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

pubs.acs.org/JPCC

Article

Effective Local Geometry Descriptor for ^{29}Si NMR Q^4 Anisotropy

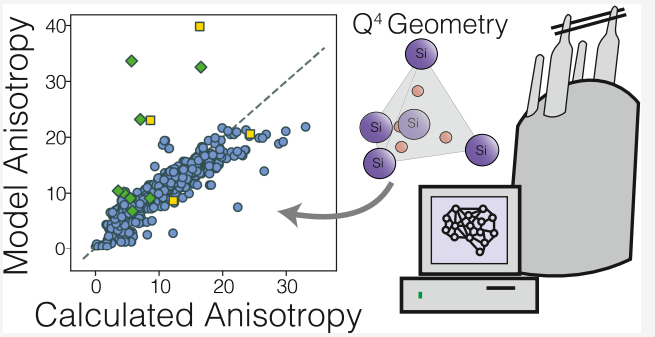
Maxwell C. Venetos, Shyam Dwaraknath, and Kristin A. Persson*

Cite This: https://doi.org/10.1021/acs.jpcc.1c04829

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The nuclear shielding anisotropy, ζ, is a useful nuclear magnetic resonance (NMR) shielding tensor parameter in describing the extent of electron cloud distortion about an atom. Despite the advantages afforded by NMR in structural characterization, the relationship between ζ and local structure of an atom in high-symmetry environments, such as Si−Q^4 sites, is poorly understood. Here, we use a data-driven approach combining random forest feature ranking and the Sure Independence Screening and Sparsifying Operator (SISSO) approach to derive a simple and accurate geometric descriptor for ζ with a root-mean-squared prediction error of 6.77 ppm and an R^2 of 0.761. We then apply this descriptor to describe the local geometric distortion of zeolites Sigma-2 and silica-ZSM-5 whose chemical shift anisotropy tensor has been reported. We envision that this geometric descriptor will allow for structural description and refinement in previously difficult-to-describe materials.



INTRODUCTION

Solid-state nuclear magnetic resonance (NMR) spectroscopy offers localized structural information that complements the long-range structural information afforded by diffraction techniques. Combining these complimentary techniques with quantum chemical calculations is a powerful strategy advancing the characterization and study of complex materials, particularly materials with unknown crystal structures and local atomic environments. In the field of local structure characterization, this is often referred to as NMR crystallography. However, while the local coordination environment and number of unique sites can be inferred from simple spectral features such as the overall chemical shift[1,2] and number of peaks,[3] more detailed structural information, particularly about the second-coordination shell, is still difficult to unravel without the aid of bespoke quantum chemical calculations.[3−6] For example, ^{29}Si chemical shielding tensors may reveal detailed structural information for materials containing a Si−O_4 backbone; there are still large gaps in the understanding of how these parameters relate to local geometry, in particular, with respect to the full shielding tensors of Q^4 species, such as Si. Q^n denotes the number of bridging oxygens where n is the number of Si−O−Si linkages, from n = 4 being 4-fold-connected and n = 0 being fully disconnected.[7] Full NMR chemical shift tensors, which are very difficult to measure, have been used in Q^3 species to correlate to local structure and chemical environment.[7−11]

To the best of our knowledge, a thorough study of the relationship between geometry and anisotropy has not been performed for Q^4 sites; however, lower Q^n species are known to show a strong bond distance correlation due to a principal component along a shortened bond axis. We extend this to Q^4 species and correlate the range of bond distances as a metric for capturing expected Si tetrahedral geometries, as well as more exotic geometries for full analysis of ^{29}Si anisotropy. As a result, we present a new, intuitive, and easily interpretable structural descriptor (Υ) for Q^4 chemical shift anisotropy, $|\zeta|$. Υ constrains the distortions of silicon tetrahedra and is defined as

$$\Upsilon = |\ln(\langle\Omega_{Si_iSi_0Si_j}\rangle) - \exp(-C_{Si})| \quad (1)$$

where $\langle\Omega_{Si_iSi_0Si_j}\rangle$ is the average of the six Si−Si−Si tetrahedral angles surrounding a central silicon atom, and C_{Si} is the distance between the central Si_0 atom and the centroid of the tetrahedron defined by the second-coordination sphere silicon atoms (where the second-coordination sphere is represented by atoms Si_i, Si_j, Si_k, and Si_l in Figure 1). These geometric features are defined in the Discussion section, and a diagram is shown in Figure 1. We derive this descriptor from a data set of 885 computed structures using a combined machine-learning- and data-driven approach. Υ takes into account both the site geometry and symmetry to give a linear correlation to anisotropy with an R^2 of 0.761 and a root-mean-squared error (RMSE) of 6.77 ppm when used to predict the

Received: June 1, 2021
Revised: August 12, 2021

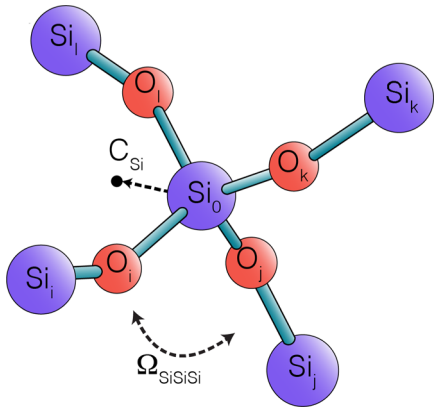


Figure 1. Si cluster out to two coordination spheres showing the $Si_iSi_0Si_j$ angle, $\Omega_{Si_iSi_0Si_j}$, and the silicon centroid, C_{Si}, calculated using $P_j = \{Si_i, Si_j, Si_k, Si_l\}$ in eq 7.

anisotropy directly from the structure. We note that the sign of the anisotropy ζ dictates the shape of the shielding tensor either oblately or prolately distorted. We find that the shape of the shielding tensor has no significant effect on the geometry of a silicon site, and thus we are able to separate the sign from the magnitude of anisotropy. Hence, the magnitude, $|\zeta|$, is used in our model.

