scientific reports

OPEN

Machine learning of ^{27}Al NMR electric field gradient tensors for crystalline structures from DFT

He Sun[1], Shyam Dwaraknath[2]✉, Handong Ling[2], Kristin A. Persson[2,3] & Sophia E. Hayes[1]✉

NMR crystallography has emerged as a promising technique for the determination and refinement of atomic coordinates in crystal structures. The crystal structure of compounds containing quadrupolar nuclei, such as ^{27}Al, can be improved by directly comparing solid-state NMR measurements to DFT computations of the electric field gradient (EFG) tensor. The non-negligible computational cost of these first-principles calculations limits the applicability of this method to all but the most well-defined structures. We developed a fast, low-cost machine learning model to predict EFG parameters based on local structural motifs and elemental parameters. We computed 8081 EFG tensors from 1681 ^{27}Al crystalline solids using DFT and benchmarked them against 105 experimentally measured ^{27}Al sites. Surprisingly, simple local geometric features dominate the predictive performance of the resulting random-forest model, yielding an R^2 value of 0.98 and an RMSE of 0.61 MHz for C_Q, the quadrupolar coupling constant. This model accuracy should enable pre-refining future structural assignments before finally validating with first-principles calculations. Such a catalogue of ^{27}Al NMR tensors can serve as a tool for researchers assigning complex NMR spectra influenced by the nuclear electric quadrupole interaction.

Keywords ^{27}Al solid-state NMR, Machine learning

Solid-state nuclear magnetic resonance (SSNMR) is a powerful tool for probing structural differences in local environments for both crystalline and amorphous materials. As a local probe of structure, SSNMR can be a highly effective characterization tool, because long-range order (often required for diffraction methods) is not needed. Consequently, there is broad applicability of NMR across chemical, biological and materials science fields, characterizing diverse systems ranging from battery anodes, to biological solids, to zeolites[1–6].

The most familiar NMR methods have focused on nuclear spin–½ ($I = 1/2$) systems, such as ^{1}H and^{13}C; however, a majority of NMR-active isotopes are quadrupolar, with $I > 1/2$, studies on which can yield exquisite details about the interaction between the nuclear spin's electric quadrupole moment and its electric field gradients (EFG) produced by the surrounding electron clouds[7–14]. Even with the revolutionary demonstration of new NMR pulse-sequence methods for quadrupolar species in the 1990's (resulting in highly resolved spectra)[15]SSNMR experiments using quadrupolar probe nuclei are still often plagued by complicated lineshapes that can overlap and become difficult to elucidate for structural features. Hence, 'NMR crystallography'—combining NMR with other experimental techniques such as diffraction and computational methods like density functional theory (DFT) to achieve a comprehensive, data-consistent picture of a material's structure[16–21] -- have been transformative to interpretation of SSNMR of quadrupolar species.

NMR crystallography relies on state-of-the-art first-principles calculations such as DFT. Despite this advantage, the computational cost of DFT is relatively large for broad adoption. More importantly, the reliability of these calculations in predicting experimental parameters has to be assessed one isotope at a time, with the literature focusing on ^{1}H,^{13}C,^{29}Si,^{31}P and^{17}O in various systems[22–25].

Literature benchmarks have provided large datasets of computed data available via community databases, such as in the Materials Project (materialsproject.org) and the Collaborative Computational Project for NMR (CCP-NC)[26,27] that can be utilized for advanced machine learning (ML) studies to reduce the computational cost. The cubic scaling[28] of DFT calculation time with respect to the number of valence electrons in the system limits these datasets to focusing on comparatively small unit cells of perfect crystalline materials modeled at a temperature of 0 K. Still, appropriately trained ML algorithms have demonstrated the ability to capture local

[1]Department of Chemistry, Washington University, St. Louis, MO 63130, USA. [2]Division of Materials Science, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [3]Department of Materials Science and Engineering, University of California, Berkeley, BERKELEY, CA 94720, USA. ✉email: shyamd@lbl.gov; hayes@wustl.edu

geometry to predict δ_{iso} with accuracy close to DFT while requiring only a fraction of computing time[28–37]. While most of the machine learning efforts have been focused on the prediction of δ_{iso}, the experimentally measured isotropic chemical shift (or σ_{iso}, the DFT computed chemical shielding)[38–41], there have been fewer studies of quadrupolar nuclei that demonstrate the ability of machine learning algorithms in predicting expressions of the electric field gradient (EFG) tensor parameters, such as C_Q[42–44]. NMR of quadrupolar nuclei often results in complex lineshapes that must be deconvoluted in order to extract chemical insights from the connection between the structure and spectroscopic lineshape. The lineshape itself is representative of the electron density distribution surrounding nuclei. The quadrupolar tensor elements provide a complementary measurement of small perturbations to local environments, especially when it is hard to distinguish different sites based on isotropic chemical shift alone[45,46]. Thus, the development of a machine learning method for the prediction of EFG tensor elements, and expressions of those elements such as C_Q, can be highly informative for NMR crystallography studies.

Herein, we present a solid-state NMR ^{27}Al benchmarking set with both DFT calculated EFG and magnetic shielding tensor elements and their experimentally-measured counterparts, reported in the literature. It is worth noting that the most common way for the magnetic shielding tensor elements to be reported is using expressions that employ the "Haeberlen" convention. We follow that convention here, for ready comparison between computed and experimentally-measured quantities, reporting both computed values and their experimental complements for isotropic chemical shielding (σ_{iso}) in the Supplementary Information Table S1 and Table S2. The full definition of NMR parameters can be found in the Supplementary Information Section V. The DFT calculations were performed by two popular DFT packages: Vienna Ab initio Simulation Package (VASP) and Cambridge Serial Total Energy Package (CASTEP)[47,48]. The reliability of DFT predictions of values for σ_{iso}, C_Q, and tensor elements (V_{ab}) for the EFG tensor V in the principal axis system, for ^{27}Al materials was confirmed. We further trained a "random-forest" machine learning model to predict the quadrupolar coupling constant C_Q as a widely-used experimental parameter for compounds containing 4-, 5- and 6-coordinate ^{27}Al sites based on a larger DFT calculated dataset with 1681 aluminum-containing crystalline solid materials. To train the model, we constructed two sets of features, structural features and elemental features (or sometimes termed "alchemical features" in the language of machine learning literature)[49]based on the crystal structure to represent the ^{27}Al local environment. We have found the ^{27}Al C_Q value is closely correlated with the geometric properties of the next-neighbor bonding environment (surprisingly, regardless of the chemical identity of the bonded species). The next-neighbor bonding environment is typically depicted for ease of visualization as a space-filling polyhedron. Distortions to the polyhedron given by variance of bond lengths and bond angles, in combination with other features denoting elemental variance, produce a simple but effective model.

Results and discussion

^{27}Al DFT benchmarking

We begin by benchmarking the ability of DFT to predict chemical shielding tensors against experimentally compiled chemical shift tensors. Unfortunately, in cases where the central transition pattern is strongly influenced by the second-order quadrupolar interaction, it can be challenging to extract information on the chemical shielding tensor[50,51]. In particular, many references do not report the anisotropy of the chemical shift and the asymmetry parameter from the shielding tensor, because these are difficult to know with precision, when quadrupolar interactions are present[52]leaving only the isotropic chemical shift to compare with our computational dataset.

Figure 1 shows the correlation plot between the experimental isotropic chemical shift (δ_{iso}) and DFT calculated isotropic shielding (σ_{iso}) with two different packages (VASP and CASTEP). Both DFT packages demonstrate the ability to accurately predict ^{27}Al isotropic chemical shifts with $R^2 = 0.98$ and RMSE = 4.0 ppm and 4.4 ppm values, respectively. Due to VASP's unique definition of the shielding tensor[53]we have inverted the sign of its output to ensure consistency in the interpretation of the plots. Further details on this issue can be found in our previous work[24]. Further, Fig. 1 (c) demonstrates a strong correlation between the two packages with $R^2 = 0.99$ and RMSE = 3.0 ppm, suggesting that future calculations with either of these codes should yield comparable results. As expected for isotropic chemical shift (shielding), there is grouping or "clustering" of the data points, shown in the correlation plots of Fig. 1 based on the local coordination numbers (4, 5 and 6). To better understand the performance of DFT within each individual coordinate environment, we also plot the correlation between DFT and experiment separately, for reference. Those data are found in the Supplementary Information Section IX. Outliers can be identified by plotting the standardized residual values against each independent variable (Supplementary Information Figure S1) and identifying those that fall outside of a given confidence interval, which for this study was set at 99%. A discussion about the possible origin for such outliers can be found in Supplementary Information Section II.

