	270.095
	2-Body Frag ϵ=8.9	10.77	-1.06922	268.508
	2-Body Frag ϵ=6.3	11.64	-1.06575	267.722
	2-Body Frag ϵ=1.4	22.90	-1.02773	264.093
	2-Body Frag No Embedding	83.44	-0.13993	252.291
	Cluster/Frag ϵ=8.9	7.95	-1.05116	271.635
	SCRMP Cluster/Frag	7.17	-1.03611	271.107

B.3 Indoline Carbanionic Substrate of Tryptophan Synthase

The indoline carbanionic intermediate cluster and various fragments are provided separately in XYZ format. Figure B.2 shows the larger fragments used in the second, less aggressive fragmentation scheme for the indoline carbanionic intermediate in tryptophan synthase. For each of the models presented in Table 1 of the main paper, Tables B.8–B.13 report the predicted chemical shifts for the Phenolic Oxygen and Schiff Base Nitrogen tautomers as well as the shifts that result from the two-site mixing. Note that in the χ_r^2 calculations, the single site calculations have 13 degrees of freedom, while the two-site exchange models have 12 degrees of freedom, since one degree of freedom is involving in fitting the optimal mixture.

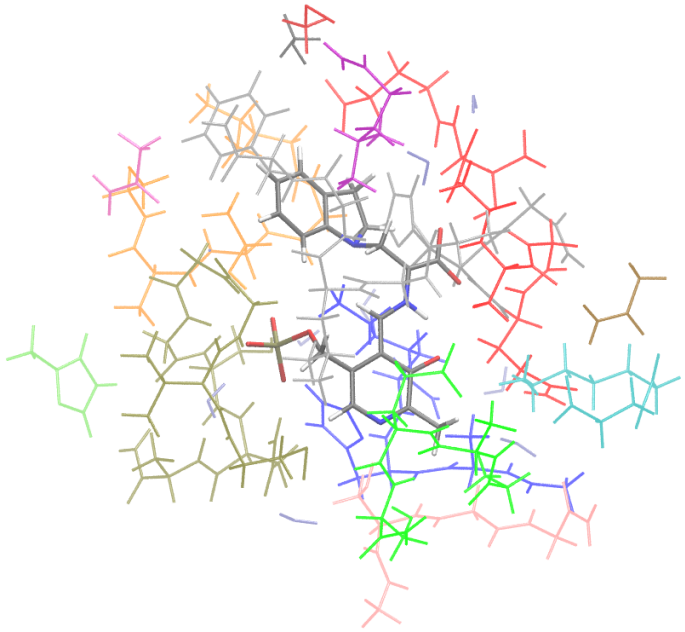


Figure B.2: Structure of the indoline substrate bound in the cluster of tryptophan synthase. Coloring indicates the fragments used in the larger fragmentation scheme.

Table B.8: Chemical shifts for the indoline carbanionic intermediate bound in tryptophan synthase using the 1+2-body fragment PBE0 model with single-amino acid fragments and a $\epsilon = 181.6$ dielectric environment.

PBE0, Single Amino Acid Fragments ($\epsilon = 181.6$)				
	Phenolic O	**Schiff Base N**	**2-Site Exchange**	**Experiment**
PLP N1	258.0	275.8	263.3	265.0
PLP C2	139.7	147.2	141.9	145.4
PLP C2′	18.5	20.5	19.1	17.0
PLP C3	147.6	162.6	152.1	154.1
PLP P	0.0	0.0	0.0	0.0
Schiff Base N	327.6	204.3	290.6	296.0
Serine Cα	108.1	94.1	103.9	103.5
Serine C′	171.7	168.5	170.7	173.0
Serine Cβ	51.1	51.3	51.2	54.1
Serine O1	268.9	254.4	262.1	243.0
Serine O2	257.1	246.1	253.8	233.0
Indoline N1	77.7	76.7	77.4	83.5
Indoline C2	49.3	49.4	49.4	50.5
Indoline C3	29.4	29.3	29.3	28.5
χ^2_r	11.76	49.14	3.15	(70% Phenolic O)

Table B.9: Chemical shifts for the indoline carbanionic intermediate bound in tryptophan synthase using the 1+2-body fragment PBE0 model with single-amino acid fragments and a $\epsilon = 78.4$ dielectric environment.

PBE0, Single Amino Acid Fragments ($\epsilon = 78.4$)				
	Phenolic O	**Schiff Base N**	**2-Site Exchange**	**Experiment**
PLP N1	257.8	275.7	263.1	265.0
PLP C2	139.7	147.3	142.0	145.4
PLP C2′	18.5	20.4	19.1	17.0
PLP C3	147.8	162.8	152.3	154.1
PLP P	0.0	0.0	0.0	0.0
Schiff Base N	327.4	204.1	290.4	296.0
Serine Cα	108.1	94.0	103.8	103.5
Serine C′	171.8	168.6	170.9	173.0
Serine Cβ	51.2	51.3	51.2	54.1
Serine O1	269.2	254.7	264.8	243.0
Serine O2	256.3	245.0	252.9	233.0
Indoline N1	78.0	77.0	77.7	83.5
Indoline C2	49.3	49.4	49.4	50.5
Indoline C3	29.4	29.3	29.4	28.5
χ^2_r	12.18	52.63	3.26	(70% Phenolic O)

Table B.10: Chemical shifts for the indoline carbanionic intermediate bound in tryptophan synthase using the 1+2-body fragment PBE0 model with single-amino acid fragments and a $\epsilon = 24.9$ dielectric environment.

PBE0, Single Amino Acid Fragments ($\epsilon = 24.9$)				
	Phenolic O	**Schiff Base N**	**2-Site Exchange**	**Experiment**
PLP N1	258.3	276.4	263.7	
PLP C2	139.5	147.3	141.9	
PLP C2′	18.6	20.5	19.1	
PLP C3	147.9	162.9	152.4	
PLP P	0.0	0.0	0.0	
Schiff Base N	328.3	204.1	291.0	
Serine Cα	107.8	93.6	103.5	
Serine C′	171.7	168.6	170.8	
Serine Cβ	51.2	51.4	51.2	
Serine O1	269.9	255.4	261.5	
Serine O2	253.5	241.8	250.0	
Indoline N1	78.8	77.9	78.6	
Indoline C2	49.3	49.4	49.3	
Indoline C3	29.4	29.4	29.4	
χ^2_r	12.04	53.17	2.92	(70% Phenolic O)

Table B.11: Chemical shifts for the indoline carbanionic intermediate bound in tryptophan synthase using the 1+2-body fragment PBE0 model with single-amino acid fragments and a $\epsilon = 8.9$ dielectric environment.

PBE0, Single Amino Acid Fragments ($\epsilon = 8.9$)				
	Phenolic O	**Schiff Base N**	**2-Site Exchange**	**Experiment**
PLP N1	259.7	278.0	265.0	265.0
PLP C2	139.6	147.8	142.0	145.4
PLP C2′	18.6	20.5	19.2	17.0
PLP C3	148.4	163.4	152.8	154.1
PLP P	0.0	0.0	0.0	0.0
Schiff Base N	329.3	203.4	292.8	296.0
Serine Cα	107.3	92.9	103.1	103.5
Serine C′	171.8	168.6	170.9	173.0
Serine Cβ	51.2	51.5	51.3	54.1
Serine O1	271.0	256.4	266.7	243.0
Serine O2	248.9	236.4	245.3	233.0
Indoline N1	79.7	78.8	79.5	83.5
Indoline C2	49.4	49.5	49.4	50.5
Indoline C3	29.6	29.5	29.6	28.5
χ^2_r	11.63	57.15	2.63	(71% Phenolic O)

Table B.12: Chemical shifts for the indoline carbanionic intermediate bound in tryptophan synthase using the 1+2-body fragment PBE0 model with larger fragments and a $\epsilon = 8.9$ dielectric environment.

