frontiers in Materials

ORIGINAL RESEARCH
published: 13 November 2017
doi: 10.3389/fmats.2017.00034

# Assessing Local Structure Motifs Using Order Parameters for Motif Recognition, Interstitial Identification, and Diffusion Path Characterization

*Nils E. R. Zimmermann*[1]*, *Matthew K. Horton*[2,3], *Anubhav Jain*[3] and *Maciej Haranczyk*[1]

[1] *Lawrence Berkeley National Laboratory, Computational Research Division, Berkeley, CA, United States,* [2] *Department of Materials Science and Engineering, University of California, Berkeley, Berkeley, CA, United States,* [3] *Lawrence Berkeley National Laboratory, Energy Technologies Area, Berkeley, CA, United States*

Structure–property relationships form the basis of many design rules in materials science, including synthesizability and long-term stability of catalysts, control of electrical and opto-electronic behavior in semiconductors, as well as the capacity of and transport properties in cathode materials for rechargeable batteries. The immediate atomic environments (i.e., the first coordination shells) of a few atomic sites are often a key factor in achieving a desired property. Some of the most frequently encountered coordination patterns are tetrahedra, octahedra, body and face-centered cubic as well as hexagonal close packed-like environments. Here, we showcase the usefulness of local order parameters to identify these basic structural motifs in inorganic solid materials by developing classification criteria. We introduce a systematic testing framework, the Einstein crystal test rig, that probes the response of order parameters to distortions in perfect motifs to validate our approach. Subsequently, we highlight three important application cases. First, we map basic crystal structure information of a large materials database in an intuitive manner by screening the Materials Project (MP) database (61,422 compounds) for element-specific motif distributions. Second, we use the structure-motif recognition capabilities to automatically find interstitials in metals, semiconductor, and insulator materials. Our Interstitialcy Finding Tool (InFiT) facilitates high-throughput screenings of defect properties. Third, the order parameters are reliable and compact quantitative structure descriptors for characterizing diffusion hops of intercalants as our example of magnesium in $MnO_2$-spinel indicates. Finally, the tools developed in our work are readily and freely available as software implementations in the pymatgen library, and we expect them to be further applied to machine-learning approaches for emerging applications in materials science.

**Keywords: materials science, crystal structure, descriptors, databases, interstitials, intercalation, diffusion**

OPEN ACCESS

***Edited by:***
*Zhenyu Li,*
*University of Science and Technology of China, China*

***Reviewed by:***
*Yi Liu,*
*Shanghai University, China*
*Zhen Zhou,*
*Nankai University, China*
*Jun Jiang,*
*University of Science and Technology of China, China*

****Correspondence:***
*Nils E. R. Zimmermann*
*nerz@lbl.gov*

***Specialty section:***
*This article was submitted to Computational Materials Science, a section of the journal Frontiers in Materials*

***Received:*** *13 July 2017*
***Accepted:*** *18 October 2017*
***Published:*** *13 November 2017*

***Citation:***
*Zimmermann NER, Horton MK, Jain A and Haranczyk M (2017) Assessing Local Structure Motifs Using Order Parameters for Motif Recognition, Interstitial Identification, and Diffusion Path Characterization. Front. Mater. 4:34. doi: 10.3389/fmats.2017.00034*

## 1. INTRODUCTION

Crystals consist of atoms that are arranged in periodic patterns in three dimensions (Sands, 1993). This regular arrangement is called the crystal structure which, together with the chemical composition, dictates the properties of a material (Morris, 2007). Typically, the crystal structure is described with approximations and abstractions (Morris, 2007). One approach is to focus on the immediate surrounding of each atom (first coordination shell) and to use the number of surrounding atoms (coordination number) and the pattern (structure motif) for structure description, the discipline of which was coined by Werner and which is today known as coordination chemistry (Werner, 1912). Among frequently occurring structure motifs are tetrahedra, octahedra, body-center and face-centered cubic as well as hexagonal close-packed motifs (**Figure 1**).

The occurrence of basic structural motifs in crystalline compounds has been shown to be important indictors for predicting materials properties in several scientific and technological contexts. Finding and quantitatively assessing primary building blocks of zeolite materials ($SiO_4$ tetrahedra) can be used to predict the feasibility of synthesizing a (hypothetical) material (Li et al., 2013; Mazur et al., 2015) and to rate their likelihood for industrial deployment—for example, as a catalyst—(Zimmermann and Haranczyk, 2016). Design rules for novel battery materials are frequently developed employing information about the coordination pattern of the migrating ion (Rong et al., 2015) and the host structure (Li et al., 2009; Wang et al., 2015). Models based on structural fragments can be used to assess influencing factors to the superconductivity critical temperature (Isayev et al., 2015). Interstitials in dense inorganic materials are frequently found in positions where the interstitials assume tetrahedrally or octahedrally coordinated positions (Decoster et al., 2008, 2009a,b, 2010a,b, 2012; Pereira et al., 2011, 2012; Amorim et al., 2013; Silva et al., 2014).

Screening large databases for structure motif occurrence has hence the potential to find new candidate materials for various emerging applications. The inherent difficulty is to develop recognition tools that allow for reliable and rapid motif identification. There are two basic steps involved in the (automatic) identification of a coordination motif around a given atom: (i) neighbor finding and (ii) pattern matching. Neighbors can be found on the basis of interatomic distances—possibly in combination with typical bond lengths (Brunner, 1977; Hoppe, 1979; O'Keeffe and Brese, 1991)—or by a topology-based approach (Dirichlet, 1850; Voronoi, 1908; Mickel et al., 2013). For pattern matching, there exist two popular conceptual approaches: (i) using Monte Carlo (MC) moves (Shetty et al., 2002) and (ii) using order parameters (Steinhardt et al., 1983; Peters, 2009; Zimmermann et al., 2015). In the MC approach, an ideal structure motif is placed onto a central atom and its neighbors, and the ideal motif's position and size are varied to yield a small root mean square deviation between the positions of the ideal motif and the neighbors of the central atom. In the systematically expandable (Santiso and Trout, 2011) order parameter approach, the bond angles of a given motif are used in mathematical functions to directly yield a measure of motif resemblance, thus, being a deterministic method. Note that the MC-based approach is expected to be much more time-consuming than the order parameter route.

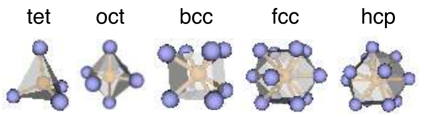

**FIGURE 1** | Basic structural motifs that frequently recur in various materials databases (from left to right): tetrahedron (tet), octahedron (oct) as well as body-centered cubic (bcc), face-centered cubic (fcc), and hexagonal close packed (hcp) motifs. The central atom and bonds are shown in orange, whereas the coordinating atoms are displayed in blue.

We here develop an effective and computationally efficient approach for finding atomic neighbors and identifying motif types in inorganic materials using order parameters (Steinhardt et al., 1983; Peters, 2009; Zimmermann et al., 2015) for pattern matching. Furthermore, we introduce a testing framework (Einstein crystal test rig) for validation of any such motif-finding effort. We then apply our approach to the database provided by the Materials Project (Jain et al., 2013), where we use well-defined materials subsets for testing. Finally, the method is used to generate crystal structure representations of the Materials Project database, to determine potential interstitial sites in several materials, and to quantitatively characterize the coordination environment change along the jump-diffusion path of an intercalating ion.

## 2. METHODS

We focus on local structural motifs that are based on a central atom and its first coordination shell. The two basic steps in identifying structural motifs are therefore:

1. finding bonded neighbors and
2. motif recognition.

More complex patterns such as those involving second-shell neighbors and cyclic motifs (rings) would require a more extensive analysis of the connectivity between atoms.

