THE JOURNAL OF PHYSICAL CHEMISTRY C
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCC

Article

# A Machine Learning Model of Chemical Shifts for Chemically and Structurally Diverse Molecular Solids

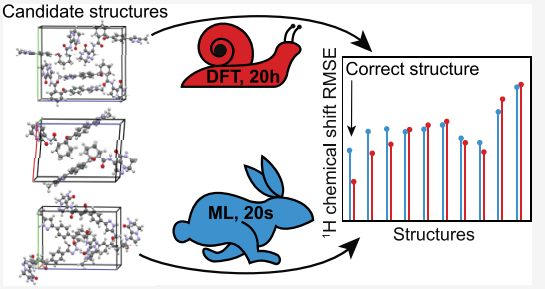
Manuel Cordova, Edgar A. Engel, Artur Stefaniuk, Federico Paruzzo, Albert Hofstetter, Michele Ceriotti, and Lyndon Emsley*

Cite This: J. Phys. Chem. C 2022, 126, 16710−16720

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Nuclear magnetic resonance (NMR) chemical shifts are a direct probe of local atomic environments and can be used to determine the structure of solid materials. However, the substantial computational cost required to predict accurate chemical shifts is a key bottleneck for NMR crystallography. We recently introduced ShiftML, a machine-learning model of chemical shifts in molecular solids, trained on minimum-energy geometries of materials composed of C, H, N, O, and S that provides rapid chemical shift predictions with density functional theory (DFT) accuracy. Here, we extend the capabilities of ShiftML to predict chemical shifts for both finite temperature structures and more chemically diverse compounds, while retaining the same speed and accuracy. For a benchmark set of 13 molecular solids, we find a root-mean-squared error of 0.47 ppm with respect to experiment for $^{1}$H shift predictions (compared to 0.35 ppm for explicit DFT calculations), while reducing the computational cost by over four orders of magnitude.




## INTRODUCTION

The atomic-level structures of solid materials are of high interest across many areas of chemistry. While X-ray diffraction (XRD) is the most well-established method for determining the structure of crystalline compounds, many materials lack the long-range order required to perform single-crystal XRD. Solid-state nuclear magnetic resonance (NMR) directly probes local atomic environments and so does not require a long-range order, making it a popular method for studying the structure of microcrystalline and disordered solids. Notably, combining solid-state NMR experiments with chemical shift calculations, a process typically referred to as NMR crystallography, allows determination of a wide range of structures,[1−4] from pharmaceuticals[5−7] to capping groups on nanoparticle surfaces[8] to the spacer layers in two-dimensional hybrid perovskite materials.[9] Striking recent examples include the determination of the structures of drug molecules in pharmaceutical formulations,[10,11] and the precise determination of the structure of active sites in enzyme reaction pathways[12,13] and of the disordered structure of an amorphous drug.[14]

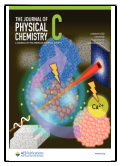
A key step in NMR crystallography is the computation of chemical shifts for candidate structures. Here, high accuracy is required in order to capture the effect of the particular conformation and packing of the molecular building blocks on the chemical shifts and to allow the identification of the correct structure among a set of potential candidates based on a comparison between computed and measured chemical shifts.[15−19] With the current best calculations, the root-mean-square error (RMSE) between the experiment and calculation can be as low as 1.5 ppm for $^{13}$C and 0.2 ppm for $^{1}$H.[2,18,20−22]

Plane-wave density functional theory (DFT) methods using the gauge including projected augmented wave (GIPAW) formalism[23−25] generally offer a good tradeoff between accuracy and computational cost for computing chemical shifts in small periodic structures. Consequently, DFT has been widely used in NMR crystallography to determine the structure of powdered solids.[1−3,26] However, the computational cost of DFT methods severely limits the size of systems accessible, preventing the study of large or disordered systems.

In recent years, machine-learning models have proven a powerful tool for supplementing and bypassing intensive quantum-mechanical calculations of molecular and atomic properties. In particular, NMR chemical shifts have been modeled using kernel methods[27−29] and neural networks.[30−35] Such approaches have proven able to yield chemical shifts to within DFT accuracy at a fraction of the computational cost, allowing applications to large ensembles of large systems.

We have previously introduced ShiftML,[15,36] a machine-learning model of chemical shifts trained on GIPAW DFT data for 3,546 structures from the Cambridge structural database

Received: June 3, 2022
Revised: August 24, 2022
Published: September 23, 2022



(CSD),[37] allowing fast and accurate predictions of chemical shifts for any molecular solid containing C, H, N, O, and S atoms. We have further demonstrated how this model can be used to enable new approaches in NMR crystallography. It has enabled structure determination of amorphous materials through chemical shift computations of molecular dynamics (MD) snapshots containing thousands of atoms[14] and enabled accounting for the effects of thermal and quantum-mechanical nuclear motion on the experimentally observable chemical shifts of crystalline solids based on path-integral molecular dynamics (PIMD) simulations.[19] Moreover, it has allowed on-the-fly chemical shift calculations in a chemical shift-driven direct structure determination protocol.[38] ShiftML also opens up the possibility to transform databases of crystal structures into databases of chemical shifts, which we have used, for example, to build a Bayesian framework to assign the NMR spectra of organic crystals.[39]

Although ShiftML constitutes a powerful method for computing chemical shifts with high accuracy and at a low computational cost, two important limitations prevent its more widespread use. First, the model is currently limited to compounds containing only C, H, N, O, and S atoms. While these elements are among the most prevalent in the CSD, numerous organic crystals contain elements outside of this set, leaving them beyond the scope of ShiftML. Second, the training set of ShiftML only contains structures that were geometry-optimized using DFT, resulting in lower accuracy for predictions on finite temperature or distorted structures, or for structures that are geometry-optimized using other methods (such as semi-empirical electronic structure calculations[40,41]).

Here, we present ShiftML2, an updated version of ShiftML, trained on GIPAW DFT chemical shifts for an extended set of over 14,000 structures containing any of 12 common elements (H, C, N, O, S, F, P, Cl, Na, Ca, Mg, and K) and composed of roughly equal amounts of relaxed and thermally perturbed structures of crystals extracted from the CSD. ShiftML2 shows slight improvements over the previous versions of ShiftML on DFT-relaxed structures ($^1$H RMSE of 0.47 ppm against 0.51 ppm for the ShiftML model described in ref 15, which we refer to as ShiftML1 here). More importantly, it effectively retains this accuracy for distorted structures, for which the performance of ShiftML1 degrades dramatically, while additionally allowing chemical shift computations for more chemically diverse structures.

## METHODS

**Configurational Sampling.** In order to construct suitable reference data for an accurate and robust ShiftML2 model, we first extracted all crystal structures from the CSD with unit cells containing no more than 200 atoms (for which high-throughput first-principles calculations are comparatively affordable) and including H and C, but no additional elements other than N, O, S, F, P, Cl, Na, Ca, Mg, and K. We note that we initially allowed the presence of Br and I atoms but later discarded the structures containing these atoms due to the need for relativistic corrections to obtain accurate shieldings for atoms in their vicinity. After extracting a random selection of 1,000 molecular crystals as a test set, the selection of the training set was performed by farthest point sampling (FPS)[42] of the remaining 140,373 structures based on the kernel-induced pairwise distances.

$$D(X_i, X_j) = k(X_i, X_i) + k(X_j, X_j) - 2k(X_i, X_j) \quad (1)$$

Here, the kernel function $k(\cdot, \cdot) = (X_i \cdot X_j)^2$ measures the similarity of the average smooth overlap of atomic positions (SOAP) power spectra[43] of the constituent atoms within a crystal structure, $X_i$, computed using the hyperparameters specified in Supplementary Table S1. The first 10,000 FPS-sorted (most structurally diverse) structures were selected as the training set.