	PBE0, Larger Fragments ($\epsilon = 8.9$)			
	Phenolic O	**Schiff Base N**	**2-Site Exchange**	**Experiment**
PLP N1	262.9	279.6	267.4	265.0
PLP C2	141.5	150.5	143.9	145.4
PLP C2′	18.8	20.6	19.3	17.0
PLP C3	149.0	163.8	153.0	154.1
PLP P	0.0	0.0	0.0	0.0
Schiff Base N	327.9	200.9	293.6	296.0
Serine Cα	107.0	92.9	103.2	103.5
Serine C′	172.1	169.1	171.3	173.0
Serine Cβ	51.1	51.2	51.1	54.1
Serine O1	257.3	244.2	253.8	243.0
Serine O2	235.5	224.2	232.4	233.0
Indoline N1	78.5	77.6	78.2	83.5
Indoline C2	49.5	49.7	49.5	50.5
Indoline C3	29.6	29.5	29.6	28.5
χ^2_r	9.18	61.01	1.66	(73% Phenolic O)

Table B.13: Chemical shifts for the indoline carbanionic intermediate bound in tryptophan synthase using PBE0 on the full cluster with no fragmentation or embedding.

	PBE0, Full Cluster—No Embedding			
	Phenolic O	**Schiff Base N**	**2-Site Exchange**	**Experiment**
PLP N1	261.7	276.3	265.1	265.0
PLP C2	142.8	153.3	145.3	145.4
PLP C2′	19.5	21.1	19.8	17.0
PLP C3	149.8	165.2	153.5	154.1
PLP P	0.0	0.0	0.0	0.0
Schiff Base N	326.4	199.7	296.6	296.0
Serine Cα	106.5	91.7	103.0	103.5
Serine C′	171.3	168.3	170.6	173.0
Serine Cβ	51.0	51.0	51.0	54.1
Serine O1	254.0	241.4	251.1	243.0
Serine O2	227.1	215.4	224.3	233.0
Indoline N1	82.6	81.8	82.4	83.5
Indoline C2	50.5	50.2	50.4	50.5
Indoline C3	26.7	26.4	26.6	28.5
χ^2_r	6.53	57.67	1.26	(77% Phenolic O)

Appendix C

Predicting Density Functional Theory-Quality Nuclear Magnetic Resonance Chemical Shifts via Δ-Machine Learning

C.1 Structures Excluded From the ANI-1 Dataset for Training

The ANI-1 data set,[222] which consists of 57,462 molecules with 1–8 heavy atoms were used in training the neural network. The following six structures from this data set

gdb11_s08-2840_110
gdb11_s08-3312_11
gdb11_s08-8831_10
gdb11_s08-8864_3
gdb11_s08-8865_16
gdb11_s08-8866_3

were excluded due to corrupt geometries (e.g. incorrect number of saturating hydrogen atoms or atom contacts that are too close):

C.2 Sample Gaussian 09 Input File

The following represents a sample Gaussian input file used to optimized the geometries and compute the target NMR PBE0/6-311+G(2d,p) chemical shieldings in this work.

```
# NMR PBE1PBE/6-311+G(2d,p)

gdb11_s01-0_3425_nmr.com

0 1
C -0.000230 0.000011 -0.000191
H -0.989833 -0.409461 -0.216012
H -0.049803 1.091247 0.002853
H 0.707056 -0.330315 -0.764535
H 0.333961 -0.351538 0.978839
```

The inexpensive shielding calculations substituted the functional and/or basis set as appropriate, e.g. for PBE/6-31G shieldings:

```
# NMR PBEPBE/6-31G
```

The geometry optimizations employed the keywords:

```
# Opt wB97X/6-31G(d)
```

C.3 GDB17 Subset for Testing

As described in the main text, 3,780 molecules were randomly sampled containing only C, N, and O heavy atoms from the GDB17 data set[200]. SMILES strings for these structures can be found at `https://pubs.acs.org/doi/suppl/10.1021/acs.jctc.0c00979/suppl_file/ct0c00979_si_002.txt`

C.4 Chemical Shift Referencing

The regression parameters used for each model (pure DFT, AEV, and Δ-ML) are summarized below, as fitted by linear regression. The molecules used for the DMSO and $CDCl_3$ sets are shown in Figures C.1 and **??**. Note that the DMSO regression is much smaller, and thus less robust to predicting chemical shifts, but can still yield insight to the predictive power of the models evaluated in the text.

Table C.1: Regression parameters generated for experimental chemical shifts in DMSO and $CDCl_3$

	DMSO			$CDCl_3$		
Model	Slope	Intercept	RMSE	Slope	Intercept	RMSE
PBE0/6-311+G(2d,p)	-0.98913	183.75493	**2.42**	-0.97576	181.61750	**1.82**
PBE0/6-31G	-1.06509	209.40611	**2.12**	-1.03636	205.07289	**2.00**
PBE0/6-31G + Δ-ML	-0.98913	183.75493	**2.35**	-0.97648	181.75494	**1.80**
PBE0/STO-3G	-1.37311	308.88292	**11.96**	-1.38253	310.24774	**11.00**
PBE0/STO-3G+ Δ-ML	-0.98761	183.68978	**2.34**	-0.98197	182.82668	**3.16**
PBE/6-31G	-1.11847	213.10406	**3.14**	-1.08683	208.06852	**2.79**
PBE/6-31G + Δ-ML	-0.99156	184.09948	**2.41**	-0.97870	182.20060	**2.46**
PBE/STO-3G	-1.44312	317.35788	**13.00**	-1.45558	318.90679	**11.30**
PBE/STO-3G + Δ-ML	-0.99008	183.94935	**2.33**	-0.98441	183.13795	**3.37**
SVWN/6-31G	-1.07519	207.81702	**3.30**	-1.04517	202.83638	**2.91**
SVWN/6-31G + Δ-ML	-0.99144	184.09397	**2.43**	-0.97877	182.15682	**2.35**
SVWN/STO-3G	-1.37583	307.14032	**13.38**	-1.08919	253.12459	**25.09**
SVWN/STO-3G + Δ-ML	-0.99039	184.07443	**2.38**	-0.98544	183.27617	**3.35**
AEV	-0.98586	184.27320	**2.27**	-0.99222	184.51469	**4.98**

C.4.1 DMSO and $CDCl_3$ Regression Set

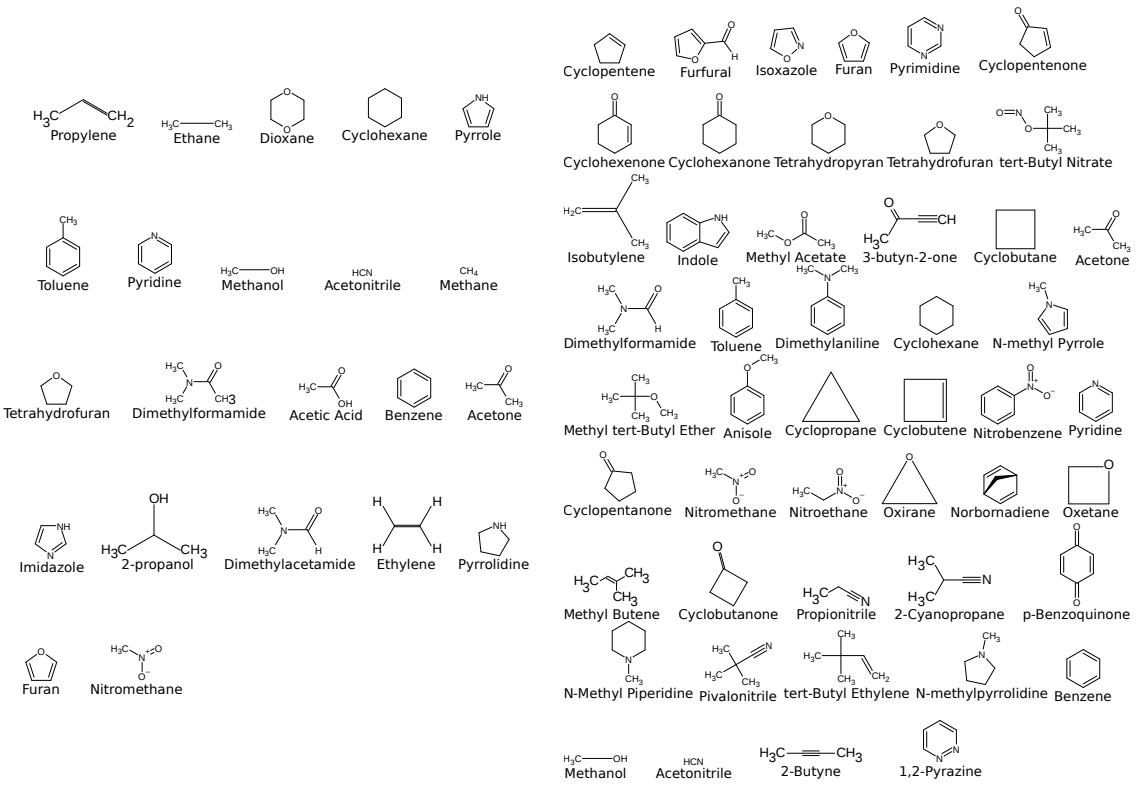


Figure C.1: (*left*) Molecules used for DMSO regression set. (*right*) Molecules used for $CDCl_3$ regression set.