The primary advantage of Υ over the current method of using density functional theory (DFT) to calculate and compare full tensor values is that Υ provides a clear dependence of anisotropy to changes in local structure and symmetry of a silicon site. Previous DFT-based methods, while accurate in calculating full tensors, act as a "black-box" and do not clearly show how the structure corresponds to the tensor. Furthermore, the obtained linear model may also be used in refinement of structurally ambiguous materials and data sets where the CSA tensor is available[12] and may be particularly useful when combined with additional models to further constrain the refinement, such as that proposed by Srivastava et al.,[13] in which additional second-coordination sphere silicon geometries were related to *J*-couplings. A combined refinement procedure using such models may greatly speed up NMR crystallographic refinements by replacing ab initio calculations with analytical models with comparable accuracy to simulation.

METHODS

Shielding Tensor Convention. In this study, we use the IUPAC[14] definition of nuclear shielding in which the isotropic nuclear shielding, σ^{iso}, is defined as the average of the trace of the nuclear shielding tensor

$$\sigma^{iso} = \frac{1}{3}(\sigma_{XX} + \sigma_{YY} + \sigma_{ZZ}) \tag{2}$$

where σ_{XX}, σ_{YY}, and σ_{ZZ} are the principal components of the nuclear shielding tensor in the principal axis frame. The principal axis labeling convention used is the Haeberlen convention in which the axes are ordered such that

$$|\sigma_{ZZ}-\sigma^{iso}| \geq |\sigma_{XX}-\sigma^{iso}| \geq |\sigma_{YY}-\sigma^{iso}| \tag{3}$$

Following the Haeberlen convention,[14] the shielding anisotropy, ζ, is defined as

$$\zeta = \sigma_{ZZ} - \sigma^{iso} \tag{4}$$

And the absolute value of the shielding anisotropy, $|\zeta|$, is the main focus of this study. Carrying on with the Haeberlen convention, shielding asymmetry, η, is defined as

$$\eta = \frac{\sigma_{YY} - \sigma_{XX}}{\zeta} \tag{5}$$

Random Forest. Random forest (RF) models were implemented using the Python package *scikit-learn* v0.24.[15] Models were created using the "RandomForestRegressor" class with "n_estimators" set to 1000 and the remaining parameters set to their default values. We adapt a feature selection methodology similar to that outlined by Hapfelmeier and Ulm,[16] in which a forest was trained over all of the features. A 5-fold cross-validation was used as a check for assessing the quality of the RF. The feature ranking metric chosen for ranking feature importance was the Gini importance.[17] During training of an RF, decision trees are generated and trained on the data, where each node in the tree is split on a feature. Each time a node is split on a feature in a forest, the Gini impurity for the resulting nodes decreases. The summed impurity for each feature over every tree in the forest gives a rapid importance metric in which important features have greater values. The least important feature was removed, and this process was repeated until there was one feature remaining. The RF was stable in its performance on a testing set down to one feature in each case indicating a dominating feature. To assess nondominant features, an arbitrary cutoff of 10 features was chosen. Feature importance was determined using the "feature_importances_" attribute, which calculates the Gini importance.

SISSO. The Sure Independence Screening and Sparsifying Operator (SISSO)[18] algorithm was used to create a nonlinear descriptor of the data. The features found to be the 10 most promising features from the RF models in the previous section were used to construct the feature sets, Φ. The feature set expansion was performed using an operator set $\hat{H}^{(m)} = \{I, +, -, |-|, *, /, ^{-1}, \exp, \ln, \sqrt{}, ^2, ^3, 1+, 1-\}$, and expansions were performed out to Φ_3. The expansion was performed using the *AutoFeat*[19] Python library, and the SISSO analysis was performed using a Python implementation of SISSO.[20] The SISSO regressor used was set to have 200 features per SIS iteration and was set to create a descriptor with one nonzero coefficient.

Additional Features from SISSO. We also considered expanding the dimensionality of the resulting SISSO descriptor to capture the effects of first-coordination sphere parameters, which may have a weak effect compared to the one-dimensional SISSO descriptor. To identify higher-dimension descriptors, the SISSO method uses the residuals from regression, with Υ as the target property for increasing the correlation to ζ. From the same set of 10 features derived from the RF and expanded by *AutoFeat*, the best-performing second-dimension descriptor found by SISSO was $|1/\langle\Omega_{O_iSi_0O_j}\rangle - 1/\langle\Omega_{Si_iSi_0Si_j}\rangle|$, which incorporates the O−Si−O tetrahedral angle, which was deemed responsible for additional anisotropy effects based on the outliers from the one-dimensional descriptor. This additional feature may increase the performance of the model in severely distorted Si−O tetrahedra. As mentioned in the main text, these types of sites are rare, and thus the increase in complexity from including a second descriptor was not included in the model.

https://doi.org/10.1021/acs.jpcc.1c04829
J. Phys. Chem. C XXXX, XXX, XXX−XXX

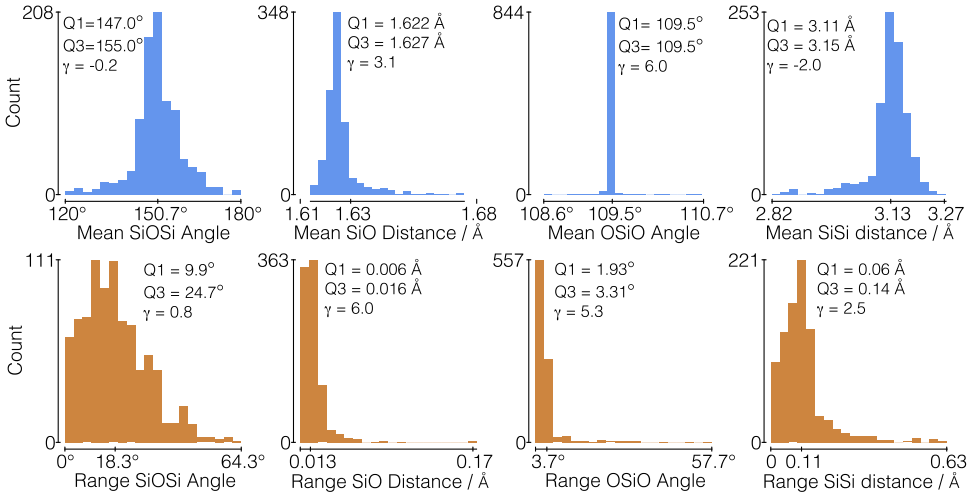


Figure 2. Distributions of simple geometric parameter mean values and ranges around a Si site. The range of the *x*-axes denotes the range of the geometric parameters, and the mean value is denoted by an additional tick. In the case of Si−O and O−Si−O ranges, the minimum of zero was not denoted due to labeling clash with the mean value. The distributions mostly resemble skewed Gaussian distributions, and thus first quartile, Q^1, third quartile, Q^3, and skewness, γ, are displayed. Raw data for this figure are provided in the SI.