All training and test structures were relaxed using DFT-fixed cell geometry optimizations using the Quantum ESPRESSO (QE) electronic structure package[44,45] with the PBE density functional,[46] a Grimme D2 dispersion correction,[47,48] wave-function and charge density energy cut-offs of 60 and 240 Ry, respectively, and ultrasoft pseudopotentials with GIPAW reconstruction.[49,50] To render this computation efficient, only the Gamma point was accounted for. Further details may be found in the SI.

Subsequently, short constant-volume MD simulations of 500 fs were performed using i-PI[51,52] to drive the dynamics, and the above QE setup to evaluate energies and forces. We used a timestep of 1 fs and a Generalized Langevin equation thermostat[53,54] to equilibrate the system at 300 K.

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

**GIPAW-DFT Chemical Shieldings.** The GIPAW NMR calculations were performed using the QE code with the same DFT parameters as for the structure relaxation above but using refined plane wave and charge density energy cut-offs of 100 and 400 Ry, respectively, a Monkhorst−Pack $k$-point grid[55] with a maximum spacing of 0.06 Å$^{-1}$, and the ultrasoft pseudopotentials with GIPAW reconstruction from the USSP pseudopotential database v1.0.0.

Finally, all structures were discarded, which displayed at least one outlier shift (defined as being outside the range of chemical shifts between the 1st and 99th percentile of all shifts of that element by at least 1.5 times that range), or where the calculation failed. Overall, 2650 structures were discarded because the self-consistent loop did not reach the high level of convergence needed for reliable GIPAW calculations, we removed 3313 additional structures containing Br or I atoms, and we discarded 24 structures that displayed outlier shieldings. This led to final training and test sets containing 14,254 and 1759 structures, respectively.

**Machine Learning Model.** We use kernel ridge regression (KRR)[56] to predict the isotropic chemical shielding of an atom based on its local atomic environment as follows:

$$\sigma(X) = \sum_i^N w_i k(X, X_i) = \sum_i^N w_i (X^T \cdot X_i)^\zeta \quad (2)$$

where $X$ and $X_i$ are symmetry-adapted descriptors, which encode the local atomic environment around the atom of interest and those in the training set, respectively, and $w_i$ denotes the regression weight associated with training sample $i$. $k(\cdot,\cdot)$ is the kernel function that defines the similarity between two atomic environments. Here, we measure the similarity between two environments as the scalar product between the vectors corresponding to their descriptor, raised to a power $\zeta$. Training a KRR model involves determining the weights $w_i$ such that eq 2 is best satisfied for the training data, with an


https://doi.org/10.1021/acs.jpcc.2c03854
J. Phys. Chem. C 2022, 126, 16710−16720

additional regularization term that reduces the magnitude of regression weights. Further information is available in the SI.

**Uncertainty Estimation.** Uncertainty estimation is performed using a resampling approach to generate a committee of $M = 32$ KRR models,[57] trained on random two-fold splits of the training data. The final prediction for a sample $i$ in the test set, $\hat{\sigma}_i$, is given by the mean of the prediction for each model, and the estimated uncertainty is defined as the standard deviation $s_i$ of the prediction of each model, rescaled by a factor $\alpha$ given by[57]

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

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

**Local Atomic Environment Descriptor.** We describe local atomic environments using smooth overlap of atomic positions (SOAP) power spectra[43] as implemented in librascal.[58] We use a sparse implementation of the SOAP descriptors, making use of the sparsity of elements in local environments around individual atoms.

The relevant hyperparameters were optimized by five-fold cross-validation performed on the $^1$H environments of a subset of 1000 training structures, selected at random other than including all training structures containing Na, Ca, Mg, or K. The latter ensures that these elements are represented during hyperparameter optimization, despite their low abundance in the training data. The structures selected for hyperparameter optimization contain a total of 27,802 $^1$H environments. In each cross-validation fold, the training data were partitioned into three equal parts, and a KRR model was trained on each part. This was done in order to reduce the computational resources required to train the models for each split. The selection of descriptor parameter values was based on the RMSE obtained on the validation data. The explored and selected hyperparameter values can be found in the SI. We note that ref 15 found almost identical hyperparameters to be optimal for H, C, N, O, and S through independent optimizations for the different elements. We therefore apply the hyperparameters optimized for $^1$H to the other elements without further optimization, except for the optimal radial basis,[59] which was constructed individually with the complete final training data for each element.

**Farthest Point Sampling of Training Environments.** The training data were sorted using FPS[42] based on distances between pairs of environments $X_i$ and $X_j$ defined as in eq 1. This serves two purposes: first, it permits the removal of duplicate environments arising from, for example, equivalent atomic sites related by the crystal symmetries in relaxed structures. Second, it identifies the most structurally diverse set of training environments.

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

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

**Atom Type Identification.** The different atom types, defined here as hybridization and formal charge, in the training and test structures were identified using the RDKit[60] Python package on the asymmetric unit of the crystals extracted using the CSD Python API.[37] The structures where RDKit failed to identify bonds and/or formal charges were discarded from the atom-type analysis. Carbon atoms identified as charged were set to a neutral charge, as well as nitrogen atoms identified with a negative charge and oxygen atoms identified as positively charged. This was done upon visual inspection of a subset of crystal structures displaying such unusual atom types, confirming that such atom types were incorrectly determined by the package. In total, atomic types of 6,960 out of the 10,593 final training structures and 1,443 out of the 1,759 test structures were identified.

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

## RESULTS AND DISCUSSION

**Training Set Selection and Model Training.** Because of the lack of large databases of experimental chemical shifts in molecular solids, we trained the model on shielding values computed by DFT, as was done previously for ShiftML1.[15,36] This ensures both consistency in the training data as well as the ability to perform high-throughput computations to obtain a substantial amount of training data in reasonable time. The training structures were chosen to be as diverse as possible through FPS. Because computed shieldings are related to chemical shifts by a simple linear relationship, we use the two terms interchangeably.

High quality of the training data is key to producing an accurate machine learning model. In addition, the kernel model framework used here has a linear time and memory



https://doi.org/10.1021/acs.jpcc.2c03854
J. Phys. Chem. C 2022, 126, 16710−16720

complexity with respect to the training set size for inference. It is thus important to reduce the amount of training data while retaining diverse atomic environments and removing outliers to obtain both fast and accurate predictions of chemical shifts. To this end, we performed an iterative, batched FPS of the chemical environments, as described in the Methods section. Figure 1A shows the first and last FPS iterations on typical batches. The significant drop in minimum distance between FPS-selected samples after selecting 30,000 of the 50,000 environments in an initial batch corresponds to symmetrically equivalent atomic sites in relaxed crystal structures. After gathering the FPS-selected environments from all batches after the final iteration, we obtained 67,535 $^{1}$H environments.