C.5 Pharmaceutical Molecules and Predicted Experimental Chemical Shifts

C.5.1 Acetaminophen (HXACAN14) in DMSO

Assignment	Experiment	PBE0 6-311+G(2d,p)	PBE0 6-31G	PBE0 6-31G$^{\Delta}$	PBE0 STO-3G	PBE0 STO-3G$^{\Delta}$	PBE 6-31G	PBE 6-31G$^{\Delta}$	PBE STO-3G	PBE STO-3G$^{\Delta}$	SVWN 6-31G	SVWN 6-31G$^{\Delta}$	SVWN STO-3G	SVWN STO-3G$^{\Delta}$
1	167.44	164.9	165.3	165.0	135.2	164.8	162.3	164.6	130.8	164.5	161.1	164.8	129.6	164.8
2	153.15	153.8	152.0	153.9	154.4	154.0	153.3	154.3	154.7	153.9	152.4	154.3	153.4	154.1
3	130.99	133.7	133.1	134.8	138.4	135.6	132.8	134.6	137.4	134.6	132.2	135.1	136.6	133.8
4	120.91	119.4	119.8	120.2	124.7	131.1	118.5	119.6	123.5	121.7	119.2	119.2	124.0	122.1
5	114.97	113.8	117.2	120.2	124.3	122.2	116.2	116.7	123.1	117.6	116.9	116.8	123.6	117.7
6	23.61	22.9	24.6	23.2	18.7	22.8	22.1	23.2	17.4	22.6	22.8	23.4	18.9	22.9
	RMSE	**1.76**	**1.74**	**2.87**	**14.28**	**5.56**	**2.55**	**2.13**	**15.81**	**2.27**	**2.87**	**2.30**	**16.16**	**2.05**

Table C.2: Acetaminophen experimental chemical shift predictions from various models using the stated density functional and basis set. Models labeled with a Δ superscript are Δ-ML models using the listed inexpensive chemical shielding prediction.

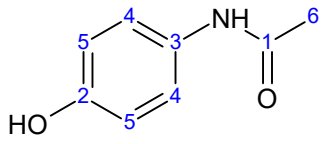


Figure C.2: Acetaminophen chemical shift assignment

C.5.2 Aspirin (ACSALA14) in DMSO

Assignment	Experiment	PBE0 6-311+G(2d,p)	PBE0 6-31G	PBE0 6-31G$^{\Delta}$	PBE0 STO-3G	PBE0 STO-3G$^{\Delta}$	PBE 6-31G	PBE 6-31G$^{\Delta}$	PBE STO-3G	PBE STO-3G$^{\Delta}$	SVWN 6-31G	SVWN 6-31G$^{\Delta}$	SVWN STO-3G	SVWN STO-3G$^{\Delta}$
1	169.06	171.3	174.3	171.2	143.7	168.9	173.1	170.8	138.7	168.0	171.3	170.9	136.7	168.3
2	165.51	166.2	169.5	166.3	146.6	164.5	168.0	166.4	142.9	164.9	166.4	166.7	140.6	164.5
3	150.08	154.6	153.5	153.7	161.5	157.3	155.8	154.3	164.2	158.3	154.7	154.1	162.7	159.5
4	133.67	136.6	136.0	138.7	141.9	144.6	136.0	139.2	141.8	143.8	136.3	138.5	141.9	143.2
5	131.27	135.7	135.5	135.3	141.0	133.1	135.0	134.9	140.6	132.9	135.3	135.1	140.6	133.0
6	125.96	126.1	126.6	126.6	135.0	128.4	126.7	125.4	134.8	128.5	127.3	125.1	135.0	127.9
7	123.99	123.9	125.9	124.7	133.1	128.2	125.9	126.3	133.2	127.6	126.7	126.3	133.4	128.0
8	123.68	123.1	125.3	124.5	131.4	125.8	125.5	124.1	130.4	127.0	125.7	124.1	130.4	127.9
9	20.74	20.4	21.8	20.4	16.0	19.4	19.3	20.5	14.9	19.4	19.7	20.5	16.2	19.7
	RMSE	**2.47**	**3.11**	**2.62**	**13.10**	**4.80**	**3.06**	**2.82**	**15.00**	**4.79**	**2.67**	**2.69**	**15.71**	**4.98**

Table C.3: Aspirin experimental chemical shift predictions from various models using the stated density functional and basis set. Models labeled with a Δ superscript are Δ-ML models using the listed inexpensive chemical shielding prediction.

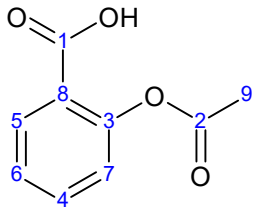


Figure C.3: Aspirin chemical shift assignment

C.5.3 Estrone (ESTRON11) in DMSO

Assignment	Experiment	PBE0 6-311+G(2d,p)	PBE0 6-31G	PBE0 6-31G$^{\Delta}$	PBE0 STO-3G	PBE0 STO-3G$^{\Delta}$	PBE 6-31G	PBE 6-31G$^{\Delta}$	PBE STO-3G	PBE STO-3G$^{\Delta}$	SVWN 6-31G	SVWN 6-31G$^{\Delta}$	SVWN STO-3G	SVWN STO-3G$^{\Delta}$
1	219.63	221.3	227.4	221.1	183.7	220.4	231.8	221.8	186.0	221.0	231.4	221.6	184.6	221.6
2	154.94	156.6	154.7	156.5	156.7	156.4	155.9	156.6	156.9	156.1	154.9	156.6	155.4	156.0
3	137	139.1	137.5	140.3	144.0	141.5	138.0	139.1	143.6	142.8	137.6	139.1	143.2	141.9
4	129.8	132.0	132.2	134.7	137.2	137.6	132.8	133.8	136.9	138.7	132.8	133.5	136.8	137.2
5	125.97	128.7	128.7	129.2	137.3	128.3	127.9	128.6	136.6	127.3	128.2	128.9	136.5	127.4
6	114.86	113.5	114.9	113.8	121.0	115.8	113.6	113.4	119.7	114.7	114.4	113.0	120.3	114.5
7	112.69	112.9	114.2	113.8	124.3	115.3	113.9	113.7	123.8	114.8	114.2	113.9	123.6	114.4
8	49.46	50.6	49.7	50.7	43.8	53.2	52.8	50.5	46.1	53.6	54.7	50.5	47.4	52.6
9	47.23	49.0	50.4	49.9	45.7	52.3	52.6	49.4	47.6	51.7	54.7	49.7	50.5	52.1
10	43.34	45.4	44.7	46.0	40.5	48.4	47.5	45.6	42.9	48.6	48.8	45.6	44.2	47.2
11	37.86	39.2	38.0	40.0	32.6	40.4	40.2	40.0	34.5	40.8	41.5	40.0	36.0	39.4
12	35.27	35.8	36.9	36.4	33.7	37.2	37.0	36.4	34.2	36.8	38.0	36.5	36.1	37.0
13	31.25	32.3	32.0	32.4	28.7	34.0	33.3	32.4	30.1	33.3	33.6	32.2	30.7	33.2
14	28.97	31.2	30.6	31.4	28.3	32.4	31.4	31.4	29.2	32.0	31.6	31.1	30.1	31.9
15	26.04	27.7	26.7	27.7	24.2	29.1	27.3	27.6	25.0	28.7	27.6	27.4	25.7	28.5
16	25.46	26.8	26.0	26.5	24.0	24.2	26.5	26.7	24.9	24.4	26.7	26.8	25.6	24.3
17	21.04	21.9	22.4	22.3	26.1	22.9	22.0	22.0	26.7	22.9	22.0	22.2	27.5	22.9
18	13.4	12.0	13.3	12.1	13.6	13.5	12.1	12.5	13.5	13.3	12.2	12.5	13.9	14.6
	RMSE	**1.65**	**2.33**	**2.20**	**10.03**	**3.38**	**3.70**	**1.94**	**9.28**	**3.52**	**4.14**	**1.95**	**9.54**	**3.05**

Table C.4: Estrone experimental chemical shift predictions from various models using the stated density functional and basis set. Models labeled with a Δ superscript are Δ-ML models using the listed inexpensive chemical shielding prediction.