### 2.1. Bonding Identification

Bonds are determined on the basis of the distance, $d_{i,j}$, between two atoms $i$ and $j$:

$$d_{i,j} = \| \mathbf{p}_i - \mathbf{p}_j \|, \tag{1}$$

where $\mathbf{p}_i$ is the position of atom $i$. We systematically investigate three different neighbor-finding methods, all of which work with a site-specific cutoff distance, $r_{\mathrm{cut},i}$. In the first method ("min_dist"), we determine the (absolute) distance to the nearest neighbor, $d_{\mathrm{min},i}$, of a given site $i$ and, subsequently, we consider all additional sites that are at maximum $r_{\mathrm{cut},i} = (1 + \delta)d_{\mathrm{min},i}$ apart from site $i$ (**Figure 2**), where $\delta$ denotes a (relative) neighbor-finding tolerance (distance). The other two approaches work similarly, except for the fact that we use dimensionless distances, $\tilde{d}_{i,j} = d_{i,j}/l_{i,j}$, where $l$ is a length being characteristic for the considered pair of atoms $i$ and $j$. The following two approaches for the characteristic length are tested: the sum of atom (or, ion) radii (Shannon (1976); "min_VIRE": $l_{i,j} = r_i^{\mathrm{atom}} + r_j^{\mathrm{atom}}$) and the typical bond length (O'Keeffe and Brese (1991); "min_OKeeffe": $l_{i,j} = l_{i,j}^{\mathrm{bond}}$). The radii are calculated with a valence-ionic radius estimator (VIRE) implemented in pymatgen (Ong et al., 2013). The estimator uses a maximum *a posteriori* estimation method of the oxidation state of each site based on bond-valence sums (O'Keeffe and Brese, 1991). Furthermore, a first estimate of the coordination number is inferred from Voronoi decomposition (Dirichlet, 1850; Voronoi, 1908) as the number of faces making up the polyhedron, weighing each face's contribution in proportion to its solid angle subtended by that face at the center (O'Keeffe, 1979). The oxidation state and coordination number estimates are subsequently used to calculate

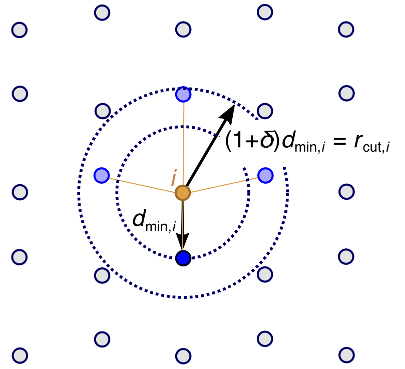

**FIGURE 2** | Atoms that are bonded (blue spheres) to a given central site $i$ (orange) are identified on the basis of the nearest-neighbor distance, $d_{\min,i}$ (absolute or dimensionless) and an *a priori* chosen tolerance, $\delta$, thus, defining a site-specific cutoff radius $r_{\mathrm{cut},i} = (1+\delta)d_{\min,i}$.

the atom radius on the basis of the ionic radius list provided by Shannon (1976). In the case when no oxidation states can be assigned, the estimator uses the atomic radii as provided by pymatgen (Ong et al., 2013).

For screening the equilibrium structures found in the Materials Project database, we will later use a single global tolerance, $\delta$, that will be optimized on the basis of a diverse structure test set. By contrast, both the interstitial finding and the ionic diffusion path characterization proceed with an increasing tolerance (from $\delta = 0.1$ in steps of 0.1 until a motif is found or $\delta = 0.8$ reached) to find neighbors in these (non-equilibrium) configurations.

Finally, note that the "min_VIRE" and "min_OKeeffe" methods may seem more reliable because they introduce well-established chemical properties (atomic/ionic radii and bond lengths, respectively) that are directly connected to the bonded atoms. But our optimization data suggest that the more *ad hoc* "min_dist" method performs in fact best.

## 2.2. Motif Assessment

Once all neighbors of a central atom $i$ are identified (**Figure 3**), the coordination pattern is evaluated. We use analytic order parameters (Steinhardt et al., 1983; Peters, 2009; Zimmermann et al., 2015) to perform the pattern recognition. The order parameters are typically designed in such a way to give a numerical value of 1 if the coordination pattern perfectly resembles the target motif and 0 if it is very different from the target motif.

### 2.2.1. Order Parameters

Order parameters (OPs), $q$, are mathematical constructs which aim to provide a numerical measure of the immediate local environment around an atom. The simplest OP, $q_{\mathrm{CN}}$, is a coordination number (CN (Sprik, 1998)),

$$q_{\mathrm{CN}} = \sum_{j \neq i} S\Big(\parallel \mathbf{p}_j - \mathbf{p}_i \parallel\Big), \tag{2}$$

obtained from counting neighbors within a cutoff radius $r_{\mathrm{cut},i}$ around a central atom $i$. $S$ is a weight that is 1 if an atom $j$ is within a distance $d_{i,j} < r_{\mathrm{cut},i}$ of atom $i$ and $S = 0$ otherwise. The next level of sophistication is a distance-weighted approach such as using the Fermi function (Sprik, 1998):

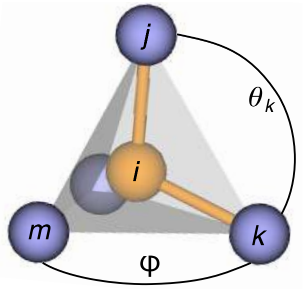

**FIGURE 3** | Definition of atom indices, $i$, $j$, $k$, and $m$, as well as angles, $\theta$ (polar) and $\varphi$ (azimuth), which are used in the computation of the order parameters introduced by Peters (2009) and Zimmermann et al. (2015).

$$q_{\mathrm{CN,Fermi}} = \frac{1}{\exp[\kappa(d_{i,j} - r_{\mathrm{cut},i})] + 1}, \tag{3}$$

where $\kappa^{-1}$ defines the transition width in which the contribution of an atom to the OP changes fastest to go from around 1 toward 0 as $d_{i,j}$ increases (Sprik, 1998). Note that neither of these approaches provides information about how closely an environment resembles a given structural motif.

Bond-orientational order parameters, introduced by Steinhardt et al. (1983), can, to some extent, be used to discern different structural motifs (Mickel et al., 2013). Thermal motion and other small distortions (e.g., caused by relaxation to the ground state of a compound from an ideal initial prototype structure) can, however, yield overlap between the OP distributions so that identification of the motif type becomes difficult. Hence, there is a need to more reliably identify structural motifs with order parameters.

Peters (2009) and Zimmermann et al. (2015) have introduced order parameters based on pattern-matching ideas put forward by Shetty et al. (2002). The OPs specifically recognize body-centered cubic-like (Peters, 2009) as well as tetrahedral and octahedral environments (Zimmermann et al., 2015). The pattern-matching ansatz places a reward on local environments that are similar to the target structure, thus, resulting in a value, $q_i$, of (close to) 1 for perfect resemblance. When the surrounding atoms are not in a configuration resembling the perfect prototype motif, penalties force the order parameter to attain values around zero. The pattern matching is achieved by setting up a spherical coordinate system around a central atom $i$ with a subset of neighboring atoms $j$ and $k$ (**Figure 3**). This allows the determination of the remaining neighbors' ($m, \ldots$) polar angles, $\theta$, and azimuth angles, $\varphi$. If a neighbor is not located at angles that are commensurate with the expected positions in the underlying structure motif, the decline in reward for being away from the perfect position follows a Gaussian function. Conversely, a full reward is given if any expected remaining position is exactly assumed. The procedure provides rotationally invariant order parameters because it is applied to each neighbor $j$ being used as the North pole position and any remaining neighbor $k$ for defining the prime meridian (cf., **Figure 3**); all possible combinations are then averaged. Below, we provide the definitions of the OPs that we use in this work.

The tetrahedral order parameter, $q_{\mathrm{tet}}$, is given by (Zimmermann et al., 2015):

$$q_{\mathrm{tet}} = \frac{1}{N_{\mathrm{ngh}}(N_{\mathrm{ngh}}-1)(N_{\mathrm{ngh}}-2)} \sum_{j\neq k}^{N_{\mathrm{ngh}}} \left\{ \exp\left[\frac{-(\theta_k - 109.47^\circ)^2}{2\,\Delta\theta^2}\right] \sum_{m\neq j,k}^{N_{\mathrm{ngh}}} \cos^2(1.5\,\varphi) \exp\left[\frac{-(\theta_m - 109.47^\circ)^2}{2\,\Delta\theta^2}\right] \right\}, \tag{4}$$

where $N_{\mathrm{ngh}}$ denotes the number of neighbors bonded to the central atom $i$ in a motif, $\theta_k$ is the polar angle formed between the bonds of neighboring atoms $j$ and $k$ with their mutually bonded atom $i$, $\varphi$ is the azimuth angle between bond $i-m$ with the plane spanned by $i$, $j$, and $k$, and $\Delta\theta = 12^\circ$ a parameter controlling the reward loss for increasingly non-ideal positions, which was optimized to distinguish tetrahedral and octahedral environments in NaCl wurtzite and conventional rocksalt (Zimmermann et al., 2015). Note that we use a slightly different variant of $q_{\mathrm{tet}}$ as in the original formulation [$\cos^2(1.5\,\varphi)$ instead of $\cos(3\,\varphi)$] to avoid negative values.

The octahedral order parameter, $q_{\mathrm{oct}}$, is given by (Zimmermann et al., 2015):

$$q_{\mathrm{oct}} = \frac{1}{N_{\mathrm{ngh}}[3+(N_{\mathrm{ngh}}-2)(N_{\mathrm{ngh}}-3)]} \left\{ \left[\sum_{j\neq k}^{N_{\mathrm{ngh}}} 3\,\mathrm{H}(\theta_k-\theta_{\mathrm{thr}}) \exp\left(\frac{-(\theta_k-180^\circ)^2}{2\,\Delta\theta_1^2}\right)\right] + \left[\sum_{m\neq j,k}^{N_{\mathrm{ngh}}} \mathrm{H}(\theta_{\mathrm{thr}}-\theta_k)\,\mathrm{H}(\theta_{\mathrm{thr}}-\theta_m) \cos^2(2\,\varphi) \exp\left(\frac{-(\theta_m-90^\circ)^2}{2\,\Delta\theta_2^2}\right)\right]\right]\right\}, \tag{5}$$

where $H(x)$ denotes the Heaviside function which is 1 if the argument $x > 0$ and 0 otherwise, $\Delta\theta_1 = 12^\circ$, $\Delta\theta_2 = 10^\circ$, and $\theta_{\mathrm{thr}}$ is a threshold angle to distinguish second neighbors that are considered to be either in a "South pole" configuration or in a "prime meridian" position; we set this threshold to 160°. Note that we change the definition of $q_{\mathrm{oct}}$ in a similar manner as we have done for $q_{\mathrm{tet}}$ to avoid negative values.