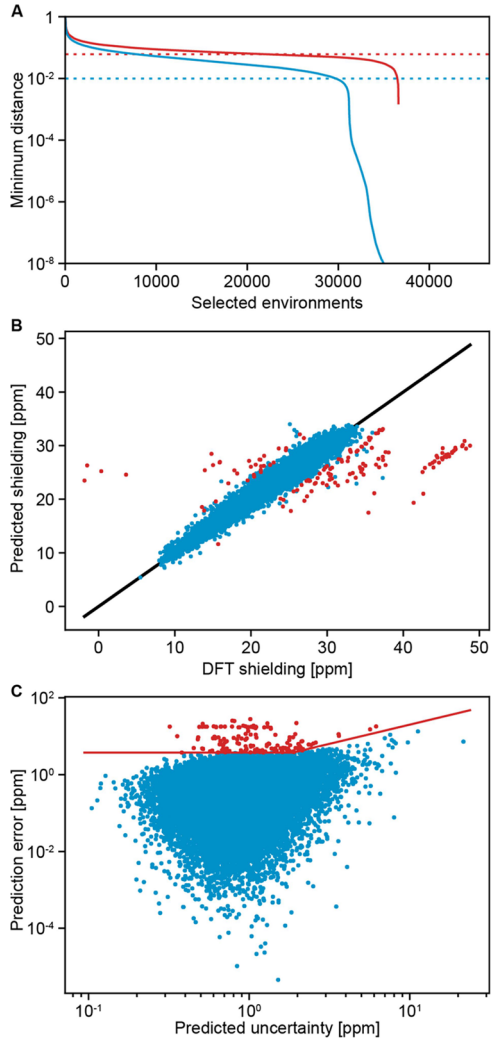



**Figure 1.** (A) First (blue) and last (red) FPS selection step for a batch of up to 50,000 $^{1}$H environments. The blue and red dashed lines show the threshold for the minimum distance between FPS-selected samples set to select environments in the first and last FPS selection steps, respectively.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

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

The final model was constructed by training 32 models on random half splits of the remaining training environments. Prediction uncertainties were estimated as the rescaled standard deviation of the 32 predictions to fit the error distribution, as described in ref 61.

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

We observe a reduced prediction uncertainty and error for shieldings above 20 ppm (see Supplementary Figure S8). This behavior is expected considering that 90% of the training data have DFT shieldings computed above 20 ppm, which corresponds to typical chemical shifts of aliphatic and aromatic CH protons (<10 ppm). The reduced density of training data at lower shieldings (corresponding to higher chemical shifts) results in increased error and uncertainty of the predictions.

To compare ShiftML1 and ShiftML2, we apply both models to the ShiftML1 test set, as well as all structures from the current test set which contain exclusively H, C, N, O, and S atoms (i.e., those for which ShiftML1 is applicable). Figure 3 shows the $^{1}$H shift predictions of the two models for the ShiftML1 test set (Figure 3A, B) and for the relaxed (Figure 3C,D) and finite temperature (Figure 3E,F) structures from the ShiftML2 test set, which only contain H, C, N, O, and S atoms. Table 1 summarizes the results obtained by both models. There are two striking conclusions that are illustrated by the figure and table. First, overall, ShiftML2 displays slight improvements over ShiftML1 for relaxed structures (0.47 ppm RMSE compared to 0.49 ppm on the ShiftML1 test set, and 0.47 ppm RMSE compared to 0.51 ppm on relaxed structures from the ShiftML2 test set), indicating that the increase in the number of training environments was sufficient to avoid deterioration of the accuracy despite the greater chemical diversity. Second, ShiftML2 is substantially more accurate for finite temperature structures (0.53 ppm RMSE for ShiftML2 compared to 0.98 ppm for ShiftML1), highlighting the greater robustness of a model trained on finite temperature structures


https://doi.org/10.1021/acs.jpcc.2c03854
J. Phys. Chem. C 2022, 126, 16710−16720

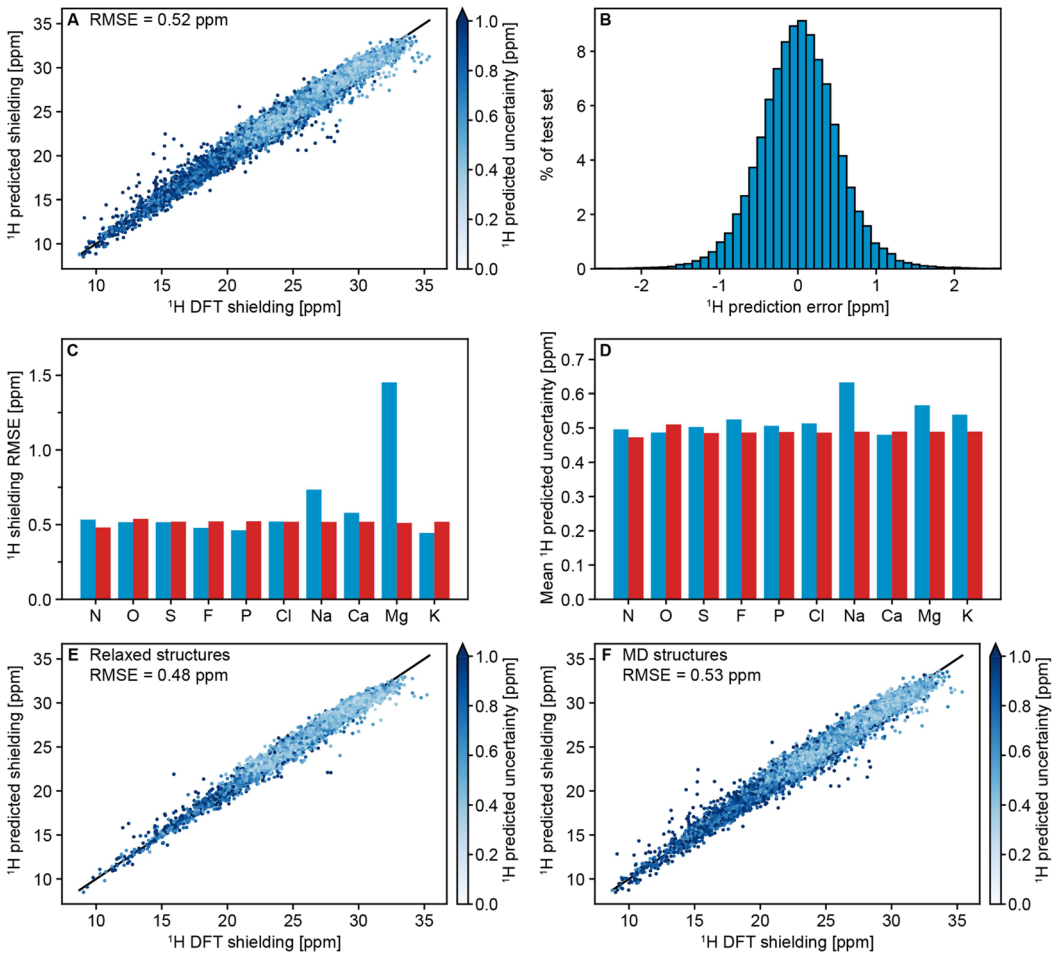



**Figure 2.** (A) Comparison of DFT-computed $^1$H shieldings and ShiftML2 predictions on the test set. (B) Histogram of the of prediction error between ShiftML2 predictions and DFT-computed shieldings for $^1$H environments. Comparison of $^1$H (C) chemical shift RMSE and (D) average prediction uncertainties on test structures containing (blue) or lacking (red) a given element. Comparison of DFT-computed $^1$H shieldings and ShiftML2 predictions on (E) relaxed and (F) MD structures in the test set. Black lines in (A, E, and F) show perfect correlations.

when predicting atomic properties for distorted structures. To confirm the robustness of ShiftML2 toward distorted structures, we evaluated the error against DFT-computed $^1$H shieldings for up to 50 snapshots taken every 100 fs from MD simulations of the crystal structures of cocaine, AZD5718 and form 4 of AZD8329. We found that the average RMSEs along the MD trajectories were only slightly above the RMSEs obtained for the relaxed structures (0.58 ppm against 0.55 ppm RMSE for AZD5718, 0.50 ppm against 0.45 ppm RMSE for form 4 of AZD8329, and 0.49 ppm against 0.42 ppm RMSE for cocaine, see Supplementary Figure S9).

This is a key improvement compared to the previous ShiftML version because it allows accurate predictions of chemical shifts beyond relaxed structures and yields a better description of shifts in (PI)MD snapshots, and for intermediate structures during structural optimization.