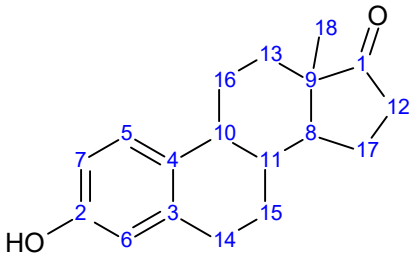


Figure C.4: Estrone chemical shift assignment

C.5.4 Mefenamic Acid (XYANAC07) in DMSO

Assign-ment	Experi-ment	PBE0 6-311+G(2d,p)	PBE0 6-31G	PBE0 6-31G$^{\Delta}$	PBE0 STO-3G	PBE0 STO-3G$^{\Delta}$	PBE 6-31G	PBE 6-31G$^{\Delta}$	PBE STO-3G	PBE STO-3G$^{\Delta}$	SVWN 6-31G	SVWN 6-31G$^{\Delta}$	SVWN STO-3G	SVWN STO-3G$^{\Delta}$
1	170.12	170.7	174.0	171.2	154.4	171.1	171.8	171.1	149.9	170.6	170.0	171.1	147.4	170.5
2	148.7	151.8	149.6	151.6	150.2	149.7	148.1	149.9	148.1	148.9	146.9	150.7	146.7	148.5
3	138.3	139.8	139.0	140.6	145.8	140.3	140.1	139.1	146.4	138.8	139.9	139.4	146.3	140.6
4	137.8	138.9	138.4	139.6	146.2	139.6	139.0	140.4	146.5	139.5	138.3	140.8	145.5	140.0
5	134.09	137.5	137.6	137.6	146.0	136.8	138.2	136.3	146.3	135.6	137.8	136.6	146.1	135.4
6	131.64	136.6	136.0	136.1	140.0	136.5	134.7	135.8	138.9	135.2	134.9	135.0	138.8	134.7
7	131.17	134.5	134.3	135.5	138.7	131.8	133.3	134.8	137.7	130.8	133.5	134.9	137.4	130.5
8	126.35	128.8	129.5	129.3	138.2	129.6	129.5	129.3	138.2	128.8	129.9	129.3	138.1	128.9
9	125.94	128.0	127.5	127.7	136.4	127.9	127.3	128.1	136.3	127.5	127.5	128.1	136.2	127.2
10	122.13	126.2	127.1	126.3	136.3	126.5	126.4	125.3	135.8	125.6	126.7	125.3	135.7	125.6
11	116.16	115.1	115.8	116.6	126.1	120.7	116.3	117.3	126.2	120.5	117.0	117.3	126.6	120.4
12	113.02	113.2	113.5	113.4	122.2	116.4	113.4	114.0	122.1	117.7	114.2	113.7	122.7	117.2
13	111.2	107.6	109.5	107.7	117.8	110.5	109.4	107.3	116.6	110.9	110.0	106.6	117.0	110.2
14	20.13	21.0	21.6	20.9	20.6	20.8	20.5	21.0	20.7	20.5	20.8	21.1	21.5	21.0
15	13.58	14.2	15.1	13.6	15.6	11.9	13.4	13.7	15.3	11.9	13.7	14.2	16.2	12.2
	RMSE	**2.62**	**2.62**	**2.73**	**9.40**	**2.70**	**2.19**	**2.41**	**9.81**	**2.34**	**2.25**	**2.51**	**10.17**	**2.32**

Table C.5: Mefanamic acid experimental chemical shift predictions from various models using the stated density functional and basis set. Models labeled with a Δ superscript are Δ-ML models using the listed inexpensive chemical shielding prediction.

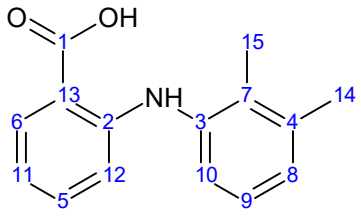


Figure C.5: Mefanamic acid chemical shift assignment

C.5.5 Nalidixic Acid (NALIDX01) in DMSO

Assign-ment	Experi-ment	PBE0 6-311+G(2d,p)	PBE0 6-31G	PBE0 6-31G$^{\Delta}$	PBE0 STO-3G	PBE0 STO-3G$^{\Delta}$	PBE 6-31G	PBE 6-31G$^{\Delta}$	PBE STO-3G	PBE STO-3G$^{\Delta}$	SVWN 6-31G	SVWN 6-31G$^{\Delta}$	SVWN STO-3G	SVWN STO-3G$^{\Delta}$
1	178	177.9	179.2	178.8	161.9	182.3	177.7	178.3	160.5	180.9	176.3	176.9	158.7	181.0
2	165.45	164.7	166.7	165.0	141.0	162.8	165.3	164.8	137.0	162.6	163.6	164.5	134.7	163.2
3	164.57	164.5	165.2	165.0	160.6	161.1	165.4	164.2	160.4	160.9	164.7	164.4	159.6	160.4
4	149.48	149.9	149.3	152.5	149.5	153.9	148.7	152.5	149.0	154.7	147.3	152.7	147.2	155.5
5	148.26	148.9	148.7	152.1	146.2	152.0	145.9	149.7	144.8	151.6	145.7	148.8	144.3	151.5
6	135.53	139.1	138.0	140.3	140.2	141.9	136.4	139.3	138.9	140.0	136.6	139.1	138.9	140.4
7	122.5	121.1	123.6	123.9	129.3	135.9	124.4	123.6	128.3	137.7	124.9	122.2	128.4	136.8
8	118.29	120.5	121.9	120.6	130.7	120.4	122.6	120.7	130.3	120.0	123.4	121.1	130.5	120.1
9	108.63	113.8	117.4	115.6	121.9	115.0	119.0	116.2	119.3	115.0	120.1	115.7	119.7	115.0
10	46.7	47.4	48.0	47.3	53.0	46.3	49.2	47.2	54.9	46.3	48.9	47.6	55.6	46.7
11	24.93	24.8	25.2	24.5	21.6	24.3	24.1	24.6	21.8	24.3	24.5	24.9	22.9	24.9
12	14.9	13.7	13.8	13.6	11.9	13.9	12.1	13.6	11.5	13.9	11.7	13.7	12.0	13.9
	RMSE	**2.03**	**2.95**	**2.96**	**10.55**	**5.31**	**3.55**	**2.78**	**11.32**	**5.50**	**4.04**	**2.67**	**12.17**	**5.43**

Table C.6: Nalidixic acid experimental chemical shift predictions from various models using the stated density functional and basis set. Models labeled with a Δ superscript are Δ-ML models using the listed inexpensive chemical shielding prediction.