The body-centered cubic order parameter, $q_{\mathrm{bcc}}$, is given by (Peters, 2009):

$$q_{\mathrm{bcc}} = \frac{1}{N_{\mathrm{ngh}}[6+(N_{\mathrm{ngh}}-2)(N_{\mathrm{ngh}}-3)]} \sum_{j\neq k}^{N_{\mathrm{ngh}}} \left\{ 6\,\mathrm{H}(\theta_k-\theta_{\mathrm{thr}}) \exp\left(\frac{-(\theta_k-180^\circ)^2}{2\,\Delta\theta_1^2}\right) + \mathrm{H}(\theta_{\mathrm{thr}}-\theta_k) \sum_{m\neq j,k}^{N_{\mathrm{ngh}}} \cos(3\,\varphi)\; 1.6\,\frac{\theta_m-90^\circ}{\Delta\theta_2} \exp\left(\frac{-(\theta_m-90^\circ)^2}{2\,\Delta\theta_1^2}\right) \mathrm{sgn}(\theta_k-90^\circ)\right\}, \tag{6}$$

where $\Delta\theta_1 = 12^\circ$, $\Delta\theta_2 = 19.47^\circ$, and $\mathrm{sgn}(\theta)$ is the signum function which is $-1$ for $\theta < 0$, 0 if $\theta = 0$, and 1 if $\theta > 0$.

The mathematical definitions of the motif-specific order parameters $q_{\mathrm{tet}}$, $q_{\mathrm{oct}}$, and $q_{\mathrm{bcc}}$ in equations (4)–(6) follow a mutual recipe:

1. The innermost sum gives the contribution of how closely neighbor $m$ is located at its expected position with respect to polar angle, $\theta_m$, and azimuth angle, $\varphi$. For the azimuth angle, squared cosine functions are preferably used to pinpoint locations around a circle to ensure that the OP strictly gives positive values.
2. The outermost sum accounts for neighbor $k$'s match with the expected polar angle, $\theta_k$. The Gaussian functions are used with the polar angles to penalize deviations from expected positions.
3. The preceding factor normalizes the sums to give values between 0 and 1 based on combinatoric considerations.

For $q_{\mathrm{oct}}$ and $q_{\mathrm{bcc}}$, there is furthermore the need to distinguish whether or not neighbor $k$ is in (approximate) South pole position via the Heaviside function $\mathrm{H}(\theta_k - \theta_{\mathrm{thr}})$. And, $q_{\mathrm{bcc}}$ also requires incorporating the alternating pattern of subsequent neighbors $m$ being above and below the equator in bcc via the term $1.6 \times \mathrm{sgn}(\theta_k - 90^\circ) \times (\theta_m - 90^\circ)/(\Delta\theta_2)$.

Finally, we use the bond-orientational order parameters (Steinhardt et al., 1983) $q_4$ and $q_6$ for identifying close-packed motifs—fcc and hcp—(Ackland and Jones, 2006):

$$q_i = \frac{1}{N_{\mathrm{ngh}}} \sum_{j=1}^{N_{\mathrm{ngh}}} Y_{im}\Big(\theta(\mathbf{p}_j), \varphi(\mathbf{p}_j)\Big), \tag{7}$$

where $Y_{im}$ are the spherical harmonics of degree $i$. Note that (i) the angles $\theta$ and $\varphi$ are here with respect to a fixed frame of reference (Jungblut et al., 2013) and that (ii), while between 0 and 1, the values of the bond-orientational order parameters are typically not close to one for any motif, which is different to the behavior of $q_{\mathrm{tet}}$, $q_{\mathrm{oct}}$, and $q_{\mathrm{bcc}}$. Despite the fact that the bond-orientational order parameters $q_4$ and $q_6$ are frequently used in nucleation studies involving bcc, fcc, and hcp environments (ten Wolde et al., 1996; Peters, 2009; Jungblut et al., 2013; Limmer and Chandler, 2013), they are not highly reliable indicators to distinguish between all of these motifs when distortions (thermal noise) are introduced (Gasser et al., 2003). For this reason, we extensively use here those order parameters that were specifically designed for a given motif type (e.g., $q_{\mathrm{tet}}$ for tetrahedra).

### 2.2.2. Motif-Recognition Criteria

Motif recognition is typically achieved on the basis of a threshold approach (Peters, 2009; Zimmermann et al., 2015): if a given order parameter, $q_i$, is larger than an appropriate threshold, $q_{i,\mathrm{thr}}$, the coordination pattern is confirmed. Because of the design of these OPs, a threshold of 0.5 is often a reasonable *a priori* choice. The motif-specific order parameters $q_{\mathrm{tet}}$, $q_{\mathrm{oct}}$, and $q_{\mathrm{bcc}}$ should then be ideally usable as stand-alone identifiers for tetrahedral, octahedral, and bcc-like coordination environments, respectively. However, **Table 1** indicates that such an approach to defining criteria for motif recognition is not effective. For example, a site in a diamond structure would be identified as having both tetrahedral and bcc-like coordination; this is not surprising because the bcc motif can be viewed as two point-symmetric tetrahedra. Therefore, slightly more complex criteria must be developed to accurately distinguish structure motifs on the basis of order parameter values. We start with following set of criteria that allows us to identify all motifs separately and unambiguously for perfect prototype structures:

$$q_{\mathrm{tet}} > 0.5 \qquad \text{tetrahedral,} \tag{8}$$

$$q_{\mathrm{oct}} > 0.5 \qquad \text{octahedral,} \tag{9}$$

$$q_{\mathrm{bcc}} > 0.5 \quad \text{and} \quad q_{\mathrm{tet}} < 0.5 \qquad \text{bcc,} \tag{10}$$

$$q_6 > 0.5 \quad \text{and} \quad q_{\mathrm{tet}}, q_{\mathrm{oct}}, q_{\mathrm{bcc}} < 0.5 \qquad \text{fcc,} \tag{11}$$

$$q_6 < 0.5 \quad \text{and} \quad q_{\mathrm{tet}}, q_{\mathrm{oct}}, q_{\mathrm{bcc}} < 0.5 \qquad \text{hcp.} \tag{12}$$

The fourth-order bond-orientational order parameter $q_4$ is hence not necessary to identify all motifs. However, as we will explain later, small modifications to these criteria are needed to more accurately distinguish non-ideal motifs. In particular, we will merge the fcc and hcp criteria into a single close-packed rule ($q_6 > 0.4$ and $q_{\mathrm{tet}}$, $q_{\mathrm{oct}}$, $q_{\mathrm{bcc}} < 0.4$), and we will decrease the threshold of the tetrahedral order parameter to 0.3.

### 2.2.3. Validation

Validation is essential for reliable quantitative structure–property relationships (QSPRs) (Tropsha et al., 2003). We follow a three-step hierarchical approach for our structure-motif assessment on the basis of order parameters using perfect prototype structures. First, the responses of the tetrahedral and octahedral order parameters are measured when a single neighbor in the corresponding prototype structure is subjected to defined perturbations. We focus here on $q_{\mathrm{tet}}$ and $q_{\mathrm{oct}}$ because those motifs are particularly important throughout materials science and in corresponding design rules. Second, we randomly, but systematically, perturb the locations of sites in all prototype structures to mimic the effect of small distortions in equilibrium structures on all order parameters. This provides a more detailed insight into the sensitivity of order parameters to motif distortion, and it provides the necessary data for the next validation level. Third, we calculate motif-recognition likelihoods based on the histograms of the second validation level, which provides a first reasonable value of the tolerance, $\delta$, that is later used for neighbor finding in materials from a database.