The ability of the model to generalize to distorted structures is key in many applications of NMR crystallography. In particular, it allows accurate computation of chemical shifts on structures that are geometry optimized with different levels of theories (e.g., force fields or DFTB), which is important for the accurate description of shifts in MD simulations of materials.[14] It also enables more confident on-the-fly computations of chemical shifts of intermediate structures during the optimization of crystal structures in chemical-shift driven structure determination protocols, resulting in a potentially more powerful driving force toward the experimental structure.[38]

Figure 4 shows the prediction error for different types of protons in the test set. The two most common proton types H−C(sp3) and H−C(aromatic), making up 90% of the test set, yielded chemical shift RMSEs below 0.5 ppm. All other proton types displayed chemical shift RMSEs below 0.9 ppm, with the exception of alkyne protons, for which the RMSE was found to be 5.3 ppm. Such a high error is explained by the presence of only two alkyne protons identified in the final training data. Interestingly, we find that protons attached to nitrogens in charged groups display a lower error compared to their neutral counterparts. Molecular salts were found to display comparable shift RMSEs to neutral compounds. H-bonded protons yielded a chemical shift RMSE of 0.79 ppm.

**Experimental Benchmark Set and Polymorphs.** We evaluate the accuracy of the model with respect to experimental $^1$H chemical shifts using a benchmark set of 13 molecular crystals made up of cocaine, form 4 of AZD8329, theophylline, uracil, naproxen, the co-crystal of 3,5-dimethylimidazole, 4,5-dimethylimidazole, AZD5718, furosemide, flutamide, the co-crystal of indomethacin and nicotinamide, flufenamic acid, the potassium salt of penicillin G, and phenylphosphonic acid.[14,26,36,62−65] Figure 5A compares the


https://doi.org/10.1021/acs.jpcc.2c03854
J. Phys. Chem. C 2022, 126, 16710−16720

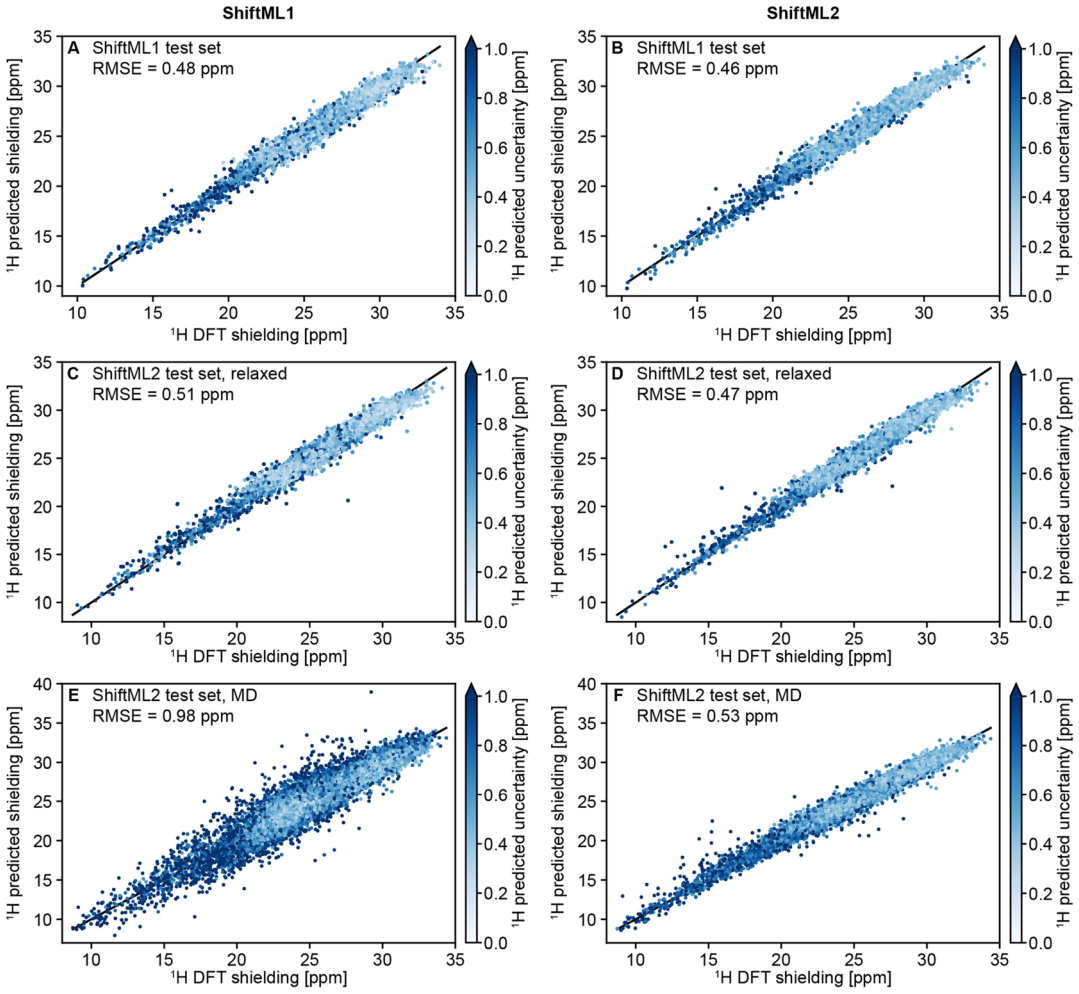



**Figure 3.** Comparison of DFT-computed $^1$H shieldings and predictions using ShiftML1 (A, C, E) or ShiftML2 (B, D, F) on (A, B) the ShiftML1 test set, (C, D) relaxed structures containing only H, C, N, O, and S in the ShiftML2 test set, and (E, F) MD structures containing only H, C, N, O, and S in the ShiftML2 test set. Black lines show perfect correlations.

**Table 1. Chemical Shift Root-Mean-Square Error (RMSE), Mean Absolute Error (MAE), and $R^2$ Coefficient of ShiftML1 and ShiftML2 Compared to DFT-Computed Shieldings**[a]

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

[a]The values are given for ShiftML1 and ShiftML2, separated by a slash.

predicted and experimentally measured shifts for this set. We obtain a RMSE of 0.47 ppm, compared to 0.35 ppm using DFT. For reference, the RMSE obtained on the experimental benchmark set for ShiftML1 (containing the six first molecular solids mentioned previously) is 0.41 ppm for ShiftML2, compared to 0.39 ppm for ShiftML1 and 0.36 for DFT. This further highlights that the accuracy of ShiftML1 for relaxed structures has been retained by ShiftML2, while extending the capabilities of the model to predict shifts for more chemically and structurally diverse structures. Notably, within the limits of the small experimental set used here, the accuracy against experimental shifts is found to decrease when including structures containing F, Cl, P, or K atoms, while DFT remained at the same level of accuracy. Because no such deterioration is observed for the structures in the test set (see Figure 2C), we attribute this to the chemical environments in the experimental set, which are not well represented in the training data.

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

The ability to determine the correct structure from among a set of candidates based on comparison between experimental and computed shifts is key to NMR crystallography. Figure 5B−D shows the RMSE between experimental and predicted $^1$H shifts for different sets of candidate structures for cocaine, form 4 of AZD8329, and AZD5718. The correct candidates systematically yielded a chemical shift RMSE below 0.6 ppm and corresponded to the lowest RMSE among the sets of candidates for form 4 of AZD8329 and AZD5718 and to the second lowest RMSE for cocaine.