We compared the computed diagonalized EFG tensor components against experimentally reported values for C_Q and η_Q. The diagonalized EFG tensor for quadrupolar nuclei also can be translated into these convenient algebraic expressions, C_Q and η_Q, that reflect the appearance of the experimentally-measured spectra. For quadrupolar species such as ^{27}Al, both C_Q and η_Q values from the EFG tensor are often reported in the literature, which enables a more sensible and direct comparison between experimental acquired and DFT computed results. (We report in the Supplementary Information, Section X, the principal components of the 2nd -rank symmetrical EFG tensor, V (i.e., Vxx, Vyy, and Vzz), and comparison to each of the DFT-computed components). Figure 2 shows the high degree of correlation between experimentally measured $|C_Q|$ and the corresponding values calculated by DFT packages (VASP and CASTEP). It is not possible to measure the sign of C_Q using 1D NMR spectrum and a special double resonance experiment must be used in this case[54,55]thus nearly all experimental papers choose to report the magnitude of C_Q. The strong correlation between DFT and experiment for both VASP and CASTEP, with $R^2 = 0.96$ for VASP and $R^2 = 0.95$ for CASTEP, demonstrates that

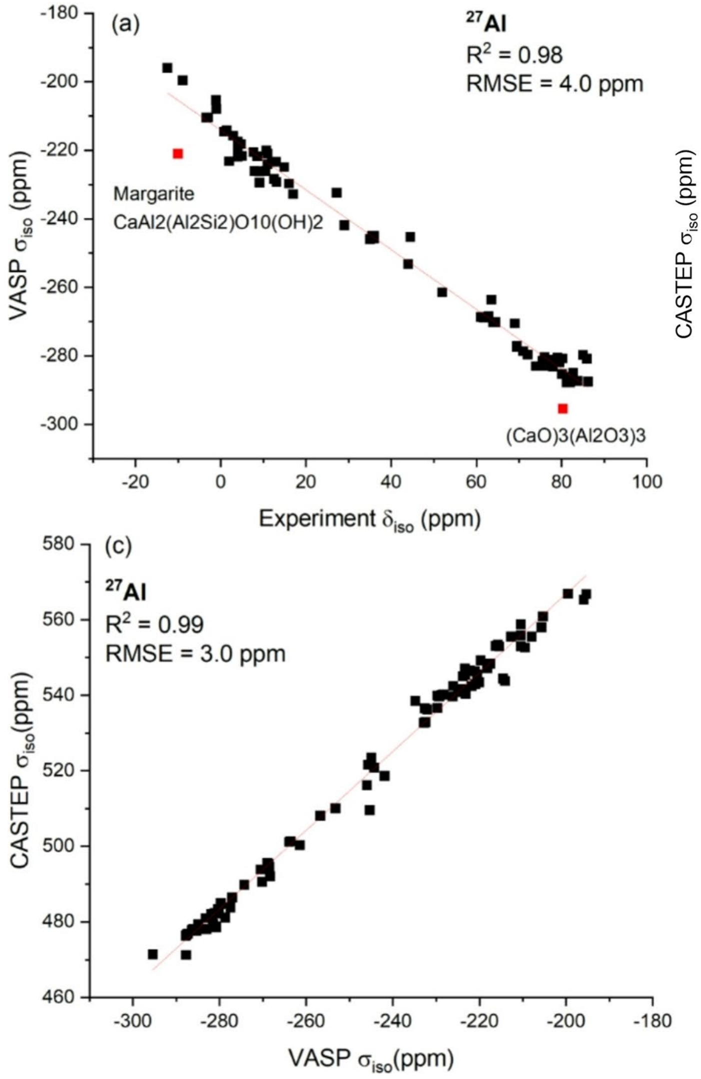


Fig. 1. (**a**) VASP calculated σ_{iso} (ppm) versus experimental δ_{iso} (ppm). (**b**) CASTEP calculated σ_{iso} (ppm) versus experimental δ_{iso} (ppm). (**c**) CASTEP calculated σ_{iso} (ppm) versus VASP calculated σ_{iso} (ppm). In plot (**a**) and (**b**), the outlier species are highlighted in red. Outliers identified by a standardized residual plot with a 99% confidence range are shown in red. (see Supplementary Information, Figure S1).

DFT has the ability to accurately predict the EFG tensor. We note two outliers using the same confidence interval sampling method used previously. Two significantly different C_Q values of β-AlF_3 were reported by previous publications with one stating the C_Q of the single ^{27}Al site in β-AlF_3 is |3.4 MHz| while the other one stating a C_Q of |0.8 MHz|.[56,57] Our calculation result (-1.31 MHz) suggests that 0.8 MHz lies closer to the computed value, and this result is supported by a more recent publication in 2014[58]. The second outlier is the previously noted $(CaO)_4(Al_2O_3)_3$ with experimentally-reported C_Q = |2.4 MHz| and VASP calculated C_Q = 4.41 MHz. It is still unclear if our idealized structural model is an accurate representation of the local structural motifs in the measured sample of $(CaO)_4(Al_2O_3)_3$ resulting in an inappropriate comparison of NMR parameters. Figure 2 (c) shows a strong correlation between the two DFT packages, with $R^2 = 0.99$, for C_Q, suggesting that future calculations with either of these codes should yield comparable results.

The correlation between experimentally-reported η_Q and DFT-computed values for η_Q is shown in Fig. 3. Both CASTEP and VASP show a strong correlation, $R^2 = 0.95$, which suggests that these codes remain self-consistent with respect to the full expression of the EFG tensor. Any correlation between computed and experimental values is tenuous, at best, with many outliers. It may not be surprising that the correlation is weak, since η_Q values, at present, offer limited utility for benchmarking EFG tensors considering the nature of η_Q's mathematical definition which makes it numerically unstable and prone to small perturbations[59,60]. Some experimentalists resort to assuming an η_Q value based on knowledge of the crystal structures, usually at one of the two extremes, as 0 or 1[61]. Consequently, we are demonstrating with these data that the experimentally-reported asymmetry parameter may not be sufficiently robust for benchmarking comparisons. Nevertheless, the high degree of correlation of the individual tensor elements (V_{xx}, V_{yy}, V_{zz}), shown in the Supplementary

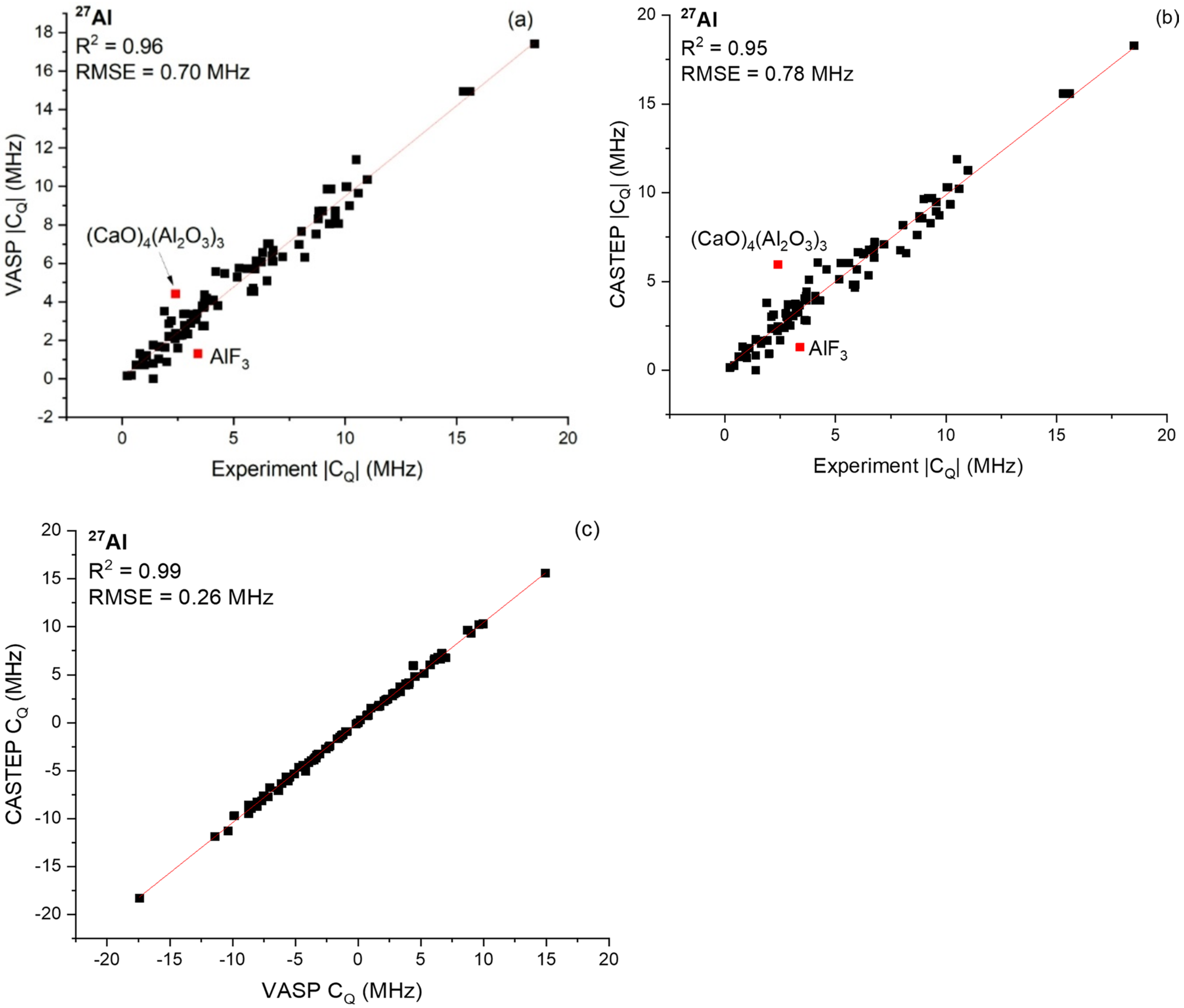


Fig. 2. (**a**) VASP calculated $|C_Q|$ (MHz) versus experimental $|C_Q|$ (MHz). (**b**) CASTEP calculated $|C_Q|$ (MHz) versus experimental $|C_Q|$ (MHz). (**c**) CASTEP calculated C_Q (MHz) versus VASP calculated C_Q (MHz). In plots (**a**) and (**b**), the outlier species are highlighted in red.