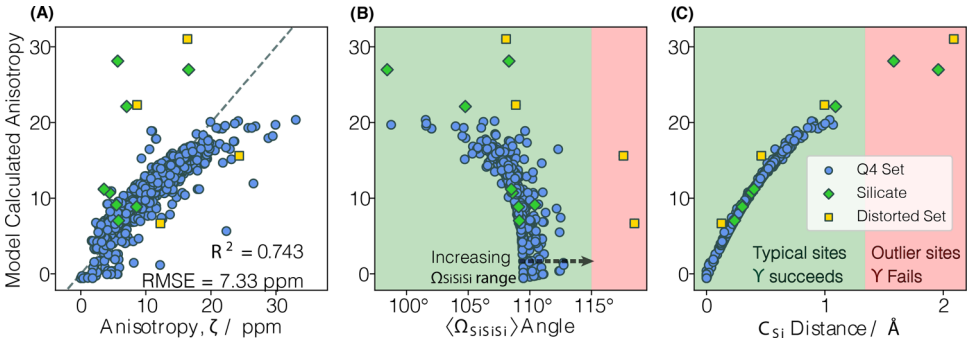


Figure 3. (A) Comparison of absolute anisotropy to the Υ-based model (eq 6) showing the overall correlation with few outliers. (B) $\langle\Omega_{Si_iSi_0Si_j}\rangle$ plotted against the Υ model showing the scattering in the feature space as structural distortion increases. (C) C_{Si} plotted against the Υ model showing the overall correlation of the centroid to the model and the little scattering over the feature space. In both panels (B) and (C), we may separate regions of low structural distortion (green) from regions of severe structural distortion (red), where Υ fails. Cutoff regions are set at the midpoint of the highest value in the "typical" population and the lowest value of the "outlier" population.

RESULTS

Here, we present the underlying ^{29}Si NMR Q^4 data set, the machine-learning-derived Υ parameter along with structural considerations and limitations of the parameter and the feature selection process for construction of the Υ parameter. The Υ parameter is then used to describe the local structure of siliceous zeolites Sigma-2 and silica-ZSM-5 (hereafter referred to simply as ZSM-5), whose CSA tensors have been reported.

Description of Data Set. The Q^4 ^{29}Si NMR data set is a subset of VASP NMR tensor calculations and relaxed structures calculated by Sun et al.[21] The data set is composed of 885 unique silicon sites from 288 $Si-O_4$ tetrahedron-containing structures, of which 282 are various forms of SiO_2 and the remaining 6 are $K_2Si_4O_9$ (mp-558603), $Na_2Si_3O_7$ (mp-556198), $Na_6Si_8O_{19}$ (mp-554033), $K_2MgSi_5O_{12}$ (mp-667292), $Na_2MgSi_5O_{12}$ (mp-560603), and $Li_2Si_3O_7$ (mp-555899). From the calculated shielding tensors, the shielding anisotropy was extracted, as described in the Shielding Tensor Convention section.

Differences in geometric parameters typically arise from structural and/or electronic variations among the silicon sites, and the data set shows a wide variety of differing environments. The distributions of mean Si−O−Si bond angles (mean = 150.7°, first quartile, Q^1 =147°, third quartile, Q^3 = 155°, ranging from 120 to 180°), Si−O bond distances (mean = 1.63 Å, Q^1 = 1.622 Å, Q^3 = 1.627 Å, ranging from 1.61 to 1.68 Å), O−Si−O tetrahedral angles (mean = 109.5°, Q^1 = 109.5°, Q^3 = 109.5°, ranging from 108.6 to 110.7°), and Si−Si distances (mean = 3.13 Å, Q^1 = 3.11 Å, Q3 = 3.15 Å, ranging from 2.82 to 3.27 Å) about a given silicon site are shown in blue in Figure 2. The mean geometric parameters all exhibit values similar to those reported in structural studies of Q^4 sites by Cruikshank (O−Si−O angles = 109.5),[22] Sen et al. (Si−O−Si angles 148 ± 12, Si−O distance = 1.59 ± 0.003, Si−Si distance = 3.07 ± 0.024),[23] and Srivastava et al. (Si−O−Si angle = 147.8 ± 4.8)[13] and more expanded SiO_2 structures such as zeolites (Si−O−Si angles = 154.3 ± 8.8, Si−O distances = 1.595 ± 0.01, O−Si−O angles = 109.47 ± 1.04, Si−Si distances = 3.10 ± 0.04)[24] to within a 2% distortion, which is commonly observed in DFT calculations of this type.[25] In addition to typical Si sites, the data set also contains more extreme local environments, allowing for a wide range of geometries to be modeled. Silicon site distortion is also well represented in the data set, which may better represent the shielding anisotropy. While standard deviation of the geometric parameters is

https://doi.org/10.1021/acs.jpcc.1c04829
J. Phys. Chem. C XXXX, XXX, XXX−XXX

sometimes used in structural investigations,[26] we argue that standard deviation is not an accurate parameter considering that only four bonds are used in its calculation. Instead, we opt to use the *range* of the geometric parameters about a site as an indication of geometric distortion, which are shown in orange in Figure 2. All parameters show positive skew and with mean values close to zero, implying that a large proportion of sites show little distortion, particularly in the first-coordination sphere with the highly skewed Si−O distances and O−Si−O angles, whereas there is higher distortion from high symmetry in the second-coordination sphere due to more flexible parameters like the Si−O−Si angle and Si−Si distance. Despite the high skew of all of the range distributions, all of the geometric parameters show some amount of spread and geometric distortion from an ideal Si−O4 tetrahedron and likely sample a wide range of local geometries.