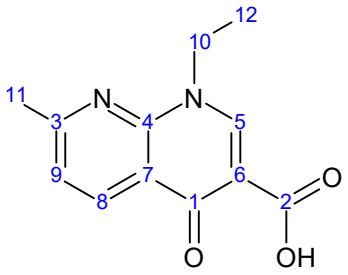


Figure C.6: Nalidixic acid chemical shift assignment

C.5.6 Nitrofurantoin (LABJON) in DMSO

Assign-ment	Experi-ment	PBE0 6-311+G(2d,p)	PBE0 6-31G	PBE0 6-31G$^{\Delta}$	PBE0 STO-3G	PBE0 STO-3G$^{\Delta}$	PBE 6-31G	PBE 6-31G$^{\Delta}$	PBE STO-3G	PBE STO-3G$^{\Delta}$	SVWN 6-31G	SVWN 6-31G$^{\Delta}$	SVWN STO-3G	SVWN STO-3G$^{\Delta}$
1	168.47	164.1	164.2	164.6	132.4	167.6	163.2	165.3	130.7	168.7	162.1	165.0	129.6	168.9
2	153.14	155.1	159.5	154.6	158.6	156.4	159.9	154.4	155.3	154.5	159.1	154.2	153.5	155.5
3	151.86	151.3	153.6	151.1	151.0	149.7	154.5	149.6	151.1	150.1	154.0	149.5	150.1	149.8
4	151.68	148.8	147.8	148.9	120.7	149.2	146.9	151.1	119.1	149.6	146.3	151.4	118.2	149.1
5	131.1	125.3	125.8	124.9	124.9	125.9	123.9	125.2	123.3	127.9	124.3	125.2	123.5	126.3
6	114.61	113.8	115.1	114.3	124.3	117.0	115.5	115.8	124.9	117.6	116.5	115.8	125.6	118.3
7	114.34	113.1	113.3	112.7	118.1	111.4	112.3	112.4	117.9	111.4	113.5	113.0	119.2	113.2
8	49.04	46.1	46.4	45.8	43.0	46.3	45.2	45.3	42.1	45.9	45.8	45.2	43.9	45.1
	RMSE	**3.07**	**3.75**	**3.10**	**17.58**	**2.98**	**4.67**	**2.98**	**18.45**	**2.41**	**4.61**	**3.00**	**18.93**	**2.95**

Table C.7: Nitrofurantoin experimental chemical shift predictions from various models using the stated density functional and basis set. Models labeled with a Δ superscript are Δ-ML models using the listed inexpensive chemical shielding prediction.

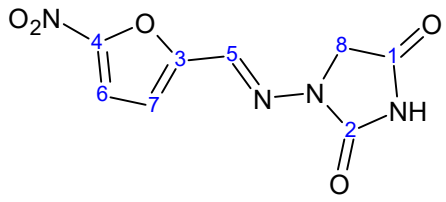


Figure C.7: Nitrofurantoin chemical shift assignment

C.5.7 Trimethoprim (AMXBPM12) in DMSO

Assign-ment	Experi-ment	PBE0 6-311+G(2d,p)	PBE0 6-31G	PBE0 6-31G$^{\Delta}$	PBE0 STO-3G	PBE0 STO-3G$^{\Delta}$	PBE 6-31G	PBE 6-31G$^{\Delta}$	PBE STO-3G	PBE STO-3G$^{\Delta}$	SVWN 6-31G	SVWN 6-31G$^{\Delta}$	SVWN STO-3G	SVWN STO-3G$^{\Delta}$
1	162.19	163.9	161.2	163.7	156.1	162.7	159.2	163.9	153.8	162.1	157.4	164.1	151.6	161.9
2	No ref													
3	155.72	158.5	156.9	157.8	154.5	157.3	156.1	157.1	153.0	157.9	155.5	157.2	151.6	158.3
4	152.69	156.0	156.9	156.9	159.3	156.0	158.2	156.5	159.8	156.0	157.5	156.5	158.4	156.0
5	135.97	138.1	140.5	139.9	138.8	142.2	141.8	140.4	138.6	143.0	142.0	140.0	138.2	144.1
6	135.73	137.9	136.4	138.1	141.3	141.2	136.3	137.8	139.6	141.3	135.6	137.1	138.6	141.2
7	105.99	102.4	104.9	104.1	117.7	107.5	103.5	102.8	116.4	106.4	104.1	102.3	116.6	105.6
8	105.75	107.0	109.1	107.7	114.8	110.9	109.8	107.7	114.0	111.3	110.6	107.8	114.7	109.5
9	59.89	56.9	57.4	57.5	54.5	57.6	58.7	57.3	56.7	58.8	58.7	57.2	57.2	58.0
10	55.83	52.1	52.5	52.1	52.7	52.8	53.5	52.2	54.9	52.6	53.1	52.0	55.1	52.5
11	32.94	37.1	37.6	36.6	36.4	36.2	39.0	36.6	37.7	35.4	39.5	36.9	38.6	35.6
	RMSE	**2.93**	**3.04**	**2.94**	**6.26**	**3.70**	**3.74**	**3.01**	**6.01**	**3.82**	**4.03**	**3.07**	**6.38**	**3.86**

Table C.8: Trimethoprim experimental chemical shift predictions from various models using the stated density functional and basis set. Models labeled with a Δ superscript are Δ-ML models using the listed inexpensive chemical shielding prediction.

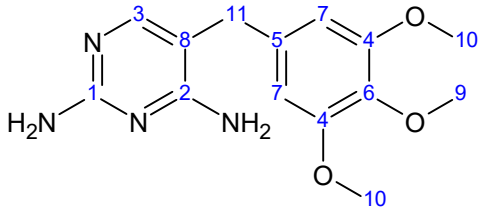


Figure C.8: Trimethoprim chemical shift assignment

C.5.8 Aspirin (ACSALA14) in $CDCl_3$

Assignment	Experiment	PBE0 6-311+G(2d,p)	PBE0 6-31G	PBE0 6-31G$^{\Delta}$	PBE0 STO-3G	PBE0 STO-3G$^{\Delta}$	PBE 6-31G	PBE 6-31G$^{\Delta}$	PBE STO-3G	PBE STO-3G$^{\Delta}$	SVWN 6-31G	SVWN 6-31G$^{\Delta}$	SVWN STO-3G	SVWN STO-3G$^{\Delta}$
1	170.2	169.4	171.1	169.2	144.2	168.0	169.3	169.1	139.0	166.4	171.4	170.1	118.4	166.6
2	169.76	164.3	166.4	164.4	147.0	163.6	164.5	164.7	143.2	163.3	166.4	165.9	121.5	162.9
3	151.28	152.9	150.8	152.0	162.0	156.5	152.6	152.8	164.6	156.8	154.6	153.4	138.9	157.8
4	134.9	135.1	133.7	137.2	142.4	143.9	133.4	137.9	142.1	142.5	136.0	137.8	122.5	141.8
5	132.51	134.2	133.3	133.8	141.5	132.4	132.5	133.6	140.8	131.8	135.0	134.5	121.5	131.7
6	126.17	124.7	124.6	125.2	135.4	127.8	124.5	124.3	135.1	127.4	127.0	124.6	117.1	126.7
7	124.01	122.6	123.9	123.3	133.5	127.6	123.7	125.2	133.4	126.5	126.4	125.8	115.8	126.7
8	122.26	121.8	123.3	123.1	131.8	125.1	123.3	123.0	130.6	125.9	125.4	123.6	113.5	126.6
9	20.99	20.5	22.4	20.4	16.0	19.5	20.4	20.7	14.5	19.6	18.6	20.7	23.3	19.8
	RMSE	**2.11**	**1.50**	**2.10**	**13.91**	**4.43**	**2.03**	**2.24**	**15.83**	**4.32**	**2.43**	**2.09**	**25.12**	**4.48**

Table C.9: Aspirin experimental chemical shift predictions from various models using the stated density functional and basis set. Models labeled with a Δ superscript are Δ-ML models using the listed inexpensive chemical shielding prediction.