**TABLE 1** | Coordination numbers and order parameter values for different prototype structures.

| | Diamond | Cubic | bcc | fcc | hcp |
|---|---|---|---|---|---|
| CN | 4 | 6 | 8 | 12 | 12 |
| $q_{\mathrm{tet}}$ | 1.0 | 0.014 | 0.143 | 0.030 | 0.030 |
| $q_{\mathrm{oct}}$ | 0.045 | 1.0 | 0.146 | 0.078 | 0.090 |
| $q_{\mathrm{bcc}}$ | 0.728 | 0.333 | 0.975 | 0.000 | −0.039 |
| $q_4$ | 0.509 | 0.764 | 0.509 | 0.191 | 0.097 |
| $q_6$ | 0.629 | 0.354 | 0.629 | 0.575 | 0.485 |

In **Figure 4**, we display the response of the tetrahedral (**Figure 4A**) and octahedral (**Figure 4B**) order parameter to perturbations of a single tagged neighbor in the respective perfect structure motif along the azimuth (orange) and polar angle (blue) as well as along an intermediate path (purple). The order parameters decline from 1 almost linearly to approximately 0.37 and 0.8 for polar and azimuth angular perturbations up to 27° and 18° for $q_{\mathrm{tet}}$ and $q_{\mathrm{oct}}$, respectively. The response along the intermediate perturbation path is similar. This indicates that the OPs can be used as (linear) measures of tetrahedrality and octahedrality for non-negligible perturbations of single neighbors, making the tetrahedral OP, for example, also attractive for structure analysis of liquid and solid water phases (Tao et al., 2017). **Figure 4** indicates that configurations with two atoms nearly overlapping score relatively high, which, however, does not pose a problem because those arrangements are very unlikely (i.e., two atoms will not be present at such close distances).

Next, we apply Gaussian-distributed random perturbations to all sites in the prototype structures via the polar form of Box–Muller transforms (Box and Muller, 1958) as implemented in numpy (van der Walt et al., 2011). The perturbations resemble the spatial distribution behavior of atoms in Einstein crystals (Einstein, 1906; Frenkel and Ladd, 1984; Aragones et al., 2012), the procedure of which we therefore call *Einstein crystal test rig*. Each lattice atom oscillates independently around its equilibrium position with a distribution width of $\sigma_{\mathrm{ES}}$. We refer the reader to the Supplementary Material (Section 1.1 in Supplementary Material) for more details on the Einstein crystal calculations.

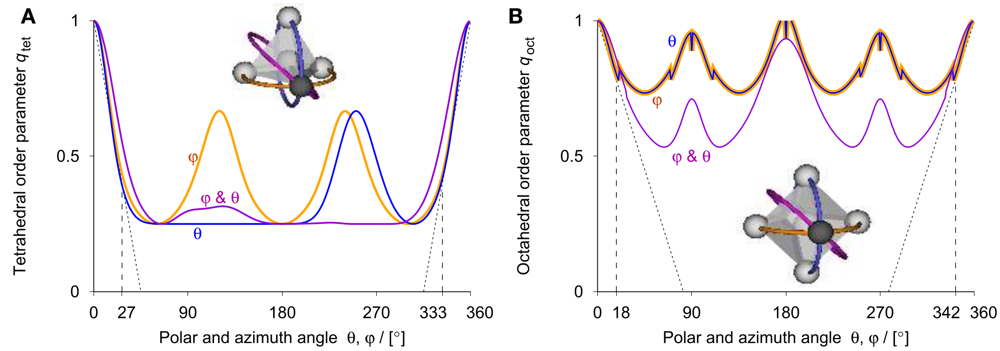

**FIGURE 4** | Responses of **(A)** tetrahedral, $q_{\mathrm{tet}}$, and **(B)** octahedral order parameter, $q_{\mathrm{oct}}$, to well-defined perturbations of a single neighbor in the corresponding perfect structure motif along the azimuth (orange) and polar (blue) angle as well as an intermediate path (purple). Peaks at intermediate angles indicate overlapping atoms.

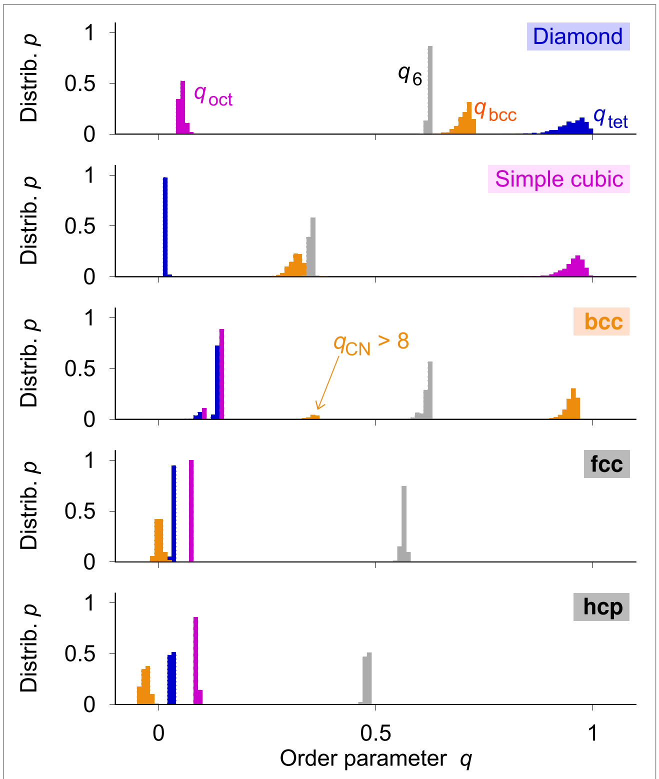

**FIGURE 5** | Relative frequency distribution, $p$, of all four order parameters $q_{tet}$, $q_{oct}$, $q_{bcc}$, and $q_6$ in all five prototype structures diamond, cubic, bcc, fcc, and hcp when the positions of atoms are subject to Einstein crystal-like perturbations away from their ideal location (here: Gaussian width of perturbations $\sigma_{ES}/d_{id} = 0.02$, and cutoff radius for neighbor finding relative to the ideal neighbor distance in the perfect prototype structure $r_{cut}/d_{id} = 1.1$). Note that the small intermediate peak at $q_{bcc} = 0.35$ obtained with the bcc prototype structure (central panel) is caused by configurations that have a non-ideal number of neighbors ($q_{CN} > 8$)—not by the order parameter definition itself.

**Figure 5** indicates that the OPs respond to Einsteinian perturbations of magnitude $\sigma_{ES}/d_{id} = 0.02$ and a relative cutoff radius of $r_{cut}/d_{id} = 1.1$ with relative frequency distributions of finite width. The distributions confirm the validity of both: (i) our proposed structure motif-recognition criteria and (ii) that $q_{tet}$ and $q_{oct}$ are in particular well-behaved measures for the degree of a given motif. The data, however, also suggest that the ability to distinguish between fcc and hcp on the basis of $q_6$ will be limited. Therefore, we redefine those structure motif criteria (equations (11) and (12)) to provide a single criterion for close-packed motifs (cp = fcc + hcp):

$$q_6 > 0.4 \quad \text{and} \quad q_{tet}, q_{oct}, q_{bcc} < 0.4 \quad \text{fcc} + \text{hcp}. \tag{13}$$

In this context, the approach by Honeycutt and Andersen (1987) is worthwhile noting, which compresses information about local environments—specifically, the number of shared near neighbors of a pair of atoms and the connectivity among the shared neighbors—into a four-integer index. This index can, however, not easily be used as an automatic motif recognition tool for hcp–fcc distinction because of two reasons. First, it requires visual inspection of the index-underlying graphs because the fourth integer is an arbitrary enumeration (cf., indices 1,421 and 1,422 occurring in fcc and hcp). Second, the index is computed for *pairs* of atoms so that the index itself cannot be used to directly characterize the entire coordination environment of a single atom. We are currently working toward solving the hcp–fcc distinction problem via definition of additional order parameters, resulting in an order parameter feature vector that might enable distinction between hcp and fcc.

To quantitatively assess our structure-motif recognition capabilities, we systematically expand the Einstein crystal sensitivity test approach to various distortion degrees, $\sigma_{ES}$, and relative cutoff radii, $r_{cut}/d_{id}$. For this purpose, we use following basic likelihood function (Sivia, 2012), $\mathcal{L}$:

$$\mathcal{L} = \prod_{i=1}^{N_{str}} \prod_{j=1}^{N_{mot}} \left[\delta_{i,j}^{Kr} N_{i,j} + (1 - \delta_{i,j}^{Kr})(1 - N_{i,j})\right] \Big/ N_{samp}, \tag{14}$$

where $N_{str}$ and $N_{mot}$ are the number of tested prototype structures and motifs, respectively (both are 4 here: tet, oct, bcc, and cp), $N_{i,j}$ is the number of times that structure $i$ is recognized as consisting of motif $j$ of all $N_{samp}$ samples for $i$ and $j$ (here: $N_{samp} = 1{,}000$), and $\delta_{i,j}^{Kr}$ is the Kronecker delta function, which is 1 if $i = j$ and 0 otherwise. Since we have merged fcc and hcp to one cp motif, we average data from fcc and hcp evaluations so that each of the four regrouped motifs (tet, oct, bcc, and cp) has an equal weight on $\mathcal{L}$. The likelihood function thus represents the joint probability to do both: to correctly identify true motifs and to correctly reject false motifs. Furthermore, the particular form of $\mathcal{L}$ as products of the separate motif and structure likelihoods more stringently requires that all motifs in all structures are correctly identified and rejected, respectively, to an acceptable degree. A mere arithmetical mean would favor compensation effects in which high recognition scores achieved with one motif and/or structure would balance entirely unsuccessful recognitions with low scores.