To compare and contrast our model results, we also include analyses on the experimental Sigma-2 and silica-ZSM-5 data sets, which are siliceous zeolites containing 4 and 24 unique Si sites, respectively. Both materials have well-known structures, derived from single-crystal XRD experiments, and both have had their CSA tensors measured and verified against *ab initio* Hartree−Fock cluster calculations. The data set of both siliceous zeolites contains structure and CSA tensor information of all 4 of the Sigma-2 sites and 15 of the 24 ZSM-5 sites, as the remaining sites had CSA tensors that were unresolved.

Descriptor for Absolute Anisotropy. To analyze the descriptor and determine where the model performs well, as well as underperforms, a linear model using Υ was constructed in the Python package *scikit-learn*.[15] The resulting linear model for the absolute anisotropy was found to have an RMSE of 7.33 ppm and an R^2 of 0.743 and is shown in Figure 3A.

Figure 3B,C shows the correlation of each individual feature, $\langle\Omega_{SiSiSi}\rangle$ and C_{Si}, respectively, with respect to the Υ descriptor. The primary descriptive feature is the centroid distance, as it is clear from Figure 3C that an increase in the distance of the C_{Si} leads to an increase in the shielding anisotropy and thus a distortion away from spherical symmetry of the shielding tensor. From Figure 3B, we can see that up to an anisotropy of approximately 10 ppm the average $\langle\Omega_{SiSiSi}\rangle$ angle does not drop below 109.5° and instead clusters very tightly around 110°, with some points that deviate toward higher angles, which are due to increases in the range of $\langle\Omega_{SiSiSi}\rangle$ angle. Above ζ = 10 ppm, the average $\langle\Omega_{SiSiSi}\rangle$ angle drops abruptly and smaller $\langle\Omega_{SiSiSi}\rangle$ angles are seen down to about 100°. This likely indicates that severe structural distortion is required to sustain such high anisotropies. Additionally, the individual geometric components in Figure 3B,C may be used to identify outliers and useful range geometric parameters in which the model fails to explain. The regions in which the model fails are denoted by the red regions in Figure 3B,C in which $\langle\Omega_{SiSiSi}\rangle$ > 115° and C_{Si} >1.34 Å were chosen as they were the midpoints between the points within the useful region and the closest outlying point.

The major factor in determining outliers appears to be the distortion of the Si−O_4 tetrahedron, which is not captured by eq 1. The outliers shown in Figure 3 belong to one of two classes: the silicates (designated by green diamonds) or SiO_2 materials with severely distorted tetrahedra (designated by yellow squares). The outliers were analyzed to reveal structural features causing the model to perform poorly. In all of the outliers, a heavily distorted O−Si−O tetrahedral angle was observed (typical ranges of a minimum tetrahedral angle of 90° up to a maximum angle of 130°). This result is not unexpected, as our descriptor only takes into account second-coordination sphere parameters and does not consider distortion to the Si−O_4 tetrahedron. The Sure Independence Screening and Sparsifying Operator (SISSO)[18] algorithm may be employed to search a higher-dimensional space and identify an additional descriptor to handle outliers. The second descriptor found by SISSO is $|1/\langle\Omega_{O_iSi_0O_j}\rangle - 1/\langle\Omega_{Si_iSi_0Si_j}\rangle|$ and is able to reduce the effect of outliers. However, we note that, in our data set of 885 Si sites, these heavily distorted sites account for a small fraction of Si−O_4-containing materials.

Additionally, the deviation seen in the silicates appears to be due to distortion of the Si−O_4 tetrahedron rather than effects from other constituent ions in the material. For example, the silicates $Na_2Si_3O_7$, $Na_6Si_8O_{19}$, and $Na_2MgSi_5O_{12}$ contain four sites showing Na ion coordination to bridging oxygens at distances ranging from 2.4 to 2.6 Å. The silicon site bonded to a bridging oxygen atom with a Na ion coordination distance of 2.6 Å does not show significant tetrahedral distortion and is not an outlier according to the model, whereas the other three sites show closer coordination distances and also distorted tetrahedra. The remaining silicates show cation−oxygen coordinations of much greater distances, where $K_2Si_4O_9$, $K_2Si_4O_9$, and $K_2MgSi_5O_{12}$ show O^{2-}−K^+ coordination distances of 3.1 and 3.3 Å, respectively, whereas the Li ion in $Li_2Si_3O_7$ does not coordinate to the bridging oxygen atom in the structure. Additionally, the three K^+-containing structures do not show significant Si−O_4 tetrahedral distortion. It is unclear how much the cation-bridging oxygen coordination distance affects the anisotropy or tetrahedral distortion of a site; however, the proximity of cations is likely to affect the electron density from the bridging oxygen and thus subsequently the Si atom. However, this effect is likely weaker than that seen in the lower Q^n species. In each case of deviation from the model, there is tetrahedral distortion, however, regardless of the coordination distance.