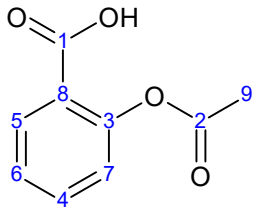


Figure C.9: Aspirin chemical shift assignment

C.5.9 Benzoic Acid (BENZAC02) in $CDCl_3$

Assignment	Experiment	PBE0 6-311+G(2d,p)	PBE0 6-31G	PBE0 6-31G$^{\Delta}$	PBE0 STO-3G	PBE0 STO-3G$^{\Delta}$	PBE 6-31G	PBE 6-31G$^{\Delta}$	PBE STO-3G	PBE STO-3G$^{\Delta}$	SVWN 6-31G	SVWN 6-31G$^{\Delta}$	SVWN STO-3G	SVWN STO-3G$^{\Delta}$
1	172.77	164.8	167.4	165.3	148.1	165.4	165.5	165.8	144.2	164.7	167.7	166.7	122.3	164.5
2	133.83	133.0	131.8	133.1	140.8	136.0	131.1	133.3	140.5	134.7	133.6	133.3	121.3	134.9
3	130.28	130.9	129.6	131.7	138.3	131.9	128.9	129.8	138.0	131.8	131.2	130.1	119.2	131.4
4	129.44	128.1	127.0	130.6	138.5	131.0	127.3	132.3	138.7	129.7	129.8	133.1	119.8	129.9
5	128.49	127.7	127.4	128.6	137.3	129.9	127.7	128.2	137.2	129.0	130.3	128.9	118.7	129.2
	RMSE	**3.65**	**2.85**	**3.45**	**13.27**	**3.63**	**3.66**	**3.39**	**14.69**	**3.71**	**2.45**	**3.17**	**24.55**	**3.78**

Table C.10: Benzoic acid experimental chemical shift predictions from various models using the stated density functional and basis set. Models labeled with a Δ superscript are Δ-ML models using the listed inexpensive chemical shielding prediction.

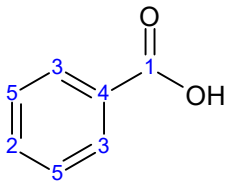


Figure C.10: Benzoic acid chemical shift assignment

C.5.10 Cortisone Acetate (ACPRET) in $CDCl_3$

Assign-ment	Experi-ment	PBE0 6-311+G(2d,p)	PBE0 6-31G	PBE0 6-31G$^{\Delta}$	PBE0 STO-3G	PBE0 STO-3G$^{\Delta}$	PBE 6-31G	PBE 6-31G$^{\Delta}$	PBE STO-3G	PBE STO-3G$^{\Delta}$	SVWN 6-31G	SVWN 6-31G$^{\Delta}$	SVWN STO-3G	SVWN STO-3G$^{\Delta}$
1	208.55	210.7	215.4	210.9	181.3	213.1	218.3	211.6	183.3	211.4	223.4	212.4	153.8	211.5
2	204.16	208.5	213.2	207.7	185.7	213.0	216.1	208.1	191.1	213.9	220.9	208.4	160.0	212.5
3	199.37	195.4	198.3	195.7	170.4	198.3	199.0	196.5	171.0	197.9	203.4	197.3	144.5	198.0
4	170.39	171.4	173.2	171.7	147.4	171.8	171.1	172.0	142.1	170.9	173.0	172.8	120.8	170.7
5	168.36	168.7	163.8	166.3	155.8	157.6	166.4	166.4	158.6	158.7	170.0	167.7	135.1	157.9
6	124.74	126.7	126.8	125.7	133.5	125.8	127.7	126.0	132.8	124.6	130.8	126.5	114.7	124.1
7	89.07	88.7	92.7	89.2	86.8	86.4	97.5	89.6	91.1	85.0	99.6	89.9	83.4	84.7
8	67.3	68.1	70.2	68.9	66.1	65.9	71.9	68.7	67.6	65.6	72.7	69.9	64.9	66.0
9	62.92	63.0	63.5	63.4	50.8	62.8	66.8	62.8	52.5	63.5	69.1	63.9	53.5	62.2
10	51.4	52.7	55.5	53.8	53.2	54.7	60.3	55.3	57.3	55.2	61.7	56.3	57.8	54.5
11	50.09	50.5	50.7	51.1	46.4	53.8	54.4	51.5	49.1	54.0	55.5	51.8	50.5	53.2
12	No ref													
13	38.47	40.5	39.7	39.7	40.1	42.7	42.5	39.9	42.6	42.1	41.9	40.1	45.9	41.4
14	36.75	37.6	37.3	38.6	34.5	40.9	40.3	38.9	36.7	41.2	40.2	39.2	40.9	40.3
15	35.17	36.9	38.6	36.8	38.7	38.1	40.5	37.4	40.4	38.2	39.4	37.4	43.2	37.7
16	35.1	34.7	34.9	34.5	30.8	34.2	36.9	33.8	32.0	34.0	36.0	33.8	36.6	34.1
17	33.86	34.0	33.9	33.8	27.3	34.1	34.3	33.7	27.3	33.8	33.4	33.9	33.6	33.9
18	32.64	33.4	33.0	33.4	27.5	32.6	35.2	34.8	28.6	32.7	34.1	35.1	33.9	31.8
19	32.4	33.4	33.5	33.3	30.2	32.4	35.1	34.0	31.2	32.5	34.0	34.1	36.1	32.3
20	23.37	23.5	25.0	23.8	28.4	24.0	25.3	24.0	28.9	23.9	23.1	23.9	34.1	23.7
21	20.32	19.2	21.1	19.3	14.5	19.1	18.8	19.3	12.8	19.1	16.9	19.3	21.9	19.1
22	17.55	15.0	16.6	14.5	14.3	14.8	15.7	14.8	13.5	14.6	13.7	14.8	21.8	15.5
23	15.55	13.7	15.1	13.0	14.6	13.3	14.5	13.9	14.3	12.8	13.1	14.1	22.7	13.0
	RMSE	**1.76**	**3.17**	**1.84**	**11.76**	**3.76**	**4.97**	**2.06**	**11.56**	**3.76**	**6.58**	**2.29**	**23.49**	**3.52**

Table C.11: Cortisone acetate experimental chemical shift predictions from various models using the stated density functional and basis set. Models labeled with a Δ superscript are Δ-ML models using the listed inexpensive chemical shielding prediction.

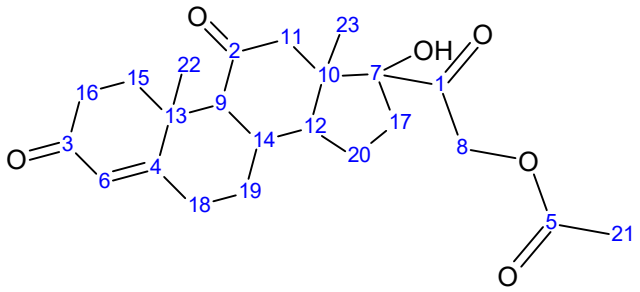


Figure C.11: Cortisone Acetate chemical shift assignment

C.5.11 Estimated Uncertainty in Drug Shieldings

The 95% confidence intervals were estimated for each the Δ-ML shieldings in the drug molecule set based on the S_{ens} ranges shown in Figure 3b in the main paper, and the results are plotted in Figure C.12. For example, if $0.25 \leq S_{ens} < 0.50$ ppm for some predicted shieilding, that shielding was assigned a confidence interval of ±1.4 ppm. While there is appreciable variation in S_{ens} across the predictions, S_{ens} generally trends upwards as the error in the ML shielding relative to the target one increases. In addition, the ML-predicted shielding lies within the estimated 95% confidence interval of the target DFT one for 108 of the 114 shieldings (94.7%), suggesting that the predicted shieldings in this set behave consistently with these confidence interval estimates. The errors for the six "outlier" shieldings exceed the confidence interval range by a mere 0.3 ppm or less. These outliers are highlighted in red in Figure C.12.