There exists a very localized optimal region of parameters where $\mathcal{L}$ is maximal (close to 1), which is found in the vicinity of small perturbations and small relative cutoff radii (**Figure 6A**). We expected a more broad distribution at vanishing degrees of distortion, which is, in fact, observable once the results from the bcc prototype structure are removed (**Figure 6B**). This is a reflection of the well-known issue that, in the bcc structure, second nearest neighbors are close to nearest neighbors (cf., Mickel et al. (2013) and **Figure 5**). As a result, the bcc issue sets certain limits to using a global tolerance for successful neighbor finding.

## 3. RESULTS

We now apply the validated order parameters based structure motif recognition criteria and the order parameters themselves (as a degree of perfect-motif resemblance) to automatically find structure motifs in a large materials database (Jain et al., 2013), determine interstitials (Broberg et al., 2016), and analyze the coordination environment along solid-state jump-diffusion paths (Rong et al., 2015).

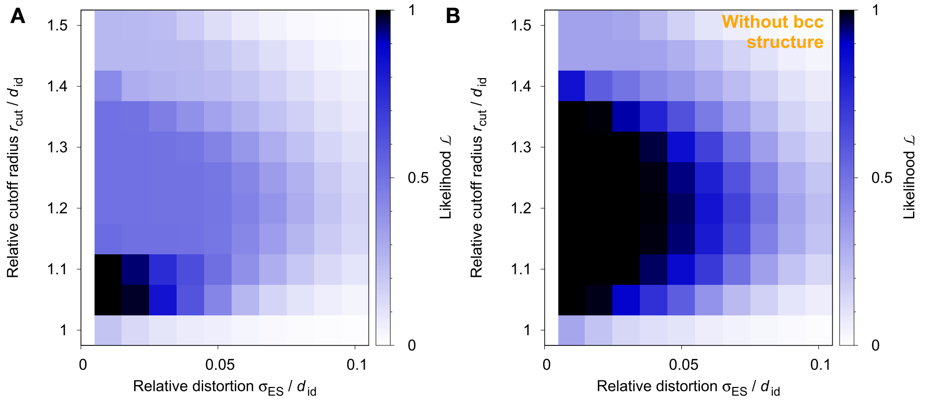

**FIGURE 6** | Likelihood, $\mathcal{L}$, to successfully identify the correct structural motif in any one of the five prototype structures (diamond, cubic, bcc, fcc, and hcp) with varying degree of relative distortion, $\sigma_{ES}/d_{id}$, and varying relative cutoff radius, $r_{cut}/d_{id}$, for finding bonded nearest neighbors: **(A)** all structures and **(B)** without bcc structure.

## 3.1. Motif Recognition in Materials Project Database

We aim to apply our structure-motif recognition approach on the entire Materials Project (Jain et al., 2013) database, which currently contains over 67,000 inorganic materials.[1] The bulk of Materials Project's (MP's) database has originated (Jain et al., 2013) from the Inorganic Crystal Structure Database (Bergerhoff et al., 1983; Belsky et al., 2002). Prior to dissemination, structures are relaxed with electronic density functional theory (Hohenberg and Kohn, 1964; Kohn and Sham, 1965), typically using the Vienna *ab initio* Simulation Package (VASP (Kresse and Hafner, 1993)) in the generalized gradient approximation (GGA) with $+\,U$ corrections for transition metal oxides.

### 3.1.1. Optimizing Neighbor Finding and Motif Criteria

Before turning to the results from the entire database, we determine the most suitable neighbor-finding method and the optimal tolerance parameter, $\delta$, given a reasonable test (sub)set from the MP database. Our initial test set, for which the number of different motif sites, $N_i$, is well known, consists of materials that are members of the structure groups listed in **Table 2**. Materials belonging to a given structure group were found by scanning the MP database for structures that are similar to a reference structure. Similarity between structures is hereby determined with a structure matcher algorithm implemented in pymatgen (Ong et al., 2013) using default parameters, except for turning off species matching (i.e., we match the frameworks of the structures). The references were ideal prototype structures in the case of unary materials (diamond-like, simple cubic, bcc, fcc, and hcp) and the canonical structure in the MP database for all other materials (zinc blende: mp-10695; rocksalt: mp-22862; CsCl-like: mp-22851; $MgAl_2O_4$-spinel: mp-3536), respectively. For the resulting 1,025 test structures, we calculate the order parameter values $q_{CN}$, $q_{tet}$, $q_{oct}$, $q_{bcc}$, and $q_6$ of all $N_{sites}$ sites in each structure. Then, we determine the numbers of different structure motifs ($N_{tet}$, $N_{oct}$, $N_{bcc}$, and $N_{cp}$) in each material on the basis of our recognition criteria (equations (8)–(10) and (13)) as well as the number of times that a site was assigned to more than a single structure motif ($N_{multi}$).

**TABLE 2** | Structure groups defining the initial test set.

| Structure group | Number of materials | Number and type of motifs |
|---|---|---|
| Diamond | 5 | $N_{tet} = N_{sites}$ |
| Simple cubic | 8 | $N_{oct} = N_{sites}$ |
| bcc | 48 | $N_{bcc} = N_{sites}$ |
| fcc | 58 | $N_{cp} = N_{sites}$ |
| hcp | 23 | $N_{cp} = N_{sites}$ |
| Zinc blende | 61 | $N_{tet} = N_{sites}$ |
| Rocksalt | 305 | $N_{oct} = N_{sites}$ |
| CsCl | 340 | $N_{bcc} = N_{sites}$ |
| Normal spinels<br>Hosseini (2008) | 177 | $N_{tet} = N_{sites}/7$<br>$N_{oct} = 2 \times N_{sites}/7$ |

The different neighbor-finding settings are compared by averaging the fraction, $p_{rec,i}$, of correctly recognized number of expected motifs $j$, in each structure $i$, $N_{j,i}$,

$$p_{rec,i} = N_{j,i}/N_{sites,i}, \tag{15}$$

overall 1,025 test structures. Note that we put a very stringent criterion on multiple assignments: if a structure contains a single site that is assigned to different motifs, the structure has no positive contribution on the recognition fraction (i.e., $p_{rec,i} = 0$). Furthermore, we address the problem that, in spinels, the target number of tetrahedral and octahedral sites, respectively, can be overpredicted by using following functional form for $p_{rec,i}^{spinel}$:

$$p_{rec,i}^{spinel} = 1 - \left|\frac{1}{3} - \frac{N_{tet,i}/3}{N_{sites,i}/7}\right| + \left|\frac{2}{3} - \frac{N_{oct,i}/3}{N_{sites,i}/7}\right|. \tag{16}$$

The results in **Figure 7** indicate that all three neighbor-finding approaches yield high recognition rates (>85%) for the chosen test set if the respective tolerance parameter is small enough ($\ll$0.1). Furthermore, the prediction quality decreases precipitously when the neighbor-finding tolerance, $\delta$, reaches a value

[1] Current numbers are available on the Materials Project website: http://materialsproject.org/.

of around 0.1 for any of the three methods. The best performing method ($\bar{p}_{\text{rec}} = 100\%$) is the minimum distance-based approach with $0.03 \leq \delta \leq 0.08$. For all following analyses, we use this method where we, however, employ a slightly larger tolerance ($\delta = 0.1$) because this tolerance yields a very low average coordination number of 3 for all sites in the entire Materials Project database.

Apart from being effective, the minimum distance-based neighbor-finding method is also computationally exceptionally efficient. Calculating all order parameters requires only 0.026 s per site on a compute node of NERSC's[2] Edison cluster (time was averaged over all sites in the MP database). That is, analyzing a structure with 100 sites should take 2–3 s, assuming the entire procedure roughly scales linearly with the number of sites. Note (i) that we decrease the tetrahedral OP threshold to 0.3 because tests on Jahn–Teller active structures suggest 0.5 being too strict and (ii) that we also require the coordination number, $q_{\text{CN}}$, to match the ideal motif value. Thus, the final set of rules that we use for motif determination across the entire Materials Project database are:

$$q_{\text{CN}} = 4 \quad \text{and} \quad q_{\text{tet}} > 0.3 \qquad \text{tetrahedral}, \tag{17}$$

$$q_{\text{CN}} = 6 \quad \text{and} \quad q_{\text{oct}} > 0.5 \qquad \text{octahedral}, \tag{18}$$

$$q_{\text{CN}} = 8 \quad \text{and} \quad q_{\text{bcc}} > 0.5 \text{ and } \quad q_{\text{tet}} < 0.5 \qquad \text{bcc}, \tag{19}$$

$$q_{\text{CN}} = 12 \quad \text{and} \quad q_6 > 0.4 \quad \text{and} \quad q_{\text{i}} < 0.4 \quad \text{fcc} + \text{hcp}. \tag{20}$$

In **Algorithm 1**, we also provide a pseudocode implementation of our motif recognition method.