**Models for Other Nuclei.** In addition to $^1$H, we constructed models for all the other nuclei present in the



https://doi.org/10.1021/acs.jpcc.2c03854
J. Phys. Chem. C 2022, 126, 16710−16720

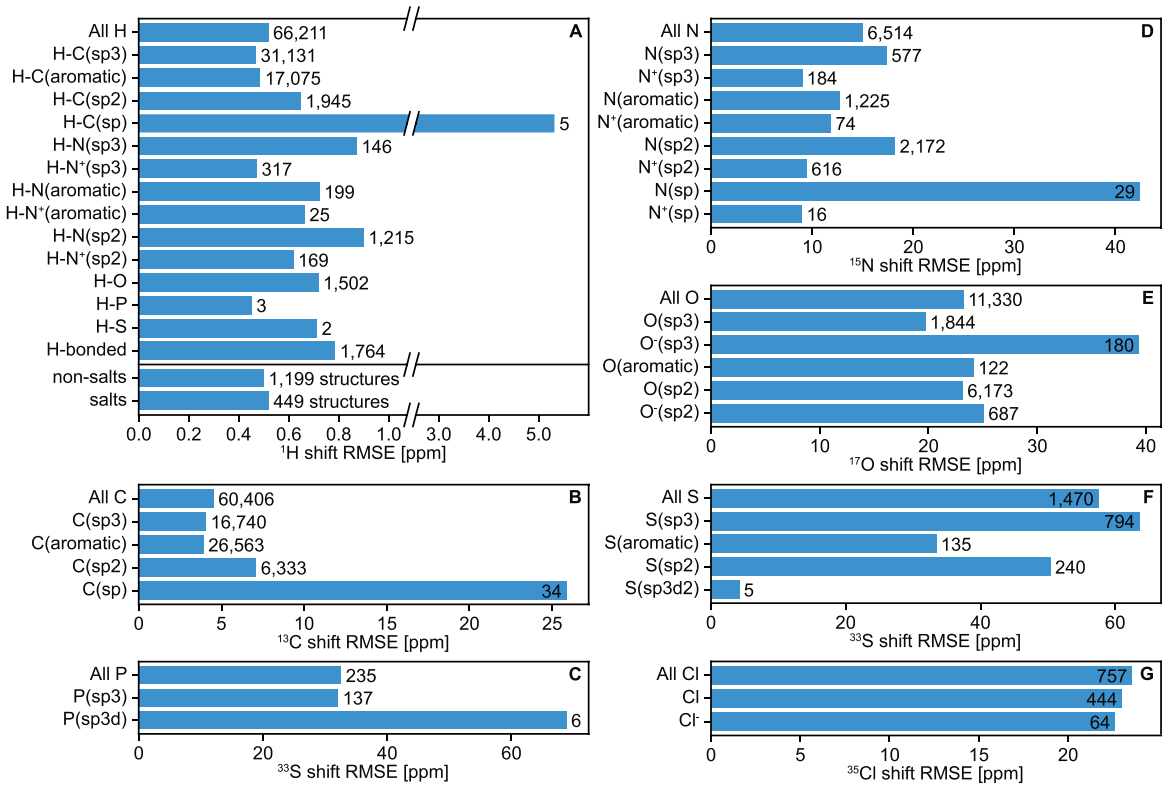



**Figure 4.** Chemical shift RMSE for different types of (A) $^{1}$H, (B) $^{13}$C, (C) $^{31}$P, (D) $^{15}$N, (E) $^{17}$O, (F) $^{33}$S, and (G) $^{35}$Cl in the test set. The number of environments (or structures) in the test set contributing to each bar is indicated next to it.

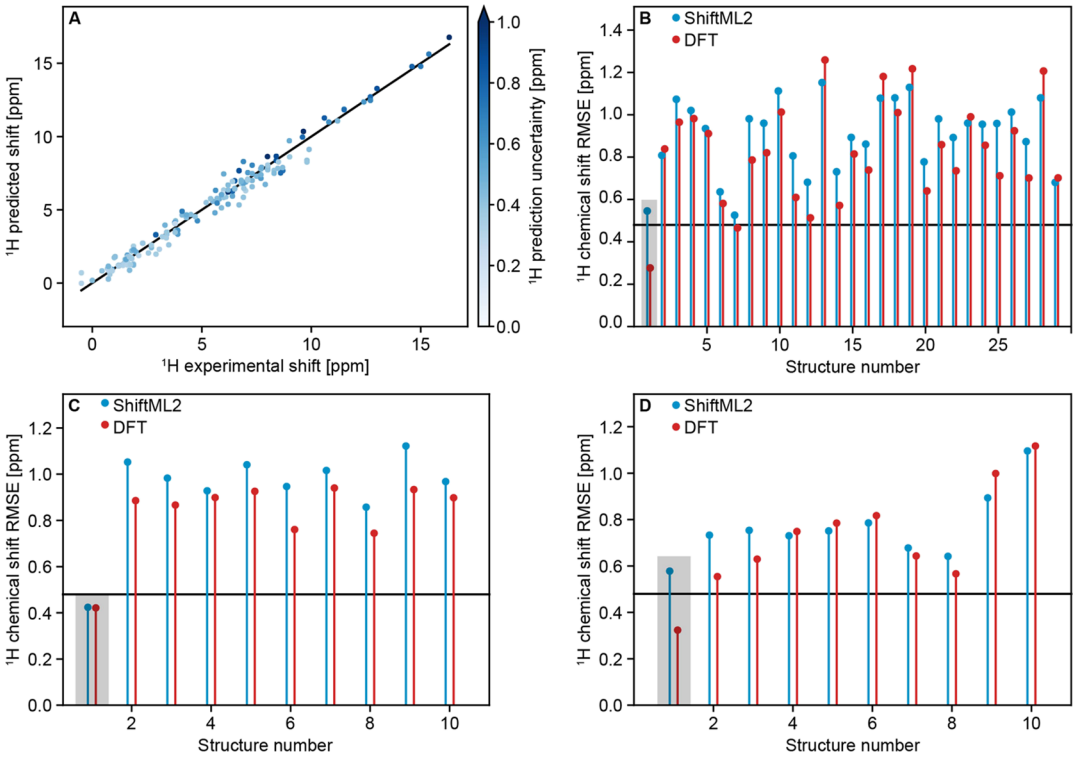



**Figure 5.** (A) Comparison between predicted and experimental $^{1}$H shifts for 13 molecular solids. Black line shows perfect correlation. Chemical shift RMSE obtained by ShiftML2 (blue) and DFT (red) against experimental shifts for candidate structures of (B) cocaine, (C) AZD8329 form 4, and (D) AZD5718. The correct crystal structure is indicated by the gray zone. The black horizontal lines indicate the expected RMSE between ShiftML2 predictions and experimental shifts (0.47 ppm).

training data. Figure 6, Supplementary Figure S7, and Table 2 compare the resulting predictions for the nuclei beyond $^{1}$H to GIPAW DFT shieldings. We note that although we refer to a particular nucleus (e.g., $^{15}$N), the isotropic chemical shift of all