We also explore the descriptor space of the individual features in eq 1, shown in Figure 3B,C, to reveal the limits of applicability of eq 1. The centroid, shown in Figure 3C, shows a tight correlation with little deviation from linearity. $\langle\Omega_{Si_iSi_0Si_j}\rangle$, shown in Figure 3B, however, shows significant scattering. Again, the points showing the most deviation are those identified previously, exhibiting significantly distorted Si−O_4 tetrahedra. At constant ordinate, moving from left to right corresponds to an increase of the $\Omega_{Si_iSi_0Si_j}$ range (dotted arrow in Figure 3). At small structural distortions, the $\langle\Omega_{Si_iSi_0Si_j}\rangle$ shows a tight nonlinear correlation to anisotropy; however, as the range of $\Omega_{Si_iSi_0Si_j}$ increases, the local structure distorts and thus anisotropy increases as well. The data set was not large enough to fully explore the dependence on these slight distortions; however, they are linked to first-coordination sphere distortions. Figure 3B shows two outliers with $\langle\Omega_{Si_iSi_0Si_j}\rangle$ greater than 115°, yet their model-calculated anisotropy is well predicted. In both cases, the $\langle\Omega_{Si_iSi_0Si_j}\rangle$ is larger than expected, while the C_{Si} distance is smaller than expected. Increasing $\langle\Omega_{Si_iSi_0Si_j}\rangle$ and decreasing C_{Si} counteract each other, which averages out the errors in both silicon sites yielding a predicted anisotropy that is within error for the model. A cutoff of

https://doi.org/10.1021/acs.jpcc.1c04829
J. Phys. Chem. C XXXX, XXX, XXX−XXX

$\langle\Omega_{Si_iSi_0Si_j}\rangle > 115°$ and $C_{Si} > 1.34$ Å (red region) is chosen to show where eq 1 performs poorly.

The outliers based on the above ranges were removed, and the model was retrained using *scikit-learn*. The resulting linear model was determined to be

$$|\zeta(\Upsilon)| = (36.6\pm0.7)\Upsilon + (-135 \pm 3) \quad (6)$$

with an RMSE of 6.77 ppm and an R^2 of 0.761. Model performance may be improved by applying $\langle\Omega_{Si_iSi_0Si_j}\rangle$ cutoff ranges to be between two standard deviations from the mean $\langle\Omega_{Si_iSi_0Si_j}\rangle$ value such that only angles in $105.8° \leq \langle\Omega_{Si_iSi_0Si_j}\rangle \leq 111.4°$ are taken. Applying cutoff regions based on the standard deviation improves the model fit to RMSE = 4.99 ppm and R^2 = 0.802. Outliers remain, however, which cannot be accounted for by using simple rules. A larger data set will be required to elucidate the geometric features that cause the Υ parameter to perform poorly in these instances.

The partial derivatives of eq 6 also show limiting behavior of Q^4 geometry and hint at the transition between Q^4 and Q^3 or lower sites. Recognizing that Υ is a strictly positive-valued function for all relevant ranges of $\Omega_{Si_iSi_0Si_j}$ and C_{Si}, we may reduce the derivatives to

$$\frac{\partial|\zeta(\Upsilon)|}{\partial\langle\Omega_{Si_iSi_0Si_j}\rangle} = \frac{36.6}{\langle\Omega_{Si_iSi_0Si_j}\rangle}$$

$$\frac{\partial|\zeta(\Upsilon)|}{\partial C_{Si}} = 36.6e^{-C_{Si}}$$

Both derivatives are decreasing positive definite functions over typical ranges of each respective parameter; however, angular distortions have a much weaker effect on anisotropy than centroid distortions. We note that a minor angular distortion, without any significant distortions to first-coordination sphere parameters, takes place through dihedral angles, which have shown little to no direct effect on anisotropy. Centroid distortions, however, require a stronger distortion of the first-coordination sphere parameters, such as bond distances and angles, and thus exhibit a greater effect on the silicon site anisotropy. Both parameters are decreasing functions indicating a limit to the distortion possible in a Q^4 site, which may indicate a transition from Q^4 to Q^3 or lower sites.

Feature Space Reduction and SISSO Analysis. To analyze feature importances and to also reduce the dimensionality of the feature space to a size manageable for SISSO, we employ random forest (RF) regression, where RF algorithms are convenient as they are able to handle high-dimensional data and nonlinear and linear relationships and they have an in-built method for ranking the importance of the variables. The RF training and feature ranking methodology are discussed in the Methods section. The least important feature was removed, and this process was repeated until there were 10 features remaining.

From the RF analysis of $|\zeta|$, the most important features were found to be C_{Si} and $\Omega_{Si_iSi_0Si_j}$ with normalized Gini importances[17] on the order of 0.8 and 0.05, respectively. The remaining features that showed importance in $|\zeta|$ were Ω^{min}_{OSiO}, Ω^{range}_{SiSiSi}, $\langle\Omega_{OSiO}\rangle$, Ω^{max}_{SiSiSi}, d^{range}_{Si-Si}, d^{max}_{Si-Si}, and d^{min}_{Si-Si}, all of which had importance scores that steadily decreased over the range of 0.016−0.009. The most important features found by RF analysis showed some instability at low importance; however, this is likely due to the features having small importance scores that are close in magnitude. The features found are also dominated by second-coordination sphere parameters, with some slight dependence on first-coordination sphere O−Si−O tetrahedral angles, as expected from the correlations obtained from the two-dimensional descriptor obtained via SISSO analysis in the Methods section.

Additionally, two tests were conducted to show the separability of sign and magnitude of ζ. An RF was trained over the ζ values and showed very little correlation between model-predicted and DFT-calculated ζ while also showing significant differences in feature importances. When a binary variable to represent the correct sign of ζ was added, then the RF was able to correctly predict magnitude and showed similar feature importances as seen with $|\zeta|$. Additionally, the positive and negative ζ values were separated and independent RF was trained over each. The feature importances were found to be similar across both positive and negative ζ down to slight instability in lesser-importance features; however, this was attributed to differences in the data set and represented geometries across the two sets. These tests suggest that the sign and the magnitude of ζ are independent of each other and are learned by different features, thus allowing the use of $|\zeta|$ in a geometric model.

The SISSO methodology was then used to identify an improved descriptor for predicting $|\zeta|$ given geometric parameters determined to be important from the RF analysis. SISSO was performed on the entire set of 885 unique silicon sites and determined that the geometric descriptor, Υ, in eq 1, depends only on the second-coordination sphere parameters $\Omega_{Si_iSi_0Si_j}$ and C_{Si}. Furthermore, Υ shows a significant increase in performance over the previously used linear dependence of simple descriptors such as the $\langle\Omega_{Si_iSi_0Si_j}\rangle$ and the "*mean deviation of Si atoms from ideal tetrahedron*".[12]

Comparing Descriptor to Experimentally Observed CSA Tensors. The model according to eq 6 was then applied to Brouwer's CSA tensor for Sigma-2 and ZSM-5 data sets. The CSA tensor for ZSM-12 was left out of consideration due to the difficulties in obtaining precise structural data, and therefore using the ZSM-12 data would result in an inaccurate analysis of the model. Additionally, the data reported by Brouwer must be converted from span to anisotropy; however, in the case of Sigma-2, the two values are equivalent.