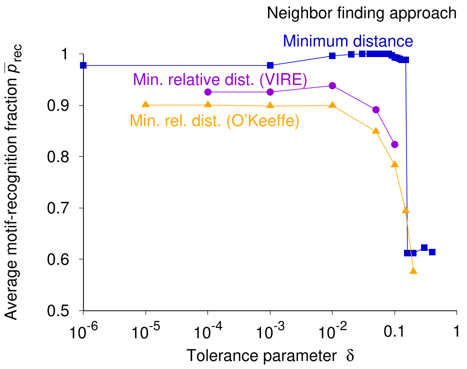

**FIGURE 7** | Average fractions of sites for which the expected motif types are correctly predicted, $\bar{p}_{\text{rec}}$, as functions of the tolerance parameter, $\delta$, involved in the respective neighbor-finding method. Three different neighbor-finding methods are tested: minimum distance (blue squares), minimum relative distance using a valence-ionic radius estimator (VIRE; purple circles) as implemented in pymatgen (Ong et al., 2013), and minimum relative distance using the bond-valence parameters according to O'Keeffe and Brese (1991), orange triangles.

### 3.1.2. Relative Motif Occurrence

The MP database screening results are presented in **Figure 8A** as relative motif occurrences for each element separately, and they agree with several previous observations (Cotton and Wilkinson, 1980; Brown, 1988). For example, Li occurs similarly often as a tetrahedral site as it occurs as an octahedral site, but bcc-like Li motifs are rare and close-packed ones are absent. These motif frequencies follow trends of Li for 4-, 6-, 8-, and 12-fold coordination that were already established by Brown (1988). Furthermore, we see comparably good agreement between motif occurrence and Brown's coordination statistics for Na, Ca, Ba, Mn, Fe, Co, B, As, and La. Examples with some differences but still overall favorable agreement include K, Sr, Pb, Ni, Cd, Sb, and Bi, whereas we obtain very different dominant motif vs coordination number distributions for Tl, Hg, Al, and In. In light of these differences, we note that our material set is far larger (and more diverse) than the set that Brown used; there is a 100-fold difference between the number of sites that we have analyzed (>15,000,000) in comparison to the previous study by Brown (1988) (14,000).

To continue the discussion of how our structure motif data relate to results from literature and common wisdom we note that (i) S and Se occur in tetrahedral coordination, (ii) S, Se, and Te can be seen in octahedra, but (iii) 8-fold coordination only occurs for Te and (iv) close-packed motifs of S, Se, and Te are not known (Cotton and Wilkinson, 1980). Silicon is known as a strong tetrahedron-former (e.g., in zeolites (Baerlocher et al., 2007)). This is confirmed when comparing both panels in **Figure 8**; **Figure 8B** gives the frequency at which a motif (here: tet) occurs relative to all motifs—recognizable and unrecognizable—with the same coordination number (here: 4). **Figure 8B** indicates that 86% of all 4-fold coordinated Si motifs are tetrahedra. Surprisingly,

**ALGORITHM 1** | Motif recognition.

```
Procedure GETMOTIFTYPE(structure, index)
    Let dmin be ← distance of closest neighbor to site with index in structure
    Let neighs be a new site list ← sites within radius of 1.1 dmin from site index
    Let ops be a new dictionary ← all order parameters obtained for site index and its neighs
    if ops["cn"] = 4 and ops["tet"] > 0.3 then
        return "tetrahedral"
    else if ops["cn"] = 6 and ops["oct"] > 0.5 then
        return "octahedral"
    else if ops["cn"] = 8 and ops["bcc"] > 0.5 and ops["tet"] < 0.5 then
        return "bcc"
    else if ops["cn"] = 12 and ops["q6"] > 0.4 and ops["tet"], ops["oct"], ops["bcc"] < 0.4 then
        return "closed packed"
    else
        return "unrecognized"
```

[2] NERSC: National Energy Research Scientific Computing Center (http://www.nersc.gov/).

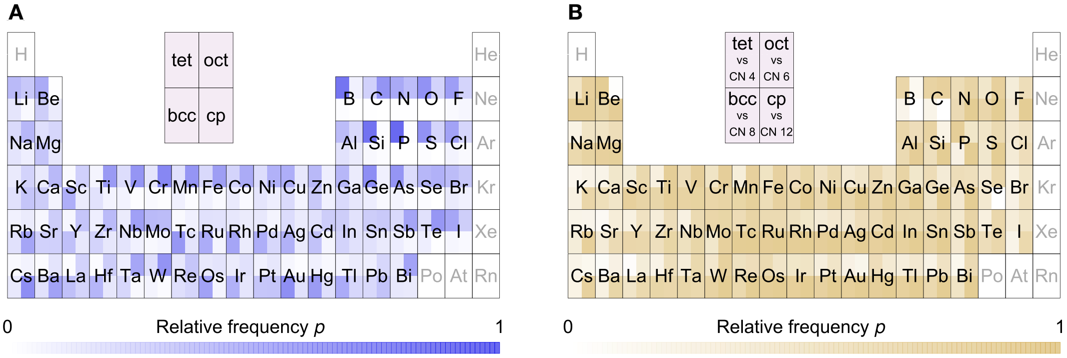

**FIGURE 8** | **(A)** Relative occurrences of motifs (tetrahedra, octahedra, bcc-like, and close packed) for each element in the Materials Project (MP) database; **(B)** relative occurrence of each recognizable motif (e.g., tet) to all local environments with the corresponding coordination number (e.g., 4). Note that each site contributes with a factor of $1/N_{\text{sites}}$, where $N_{\text{sites}}$ denotes the total number of sites (atoms) found in the primitive unit cell of the material in question (i.e., the frequencies are materials weighted, not site-weighted). Also, note that we filtered out 9% of the MP structures because we discarded noble gas-containing materials and materials that seem very unstable (energy above the convex hull of 0.2 eV per atom or more); in addition, we exclude the results for hydrogen sites. Section 1.2 in Supplementary Material lists how many sites of a given element were used for generating these figures.

the ratio at which lithium occurs in 4-fold coordination as a tetrahedron is low (47%). By contrast, the low tetrahedral coordination fraction of Ag and Cu motifs (both 41%) was expected because these elements are frequently found in square-planar environments.

An intriguing observation for us is that the known decline in tetrahedral site preference (Navrotsky and Kleppa, 1967; Burdett et al., 1982; Rong et al., 2015) as we go from Li over Mg to Ca in spinel is reflected in the MP database as a whole. The relative occurrence of Li as a tetrahedron among all 4 recognizable motifs is is 42%, Mg 20%, and Ca 13%. Zn typically takes a place between Li and Mg in spinels (Rong et al., 2015), which is, however, different in the MP database (46%).

The unexpected motif prevalences that are observed for some elements may hint at the fact that our neighbor finding method could benefit from further study and improvement. However, the overall approach represents a fast, fully automatic method to determine coordination environment motifs over large crystal databases that is relatively trustworthy and intuitive.

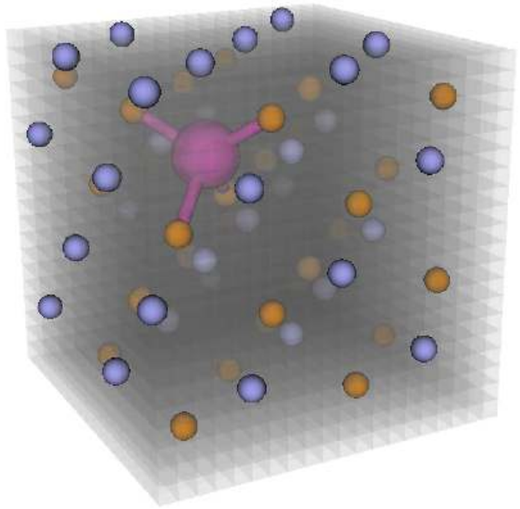

**FIGURE 9** | The Interstitialcy Finding Tool (InFiT) performs structure-motif recognition on a regular grid in the unit cell.

## 3.2. Interstitial Finding

Interstitials in dense inorganic materials are frequently found at sites that resemble basic structural motifs. Tetrahedral and octahedral coordination environments are particularly prevailing in isolated (i.e., non-complex) interstitials. This is evidenced by a series (Decoster et al., 2008, 2009a,b, 2010a,b, 2012; Pereira et al., 2011, 2012; Amorim et al., 2013; Silva et al., 2014) of $\beta^-$ emission channeling measurements (Hofsäss and Lindner, 1991; Silva et al., 2013) which were conducted at CERN's ISOLDE beamline. Those measurements inspired us to develop our Interstitialcy Finding Tool (InFiT), which is already used by the Python Charged Defect Tools (PyCDT)—a Python package for automatic setup and analysis of isolated charged defect calculations (Broberg et al., 2016).