https://doi.org/10.1021/acs.jpcc.2c03854
J. Phys. Chem. C 2022, 126, 16710−16720

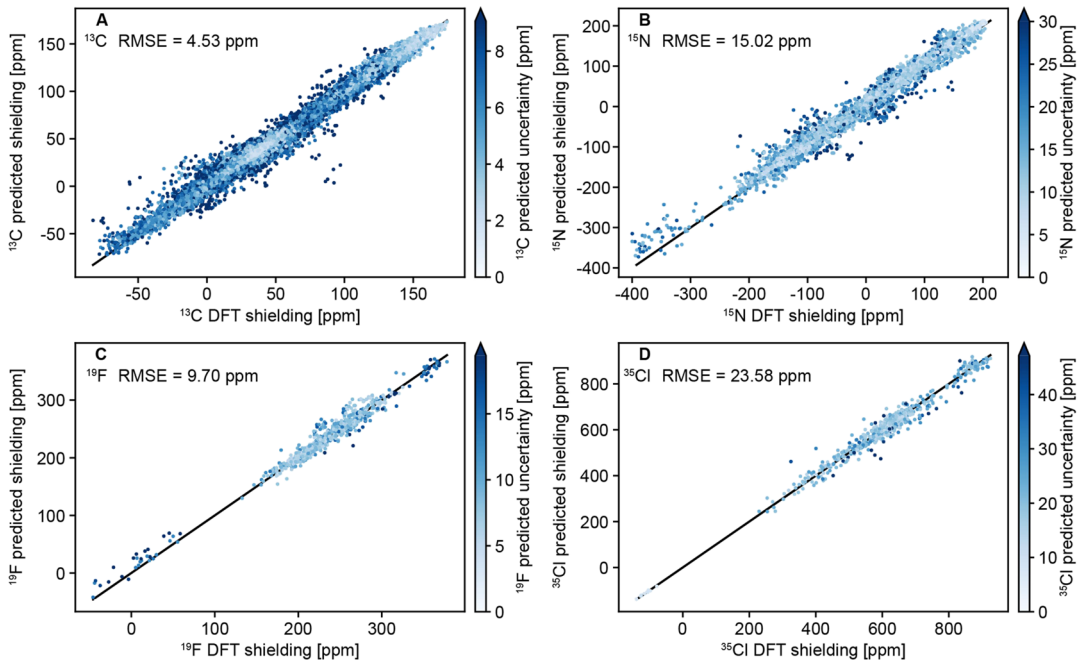



**Figure 6.** Comparison of DFT-computed and predicted (A)$^{13}$C, (B) $^{15}$N, (C) $^{19}$F, and (D) $^{35}$Cl chemical shifts in the test set. Black lines show perfect correlation.

**Table 2. Training and Test Size, Chemical Shift RMSE, MAE, and $R^2$ Coefficient for ShiftML2 Models Trained on Nuclei beyond $^1$H**

| nucleus | training set size | test set size | RMSE [ppm] | MAE [ppm] | $R^2$ |
|---|---|---|---|---|---|
| $^{13}$C | 65,498 | 60,406 | 4.53 | 3.12 | 0.99 |
| $^{15}$N | 65,506 | 6514 | 15.02 | 9.99 | 0.98 |
| $^{17}$O | 65,488 | 11,330 | 23.18 | 16.21 | 0.98 |
| $^{19}$F | 23,958 | 865 | 9.70 | 6.85 | 0.97 |
| $^{33}$S | 18,509 | 1470 | 57.53 | 35.12 | 0.87 |
| $^{31}$P | 5337 | 235 | 32.61 | 17.64 | 0.70 |
| $^{35}$Cl | 15,780 | 757 | 23.58 | 17.02 | 0.97 |
| $^{23}$Na | 728 | 14 | 5.77 | 4.58 | 0.57 |
| $^{43}$Ca | 386 | 8 | 13.01 | 10.77 | 0.99 |
| $^{25}$Mg | 186 | 10 | 12.27 | 8.21 | 0.94 |
| $^{39}$K | 632 | 9 | 9.33 | 7.07 | 0.39 |

NMR-active isotopes of a particular element can be predicted with the same accuracy, by adapting the offset (and slope) used to convert computed shieldings into chemical shifts. We obtain strong correlations ($R^2$ > 0.95) for $^{13}$C, $^{15}$N, $^{17}$O, $^{19}$F, and $^{35}$Cl. This indicates that ShiftML2 can accurately predict chemical shifts for these elements, although the absolute error is higher than that for $^1$H because of the larger chemical shift ranges for these nuclei (see Table 2). The lower number of training environments for $^{31}$P, $^{23}$Na, $^{43}$Ca, $^{25}$Mg, and $^{39}$K was found to lead to lower correlation with DFT-computed shifts. While we still provide models for these nuclei, we acknowledge that more accurate models based on more extensive training data would be required to obtain more accurate predictions for these elements. We reiterate that the main purpose of including these elements in the training data was to allow prediction of $^1$H, $^{13}$C, or $^{15}$N chemical shifts for structures containing such elements. Detailed ShiftML2 prediction accuracies for different types of $^{13}$C, $^{15}$N, $^{17}$O, $^{31}$P, $^{33}$S, and $^{35}$Cl nuclei are shown in Figure 4B–G. As for $^1$H, we observe a loss of accuracy for sp-hybridized $^{13}$C and $^{15}$N. The other nuclei ($^{19}$F, $^{23}$Na, $^{43}$Ca, $^{25}$Mg, and $^{39}$K) each displayed a unique atomic type across the test set.

## CONCLUSIONS

We have presented a machine learning model of chemical shifts that improves on our previously published model[16] in two key ways. First, the chemical diversity covered by the model has been extended from 5 to 12 elements, meaning that shifts for a much larger space of compounds can now be accessed. Second, finite temperature structures have been included in the training data, allowing reliable chemical shift predictions for distorted structures.

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

The capacity to calculate shifts for distorted structures is important for two reasons. First, it allows reliable shifts to be calculated for structures that are not geometry-optimized using DFT, such as structures optimized using (semi-)empirical approaches such as DFTB and for structures from MD simulations. Second, it means that shifts calculated for structures generated in a simulated annealing structure determination protocol[38] will be accurate even when the trial structure is not in an energy minimum, potentially providing a much more efficient driving force toward the correct structures, and this will be the subject of future studies. The model presented here scales linearly with respect to the number of local atomic environments in a structure of interest, making shifts for large ensembles of large structures accessible.



https://doi.org/10.1021/acs.jpcc.2c03854
J. Phys. Chem. C 2022, 126, 16710−16720

The new model will thus accelerate NMR crystallography by allowing large-scale computations for candidate structures, either from MD trajectories or in direct optimization methods.

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

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcc.2c03854.

SOAP hyperparameters; GIPAW pseudopotentials; details on model training; and training set selection for all elements beyond $^{1}$H and number of training samples in different $^{1}$H DFT shielding ranges (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Lyndon Emsley** − *Laboratory of Magnetic Resonance, Institute of Chemical Sciences and Engineering and National Centre for Computational Design and Discovery of Novel Materials MARVEL, Ecole Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland;* orcid.org/0000-0003-1360-2572; Email: lyndon.emsley@epfl.ch

### Authors

**Manuel Cordova** − *Laboratory of Magnetic Resonance, Institute of Chemical Sciences and Engineering and National Centre for Computational Design and Discovery of Novel Materials MARVEL, Ecole Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland;* orcid.org/0000-0002-8722-6541

**Edgar A. Engel** − *Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, U.K.*

**Artur Stefaniuk** − *Laboratory of Magnetic Resonance, Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland*

**Federico Paruzzo** − *Laboratory of Magnetic Resonance, Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland;* Present Address: Bruker Switzerland AG, 8117 Fällanden, Switzerland

**Albert Hofstetter** − *Laboratory of Magnetic Resonance, Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland;* Present Address: Laboratory of Physical Chemistry, ETH, Zurich, CH-8093 Zurich, Switzerland

**Michele Ceriotti** − *National Centre for Computational Design and Discovery of Novel Materials MARVEL and Laboratory of Computational Science and Modelling, Institute of Materials, Ecole Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland;* orcid.org/0000-0003-2571-2832

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcc.2c03854


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation Grant No. 200020_178860 and by the NCCR MARVEL, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant number 182892). E.A.E. acknowledges financial support through a Junior Research Fellowship by Trinity College, Cambridge.

## ■ REFERENCES


(1) Reif, B.; Ashbrook, S. E.; Emsley, L.; Hong, M. Solid-state NMR spectroscopy. *Nat. Rev. Methods Primers* **2021**, *1*, 2.

(2) Hodgkinson, P. NMR crystallography of molecular organics. *Prog. Nucl. Magn. Reson. Spectrosc.* **2020**, *118-119*, 10−53.

