Predicting density functional theory-quality nuclear magnetic resonance chemical shifts via Δ-machine learning

Pablo A. Unzueta, Chandler S. Greenwell, and Gregory J. O. Beran*

Department of Chemistry, University of California, Riverside, California 92521 USA

E-mail: gregory.beran@ucr.edu
Phone: +1 (951) 827-7869

Abstract

First-principles prediction of nuclear magnetic resonance chemical shifts plays an increasingly important role in the interpretation of experimental spectra, but the required density functional theory (DFT) calculations can be computationally expensive. Promising machine learning models for predicting chemical shieldings in general organic molecules have been developed previously, though the accuracy of those models remains below that of DFT. The present study demonstrates how much higher accuracy chemical shieldings can be obtained via Δ-machine learning approach, with the result that the errors introduced by the machine learning model are only one-half to one-third the errors expected for DFT chemical shifts relative to experiment. Specifically, an ensemble of neural networks is trained to correct PBE0/6-31G chemical shieldings up to the target level of PBE0/6-311+G(2d,p). It can predict \(^1\)H, \(^13\)C, \(^15\)N, and \(^17\)O chemical shieldings with root-mean-square errors of 0.11, 0.70, 1.69, and 2.47 ppm, respectively. At the same time, the Δ-machine learning approach is 1–2 orders of magnitude faster.
than the target large-basis calculations. It is also demonstrated that the machine learning model predicts experimental solution-phase NMR chemical shifts in drug molecules with only modestly worse accuracy than the target DFT model. Finally, the ability to estimate the uncertainty in the predicted shieldings based on variations within the ensemble of neural network models is also assessed.

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).\textsuperscript{1–9} Chemical shift predictions can help assign peaks in an experimental NMR spectrum, refine structures, or even discriminate among multiple plausible structures.\textsuperscript{10,11} 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.\textsuperscript{2,12–16}

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\textsuperscript{17–29} and in other more specialized systems such as acrylonitrile copolymers\textsuperscript{30} or amorphous silicon oxides.\textsuperscript{31} 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\textsuperscript{32} built ML
models based on the Coulomb matrix descriptor and kernel ridge regression that predict DFT chemical shieldings in organic molecules 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.$^{33}$ In 2018, Paruzzo et al.$^{34}$ developed an ML model for solid-state organic molecule chemical shieldings based on Gaussian process regression and the smooth overlap of atomic positions (SOAP)$^{35}$ 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.$^{36-38}$ Nevertheless, they showed that the ML chemical shift predictions could aid discrimination between candidate structures in the context of NMR crystallography.$^{34}$ Liu et al.$^{39}$ 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 34: 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,$^{40-42}$ 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 \( \Delta \)-ML.\textsuperscript{43–50} Specifically, we perform an inexpensive, low-accuracy calculation to obtain an initial approximate isotropic chemical shielding \( \sigma_{\text{cheap}} \) and utilize a trained neural network (NN) to correct it \( (\Delta_{\text{ML}}) \) up to the accuracy of a much more demanding, higher-accuracy “target” chemical shielding prediction, \( \sigma_{\text{target}} \):

\[
\sigma_{\text{target}} = \sigma_{\text{cheap}} + \Delta_{\text{ML}}
\]

The \( \Delta \)-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 \( \Delta \)-ML approach simplifies the learning problem to that of learning only the residual correction \( \Delta_{\text{ML}} \), 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\textsuperscript{33} 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 \( \Delta \)-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 \(^1\text{H}\), 1.44 ppm for \(^{13}\text{C}\), 3.86 ppm for \(^{15}\text{N}\), and 7.47 ppm for \(^{17}\text{O}\).\textsuperscript{51,52} We then investigate several potential models for the “cheap” chemical shieldings, including the
local density approximation functional SVWN, the generalized gradient approximation (GGA) functional PBE, or the hybrid functional PBE0. 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 using locally dense basis sets 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, 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 $\Delta$-ML-based chemical shielding models can potentially seamlessly replace much more expensive DFT calculations without sacrificing quantum mechanical accuracy.