The key idea of InFiT is to perform a systematic structure-motif search on a regular grid ($\Delta l \approx 0.2$ Å) that is spanned in the unit cell of a periodic material (**Figure 9**). For each point of the grid that is not closer than 1 Å to any crystal atom, the algorithm (Broberg et al., 2016) goes through following steps:

- Place an atom of the target interstitial type at this trial point.
- Perform a loop of increasing neighbor finding tolerance, $\delta$, starting from 0.1 in steps of 0.1 up to $\delta = 0.8$:
  - Get all neighbors and determine motif type.
  - If the motif type is recognized, consider the trial position for further evaluation, store the corresponding order parameter value, $q_i$, and stop the $\delta$-loop.

After a list of tentative interstitial sites is thus created, two pruning measures are taken. First, a distance-based clustering of the trial sites is performed. From each resulting motif-specific cluster, only one site is retained: the one with the highest order parameter value for the given motif type. Second, the (typically

**TABLE 3** | Interstitials found in primitive unit cells.

| Host | MP ID | Volume $V$/[Å$^3$] | Grid size $N_a \times N_b \times N_c$ | Initial interstitals $N_{\text{init}}^{\text{inter}}$ | After clustering $N_{\text{clust}}^{\text{inter}}$ | After symmetry pruning $N_{\text{sym}}^{\text{inter}}$ | Computation time[a] $t$/[s] |
|---|---|---|---|---|---|---|---|
| Ge | mp-32 | 47.9 | 20 × 20 × 20 | 2 | 2 | 1 | 90 |
| GaAs | mp-2534 | 47.5 | 20 × 20 × 20 | 20 | 2 | 2 | 96 |
| Fe | mp-13 | 11.6 | 12 × 12 × 12 | 156 | 9 | 2 | 35 |
| Cu | mp-30 | 11.8 | 12 × 12 × 12 | 120 | 3 | 2 | 59 |
| Ti | mp-46 | 34.7 | 14 × 14 × 23 | 270 | 9 | 5 | 72 |

*Details of the detected interstitial locations are presented in* ***Figure 10****.*

[a] *The computation time was determined on an Intel Core i7-4578U CPU at 3 GHz of a 2014 Macbook Pro.*

few) remaining sites are tested for (and pruned by) symmetrical equivalence.

There exists one symmetrically distinct tetrahedral interstitial site in diamond-like materials (Decoster et al., 2008), bcc (Seletskaia et al., 2005), fcc (Rosato et al., 1989), as well as hcp (Igarashi et al., 1991), where the latter three prototype structures also have each one symmetrically distinct octahedral site. Furthermore, 2 tetrahedral sites are (geometrically) possible in zinc blende-like materials, each one with different (host) atom coordination. Therefore, we use Er in diamond-like Ge (cf., Decoster et al. (2008)), Mn in sphalerite-like GaAs (cf., Pereira et al. (2011)), He in iron (bcc structure; cf., Seletskaia et al. (2005)), self-interstitials in Cu (fcc structure; cf., Rosato et al. (1989)), as well as O in $\alpha$-Ti (hcp structure; cf., Yu et al. (2015)) to test our interstitial finding tool InFiT. The results in **Table 3** and **Figure 10** indicate that, apart from hcp-Ti, we always find the correct number and type of sites. For hcp-Ti, we find 2 similarly looking tetrahedral interstitials, which only get identified as symmetrically equivalent when the corresponding symmetry threshold is increased significantly (i.e., to three times the largest distance between any two face-connected grid points). As for the octahedral interstitials in Ti, we find the expected interstitial with a usual value of the neighbor-finding threshold parameter ($\delta = 0.2$). However, we also find two very distorted octahedra with as large a $\delta$ as 0.8, the observation of which is used here to define a maximum reasonable value of $\delta$ (<0.8). A relative high neighbor-finding threshold (0.7) is desirable in some case, as the snapshots and complementary data for Fe in **Figure 10** highlight (cf., octahedral site). That example also underlines that a low tetrahedral OP threshold (0.3) can be necessary to find interstitials.

Finally, we highlight three additional points with respect to InFiT. First, **Table 3** shows the effectiveness of the clustering prune step, especially for the dense metals (from several 100 points to less than 20). Second, on an Intel Core i7-4578U CPU at 3 GHz, the interstitial finding took between 35 and 96 s for the five test systems (cf., **Table 3**), which translates into 0.011–0.034 s per grid point. And third, we successfully tested 19 additional diamond and sphalerite-like structures that are frequently investigated in the context of charged defects, and we always obtained the expected number of tetrahedral interstitials (1 and 2 for diamond and sphalerite-like materials, respectively; cf., Section 1.3 in Supplementary Material).

## 3.3. Diffusion Path Characterization

The ease of ion migration through an intercalant host can be often correlated with specific coordination environments—CN and pattern—(Rong et al., 2015). Since both tetrahedral and octahedral coordination play particularly important roles in predicting ion transport through promising new cathode materials for rechargeable batteries (Rong et al., 2015), we show the usefulness of the two order parameters $q_{\text{tet}}$ and $q_{\text{oct}}$ on the example of magnesium jump-diffusion hops in (empty) spinel manganese oxide (mvc-15009). The jump-diffusion path is obtained from a method that Zimmermann, Haranczyk, and co-workers are currently aiming to publish: the potential of electrostatics finite ion size (PfEFIS) method. It reliably estimates migration barriers deduced from nudged-elastic band (NEB; Mills et al. (1995)) calculations on the basis of electrostatic data (estimate SE around 50 meV). For our particular example, the estimate gives 711 meV (**Figure 11**; black circles and line), which agrees well with the NEB barrier (776 meV; cf., Rong et al. (2015)) while achieving a speed-up factor of 10,000. The diffusion path for Mg in spinel goes from a tetrahedral site over an intermediate octahedral site back to a tetrahedral site (Rong et al., 2015). The solid blue line in **Figure 11** indicates the change in $q_{\text{tet}}$ along the path, whereas the dotted orange line depicts the

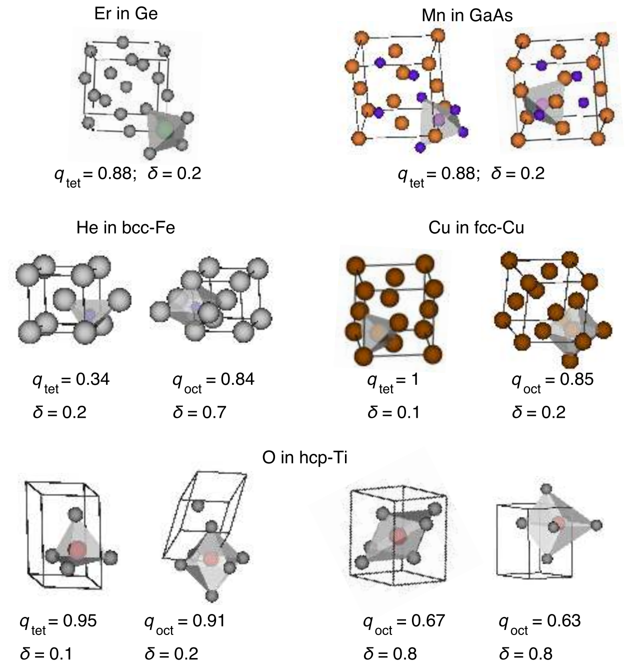

**FIGURE 10** | Interstitials found in five different materials with our tool InFiT (here: conventional unit cell for clarity).

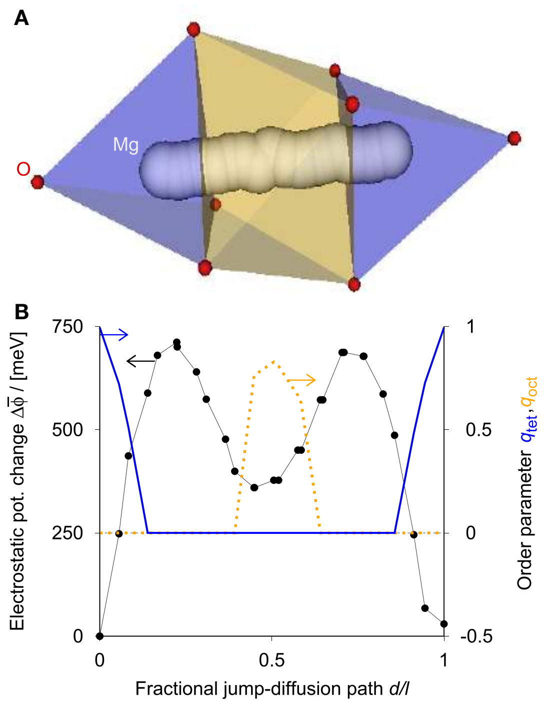

**FIGURE 11 | (A)** A magnesium ion (white spheres) hops in $Mn_2O_4$-spinel from a site at which it is coordinated by four oxygen atoms (red spheres) in a tetrahedral manner (blue polyhedron) through an intermediate O-octahedron (orange) back into a tetrahedral site. **(B)** Change in electrostatic potential, $\Delta\bar{\varphi}$ (left abscissa, black circles and line) and tetrahedral (blue solid line), $q_{\text{tet}}$, as well as octahedral order parameter (orange dotted line), $q_{\text{oct}}$ (both: right abscissa) as functions of the fractional jump-diffusion path, $d/l$. Note that the electrostatic potential is averaged over a spherical volume of radius 1.475 Å; Zimmermann and co-workers are currently aiming to publish the underlying method. Furthermore, our diffusion path characterization method is robust as the additional spinel examples in the Supplemental Material (Figure S2 in Supplementary Material) underscore.

change of the octahedral OP, $q_{\text{oct}}$. Clearly, the order parameters help visualize the change in coordination environment along the diffusion path—in a quantitative and physically meaningful (Wu et al., 2017) way.