Information, Section X, shows promise, even when the expression of these elements, η_Q, reflecting the measured lineshape is less robust.

Fast prediction of ^{27}Al C_Q with machine learning

As shown above, C_Q is specified well (predicted well) by DFT and therefore could be a good target for machine learning. In the next section we will be focusing on constructing a machine learning model to predict the C_Q value measured using ^{27}Al SSNMR experimental data based on crystal structures, in order to enable fast computation of this informative experimental NMR parameter. We chose C_Q here for our model training target because it is a frequently reported parameter to represent aspects of the EFG, and its values appear to be robust (in contrast with, for example, η_Q.) Hence, this parameter represents an effective way to compare experiments with calculated values, and therefore an indicator of accurate machine learning predictions.

By leveraging the strengths of both DFT calculations and machine learning algorithms, we aim to develop a powerful predictive tool that bridges the gap between experimental observations and theoretical predictions in solid-state NMR spectroscopy of quadrupolar nuclei.

DFT calculated ^{27}Al database

To predict C_Q values for ^{27}Al with machine learning, we constructed a VASP-calculated database with 1681 aluminum-containing solid crystalline materials utilizing the high-throughput DFT framework of the Materials Project[27]. The sites in the database can be classified as belonging to three types of local coordination environments: 4-coordinate tetrahedral (termed “T:4”), 5-coordinate trigonal bipyramidal (“T:5”) and 6-coordinate octahedral

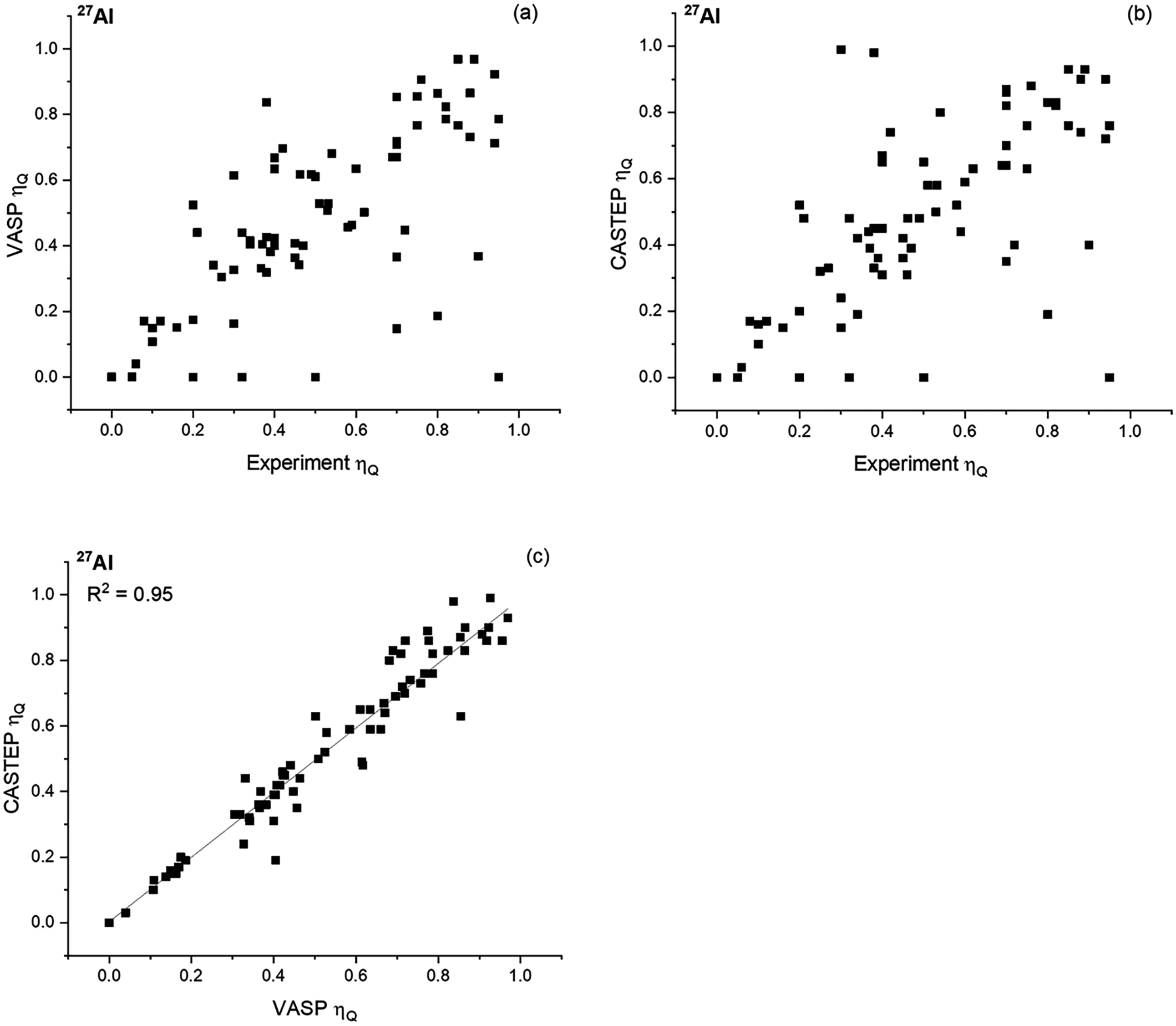


Fig. 3. Correlation of experimentally reported values for η_Q with DFT calculated η_Q. (**a**) VASP calculated versus experimentally-reported values; (**b**) CASTEP-calculated versus experimentally-reported values; and (**c**) correlation of CASTEP versus VASP values for η_Q.

("O:6"). Unusual geometries such as 2-coordinate linear or bent geometries, 3-coordinate trigonal planar, or 4-coordinate square planar were excluded.

Feature engineering

One of the most critical aspects of a successful machine learning model lies in "feature engineering." In terms of materials science, features are usually properties related to the materials or values that can be derived or calculated based on materials' structural or chemical information[62,63]. In terms of these chemical entities, our effort is to select features that provide a means for recognizing patterns in the data, and to correlate an NMR measurement with one or more specific chemical (or structural) properties. When successfully identified, features, either singly or combined, can form a numerical representation of the material, usually expressed in the form of a 1D vector. For machine learning prediction, these numerical representations need to capture the variance of the target parameter across different materials to be successful. The process of feature engineering can be as simple as collecting the atomic numbers (i.e., the chemical identity of an atom participating in a bond), while for many data sets, more complex constructed features are needed[64].

There has been considerable research on feature engineering for materials science to predict NMR parameters such as the use of smooth overlap of atomic positions (SOAP) descriptors[49,65]Coulomb matrix[34] and Behler – Parrinello symmetrical functions (BFPS)[66]. While these features are capable of describing the variance of geometries for structures with different NMR parameters such as isotropic chemical shifts/shieldings, they were designed to be general in order to be useful in many different types of applications (beyond NMR)[67]. For specific targets which aim to extract highly localized perturbations (as in NMR spectroscopy), these features may yield suboptimal results. For example, the size of a SOAP kernel scales quadratically with the number of elemental species considered, which makes it slow to process when applied to datasets with a variety of elements. Instead of using complex descriptors like SOAP, we can employ customized, NMR-specific features to streamline and optimize our feature set, for instance, based on the local environment of the target nucleus (e.g ^{27}Al) under study. This approach not only enhances the model's performance but also improves computational efficiency.

Here we propose two types of customized features to predict the C_Q of the EFG tensor: structural features, and elemental features. Structural features are extracted from the local geometry of the target nucleus alone, without taking into consideration any difference between neighboring atomic species. Significant research in solid-state ^{27}Al NMR of aluminum-containing materials has focused on the empirical correlation between NMR measurable parameters such as C_Q and simple descriptive parameters derived from the local geometry[68–72]. It appears that many of these empirical correlations are particularly useful for the recent efforts of building computational predictive models for NMR spectroscopy. For example, Ghose and Tsang[68] defined the longitudinal strain and the shear strain to quantify the distortion of the local polyhedron from the Platonic solid-like forms (i.e., with identical faces of the geometric solid). Later Baur et al. [69] suggested a distortion index (DI) to measure the angular distortion of the local geometry. These parameters were shown to have a high level of correlation with C_Q value.