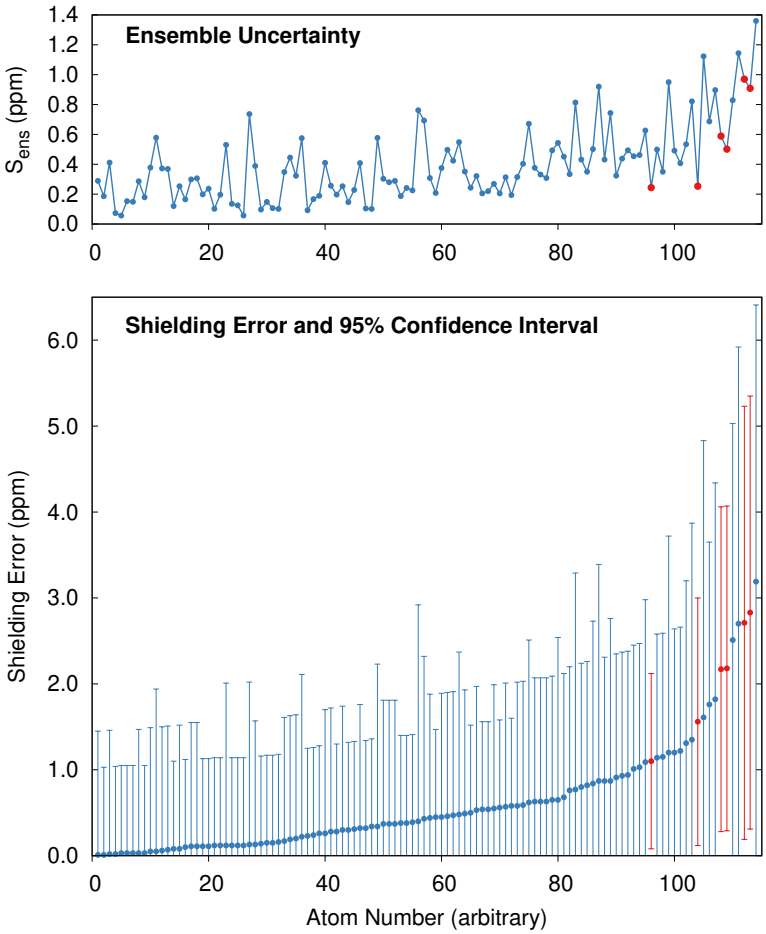


Figure C.12: Uncertainty estimation for the chemical shieldings in the drug molecule set. (a) The standard deviation S_{ens} in the shielding prediction among the ten members of the NN ensemble. (b) Error in the PBE0/6-31G + Δ-ML shielding relative to the target shielding and the associated confidence intervals. The six data points in red are those for which the confidence interval range lies outside the actual target shielding.

C.6 Neural Network Hyperparameter Optimization

A search of the optimal neural network hyperparamters was performed using a Bayesian search algorithm with Gaussian processes as implemented in the `scikit-optimize` package. Bayesian optimization provides an alternative to the popular grid search method of hyperparameter optimization when the time to train the model prohibits the use of an extensive grid search. Hyperparameter searches exploring the appropriate number of hidden layers and neurons per layer were performed for the PBE0/6-31G Δ-ML model for ^{13}C as a representative test case. For each optimization, the number of hidden layers was set to 1, 2, 3, 4, or 5 layers while the number of neurons per layer was varied in the range (32, 128) or (128, 500). Restricting the optimization to vary only the number of neurons within a fixed number of layers increases the number of individual optimizations that had to be performed but dramatically reduces the hyperparameter search space for each particular choice of the number of hidden layers.

The objective function for the optimization procedure was computed as the RMSE over the 10-fold cross-validation model trainings for a given number of hidden layers and range of neurons per layer. The models were trained on the same small-molecule training set as all the other trainings reported in the paper. The best-performing network ensemble from each number of hidden layers and range of neurons was then tested on the molecules from GDB17. As shown in Figure C.13, increasing the number of hidden layers and/or neurons has very little impact on the performance of the model. Upon increasing from 1 to 5 hidden layers, the performance on the GDB17 testing set improves by only 0.03 ppm. Increasing the number of neurons only reduces the errors by 0.01 ppm or less. Accordingly,

the NN architecture with 3 hidden layers and 128 neurons per layer described in the main paper was adopted throughout.

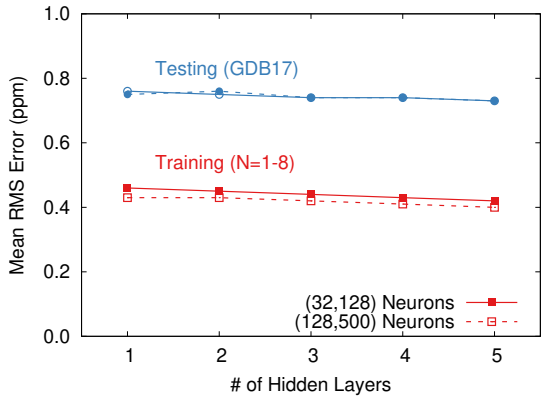


Figure C.13: Results of hyperparameter search showing the mean RMS error for the search over a range of either 32–128 or 128–500 neurons with 1–5 hidden layers. The performance on the training and testing sets vary by only a few hundredths of a ppm across the range of options considered.

C.7 Full Comparison of Δ-ML Models

Table 2 in the main paper summarize the performance of key NN models for the four nuclei. This section presents complete results for the AEV-only and all six Δ-ML functional and basis set combinations for each nucleus type. The training and testing RMSEs listed below are computed from each respective ensemble NN.

C.7.1 ^{13}C Model Comparison

Table C.12: Summary of ^{13}C RMSE (ppm) by Δ-ML model.

	Training: N=1–8		Testing: GDB17	
Model	No Δ-ML	w/ Δ-ML	No Δ-ML	w/ Δ-ML
AEV		2.15		4.74
LDA/STO-3G	9.97	1.34	8.54	2.44
PBE/STO-3G	9.62	1.33	8.2	2.51
PBE0/STO-3G	9.00	1.39	7.18	2.49
LDA/6-31G	2.99	0.52	3.31	0.93
PBE/6-31G	2.75	0.45	3.01	0.82
PBE0/6-31G	1.77	0.38	1.54	0.70

C.7.2 ^{1}H Model Comparison

Table C.13: Summary of ^{1}H RMSE (ppm) per Δ-ML model.

	Training: N=1–8		Testing: GDB17	
Model	No Δ-ML	w/ Δ-ML	No Δ-ML	w/ Δ-ML
AEV		0.225		0.360
LDA/STO-3G	0.626	0.114	0.631	0.216
PBE/STO-3G	0.628	0.113	0.643	0.211
PBE0/STO-3G	0.651	0.110	0.675	0.214
LDA/6-31G	0.262	0.068	0.267	0.122
PBE/6-31G	0.263	0.063	0.256	0.113
PBE0/6-31G	0.247	0.060	0.23	0.110

C.7.3 ^{15}N Model Comparison

Table C.14: Summary of ^{15}N RMSE (ppm) per Δ-ML model.

	Training: N=1–8		Testing: GDB17	
Model	No Δ-ML	w/ Δ-ML	No Δ-ML	w/ Δ-ML
AEV		4.85		13.86
LDA/STO-3G	21.67	3.63	19.89	6.09
PBE/STO-3G	21.25	3.15	19.95	6.04
PBE0/STO-3G	21.65	3.14	20.78	5.79
LDA/6-31G	6.77	1.43	5.54	2.69
PBE/6-31G	6.19	1.19	5.42	2.30
PBE0/6-31G	5.63	0.84	5.40	1.69

C.7.4 ^{17}O Model Comparison

Table C.15: Summary of ^{17}O RMSE (ppm) per Δ-ML model.

Model	Training: N=1–8		Testing: GDB17	
	No Δ-ML	w/ Δ-ML	No Δ-ML	w/ Δ-ML
AEV		8.09		18.22
LDA/STO-3G	29.04	5.55	26.30	9.31
PBE/STO-3G	29.18	5.09	27.05	9.06
PBE0/STO-3G	31.95	4.50	30.01	8.34
LDA/6-31G	8.30	2.71	7.13	4.24
PBE/6-31G	7.66	2.11	6.71	3.42
PBE0/6-31G	7.10	1.39	6.68	2.47

C.7.5 Uncertainty Analysis for ^{1}H, ^{15}N, and ^{17}O

Section 3.3 of the main paper describes the uncertainty analysis for ^{13}C chemical shieldings computed with PBE0/6-31G + Δ-ML based on the standard deviation among members of the ensemble. This section provides analogous uncertainty estimates for the other three nuclei. The S_{ens} windows were chosen based on the distribution of S_{ens} values observed for that nucleus type in the GDB17 data set while also maintaining "user-friendly" windows. For example, if nitrogen shielding prediction has an ensemble standard deviation $S_{ens} = 0.4$ ppm, we estimate the shielding uncertainty at ± 2.2 ppm with 95% confidence.