2 Computational Details

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,\textsuperscript{61} which uses the GDB11 database\textsuperscript{62} as a starting point. The minimum-energy geometry for each molecule at the $\omega$B97X/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 training/validation 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.\textsuperscript{63} The specific molecules are listed in SI Section S4. 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 force field geometry optimizations. Finally, the molecular geometries were optimized in Gaussian 09\textsuperscript{65} at the same $\omega$B97X/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 $\sim$60,000 molecules contain over one million chemical shieldings (Table 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.\textsuperscript{38} The inexpensive chemical shielding models
used in the Δ-ML approach will be discussed in Section 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 SI Section S2. Computational timings for the ML workflow are reported using ORCA v4.2.1 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 for the pure and hybrid functionals. NMR calculations without density fitting utilized analytic integrals.

Table 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.

<table>
<thead>
<tr>
<th></th>
<th>Training/Validation GDB11 ((N=1\text{--}8))</th>
<th>Testing GDB17 ((N=12\text{--}17))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecules</td>
<td>57,456</td>
<td>3,780</td>
</tr>
<tr>
<td>C atoms</td>
<td>298,081</td>
<td>44,146</td>
</tr>
<tr>
<td>H atoms</td>
<td>496,275</td>
<td>65,524</td>
</tr>
<tr>
<td>N atoms</td>
<td>89,010</td>
<td>9,961</td>
</tr>
<tr>
<td>O atoms</td>
<td>60,403</td>
<td>7,549</td>
</tr>
</tbody>
</table>

2.2 Feature Representation and Neural Network Architecture

Geometric information is encoded in the NN input descriptor via the atomic environment vector (AEV).\(^70\) The AEV builds on the Behler and Parrinello atomic symmetry functions,\(^71\) and it is one of several atomic descriptors\(^35,72,73\) 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\(^70\) and charges\(^74\) 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.\(^75\)
The AEV has been described in detail previously,\textsuperscript{70} 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 \( G_{a,s}^R \) are given as,

\[
G_{a,s}^R = \sum_{j \neq i} e^{\eta(R_{ij} - R_s)^2} f_C(R_{ij})
\]

where \( \eta \) equals 16 (see ref 70) 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 \( R_s = 0.9 + n a_0^2 \), where \( a_0 \) is the Bohr radius (0.529177 Å) and \( n \) ranges 0–15.\textsuperscript{70} The use of Gaussian functions in Eq 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,

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

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,

\[
G_{a,b,m,n}^{\text{mod}} = 2^{1-\xi} \sum_{j,k \neq i} (1+\cos(\theta_{ijk} - \theta_m))^\xi \exp \left[ -\eta \left( \frac{R_{ij} + R_{ik}}{2} - R_n \right)^2 \right] f_C(R_{ij}) f_C(R_{ik})
\]

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 ($\theta_m$) 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 4. 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 2) and angular triplets (Eq 4). It provides a fingerprint for chemical environment that is fed into the NN for the purpose of predicting the isotropic chemical shielding or $\Delta$-ML shielding correction. After generating the AEV and isotropic shieldings for each atom, a pandas 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 and the keras backend version 2.2.4 (www.keras.io). The NN architecture is depicted in Figure 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 standardized 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
Figure 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.

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 SI Section S8 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 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.\textsuperscript{70,76,80}

Once the cross-validation training was complete, the final ensemble model was tested on molecules sampled randomly from GDB17. Table 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 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 1. Six different possible
inexpensive chemical shielding model chemistries are considered: the SVWN, PBE, PBE0 or PBE0 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 ∆-ML approach. Generally speaking, more accurate baseline shielding models will be easier to correct with the ∆-ML approach, but the greater computational expense will also reduce the efficiency advantages of the ∆-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, 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 ^15N and ^17O 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 ∆-ML context for capturing the long-range contributions missing from the AEV. For example, the widely used and successful locally dense basis set approach 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.

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 1) appears to be well-converged with respect to the hyperparameter choices. See SI Section S6 for more details.

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 $\sigma_i$ must be referenced appropriately. Multiple referencing strategies exist; here we adopt the linear regression approach in which the final chemical shift $\delta_i$ is given as,

$$\delta_i = a\sigma_i + b$$

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 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.$^3$ 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 81 and 82. Details of these experimental data sets are provided in SI Section S5.