It is important to highlight that this is not the only approach to such EFG predictions. Autschbach and coworkers developed a method employing atomic orbitals (AOs) and localized molecular orbitals (MOs) for multiple systems such as^{13}C, ^{33}S, ^{14}N, ^{27}Al, ^{93}Nb and ^{99}Ru. In addition to empirical correlations, Autschbach et al.[13] analyzed the AO contributions to the EFG through a semi-quantitative exploration using an AO contribution model and quantitatively with first-principles computations accompanying analyses of the EFG tensor in terms of localized MOs. Determining ways to capture features via molecular orbitals would be an interesting comparison to the method we are employing here, but ultimately is entirely separate and beyond the scope of what we are presenting here.

Using the DI parameter introduced by Baur as motivation, we implemented this DI parameter in Python along with eight other features derived from local polyhedral geometry: namely the maximum, minimum, standard deviation and mean of the first-order bond lengths (fbl) and bond angles (fba). A full list of structural features and their corresponding abbreviations can be found in Supplementary Information Table S6. Figure 4 shows a correlation "heat map" between the DFT-calculated NMR parameter C_Q and these structural features. (Details on "feature importance" can be found in the Supplementary Information, Section VII). The standard deviation of the first-order bond length "std(fbl)" has a high level of correlation with C_Q, which illustrates the power of such simple features when used for the right target. The distortion index (DI) has the second-largest correlation with C_Q. The std(fbl) and DI characterize the distortion of the local polyhedron from its ideal (Platonic) form (e.g. a perfect octahedron or tetrahedron) in terms of bond length and bond angle, respectively. We found these two features are complementary to each other in the prediction of C_Q. More details about feature complementarity can be found in the Supplementary Information, Section IV. The correlation matrix also reveals strong interrelationships among the structural features themselves, suggesting a potential redundancy in the information they convey—commonly referred to as multicollinearity in machine learning. While multicollinearity may detrimentally affect the performance of linear models, it typically does not exert a significant influence on the performance of tree-based models, such as random forests.

Using just the structural features, we trained a random forest model for ^{27}Al C_Q which derives the target value by performing data segmentation with an ensemble of decision trees[73]. Figure 5 shows the correlation between the calculated DFT ^{27}Al C_Q and the model-predicted C_Q. The plot shows that the set of simple structural features can already predict C_Q with a R^2 of 0.95 and RMSE of 0.77 MHz. We do note that there are still a number of outliers (better depicted in Supplementary Information Section II) suggesting characteristics other than structural features can play a significant role in dictating NMR properties. Also, since the majority of the ^{27}Al sites in the dataset are only coordinated with oxygen, to further test the predictive performance of the model based on the structural features with more atomic variance, we rebalanced the data using the SMOTE (Synthetic Minority Oversampling) technique[74] to oversample the minority group (sites with non-oxygen neighbors) and undersample the majority group (sites with pure oxygen neighbors), then compare the model before and after the rebalance. More details on implementation of SMOTE can be found in the Supplementary Information Section VIII.

Expanding this analysis further, it is expected that any EFG tensor is not only related to the geometry of the local environment but is also strongly influenced by the properties of surrounding atomic species because it is derived from the electron density distribution. To further improve the prediction of C_Q, we therefore need to represent the variation in local chemical composition. We selected twelve elemental properties such as atomic number, electron affinity, and other properties (Supplementary Information Table S7) and utilized three treatments of those elemental features, grouping them into 3 features sets, shown schematically in Fig. 6 (e.g., "Simple statistics…", "Distance normalized deviation…" and "Pairwise atomic properties…").

We first obtain the twelve elemental properties for each atom in the first coordination shell around the ^{27}Al sites. The first set of features is represented by simple statistics of each of the elemental properties: its maximum, its minimum, standard deviation, and average value. The second set of features measures the differences between the neighbor atoms and the core atom (aluminum in our case).

$$\sum_{n} \frac{|\boldsymbol{p_c} - \boldsymbol{p_n}|}{\boldsymbol{N} \cdot \boldsymbol{r_{cn}}} \tag{1}$$

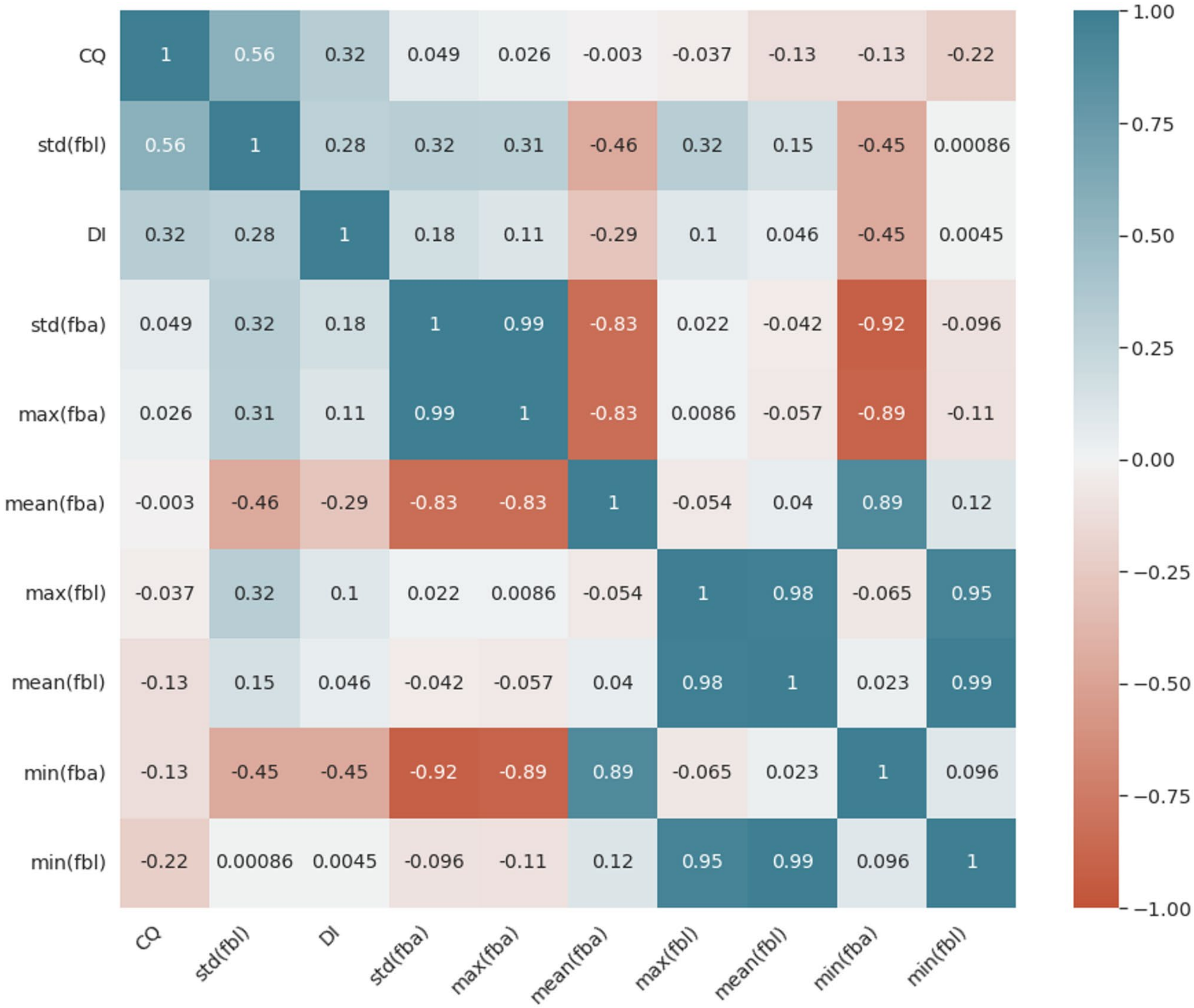


Fig. 4. Correlation heat map across C_Q and structural features. "fbl" and "fba" here are abbreviations of the first-order bond length and the first-order bond angle. The entries are arranged to match the color gradient of the C_Q row/column. The number in each block is Pearson's correlation coefficient (PCC).

Here p_c and p_n are the atomic properties of the central atom (c) and coordinate atoms (n); N is the coordination number; r_{cn} is the corresponding bond length.