The data shown in Figure 4 show a significantly better correlation to the model (R^2 = 0.581) than the simple geometric parameters identified by Brouwer and Enright (R^2 = 0.410), including accounting for the outlier Si-19 in the ZSM-5 from the mean deviation of Si atoms from ideal tetrahedron. This indicates that the Υ-based model for anisotropy (eq 6) may provide a useful probe into the structural analysis and NMR crystallography-based refinements of silicon materials.

At present, we do not have enough experimental data to confidently propose coefficients for an experimental model; we instead opt to use Υ to identify the distorted sites and show how Υ may be used for the assessment of relative distortion. In Figure 4, we show the experimental Sigma-2 and ZSM-5 Υ values in comparison to the modeled set of Q^4 geometries where the dashed lines show the first and third quartiles (Q^1 and Q^3, respectively) and the dotted line shows the mean Υ-value (μ). We may assess the relative amount of distortion of the Q^4 sites from an ideal second-coordination shell tetrahedron via where the site Υ values fall with respect to the Q^4 data set, i.e., high-symmetry sites have $\Upsilon < Q^1 = 3.92$,

https://doi.org/10.1021/acs.jpcc.1c04829
J. Phys. Chem. C XXXX, XXX, XXX−XXX

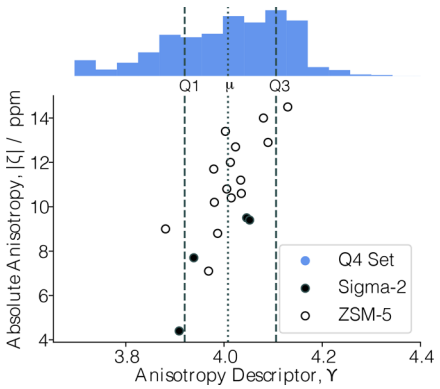


Figure 4. Correlation between absolute anisotropy and ϒ using CSA tensor values reported by Brouwer and Enright along with the distribution of ϒ seen in the Q^4 set. The mean ϒ, μ, first and third quartiles, Q^1 and Q^3, respectively, are used to segment the zeolite data to compare sites to typical ranges of structural distortion. Two bins at ϒ > 4.4 are not shown.

average sites have ϒ near μ = 4.01 between Q^1 and Q^3, and highly distorted sites have ϒ > 4.11. In the ZSM-5 sample, the anisotropy ranges from 7.1 ppm up to 14.5 ppm, with all sites exhibiting a Si−O tetrahedral angle of 109.5° (Si-5 shows considerable tetrahedral range; however, it averages out to 109.5°). In comparing the lowest-ζ site to the computational data, we observe that site 11 is on the lower quartile, Q^1, of the silicon site distortion in terms of $\langle \Omega_{Si_iSi_0Si_j} \rangle$ and between the Q^1 and median value for C_{Si}. Furthermore, the ϒ descriptor (eq 1) places this site as a low-distortion site in comparison to the computational set as it is below Q^1. Our model (eq 6), however, overpredicts the site anisotropy by 2.3 ppm, which is still within the fitting error of 6.77 ppm reported in section Descriptor for Absolute Anisotropy. Examining the highest-ζ site, we find that the corresponding Si site has an angle, centroid, and ϒ between the upper quartile, Q^3, and the maximum value in comparison to the experimental data indicating a relatively distorted $Si-O_4$ site. Additionally, the model predicts the anisotropy well, showing an error of only 0.8 ppm.

DISCUSSION

Q^4 CSA tensors are seldom reported in the literature due to experimental difficulties in fitting the tensor to the spectra, and as a result, only a handful of full tensors have been reported or studied in great detail. Brouwer and Enright developed a pulse sequence allowing for the accurate measurement of CSA tensors of individual silicon sites in crowded spectra and provides a means to grow the set of Si Q^4 sites with reported tensors.[12] *Ab initio* calculations have been shown to accurately calculate the NMR shielding tensors[21] and thus provide a useful means to study the structural dependence of these difficult-to-measure sites. While experimental data are lacking, our computational set of structurally diverse Q^4 local geometries enables us to more fully explore the feature space and find correlating features to the anisotropy.

The choice of ϒ is inspired by previous work on Q^4 anisotropy, namely, the brief correlation investigation by Brouwer and Enright and a symmetry-based investigation by Avnir and colleagues.[27,28] A feature space of 43 geometric and symmetry parameters about the first- and second-coordination spheres was constructed to investigate the structural dependence of anisotropy. The full listing of features considered in the study can be found in the Supporting Information (SI). Beyond typical parameters, such as bond distances and angles, the distance between the central Si atom and the centroid of the tetrahedron composed of the atoms in the first- or second-coordination sphere was included in the data set. We do so by

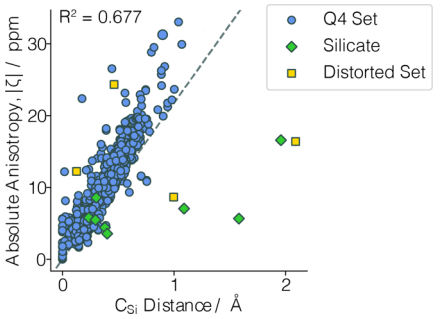


Figure 5. Correlation between the absolute anisotropy and the Si centroid.

calculating the centroid, P_{C_i}, shown in Figure 5, as the average of the coordinates of the atoms in a given coordination sphere i

$$P_{C_i} = \frac{1}{N}\sum_{j=0}^{N} P_j,\ j \in \{0, 1, ..., N\} \tag{7}$$

where $\{0, 1, ..., N\}$ is the set of atoms in the first- or second-coordination sphere, P_j is the coordinate of atom j in the coordination shell being analyzed, and N is the number of atoms in $\{0, 1, ..., N\}$. The distance between the central Si atom and the centroid, C_i, is then calculated

$$C_i = |P_{C_i} - P_{Si_0}| \tag{8}$$

where P_{Si_0} is the coordinate of the central Si atom.