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 optimization and chemical shielding calculations for all species was performed at the same levels of theory as described for the data sets in Section 2.1.
## 3 Results and Discussion

Table 2: Summary of RMSE (in ppm) per ∆-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. See SI Section S7 for a full comparison.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training: N=1–8</th>
<th>Testing: GDB17</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No ∆-ML$^a$</td>
<td>w/ ∆-ML</td>
</tr>
<tr>
<td></td>
<td>No ∆-ML$^a$</td>
<td>w/ ∆-ML</td>
</tr>
<tr>
<td><strong>$^{13}$C Chemical Shieldings</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AEV only (No ∆-ML)$^b$</td>
<td>2.15</td>
<td>4.74</td>
</tr>
<tr>
<td>SVWN/STO-3G</td>
<td>9.97</td>
<td>1.34</td>
</tr>
<tr>
<td>PBE/STO-3G</td>
<td>9.62</td>
<td>1.33</td>
</tr>
<tr>
<td>PBE0/STO-3G</td>
<td>9.00</td>
<td>1.39</td>
</tr>
<tr>
<td>SVWN/6-31G</td>
<td>2.99</td>
<td>0.52</td>
</tr>
<tr>
<td>PBE/6-31G</td>
<td>2.75</td>
<td>0.45</td>
</tr>
<tr>
<td>PBE0/6-31G</td>
<td>1.77</td>
<td>0.38</td>
</tr>
<tr>
<td><strong>$^1$H Chemical Shieldings</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AEV only (No ∆-ML)$^b$</td>
<td>0.225</td>
<td>0.360</td>
</tr>
<tr>
<td>PBE0/STO-3G</td>
<td>0.651</td>
<td>0.110</td>
</tr>
<tr>
<td>PBE0/6-31G</td>
<td>0.247</td>
<td>0.060</td>
</tr>
<tr>
<td><strong>$^{15}$N Chemical Shieldings</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AEV only (No ∆-ML)$^b$</td>
<td>4.85</td>
<td>13.86</td>
</tr>
<tr>
<td>PBE0/STO-3G</td>
<td>21.65</td>
<td>3.14</td>
</tr>
<tr>
<td>PBE0/6-31G</td>
<td>5.63</td>
<td>0.84</td>
</tr>
<tr>
<td><strong>$^{17}$O Chemical Shieldings</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AEV only (No ∆-ML)$^b$</td>
<td>8.09</td>
<td>18.22</td>
</tr>
<tr>
<td>PBE0/STO-3G</td>
<td>31.95</td>
<td>4.50</td>
</tr>
<tr>
<td>PBE0/6-31G</td>
<td>7.10</td>
<td>1.39</td>
</tr>
</tbody>
</table>

$^a$ 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.

### 3.1 ∆-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 ∆-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 1). Table 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 $\Delta$-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. 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 $\Delta$-ML approach approximates those missing longer-range contributions via the inexpensive chemical shielding calculation on the entire molecule. As shown in Table 2, the SVWN/STO-3G $\Delta$-ML model training set RMSE of 1.3 ppm already represents substantial improvement over using the AEV descriptor alone. More importantly, the $\Delta$-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 ∆-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 density functional has little impact when a minimal basis set is used.

Switching to the larger 6-31G basis set for the ∆-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 ∆-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. To our knowledge, the ∆-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. The trade-off, of course, is that the ∆-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 ∆-ML models can be gained by investigating the difficulty of learning the chemical shieldings. Figure 2a plots the kernel density estimate (KDE) for the distribution of errors in the PBE0/6-31G chemical shieldings before and after adding the ∆-ML correction as a function of chemical shielding. The highest density of points
lies in the \(\sim 100-175\) 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 \(\sim 0-50\) 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 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 \(\Delta\)-ML correction provides, we mapped the small-basis shieldings onto the larger-basis ones via a linear regression. The resulting RMSEs for all \(\Delta\)-ML model combinations are listed in the “No \(\Delta\)-ML” columns of Table 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 \(\Delta\)-ML model gives. Moreover, the performance of the PBE0/6-31G-based \(\Delta\)-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 (\(\sim 1.8-2.8\) ppm) is even more dramatic. The simpler linear mapping performs even worse than the AEV-only model with no \(\Delta\)-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 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 \(\Delta\)-ML model. Improving the quality of the density functional has a smaller effect (\(\sim 0.1-0.2\) 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 4.3, the small basis shielding calculation requires only a small fraction of the computational time required to optimize the geometry, making the cost differences between functionals a relatively minor factor.
Figure 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.
Accordingly, the remainder of the paper focuses on the performance 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.

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 2 summarizes the key results for these atom types; results for all possible Δ-ML functional and basis set combinations are provided in SI Section S7. 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. This error range is also comparable to the expected ∼0.3–0.4 ppm accuracy for large-basis DFT relative to experiment.