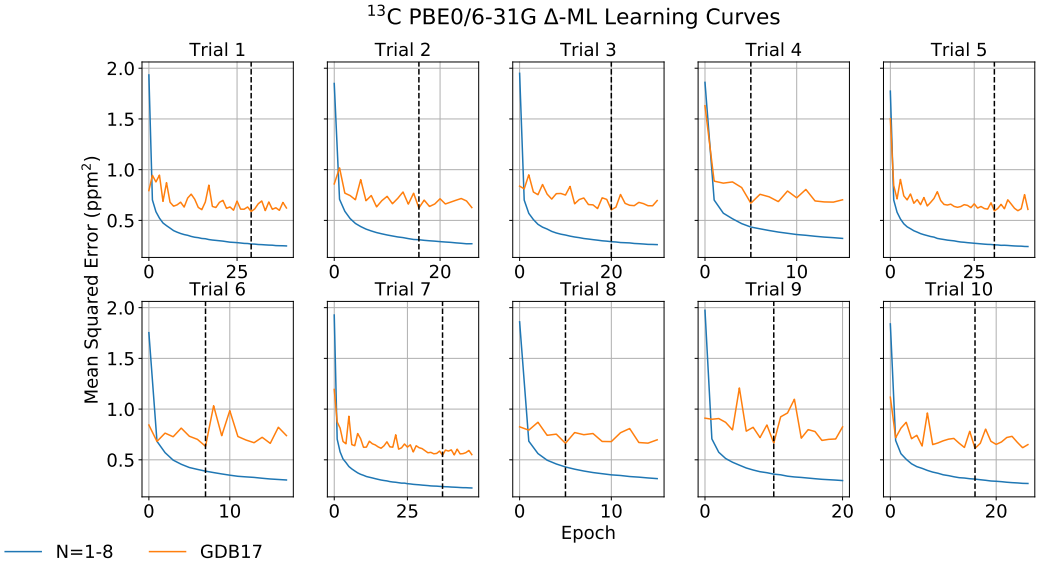
Hydrogen		Nitrogen		Oxygen	
S_{ens} (ppm)	95% CI (ppm)	S_{ens} (ppm)	95% CI (ppm)	S_{ens} (ppm)	95% CI (ppm)
$S_{ens} < 0.025$	0.15	$S_{ens} < 0.25$	1.5	$S_{ens} < 0.50$	3.0
$0.025 < S_{ens} < 0.050$	0.20	$0.25 < S_{ens} < 0.50$	2.2	$0.50 < S_{ens} < 0.75$	3.7
$0.050 < S_{ens} < 0.075$	0.26	$0.50 < S_{ens} < 0.75$	3.0	$0.75 < S_{ens} < 1.00$	5.0
$0.075 < S_{ens} < 0.100$	0.31	$0.75 < S_{ens} < 1.00$	4.0	$1.00 < S_{ens} < 1.50$	5.8
$0.100 < S_{ens} < 0.125$	0.35	$1.00 < S_{ens} < 1.50$	5.2	$1.50 < S_{ens} < 2.00$	7.7
$0.125 < S_{ens}$	0.44	$1.5 < S_{ens}$	8.0	$2.0 < S_{ens}$	9.9

Table C.16: Uncertainty, expressed in terms of 95% confidence intervals (CI), associated with PBE0/6-31G + Δ-ML chemical shieldings based on the standard deviation among the ensemble member predictions, S_{ens}, as computed from the GDB17 species.

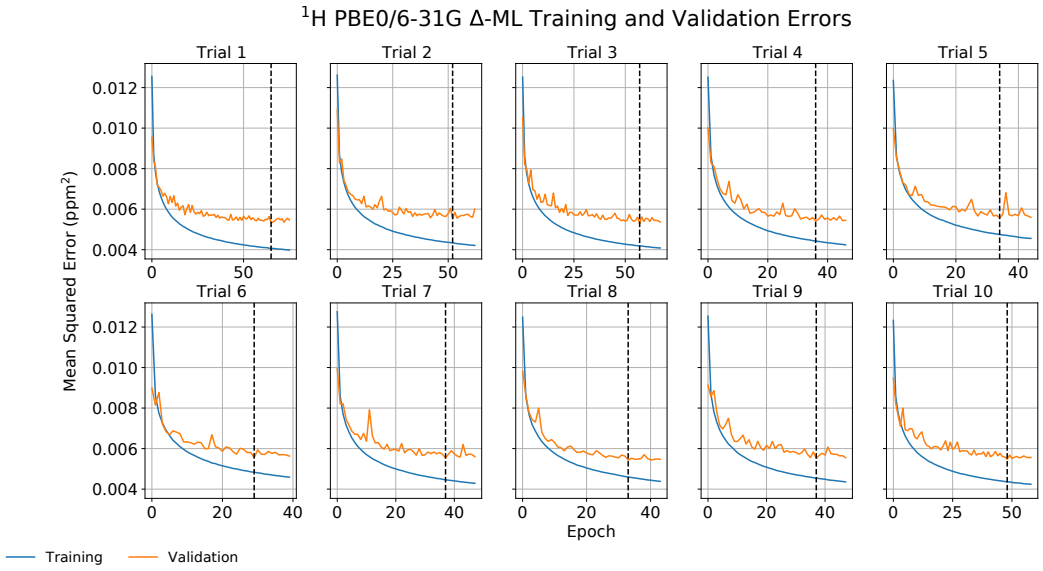
C.8 Sample Training and Validation Errors

The learning curves below represent the training and testing instances of the individual neural networks used to construct the ensemble model for each respective nuclei. As indicated in the main text, an early stopping protocol of 10 epochs was used for each training instance to avoid over-fitting. The dotted line for each plot indicates the epoch that was used to save the optimal training weights such that the testing mean-squared-error was smallest.

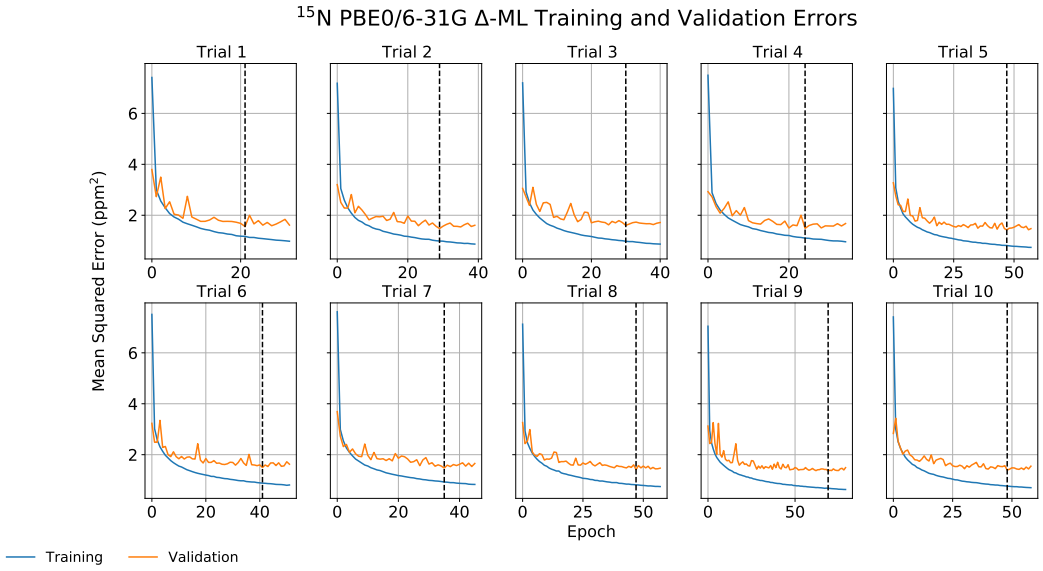
C.8.1 ^{13}C PBE0/6-31G Training and Validation Errors



C.8.2 ^{1}H PBE0/6-31G Training and Validation Errors



C.8.3 ^{15}N PBE0/6-31G Training and Validation Errors



C.8.4 ^{17}O PBE0/6-31G Training and Validation Errors