For the third set of features, we draw inspiration from the classic Coulomb matrix. For each of the twelve elemental properties, a matrix considering all the atoms within the first neighbor shell was generated.

$$\mathrm{M_{ij}} = \begin{cases} 1, & i = j \\ \frac{\mathrm{p_i p_j}}{\mathrm{r_{ij}}^2}, & i \neq j \end{cases} \quad (2)$$

Like a Coulomb matrix, this feature also considers the pairwise comparison of the selected properties between two atoms in the lattice. One challenge is that when the number of atoms considered is different, the size of the resultant matrix will also be different. In our specific case, the size of the matrix for 4-, 5- and 6-coordinated Al sites will be different. This is troublesome for machine learning predictions because most algorithms require the dimensionality of the feature space to be uniform across all the samples. To solve the problem, we decompose the matrix with singular value decomposition (SVD) and use 5 singular values, the maximum number of possible singular values for our system, as our features instead of the whole matrix.

We retrained the random forest model with both structural features and elemental features (structural + elemental) which improves the model accuracy to $R^2 = 0.98$ and RMSE = 0.61 MHz for C_Q. To further assess the performance, we also compared our models with a benchmark using the SOAP features[49,65]. The SOAP model was also trained with a random forest algorithm based on the same set of data as the other two shown. The only difference is the features used for training. Instead of using our structural and chemical features, we use SOAP features generated by an open-source package Dscribe[63]which results in a big feature

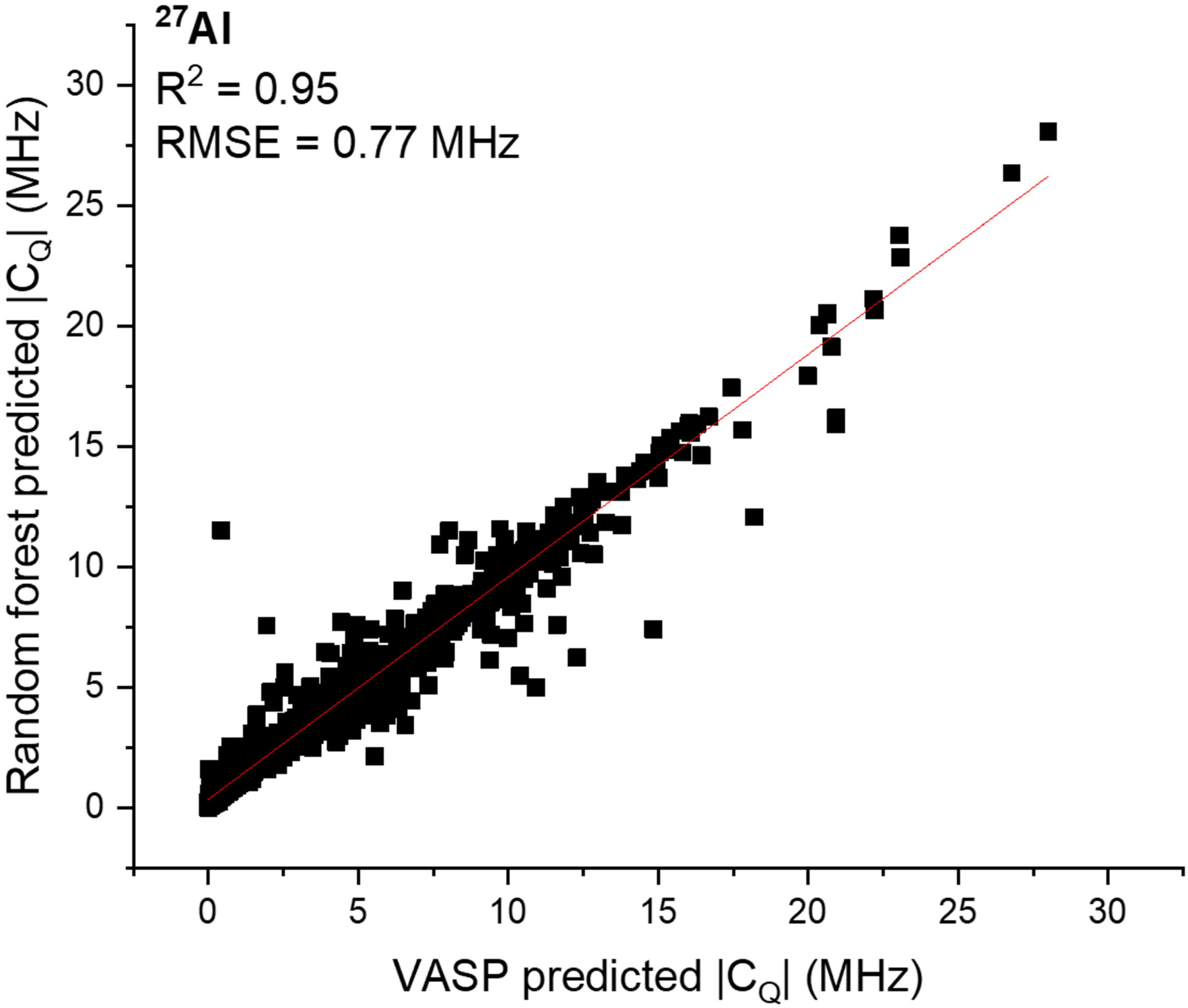


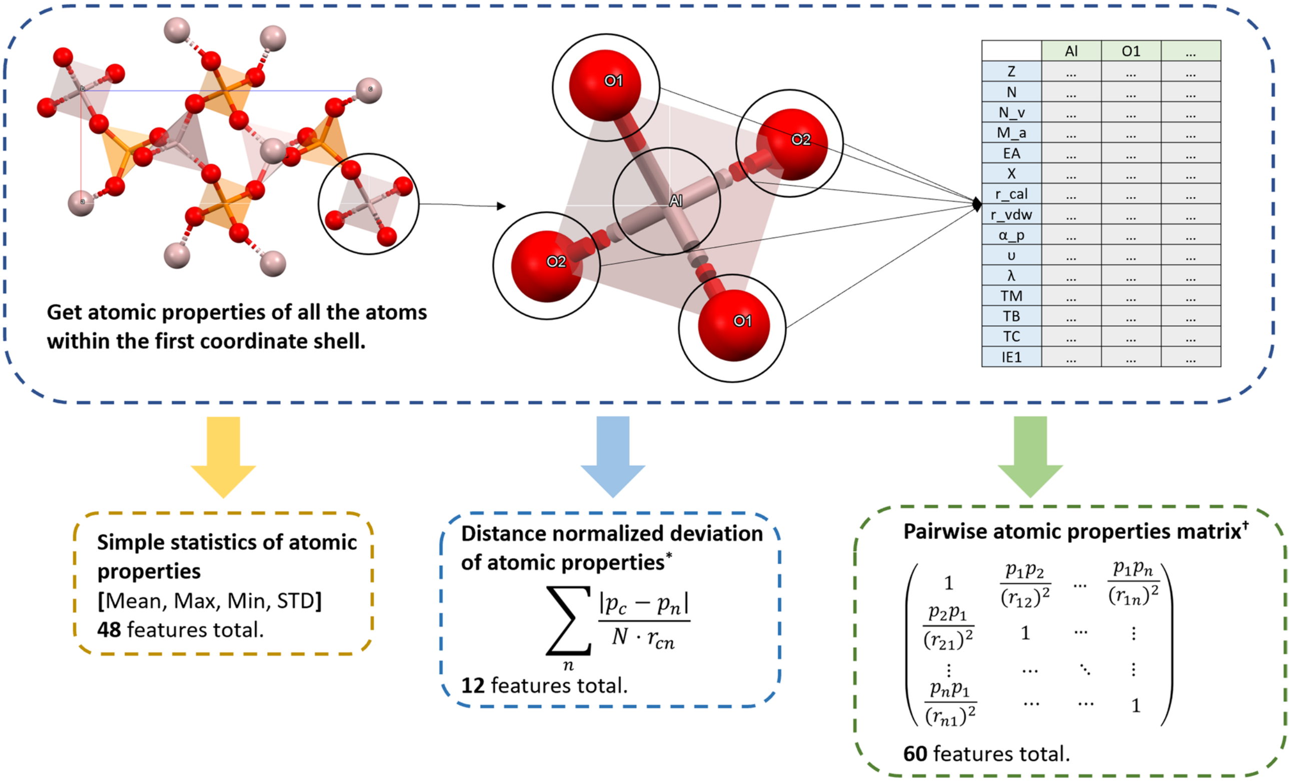
Fig. 5. Comparison between random forest-predicted ^{27}Al C_Q with DFT-calculated ^{27}Al C_Q for aluminum-containing compounds. The random forest model was trained with structural features only (i.e., not with elemental properties). The size of the test set is 1171 individual ^{27}Al sites.

set of 4,163,280. Figure 7 shows a performance comparison between models based on our proposed features (structural, structural + elemental) and that based on SOAP. As shown in Fig. 7a, both of our proposed features perform significantly better than SOAP, irrespective of the size of the sample. Structural + elemental features perform better than structural features alone when the sample size gets larger, which gives confidence for using this combination of features for very large datasets. Figure 7b and d show the correlation plots between the VASP-calculated and the machine learning-predicted $|C_Q|$ values based on the three models. The structural + elemental model significantly reduces the number of extreme outliers, evident in both Fig. 7c and d. The SOAP model achieves a usable performance benchmark of $R^2 = 0.92$ and RMSE = 0.97 MHz.