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 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 $\sim 4$ ppm for $^{15}$N and $\sim 7$–8 ppm for $^{17}$O, versus $\sim 1.2$–1.5 ppm for $^{13}$C.$^{51,52}$ Second, the molecular data sets used here contain far fewer data samples for $^{15}$N and $^{17}$O (Table 1). As shown in Figures 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. $^1$° amine functional groups), while the ones for oxygen are concentrated in the 200-350 ppm shielding range (hydroxyl and ether groups). Third, Figures 2c–d also emphasize the highly non-linear relationships between 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 $\sim 20$–30 ppm (Table 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 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 2c–d highlights how effectively the NN
learns the correction to the small basis shieldings. These RMSEs for the PBE0/6-31G $\Delta$-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 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, $\Delta$-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 $\Delta$-ML approach is effectively capturing the long-range contributions to the chemical shielding that are absent in the AEV.

### 3.3 Uncertainty Quantification

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 $\Delta$-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. 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 $S_{\text{ens}}$ of the ensemble mean can inform about the uncertainty inherent in the prediction.

Here, we examine how the standard deviation $S_{\text{ens}}$ relates to the fidelity of the prediction.
Figure 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.

to the target DFT chemical shielding in the GDB17 testing data set. Figure 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 $S_{ens} > 1.00$. 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 3b plots the resulting probability curves. These probability curves highlight that smaller stan-
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 SI Section S7.5.

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 $\sim$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.
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, while the same functional and basis set give an RMSE of 3.3 ppm for molecules in solution (when the solvent environment is neglected). 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.

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 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. 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 ∆-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 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 SI Section S4 and S5. 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.

Figure 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 + ∆-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 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 + ∆-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.$^{3,38}$
Figure 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 + ∆-ML (blue). The data and regression lines for the two models in the upper panels are nearly indistinguishable.
Table 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.

<table>
<thead>
<tr>
<th></th>
<th>CDCl$_3$ ($N=163$)</th>
<th>DMSO ($N=44$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw +Δ-ML</td>
<td>Raw +Δ-ML</td>
</tr>
<tr>
<td>AEV</td>
<td>4.82</td>
<td>2.27</td>
</tr>
<tr>
<td>PBE0/STO-3G</td>
<td>11.0</td>
<td>12.0</td>
</tr>
<tr>
<td>PBE0/6-31G</td>
<td>2.00</td>
<td>2.12</td>
</tr>
<tr>
<td>PBE0/6-311+G(2d,p)</td>
<td>1.82</td>
<td>2.42</td>
</tr>
</tbody>
</table>

Table 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 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 3 to those against the target shieldings in Table 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.
Table 4: Root-mean-square errors (ppm) from the linear regressions of the “cheap” predicted shifts versus the PBE0/6-31+G(2d,p) ones after the regression against experiment has been performed.

<table>
<thead>
<tr>
<th></th>
<th>CDCl₃ (N=163)</th>
<th>DMSO (N=44)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw +∆-ML</td>
<td>10.4</td>
<td>11.4</td>
</tr>
<tr>
<td>PBE0/STO-3G</td>
<td>1.63</td>
<td>1.25</td>
</tr>
<tr>
<td>PBE0/6-31G</td>
<td>0.55</td>
<td>0.29</td>
</tr>
</tbody>
</table>

### 4.2 Predicting Experimental Chemical Shifts For Pharmaceutical Molecules

After establishing regression parameters for DMSO and CDCl₃, 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 CDCl₃ 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 compares the performance of PBE0/6-31G and PBE0/6-31G + ∆-ML $^{13}$C chemical shifts against the target PBE0/6-311+G(2d,p) shifts. The ∆-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 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 ∆-ML correction reduces these largest errors, tightening the error distribution appreciably (bottom-left panel). In other words, the ∆-ML correction is behaving as expected for these drugs, bringing the small-basis PBE0 shifts into better agreement with the large-basis target ones.
Figure 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.

Using the uncertainty estimates from the GDB17 set in Figure 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 3.3. See SI Section S5.11 for more details.

Next, Figure 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. 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.
Figure 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 $\Delta$-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).