Both first-coordination shell oxygen centroid and second-coordination shell silicon centroid distances were analyzed, all measured from the central Si atom to the centroid, as shown in Figure 5. The oxygen centroid shows no correlation to anisotropy and reflects the rigidity of the $Si-O_4$ tetrahedron as a majority of points were centered near 0. The Si centroid showed a strong correlation to |ζ| with R^2 = 0.677. From Figure 5, it can be seen that the majority of the data follows a linear correlation between C_{Si} and |ζ| with some outliers. The majority of the data is clustered in the range of ζ = 0−20 ppm corresponding to C_{Si} between 0 and 0.8 Å, and the linear correlation continues beyond the majority of the data as C_{Si} increases beyond 0.8 Å, except the outlying points. It is likely that C_{Si} cannot explain the anisotropy alone, and a discussion on the flaws of the parameter will be discussed below.

Similarly to the correlations reported by Brouwer and Enright, there were no correlations of anisotropy (in their case, span, Ω, a similar measure to anisotropy but using a different tensor convention) to simple first-coordination sphere geometric parameters. While Brouwer and Enright reported weak correlations to some second-coordination sphere parameters, a wider range of geometries present in our data set revealed these weak correlations to be noncorrelations.

An additional geometric parameter analyzed by Brouwer and Enright that showed a weak correlation was the "*mean deviation of Si atoms from ideal tetrahedron*". This metric for spherical

https://doi.org/10.1021/acs.jpcc.1c04829
J. Phys. Chem. C XXXX, XXX, XXX−XXX

symmetry of the tensor relies on a comparison of geometries to an ideal tetrahedral point group and provides a better representation of deviation from symmetry rather than deviation of geometric parameters. This symmetry-based parameter and others from Avnir and colleagues[27,28] have shown relatively strong correlation to tensor parameters but offer weak structural and geometric insight. Avnir's continuous symmetry model offers very little geometric insight and does not perform well in predicting anisotropy for tetrahedra with arbitrary distortions. Brouwer and Enright's ideal tetrahedron model makes strong assumptions on the tetrahedral configuration of the second-coordination sphere and also assumes that the isometry center is placed on the central Si atom and that the deviation is measured from the second-coordination sphere Si atoms to the ideal tetrahedron vertices. Rather than requiring a fixed center of isometry, we will relax this constraint and instead compare the isometry center of the Si tetrahedron to the central Si atom as a geometric measure of ideal symmetry.

The centroid alone has one major flaw, however. Any highly symmetric configuration of four atoms around a central atom may deviate strongly from the T_d symmetry, yet have the centroid remain on the central atom. For example, a configuration with the T_d symmetry may be transformed to one with D_{4h} while maintaining a fixed centroid as done by Avnir et al.[27] This consideration is, however, seldom observed in our data set, which we believe is largely encompassing silicon tetrahedral geometries in naturally and synthetically occurring samples. Additionally, despite the incomplete picture of geometric dependence of $|\zeta|$, the centroid metric provides additional geometric features that may be included in future, improved models.

Finally, while the focus of this study was on anisotropy, the asymmetry tensor parameter was analyzed in the same manner. Unfortunately, no correlations or descriptors were found for asymmetry and will be the subject of future investigations.

CONCLUSIONS

We report a new local structure descriptor, Υ, that enables the study of Si local structure and significantly improved correlation as compared to simple geometric parameters and enables the prediction and interpretation of Si Q^4 anisotropy. For a data set of 885 structurally diverse silicon sites, Υ takes into account both the site geometry and symmetry to give an unprecedented correlation to anisotropy with an R^2 of 0.761 and a root-mean-squared error of 6.77 ppm. The Υ descriptor was created using a data-driven approach to search for an intuitive descriptor space yielding a descriptor that allows for a facile interpretation of the relationship between environment and shielding tensor. We use this model to interpret the structural distortion in the siliceous zeolite ZSM-5 to show that this descriptor can aid in structural analysis of Si-based materials.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcc.1c04829.

Additional information for modeling; list of all features generated and considered in random forest analysis (Table S1) (PDF)

Structures and NMR tensors used in the study are hosted on https://contribs.materialsproject.org/projects/lsdivaspsi29/ (ZIP)

AUTHOR INFORMATION

Corresponding Author

Kristin A. Persson − *Department of Materials Science and Engineering, University of California, Berkeley, California 94720, United States; Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States;* orcid.org/0000-0003-2495-5509; Email: kapersson@lbl.gov

Authors

Maxwell C. Venetos − *Department of Materials Science and Engineering, University of California, Berkeley, California 94720, United States;* orcid.org/0000-0003-3468-2006

Shyam Dwaraknath − *Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States;* orcid.org/0000-0003-0289-2607

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcc.1c04829

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The author thanks Dr. Philip Grandinetti and Dr. Deepansh Srivastava for helpful discussions. This work was supported by the U.S. National Science Foundation under Grant No. DIBBS OAC 1640899. 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).

REFERENCES

(1) Brouwer, D. H.; Cadars, S.; Eckert, J.; Liu, Z.; Terasaki, O.; Chmelka, B. F. A General Protocol for Determining the Structures of Molecularly Ordered but Noncrystalline Silicate Frameworks. *J. Am. Chem. Soc.* **2013**, *135*, 5641−5655.

(2) Zhang, P.; Grandinetti, P. J.; Stebbins, J. F. Anionic Species Determination in $CaSiO_3$ Glass Using Two-Dimensional ^{29}Si NMR. *J. Phys. Chem. B* **1997**, *101*, 4004−4008.

(3) Brouwer, D.; Huizen, J. V. NMR crystallography of zeolites: How far can we go without diffraction data? *Magn. Reson. Chem.* **2019**, *57*, 167−175.

(4) Brouwer, D. H. Structure solution of network materials by solid-state NMR without knowledge of the crystallographic space group. *Solid-State NMR* **2013**, *51*−*52*, 37−45.