(3) Baias, M.; Dumez, J. N.; Svensson, P. H.; Schantz, S.; Day, G. M.; Emsley, L. De novo determination of the crystal structure of a large drug molecule by crystal structure prediction-based powder NMR crystallography. *J. Am. Chem. Soc.* **2013**, *135*, 17501−17507.

(4) Southern, S. A.; Bryce, D. L. In *Annual Reports on Nmr Spectroscopy*, Webb, G. A., Ed., 2021; vol *102*; pp 1−80.

(5) Czernek, J.; Brus, J. Polymorphic Forms of Valinomycin Investigated by NMR Crystallography. *Int. J. Mol. Sci.* **2020**, *21*, 4907.

(6) Dudek, M. K.; Paluch, P.; Sniechowska, J.; Nartowski, K. P.; Day, G. M.; Potrzebowski, M. J. Crystal structure determination of an elusive methanol solvate - hydrate of catechin using crystal structure prediction and NMR crystallography. *CrystEngComm* **2020**, *22*, 4969−4981.

(7) Nilsson Lill, S. O.; Widdifield, C. M.; Pettersen, A.; Svensk Ankarberg, A.; Lindkvist, M.; Aldred, P.; Gracin, S.; Shankland, N.; Shankland, K.; Schantz, S.; et al. Elucidating an Amorphous Form Stabilization Mechanism for Tenapanor Hydrochloride: Crystal Structure Analysis Using X-ray Diffraction, NMR Crystallography, and Molecular Modeling. *Mol. Pharmaceutics* **2018**, *15*, 1476−1487.

(8) Al-Johani, H.; Abou-Hamad, E.; Jedidi, A.; Widdifield, C. M.; Viger-Gravel, J.; Sangaru, S. S.; Gajan, D.; Anjum, D. H.; Ould-Chikh, S.; Hedhili, M. N.; et al. The structure and binding mode of citrate in the stabilization of gold nanoparticles. *Nat. Chem.* **2017**, *9*, 890−895.

(9) Hope, M. A.; Nakamura, T.; Ahlawat, P.; Mishra, A.; Cordova, M.; Jahanbakhshi, F.; Mladenovic, M.; Runjhun, R.; Merten, L.; Hinderhofer, A.; et al. Nanoscale Phase Segregation in Supramolecular pi-Templating for Hybrid Perovskite Photovoltaics from NMR Crystallography. *J. Am. Chem. Soc.* **2021**, *143*, 1529−1538.

(10) Ni, Q. Z.; Yang, F.; Can, T. V.; Sergeyev, I. V.; D'Addio, S. M.; Jawla, S. K.; Li, Y.; Lipert, M. P.; Xu, W.; Williamson, R. T.; et al. In Situ Characterization of Pharmaceutical Formulations by Dynamic Nuclear Polarization Enhanced MAS NMR. *J. Phys. Chem. B* **2017**, *121*, 8132−8141.

(11) Brus, J.; Czernek, J.; Hruby, M.; Svec, P.; Kobera, L.; Abbrent, S.; Urbanova, M. Efficient Strategy for Determining the Atomic-Resolution Structure of Micro- and Nanocrystalline Solids within Polymeric Microbeads: Domain-Edited NMR Crystallography. *Macromolecules* **2018**, *51*, 5364−5374.

(12) Mueller, L. J.; Dunn, M. F. NMR crystallography of enzyme active sites: probing chemically detailed, three-dimensional structure in tryptophan synthase. *Acc. Chem. Res.* **2013**, *46*, 2008−2017.

(13) Klein, A.; Rovo, P.; Sakhrani, V. V.; Wang, Y.; Holmes, J. B.; Liu, V.; Skowronek, P.; Kukuk, L.; Vasa, S. K.; Guntert, P.; et al. Atomic-resolution chemical characterization of (2x)72-kDa tryptophan synthase via four- and five-dimensional (1)H-detected solid-state NMR. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119*, No. e2114690119.

(14) Cordova, M.; Balodis, M.; Hofstetter, A.; Paruzzo, F.; Nilsson Lill, S. O.; Eriksson, E. S. E.; Berruyer, P.; Simoes de Almeida, B.; Quayle, M. J.; Norberg, S. T.; et al. Structure determination of an amorphous drug through large-scale NMR predictions. *Nat. Commun.* **2021**, *12*, 2964.

(15) Engel, E. A.; Anelli, A.; Hofstetter, A.; Paruzzo, F.; Emsley, L.; Ceriotti, M. A Bayesian approach to NMR crystal structure determination. *Phys. Chem. Chem. Phys.* **2019**, *21*, 23385−23400.

(16) Dracinsky, M.; Unzueta, P.; Beran, G. J. O. Improving the accuracy of solid-state nuclear magnetic resonance chemical shift