Sorting the errors by chemical shift (three top-right panels of Figure 6) shows some of the same trends as were observed in Figure 5. PBE0/6-31G generally exhibits larger errors relative to experiment for larger chemical shifts, which again reflects the inadequacies 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 + $\Delta$-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 + $\Delta$-ML model.
Overall, the ∆-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 4 and Figure 5 highlight how much the ∆-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 2). Given the highly non-linear relationships between the PBE0/6-31G and target shieldings for $^{15}$N or $^{17}$O in Figure 2, one would expect much greater differences in the experimental accuracy of PBE0/6-31G with and without the ∆-ML correction for those nuclei.

4.3 Computational Timings

Finally, to give some perspective on the computational costs of the ∆-ML models, Table 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 ∆-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 ∆-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 ∆-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 ∆-ML model (Table 2) is arguably out-weighed by the computational savings. On the other hand, the cost of either ∆-ML shielding calculation is trivial compared to the geometry optimization. In the end, the most appropriate low-cost ∆-ML shielding calculation will depend on the application: When obtaining the geometry represents the computational bottleneck, the more accurate ∆-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 ∆-ML models become more attractive.

Table 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.

<table>
<thead>
<tr>
<th>Species</th>
<th>Geom Opt ωB97X 6-31G(d)</th>
<th>NMR Shielding calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ωB97X 6-311+G(2d,p)</td>
<td>PBE0 6-31G</td>
</tr>
<tr>
<td>Acetaminophen (C8H9NO2)</td>
<td>19.8</td>
<td>6.0 (61.1)</td>
</tr>
<tr>
<td>Aspirin (C9H8O4)</td>
<td>20.0</td>
<td>11.2 (106)</td>
</tr>
<tr>
<td>Nitrofurantoin (C8H6N4O5)</td>
<td>32.2</td>
<td>18.7 (165)</td>
</tr>
<tr>
<td>Nalidixic Acid (C12H12N2O3)</td>
<td>38.8</td>
<td>28.4 (257)</td>
</tr>
<tr>
<td>Cortisone Acetate (C23H30O6)</td>
<td>461</td>
<td>205 (1540)</td>
</tr>
<tr>
<td>Mean ∆-ML NMR Savings Factor</td>
<td>11 (33)</td>
<td>50 (42)</td>
</tr>
</tbody>
</table>
5 Conclusion

We have developed ∆-ML models using various combinations of density functionals and basis sets to predict isotropic chemical shieldings quickly. We first assessed the performance ∆-ML models for $^{13}$C data sets to establish trends regarding the roles of the functional and basis set. The ∆-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 ∆-ML models tested performed only slightly worse. The PBE0/6-31G + ∆-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 are a fraction of the errors expected for the target DFT model relative to experiment.

Using the PBE0/6-31G ∆-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 ∆-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 ∆-ML approach predicts shifts with accuracy that is only marginally worse. In other words, the ∆-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 ∆-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.

**Acknowledgement**

Funding for this work from the National Science Foundation (CHE-1665212) and supercomputer time from XSEDE (TG-CHE110064) are gratefully acknowledged. The authors thank Leonard Mueller and Jacob Holmes for providing experimental chemical shifts and geometries for the DMSO small molecule regression set.

**Supporting Information Available**

Lists of molecular structures used, sample job inputs, predicted and experimental chemical data for drug molecules, linear regression referencing model details, and additional machine learning details regarding hyperparameter optimization, training/validation errors, and com-

References


(7) Barone, G.; Gomez-Paloma, L.; Duca, D.; Silvestri, A.; Riccio, R.; Bifulco, G.


(31) Cuny, J.; Xie, Y.; Pickard, C. J.; Hassanali, A. A. Ab Initio Quality NMR Parameters


(46) Shen, L.; Wu, J.; Yang, W. Multiscale Quantum Mechanics/Molecular Mechanics Sim-


(62) Fink, T.; Reymond, J.-L. 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


(67) Neese, F.; Munzarová, M. L. In *Calculation of NMR and EPR Parameters*; Kaupp, M.,


(75) McDonagh, D.; Skylaris, C.-K.; Day, G. M. Machine-Learned Fragment-Based Energies


(82) Fulmer, G. R.; Miller, A. J. M.; Sherden, N. H.; Gottlieb, H. E.; Nudelman, A.; Stoltz, B. M.; Bercaw, J. E.; Goldberg, K. I. NMR Chemical Shifts of Trace Impurities: Common Laboratory Solvents, Organics, and Gases in Deuterated Solvents Rel-


Graphical TOC Entry