(5) Brouwer, D.; Horvath, M. A simulated annealing approach for solving zeolite crystal structures from two-dimensional NMR correlation spectra. *Solid State Nucl. Magn. Reson.* **2015**, *65*, 89−98.

(6) Brouwer, D. H. A structure refinement strategy for NMR crystallography: An improved crystal structure of silica-ZSM-12 zeolite from ^{29}Si chemical shift tensors. *J. Magn. Reson.* **2008**, *194*, 136−146.

(7) Smith, K. A.; Kirkpatrick, R. J.; Oldfield, E.; Henderson, D. M. High-Resolution Silicon-29 Nuclear Magnetic Resonance Spectroscopic Study of Rock-Forming Silicates. *Am. Mineral.* **1983**, *68*, 1206−1215.

(8) Grimmer, A.-R.; Peter, R.; Gechner, E. F.; Molgedey, G. High Resolution ^{29}Si NMR in Solid Silicates. Correlations between

Shielding Tensor and Si-O Bond Length. *Chem. Phys. Lett.* **1981**, *77*, 331−335.
(9) Grimmer, A.-R. Correlation between Individual Si-O Bond Lengths and the Principal Values of the ^{29}Si Chemical-Shift Tensor in Solid Silicates. *Chem. Phys. Lett.* **1985**, *119*, 416−420.
(10) Jardón-Álvarez, D.; Sanders, K. J.; Phyo, P.; Baltisberger, J. H.; Grandinetti, P. J. Cluster formation of network-modifier cations in cesium silicate glasses. *J. Chem. Phys.* **2018**, *148*, No. 094502.
(11) Baltisberger, J. H.; Florian, P.; Keeler, E. G.; Phyo, P. A.; Sanders, K. J.; Grandinetti, P. J. Modifier cation effects on ^{29}Si nuclear shielding anisotropies in silicate glasses. *J. Magn. Reson.* **2016**, *268*, 95−106.
(12) Brouwer, D. H.; Enright, G. D. Probing Local Structure in Zeolite Frameworks: Ultrahigh-Field NMR Measurements and Accurate First-Principles Calculations of Zeolite 29Si Magnetic Shielding Tensors. *J. Am. Chem. Soc.* **2008**, *130*, 3095−3105.
(13) Srivastava, D. J.; Florian, P.; Baltisberger, J. H.; Grandinetti, P. J. Correlating geminal $^{2}J_{Si\text{-}O\text{-}Si}$ couplings to structure in framework silicates. *Phys. Chem. Chem. Phys.* **2018**, *20*, 562−571.
(14) Harris, R. K.; Becker, E. D.; De Menezes, S. M. C.; Grangerd, P.; Hoffman, R. E.; Zilm, K. W. Further conventions for NMR shielding and chemical shifts, IUPAC recommendations 2008. *Solid State Nucl. Magn. Reson.* **2008**, *33*, 41−56.
(15) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(16) Hapfelmeier, A.; Ulm, K. A new variable selection approach using Random Forests. *Comput. Stat. Data Anal.* **2013**, *60*, 50−69.
(17) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(18) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.* **2018**, *2*, No. 083802.
(19) Horn, F.; Pack, R.; Rieger, M. In *The Autofeat Python Library for Automated Feature Engineering and Selection*, Joint European Conference on Machine Learning and Knowledge Discovery in Databases; Springer: Cham., Würzburg, Germany, Sept 16−20, 2019; pp 111−120.
(20) Ahmetcik, E.; Ziletti, A.; Ouyang, R.; Sbailó, L.; Ghiringhelli, L.; Scheffler, M. Compressed Sensing 4 Materials Science. https://gitlab.mpcdf.mpg.de/nomad-lab/analytics-compressed-sensing/-/tree/master/ (accessed Nov 12, 2020).
(21) Sun, H.; Dwaraknath, S.; Ling, H.; Qu, X.; Huck, P.; Persson, K.; Hayes, S. Enabling materials informatics for 29Si solid-state NMR of crystalline materials. *npj Comput. Mater.* **2020**, *6*, No. 53.
(22) Cruickshank, D. W. J. The Role of 3d-Orbitals in π-Bonds between (a) Silicon, Phosphorus, Sulphur, or Chlorine and (b) Oxygen or Nitrogen. *J. Chem. Soc.* **1961**, *1077*, No. 5486.
(23) Trease, N. M.; Clark, T. M.; Grandinetti, P. J.; Stebbins, J. F.; Sen, S. Bond length-bond angle correlation in densified silica-Results from ^{17}O NMR spectroscopy. *J. Chem. Phys.* **2017**, *146*, No. 184505.
(24) Brouwer, D. H.; Brouwer, C. C.; Mesa, S.; Semelhago, C. A.; Steckley, E. E.; Sun, M. P.; Mikolajewski, J. G.; Baerlocher, C. Solid-state 29Si NMR spectra of pure silica zeolites for the International Zeolite Association Database of Zeolite Structures. *Microporous Mesoporous Mater.* **2020**, *297*, No. 110000.
(25) Jones, R. O.; Gunnarsson, O. The density functional formalism, its applications and prospects. *Rev. Mod. Phys.* **1989**, *61*, 689−746.
(26) Dawson, D. M.; Moran, R. F.; Ashbrook, S. E. An NMR Crystallographic Investigation of the Relationships between the Crystal Structure and ^{29}Si Isotropic Chemical Shift in Silica Zeolites. *J. Phys. Chem. C* **2017**, *121*, 15198−15210.
(27) Steinberg, A.; Karni, M.; Avnir, D. Continuous Symmetry Analysis of NMR Chemical Shielding Anisotropy. *Chem. - Eur. J.* **2006**, *12*, 8534−8538.
(28) Zabrodsky, H.; Peleg, S.; Avnir, D. Continuous symmetry measures. 2. Symmetry groups and the tetrahedron. *J. Am. Chem. Soc.* **1993**, *115*, 8278−8289.

https://doi.org/10.1021/acs.jpcc.1c04829
J. Phys. Chem. C XXXX, XXX, XXX−XXX