It is noteworthy, despite the significantly increased computational cost of the SOAP features, this method lacks the same degree of accuracy in comparison to our straightforward feature set. Most importantly, the SOAP features produce some strong outliers. Consequently, we show that a simple set of features that are customized for a specific problem, such as NMR parameter predictions, can outperform universal features because this method excludes unnecessary information that could significantly decrease the performance of the model in terms of both efficiency and accuracy.

Conclusions

By studying the correlation between experimentally measured ^{27}Al NMR parameters and DFT calculated values with a relatively large benchmarking set, we can confirm that DFT calculations are accurate in predicting isotropic chemical shielding σ_{iso} and quadrupolar coupling constant $|C_Q|$ for crystalline materials that contain aluminum species. Similar to our previous benchmarking effort on spin ½ nuclei for ^{29}Si, DFT predictions of asymmetry


Get atomic properties of all the atoms within the first coordinate shell.
O1
O2
Al
O2
O1
Simple statistics of atomic properties
[Mean, Max, Min, STD]
48 features total.
Distance normalized deviation of atomic properties*
$\sum_{n} \frac{|p_c - p_n|}{N \cdot r_{cn}}$
12 features total.
Pairwise atomic properties matrix†
$\begin{pmatrix} 1 & \frac{p_1 p_2}{(r_{12})^2} & \cdots & \frac{p_1 p_n}{(r_{1n})^2} \\ \frac{p_2 p_1}{(r_{21})^2} & 1 & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ \frac{p_n p_1}{(r_{n1})^2} & \cdots & \cdots & 1 \end{pmatrix}$
60 features total.

	Al	O1	...
Z	...	...	...
N	...	...	...
N_v	...	...	...
M_a	...	...	...
EA	...	...	...
X	...	...	...
r_cal	...	...	...
r_vdw	...	...	...
α_p	...	...	...
υ	...	...	...
λ	...	...	...
TM	...	...	...
TB	...	...	...
TC	...	...	...
IE1	...	...	...

Fig. 6. Illustration of the feature engineering process for element-specific features. A list of atomic properties for each atom within the first coordination shell was collected and then transferred into 3 sets of features: simple statistics of atomic properties, distance-normalized deviation of those atomic properties, and pairwise atomic properties matrices. * p_c and p_n are the atomic properties of the central atom, *c*, and coordinate atom, *n*; *N* is the coordination number; r_{cn} is the corresponding bond length. †p_n are the atomic properties of the atoms within the first coordination shell; r_{mn} are the inter-atomic distances between atom *m* and atom *n*.

parameters (both η_{CS} and η_Q) are shown to be more prone to error due to the sensitivity of this parameter to slight variations in local geometry and the difficulty of determining η experimentally with precision.

Having shown DFT's accuracy at predicting ^{27}Al NMR parameters, we built a simple machine learning model to predict ^{27}Al C_Q values based on a large VASP-calculated NMR dataset of 1681 aluminum-containing solid materials. The structural and elemental features that we selected were proven to be effective in predicting C_Q, likely by capturing the variation of local environments to which experimentally-measured NMR parameters are very sensitive. It is surprising for us to find that among all the features, the pure geometrical variations such as that of bond lengths are the dominant features for C_Q prediction. This demonstration shows the possibility of building simple but effective features for the prediction of materials' properties, instead of using larger universal features.

Also, we can get a better understanding of the relationship between local geometry and SSNMR spectra that, specifically, SSNMR spectra for quadrupolar nuclei are determined primarily by local geometry distortions. These data are publicly available for further investigation, via Materials Project. Our final model was proven to be effective in predicting $|C_Q|$ for 4-, 5- and 6-coordinate aluminum sites with $R^2 = 0.98$ and RMSE = 0.61 MHz. This accuracy is comparable with the accuracy of DFT calculations versus experiment (RMSE = 0.70 MHz for VASP), thus making this machine learning method a fast and agile complement to DFT calculations.

Methods

Data sets

Benchmarking Data: We have collected experimental ^{27}Al NMR parameters from the literature on 56 different crystalline materials, accounting for 105 unique sites, including a few repeated structures with independent measurements. The distribution of coordination number of the ^{27}Al sites are: 41 for 4-coordinate, 9 for 5-coordinate, and 55 for 6-coordinate. All the parameters reported were collected via SSNMR employing either magic-angle spinning (MAS) or multiple-quantum MAS (MQMAS) in the experiments. All of the structures were calculated with both VASP and CASTEP.

Machine learning data: For machine learning model training, a larger dataset of DFT-computed ^{27}Al NMR parameters was constructed by VASP calculation. The dataset is composed of 1681 aluminum-containing structures which correspond to 8081 ^{27}Al sites (5852 after removing duplicates). The coordinating environment of the ^{27}Al sites was confined to 4-coordinate (4696 sites), 5-coordinate (202 sites), or 6-coordinate aluminum (3183 sites). There are 104 different compositions in terms of the first coordination sphere (e.g., neighboring

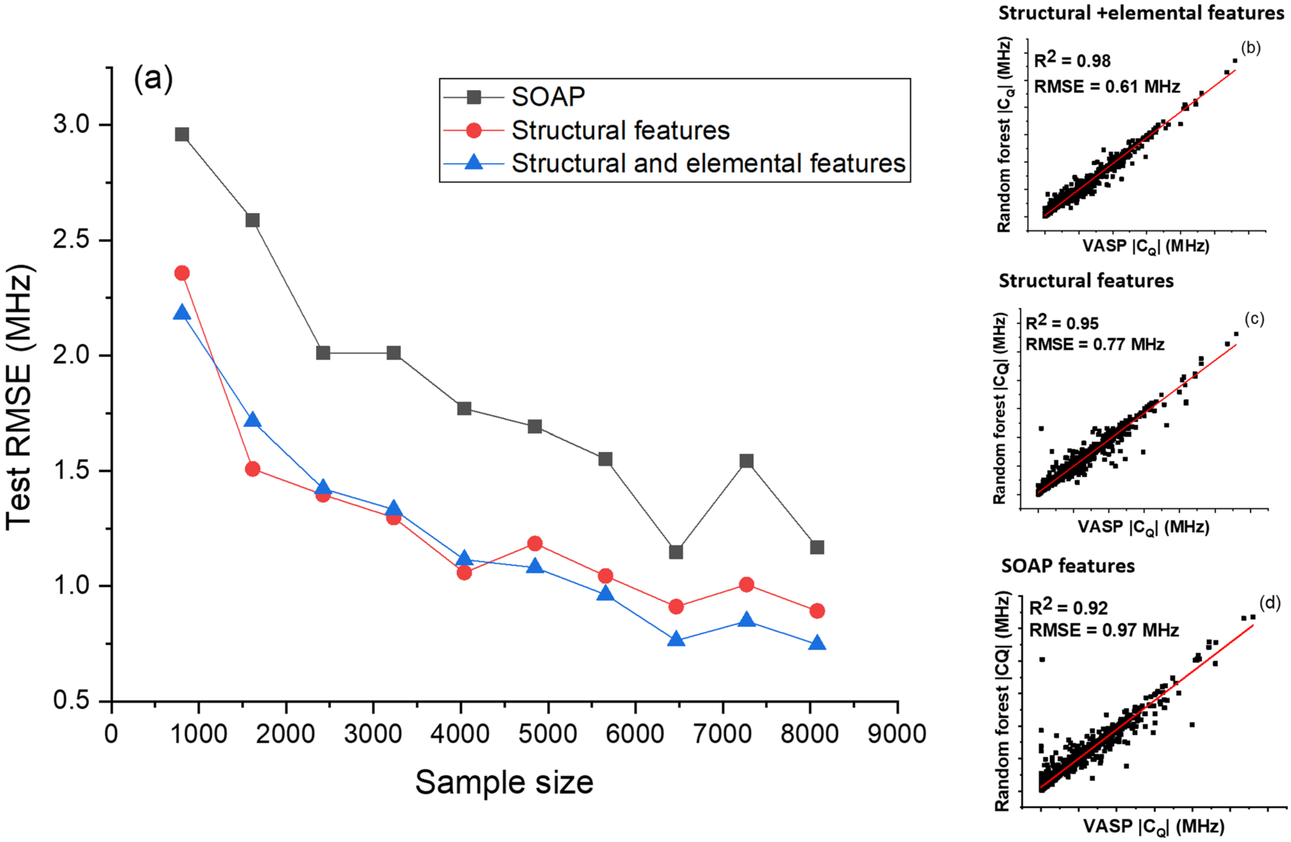


Fig. 7. Comparison of the random forest models trained based on three different feature sets (structural + elemental, structural and SOAP features). (**a**) The learning curve plot of model performance (Test RMSE) over sample size for all three models. (**b**–**d**) Correlations between random forest-predicted and VASP-calculated ^{27}Al $|C_Q|$ values for aluminum-containing compounds. The random forest model was trained with: (**b**) structural and elemental features, (**c**) structural features only, and (**d**) SOAP features.

heteronuclear species) of ^{27}Al sites in the dataset. The histogram of the 10 most commonly-found compositions is reported in the Supplementary Information Figure S6(b). All the crystal structures were obtained from the Materials Project and were geometry optimized before NMR calculations.