prediction with a simple molecular correction. *Phys. Chem. Chem. Phys.* **2019**, *21*, 14992−15000.
(17) Dracinsky, M.; Moller, H. M.; Exner, T. E. Conformational Sampling by Ab Initio Molecular Dynamics Simulations Improves NMR Chemical Shift Predictions. *J. Chem. Theory Comput.* **2013**, *9*, 3806−3815.
(18) Hartman, J. D.; Kudla, R. A.; Day, G. M.; Mueller, L. J.; Beran, G. J. Benchmark fragment-based (1)H, (13)C, (15)N and (17)O chemical shift predictions in molecular crystals. *Phys. Chem. Chem. Phys.* **2016**, *18*, 21686−21709.
(19) Engel, E. A.; Kapil, V.; Ceriotti, M. Importance of Nuclear Quantum Effects for NMR Crystallography. *J. Phys. Chem. Lett.* **2021**, *12*, 7701−7707.
(20) Hartman, J. D.; Day, G. M.; Beran, G. J. Enhanced NMR Discrimination of Pharmaceutically Relevant Molecular Crystal Forms through Fragment-Based Ab Initio Chemical Shift Predictions. *Cryst. Growth Des.* **2016**, *16*, 6479−6493.
(21) Beran, G. J. O. Calculating Nuclear Magnetic Resonance Chemical Shifts from Density Functional Theory: A Primer. *eMagRes* **2019**, *8*, 215−226.
(22) Dracinsky, M.; Vicha, J.; Bartova, K.; Hodgkinson, P. Towards Accurate Predictions of Proton NMR Spectroscopic Parameters in Molecular Solids. *ChemPhysChem* **2020**, *21*, 2075−2083.
(23) Yates, J. R.; Pickard, C. J.; Mauri, F. Calculation of NMR chemical shifts for extended systems using ultrasoft pseudopotentials. *Phys. Rev. B* **2007**, *76*, No. 024401.
(24) Harris, R. K.; Hodgkinson, P.; Pickard, C. J.; Yates, J. R.; Zorin, V. Chemical shift computations on a crystallographic basis: some reflections and comments. *Magn. Reson. Chem.* **2007**, *45*, S174−S186.
(25) Pickard, C. J.; Mauri, F. All-electron magnetic response with pseudopotentials: NMR chemical shifts. *Phys. Rev. B* **2001**, *63*, No. 245101.
(26) Baias, M.; Widdifield, C. M.; Dumez, J. N.; Thompson, H. P.; Cooper, T. G.; Salager, E.; Bassil, S.; Stein, R. S.; Lesage, A.; Day, G. M.; Emsley, L. Powder crystallography of pharmaceutical materials by combined crystal structure prediction and solid-state 1H NMR spectroscopy. *Phys. Chem. Chem. Phys.* **2013**, *15*, 8069−8080.
(27) Gupta, A.; Chakraborty, S.; Ramakrishnan, R. Revving up C-13 NMR shielding predictions across chemical space: benchmarks for atoms-in-molecules kernel machine learning with new data for 134 kilo molecules. *Mach. Learn.: Sci. Technol.* **2021**, *2*, No. 035010.
(28) Gerrard, W.; Bratholm, L. A.; Packer, M. J.; Mulholland, A. J.; Glowacki, D. R.; Butts, C. P. IMPRESSION - prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chem. Sci.* **2020**, *11*, 508−515.
(29) Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O. A. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J. Phys. Chem. Lett.* **2015**, *6*, 3309−3313.
(30) Yang, Z.; Chakraborty, M.; White, A. D. Predicting chemical shifts with graph neural networks. *Chem. Sci.* **2021**, *12*, 10802−10809.
(31) Guan, Y.; Shree Sowndarya, S. V.; Gallegos, L. C.; St John, P. C.; Paton, R. S. Real-time prediction of (1)H and (13)C chemical shifts with DFT accuracy using a 3D graph neural network. *Chem. Sci.* **2021**, *12*, 12012−12026.
(32) Liu, S.; Li, J.; Bennett, K. C.; Ganoe, B.; Stauch, T.; Head-Gordon, M.; Hexemer, A.; Ushizima, D.; Head-Gordon, T. Multi-resolution 3D-DenseNet for Chemical Shift Prediction in NMR Crystallography. *J. Phys. Chem. Lett.* **2019**, *10*, 4558−4565.
(33) Cobas, C. NMR signal processing, prediction, and structure verification with machine learning techniques. *Magn. Reson. Chem.* **2020**, *58*, 512−519.
(34) Meiler, J. PROSHIFT: protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR* **2003**, *26*, 25−37.
(35) Unzueta, P. A.; Greenwell, C. S.; Beran, G. J. O. Predicting Density Functional Theory-Quality Nuclear Magnetic Resonance Chemical Shifts via Delta-Machine Learning. *J. Chem. Theory Comput.* **2021**, *17*, 826−840.
(36) Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L. Chemical shifts in molecular solids by machine learning. *Nat. Commun.* **2018**, *9*, 4501.
(37) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr., B* **2016**, *72*, 171−179.
(38) Balodis, M.; Cordova, M.; Hofstetter, A.; Day, G. M.; Emsley, L. De Novo Crystal Structure Determination from Machine Learned Chemical Shifts. *J. Am. Chem. Soc.* **2022**, *144*, 7215−7223.
(39) Cordova, M.; Balodis, M.; Simoes de Almeida, B.; Ceriotti, M.; Emsley, L. Bayesian probabilistic assignment of chemical shifts in organic solids. *Sci. Adv.* **2021**, *7*, No. eabk2341.
(40) Gaus, M.; Cui, Q.; Elstner, M. DFTB3: Extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB). *J. Chem. Theory Comput.* **2012**, *7*, 931−948.
(41) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **1998**, *58*, 7260−7268.
(42) Eldar, Y.; Lindenbaum, M.; Porat, M.; Zeevi, Y. Y. The farthest point strategy for progressive image sampling. *IEEE Trans. Image Process.* **1997**, *6*, 1305−1315.
(43) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, No. 184115.
(44) Giannozzi, P.; Andreussi, O.; Brumme, T.; Bunau, O.; Buongiorno Nardelli, M.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Cococcioni, M.; et al. Advanced capabilities for materials modelling with Quantum ESPRESSO. *J. Phys.: Condens. Matter* **2017**, *29*, 465901.
(45) Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys.: Condens. Matter* **2009**, *21*, No. 395502.
(46) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(47) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006**, *27*, 1787−1799.
(48) Barone, V.; Casarin, M.; Forrer, D.; Pavone, M.; Sambi, M.; Vittadini, A. Role and effective treatment of dispersive forces in materials: Polyethylene and graphite crystals as test cases. *J. Comput. Chem.* **2009**, *30*, 934−939.
(49) Dal Corso, A. Pseudopotentials periodic table: From H to Pu. *Comput. Mater. Sci.* **2014**, *95*, 337−350.
(50) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **1999**, *59*, 1758−1775.
(51) Kapil, V.; Rossi, M.; Marsalek, O.; Petraglia, R.; Litman, Y.; Spura, T.; Cheng, B. Q.; Cuzzocrea, A.; Meissner, R. H.; Wilkins, D. M.; et al. i-PI 2.0: A universal force engine for advanced molecular simulations. *Comput. Phys. Commun.* **2019**, *236*, 214−223.
(52) Ceriotti, M.; More, J.; Manolopoulos, D. E. i-PI: A Python interface for ab initio path integral molecular dynamics simulations. *Comput. Phys. Commun.* **2014**, *185*, 1019−1026.
(53) Ceriotti, M.; Bussi, G.; Parrinello, M. Colored-Noise Thermostats a la Carte. *J. Chem. Theory Comput.* **2010**, *6*, 1170−1180.
(54) Ceriotti, M.; Bussi, G.; Parrinello, M. Langevin equation with colored noise for constant-temperature molecular dynamics simulations. *Phys. Rev. Lett.* **2009**, *102*, No. 020601.
(55) Monkhorst, H. J.; Pack, J. D. Special Points for Brillouin-Zone Integrations. *Phys. Rev. B* **1976**, *13*, 5188−5192.
(56) Murphy, K. P. *Machine learning: a probabilistic perspective*; MIT Press, 2012.
(57) Musil, F.; Willatt, M. J.; Langovoy, M. A.; Ceriotti, M. Fast and Accurate Uncertainty Estimation in Chemical Machine Learning. *J. Chem. Theory Comput.* **2019**, *15*, 906−915.


https://doi.org/10.1021/acs.jpcc.2c03854
*J. Phys. Chem. C* 2022, 126, 16710−16720

(58) Musil, F.; Veit, M.; Goscinski, A.; Fraux, G.; Willatt, M. J.; Stricker, M.; Junge, T.; Ceriotti, M. Efficient implementation of atom-density representations. *J Chem Phys* **2021**, *154*, 114109.
(59) Goscinski, A.; Musil, F.; Pozdnyakov, S.; Nigam, J.; Ceriotti, M. Optimal radial basis for density-based atomic representations. *J. Chem. Phys.* **2021**, *155*, 104106.
(60) *RDKit: open-source cheminformatics*, version 2022.03.4; http://www.rdkit.org.
(61) Imbalzano, G.; Zhuang, Y.; Kapil, V.; Rossi, K.; Engel, E. A.; Grasselli, F.; Ceriotti, M. Uncertainty estimation for molecular dynamics and sampling. *J. Chem. Phys.* **2021**, *154*, No. 074102.
(62) Widdifield, C. M.; Robson, H.; Hodgkinson, P. Furosemide's one little hydrogen atom: NMR crystallography structure verification of powdered molecular organics. *Chem. Commun.* **2016**, *52*, 6685−6688.
(63) Maruyoshi, K.; Iuga, D.; Antzutkin, O. N.; Alhalaweh, A.; Velaga, S. P.; Brown, S. P. Identifying the intermolecular hydrogen-bonding supramolecular synthons in an indomethacin-nicotinamide cocrystal by solid-state NMR. *Chem. Commun.* **2012**, *48*, 10844−10846.
(64) Mifsud, N.; Elena, B.; Pickard, C. J.; Lesage, A.; Emsley, L. Assigning powders to crystal structures by high-resolution (1)H-(1)H double quantum and (1)H-(13)C J-INEPT solid-state NMR spectroscopy and first principles computation. A case study of penicillin G. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3418−3422.
(65) Gervais, C.; Profeta, M.; Lafond, V.; Bonhomme, C.; Azais, T.; Mutin, H.; Pickard, C. J.; Mauri, F.; Babonneau, F. Combined ab initio computational and experimental multinuclear solid-state magnetic resonance study of phenylphosphonic acid. *Magn. Reson. Chem.* **2004**, *42*, 445−452.


https://doi.org/10.1021/acs.jpcc.2c03854
*J. Phys. Chem. C* 2022, 126, 16710−16720