DFT details

DFT calculations with CASTEP were performed within the Perdew-Burke-Enzerhof (PBE) Generalized Gradient Approximation (GGA) formulation of the exchange-correlation functional. These were performed in two steps: an initial geometry optimization where the lattice was allowed to adjust, followed by an NMR calculation on the relaxed structure. On-the-fly ultra-soft pseudopotentials were used as an approximation of nuclear and core electron interactions. Convergence tests were performed on γ-$LiAlO_2$ to find optimal energy cutoffs and k-points. See Supplementary Information for more details. It was determined that 750 eV as an energy cutoff with Monkhorst-Pack grid of 5 × 4 × 4 was enough to converge the NMR calculations to a single value.

DFT calculations were also performed using the projector augmented wave (PAW) method[48,75] as implemented in the Vienna Ab Initio Simulation Package (VASP)[76–78] within the PBE-GGA) formulation of the exchange-correlation functional[79]. A cut-off for the plane waves of 520 eV was used and a uniform k-point density of approximately 1,000/atom was employed. We note that the computational and convergence parameters were chosen in compliance with the settings used in the Materials Project[27] to enable direct comparisons with the large set of available Materials Project data.

Machine learning details

The RandomForestRegressor object from the Scikit-Learn library (https://scikit-learn.org/stable/) was employed to construct the random forest model. A subset comprising 20% (1171 sites) of the overall dataset was randomly allocated to serve as the test set, while the remaining 80% (4681) was utilized for training purposes. The model underwent training utilizing this training dataset and subsequently, its performance was assessed using the designated 20% test dataset. Hyperparameter optimization was conducted via the RandomizedSearchCV method provided by Scikit-Learn, incorporating a 5-fold cross-validation strategy across 100 iterations. For additional information regarding the hyperparameter search range, please refer to the code repository at github.com/wushanyun64/27Al_CQ_prediction.

Data availability

All the geometry optimized structures used for NMR calculation in this paper are included in the Supplementary Information as Supplementary Data. In addition, all data for the benchmarking of the computed NMR tensors including structures, spectra, computed, and experimental tensors are available via the MPContribs platform on the Materials Project at https://mpcontribs.org/. The larger database of computed NMR tensors are available via the Materials Project at https://materialsproject.org/.

Received: 18 March 2025; Accepted: 1 July 2025
Published online: 21 July 2025

References

1. Xu, J., Wang, Q. & Deng, F. Metal active sites and their catalytic functions in zeolites: insights from Solid-State NMR spectroscopy. *Acc. Chem. Res.* **52**, 2179–2189 (2019).
2. Brouwer, D. H. et al. Solid-state 29Si NMR spectra of pure silica zeolites for the international zeolite association database of zeolite structures. *Microporous Mesoporous Mater.* **297**, 110000 (2020).
3. Castellani, F. et al. Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. *Nature* **420**, 99–102 (2002).
4. Otting, G. Protein NMR using paramagnetic ions. *Annual Rev. Biophys.* **39**, 387–405 (2010).
5. Wickramasinghe, N. P. et al. Nanomole-scale protein solid-state NMR by breaking intrinsic 1 H T1 boundaries. *Nat. Methods.* **6**, 215–218 (2009).
6. Bhattacharyya, R. et al. In situ NMR observation of the formation of metallic lithium microstructures in lithium batteries. *Nat. Mater.* **9**, 504–510 (2010).
7. Ashbrook, S. E. Recent advances in solid-state NMR spectroscopy of quadrupolar nuclei. *Phys. Chem. Chem. Phys.* **11**, 6892–6905 (2009).
8. Ashbrook, S. E. & Sneddon, S. New methods and applications in solid-state NMR spectroscopy of quadrupolar nuclei. *J. Am. Chem. Soc.* **136**, 15440–15456 (2014).
9. Holmes, S. T. & Schurko, R. W. Refining crystal structures with quadrupolar NMR and Dispersion-Corrected density functional theory. *J. Phys. Chem. C.* **122**, 1809–1820 (2018).
10. Widdifield, C. M. & Bryce, D. L. Crystallographic structure refinement with quadrupolar nuclei: A combined solid-state NMR and GIPAW DFT example using MgBr2. *Phys. Chem. Chem. Phys.* **11**, 7120–7122 (2009).
11. Perras, F. A., Korobkov, I. & Bryce, D. L. NMR crystallography of sodium diphosphates: combining dipolar, shielding, quadrupolar, diffraction, and computational information. *CrystEngComm* **15**, 8727–8738 (2013).
12. Bryce, D. L. NMR crystallography: structure and properties of materials from solid-state nuclear magnetic resonance observables. *IUCrJ* **4**, 350–359 (2017).
13. Autschbach, J., Zheng, S. & Schurko, R. W. Analysis of electric field gradient tensors at quadrupolar nuclei in common structural motifs. *Concepts Magn. Reson. Part. Bridg Educ. Res.* **36**, 84–126 (2010).
14. Akitt, J. W. & McDonald, W. S. Arrangements of ligands giving low electric field gradients. *Journal of Magnetic Resonance ()* 58, 401–412 (1984).) 58, 401–412 (1984). (1969).
15. Medek, A., Harwood, J. S. & Frydman, L. Multiple-Quantum Magic-Angle spinning NMR: A new method for the study of quadrupolar nuclei in solids. *J. Am. Chem. Soc.* **117**, 12779–12787 (1995).
16. Hodgkinson, P. NMR crystallography of molecular organics. *Progress Nucl. Magn. Reson. Spectrosc.* **118–119**, 10–53 (2020).
17. Ashbrook, S. E. & McKay, D. Combining solid-state NMR spectroscopy with first-principles calculations-a guide to NMR crystallography. *Chem. Commun.* **52**, 7186–7204 (2016).
18. Martineau, C. NMR crystallography: applications to inorganic materials. *Solid State Nucl. Magn. Reson.* **63–64**, 1–12 (2014).
19. Falls, Z., Zurek, E. & Autschbach, J. Computational prediction and analysis of the 27Al solid-state NMR spectrum of Methylaluminoxane (MAO) at variable temperatures and field strengths. *Phys. Chem. Chem. Phys.* **18**, 24106–24118 (2016).
20. Harris, R. K., Wasylishen, R. E. & Duer, M. J. *NMR Crystallography* (Wiley, 2012).
21. Taulelle, F. Fundamental principles of NMR crystallography. *eMagRes* **2009**, 245–262 (2009).
22. Hartman, J. D., Kudla, R. A., Day, G. M., Mueller, L. J. & Beran, G. J. O. Benchmark fragment-based 1H, 13 C, 15 N and 17O chemical shift predictions in molecular crystals. *Phys. Chem. Chem. Phys.* **18**, 21686–21709 (2016).
23. Shenderovich, I. G. Experimentally established benchmark calculations of 31P NMR quantities. *Chemistry–Methods* **1**, 61–70 (2021).
24. Sun, H. et al. Enabling materials informatics for 29Si solid-state NMR of crystalline materials. *NPJ Comput. Mater* **6**, (2020).
25. Benassi, E. Benchmarking of density functionals for a soft but accurate prediction and assignment of 1H and 13 C NMR chemical shifts in organic and biological molecules. *J. Comput. Chem.* **38**, 87–92 (2017).
26. Hodgkinson, P., Ashbrook, S. E., Morris, A. & Yates, J. R. Collaborative Computational Project for NMR Crystallography. (2013). www.ccpnc.ac.uk
27. Jain, A. et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, (2013).
28. Paruzzo, F. M. et al. Chemical shifts in molecular solids by machine learning. *Nat. Commun.* **9**, 4501 (2018).
29. Cordova, M. et al. Structure determination of an amorphous drug through large-scale NMR predictions. *Nat. Commun.* **12**, 2964 (2021).
30. Gerrard, W. et al. IMPRESSION-prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chem. Sci.* **11**, 508–515 (2020).
31. Venetos, M. C., Dwaraknath, S. & Persson, K. A. Effective local geometry descriptor for29Si NMR Q4Anisotropy. *J. Phys. Chem. C.* **125**, 19481–19488 (2021).
32. Chaker, Z., Salanne, M., Delaye, J. M. & Charpentier, T. NMR shifts in aluminosilicate glasses via machine learning. *Phys. Chem. Chem. Phys.* **21**, 21709–21725 (2019).
33. Liu, S. et al. Multiresolution 3D-DenseNet for chemical shift prediction in NMR crystallography. *J. Phys. Chem. Lett.* **10**, 4558–4565 (2019).
34. Rupp, M., Tkatchenko, A. & Müller, K. R. Von lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
35. Li, D. W., Hansen, A. L., Bruschweiler-Li, L., Yuan, C. & Brüschweiler, R. Fundamental and practical aspects of machine learning for the peak picking of biomolecular NMR spectra. *J. Biomol. NMR.* **76**, 49–57 (2022).
36. Cobas, C. NMR signal processing, prediction, and structure verification with machine learning techniques. *Magn. Reson. Chem.* **58**, 512–519 (2020).
37. Gao, P., Zhang, J., Peng, Q., Zhang, J. & Glezakou, V. A. General protocol for the accurate prediction of molecular 13 C/1H NMR chemical shifts via machine learning augmented DFT. *J. Chem. Inf. Model.* **60**, 3746–3754 (2020).

38. Gaumard, R. et al. Regression machine learning models used to predict DFT-Computed NMR parameters of zeolites. *Computation* **10**, 74 (2022).
39. Cordova, M. et al. A machine learning model of chemical shifts for chemically and structurally diverse molecular solids. *J. Phys. Chem. C.* **126**, 16710–16720 (2022).
40. Lin, M. et al. A machine learning protocol for revealing ion transport mechanisms from dynamic NMR shifts in paramagnetic battery materials. *Chem. Sci.* **13**, 7863–7872. https://doi.org/10.1039/d2sc01306a (2022).
41. Venetos, M. C., Wen, M. & Persson, K. A. Machine learning full NMR chemical shift tensors of silicon oxides with equivariant graph neural networks. *J. Phys. Chem. A.* **127**, 2388–2398 (2023).
42. Charpentier, T. First-principles NMR of oxide glasses boosted by machine learning. *Faraday Discuss.* **255**, 370–390 (2025).
43. Gu, X., Myung, Y., Rodrigues, C. H. M. & Ascher, D. B. EFG-CS: predicting chemical shifts from amino acid sequences with protein structure prediction using machine learning and deep learning models. *Protein Sci.* **33**, e5096 (2024).
44. Shakiba, M., Philips, A. B., Autschbach, J. & Akimov, A. V. Machine learning mapping approach for computing spin relaxation dynamics. *J. Phys. Chem. Lett.* **16**, 153–162 (2024).
45. Sun, H. et al. Structural investigation of silver vanadium phosphorus oxide (Ag2VO2PO4) and its reduction products. *Chem. Mater.* **33**, 4425–4434 (2021).
46. Kobera, L. et al. The nature of chemical bonding in Lewis adducts as reflected by 27Al NMR quadrupolar coupling constant: combined Solid-State NMR and quantum chemical approach. *Inorg. Chem.* **57**, 7428–7437 (2018).
47. Segall, M. D. et al. First-principles simulation: ideas, illustrations and the CASTEP code. *J. Phys.: Condens. Matter*. **14**, 2717–2744 (2002).
48. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B*. **59**, 1758–1775 (1999).
49. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).
50. Bryce, D. L. *Tensor Interplay* (John Wiley and Sons Ltd.: Chichester,, 2009).
51. Mueller, L. J. Tensors and rotations in NMR. *Concepts Magn. Reson. Part. Bridg Educ. Res.* **38 A**, 221–235 (2011).
52. Schurko, R. W., Wasylishen, R. E. & Phillips, A. D. A definitive example of aluminum-27 chemical shielding anisotropy. *J. Magn. Reson.* **133**, 388–394 (1998).
53. LCHIMAG - Vaspwiki. https://cms.mpi.univie.ac.at/wiki/index.php/LCHIMAG
54. Perras, F. A. & Bryce, D. L. Residual dipolar coupling between quadrupolar nuclei under magic-angle spinning and double-rotation conditions. *J. Magn. Reson.* **213**, 82–89 (2011).
55. Harris, R. K. & Olivieri, A. C. Quadrupolar effects transferred to spin-12 magic-angle spinning spectra of solids. *Progress Nucl. Magn. Reson. Spectrosc.* **24**, 435–456 (1992).
56. Dirken, P. J., Jansen, J. B. H. & Schuiling, R. D. Influence of octahedral polymerization on 23Na and 27Al MAS NMR in alkali fluoroaluminates. *Am. Mineral.* **77**, 718–724 (1992).
57. Chupas, P. J., Ciraolo, M. F., Hanson, J. C. & Grey, C. P. In situ X-ray diffraction and solid-state NMR study of the fluorination of γ-Al2O3 with HCF2Cl. *J. Am. Chem. Soc.* **123**, 1694–1702 (2001).
58. Sadoc, A. et al. NMR parameters in column 13 metal fluoride compounds (AlF3, GaF3, InF3 and TlF) from first principle calculations. *Solid State Nucl. Magn. Reson* 59–60, (2014).
59. Pooransingh, N. et al. 51V solid-state magic angle spinning NMR spectroscopy and DFT studies of oxovanadium(V) complexes mimicking the active site of vanadium haloperoxidases. *Inorg. Chem.* **42**, 1256–1266 (2003).
60. Schweitzer, A. et al. 51V solid-state NMR investigations and DFT studies of model compounds for vanadium haloperoxidases. *Solid State Nucl. Magn. Reson.* **34**, 52–67 (2008).
61. Hovis, G. L., Spearing, D. R., Stebbins, J. F., Roux, J. & Clare, A. X-ray powder diffraction and 23Na, 27Al, and 29Si MAS-NMR investigation of nepheline-kalsilite crystalline solutions. *Am. Mineral.* **77**, 19–29 (1992).
62. Ward, L. et al. Matminer: an open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
63. Himanen, L. et al. DScribe: library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
64. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *NPJ Comput. Mater.* **2**, 1–7 (2016).
65. Jäger, M. O. J., Morooka, E. V., Canova, F., Himanen, F., Foster, A. S. & L. & Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *NPJ Comput. Mater.* **4**, 37 (2018).
66. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *Journal Chem. Physics* **134**, (2011).
67. Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput. Materials* **5**, (2019).
68. Ghose & Tsang, T. Structural dependence of quadrupole coupling constant e2qQ/h for 27Al and crystal field parameter D for Fe3 + in aluminosilicates. *Am. Mineralogist: J. Earth Planet. Mater.* **58**, 748–755 (1973).
69. Baur, W. H. The geometry of polyhedral distortions. Predictive relationships for the phosphate group. *Acta Crystallogr. B*. **30**, 1195–1215 (1974).
70. Cumby, J. & Attfield, J. P. Ellipsoidal analysis of coordination polyhedra. *Nat. Commun.* **8**, 14235 (2017).
71. Padro, D. et al. Variations of titanium interactions in solid state NMR-correlations to local structure. *J. Phys. Chem. B*. **106**, 13176–13185 (2002).
72. MacKenzie, K. J. D. & Smith, M. E. *Multinuclear Solid-State Nuclear Magnetic Resonance of Inorganic Materials. Elsevier* vol. 6 (2002).
73. Breiman, L. Random forests. *Mach Learn* **45**, (2001).
74. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
75. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B*. **50**, 17953–17979 (1994).
76. Kresse, G. & Hafner, J. *Ab Initio Molecular Dynamics for Liquid Metals*. vol. 47.
77. Kresse, G. & Furthmüller, J. Efficient iterative schemes for *Ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B*. **54**, 11169–11186 (1996).
78. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
79. Perdew, J. P., Burke, K. & Ernzerhof, M. *Generalized Gradient Approximation Made Simple* 3865 (1996).

Acknowledgements

This work is supported by the U.S. National Science Foundation (NSF), Award 1640899 and Award 2004915. This work made use of computational resources and software infrastructure provided through the Materials Project, which is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under Contract No. DE-AC02-05-CH11231 (Materials Project program KC23MP). Additional computational resources were provided by the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the NSF (ACI-1053575), as well as the Savio computational cluster resource provided by the Berkeley Research Computing program at the University of California, Berkeley. We extend our gratitude to Michael E. West for his contributions to CASTEP calculations during the early stages of this project.

Author contributions

Hayes and Dwaraknath led the project examining NMR tensors and conducting DFT calculations from Materials Project datasets, and both co-authored the manuscript. Sun (first author) conducted CASTEP calculations, data curation of solid-state NMR data, construction and testing of the machine learning models, and he wrote large portions of the manuscript; Lin contributed part of the DFT (VASP/CASTEP) computations. Persson is the lead of the Materials Project and contributed time and effort to direct computational efforts.

Declarations

Competing interests

The authors declare no competing interests.

Inclusion & ethics statement

This research was conducted in alignment with the principles outlined in the Global Code of Conduct for Equitable Research Partnerships. The collaboration involved contributors from institutions in diverse geographic and socioeconomic contexts, ensuring an equitable distribution of intellectual and practical contributions. Each author has actively participated in the research process, from study design to manuscript preparation, in a manner that respects their expertise and institutional responsibilities.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-10017-x.

Correspondence and requests for materials should be addressed to S.D. or S.E.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.