## 4. CONCLUSION

We have shown here that order parameters (Steinhardt et al., 1983; Peters, 2009; Zimmermann et al., 2015), when paired with efficient and effective neighbor finding methods, can be reliably used as fast automatic structure-motif finding and coordination environment assessment tools, regardless of a material's chemistry. We introduced an effective validation framework—the Einstein crystal test rig—which subjects all atoms in a (prototype) structure to well-defined (random) distortions, thus, systematically sounding out the robustness of any motif recognition approach. We then applied our approach successfully to three important applications in (computational) materials science: (i) mapping the structural character of a materials database via element-specific relative structure-motif occurrence plots, (ii) effective interstitial finding (InFiT tool developed here; cf., Broberg et al. (2016)), and ion jump-diffusion path characterization (Rong et al., 2015). Our effective and efficient motif-recognition and assessment capabilities are freely available through the Python package pymatgen (Ong et al., 2013).[3] We ultimately emphasize that materials science is currently undergoing a "change of paradigm: from description to prediction" (Heine, 2014). Thus, we expect these tools to be useful in future machine-learning (Jain et al., 2016; Ward and Wolverton, 2017) applications as descriptors that capture much of the most basic—but essential (Wagner and Rondinelli, 2016)—information of a given material: the crystal structure.

## AUTHOR CONTRIBUTIONS

NZ, with continuous advice from both AJ and MH, developed the main concept of the study, conducted the calculations as well as the (majority of) tests and analyses, and prepared the manuscript. MKH performed additional tests of structure motif recognition. All authors discussed the results and implications and commented extensively on the manuscript at all stages.

## ACKNOWLEDGMENTS

We thank Alireza Faghaninia for helpful comments on the symmetry pruning implementation of InFiT.

## FUNDING

This work was intellectually led by the U.S. Department of Energy (DOE) Basic Energy Sciences (BES) program—the Materials Project—under Grant No. EDCBEE. MKH acknowledges support of BASF Corporation through the California Research Alliance. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DEAC02-05CH11231. Lawrence Berkeley National Laboratory is funded by the DOE under award DE-AC02-05CH11231.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/article/10.3389/fmats.2017.00034/full#supplementary-material.

[3] The pymatgen library is, for example, available via GitHub: https://github.com/materialsproject/pymatgen/.

## REFERENCES

Ackland, G. J., and Jones, A. P. (2006). Applications of local crystal structure measures in experiment and simulation. *Phys. Rev. B* 73, 054104. doi:10.1103/PhysRevB.73.054104

Amorim, L. M., Wahl, U., Pereira, L. M. C., Decoster, S., Silva, D. J., da Silva, M. R., et al. (2013). Precise lattice location of substitutional and interstitial Mg in AlN. *Appl. Phys. Lett.* 103, 262102. doi:10.1063/1.4858389

Aragones, J. L., Sanz, E., and Vega, C. (2012). Solubility of NaCl in water by molecular simulation revisited. *J. Chem. Phys.* 136, 244508. doi:10.1063/1.4728163

Baerlocher, C., McCusker, L. B., and Olsen, D. H. (2007). *Atlas of Zeolite Framework Types*, 6th Edn. Amsterdam, The Netherlands: Elsevier.

Belsky, A., Hellenbrandt, M., Karen, V. L., and Luksch, P. (2002). New developments in the inorganic crystal structure database (ICSD): accessibility in support of materials research and design. *Acta Cryst. B* 58, 364–369. doi:10.1107/S0108768102006948

Bergerhoff, G., Hundt, R., Sievers, R., and Brown, I. D. (1983). The inorganic crystal structure data base. *J. Chem. Inf. Comput. Sci.* 23, 66–69. doi:10.1021/ci00038a003

Box, G. E. P., and Muller, M. E. (1958). A note on the generation of random normal deviates. *Ann. Math. Stat.* 29, 610–611. doi:10.1214/aoms/1177706645

Broberg, D., Medasani, B., Zimmermann, N. E. R., Canning, A., Haranczyk, M., Asta, M., et al. (2016). PyCDT: A python toolkit for modeling point defects in semiconductors and insulators. *arXiv: 1611.07481*.

Brown, I. D. (1988). What factors determine cation coordination numbers? *Acta Cryst. B Struct. Sci.* 44, 545–553. doi:10.1107/S0108768188007712

Brunner, G. O. (1977). Definition of coordination and its relevance in structure types $AlB_2$ and NiAs. *Acta Cryst.* A33, 226–227. doi:10.1107/S0567739477000461

Burdett, J. K., Price, G. D., and Price, S. L. (1982). Role of the crystal-field theory in determining the structures of spinels. *J. Am. Chem. Soc.* 104, 92–95. doi:10.1021/ja00365a019

Cotton, F. A., and Wilkinson, G. (1980). *Advance Inorganic Chemistry – A Comprehensive Text*, 4 Edn. New York, USA: John Wiley & Sons.

Decoster, S., Cottenier, S., De Vries, B., Emmerich, H., Wahl, U., Correia, J. G., et al. (2009a). Transition metal impurities on the bond-centered site in germanium. *Phys. Rev. Lett.* 102, 065502. doi:10.1103/PhysRevLett.102.065502

Decoster, S., De Vries, B., Wahl, U., Correia, J. G., and Vantomme, A. (2009b). Lattice location study of implanted In in Ge. *J. Appl. Phys.* 105, 083522. doi:10.1063/1.3110104

Decoster, S., Cottenier, S., Wahl, U., Correia, J. G., Pereira, L. M. C., Lacasta, C., et al. (2010a). Diluted manganese on the bond-centered site in germanium. *Appl. Phys. Lett.* 97, 151914. doi:10.1063/1.3501123

Decoster, S., Cottenier, S., Wahl, U., Correia, J. G., and Vantomme, A. (2010b). Lattice location study of ion implanted Sn and Sn-related defects in Ge. *Phys. Rev. B* 81, 155204. doi:10.1103/PhysRevB.81.155204

Decoster, S., De Vries, B., Wahl, U., Correia, J. G., and Vantomme, A. (2008). Experimental evidence of tetrahedral interstitial and bond-centered Er in Ge. *Appl. Phys. Lett.* 93, 141907. doi:10.1063/1.2996280

Decoster, S., Wahl, U., Cottenier, S., Correia, J. G., Mendonça, T., Amorim, L. M., et al. (2012). Lattice position and thermal stability of diluted As in Ge. *J. Appl. Phys.* 111, 053528. doi:10.1063/1.3692761

Dirichlet, G. L. (1850). Über die Reduction der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen. *J. Reine Angew. Math.* 40, 209–227. doi:10.1515/crll.1850.40.209

Einstein, A. (1906). Die Plancksche Theorie der Strahlung und die Theorie der spezifischen Wärme. *Ann. Phys.* 22, 180–190. doi:10.1002/andp.19063270110

Frenkel, D., and Ladd, A. J. C. (1984). New Monte Carlo method to compute the free energy of arbitrary solids. Application to the fcc and hcp phases of hard spheres. *J. Chem. Phys.* 81, 3188–3193. doi:10.1063/1.448024

Gasser, U., Schofield, A., and Weitz, D. A. (2003). Local order in a supercooled colloidal fluid observed by confocal microscopy. *J. Phys. Condens. Matter* 15, S375–S380. doi:10.1088/0953-8984/15/1/351

Heine, T. (2014). Grand challenges in computational materials science: from description to prediction at all scales. *Front. Mater.* 1:7. doi:10.3389/fmats.2014.00007

Hofsäss, H., and Lindner, G. (1991). Emission channeling and blocking. *Phys. Rep.* 201, 121–183. doi:10.1016/0370-1573(91)90121-2

Hohenberg, P., and Kohn, W. (1964). Inhomogeneous electron gas. *Phys. Rev. B* 136, B864–B871. doi:10.1103/PhysRev.136.B864

Honeycutt, J. D., and Andersen, H. C. (1987). Molecular dynamics study of melting and freezing of small Lennard–Jones clusters. *J. Phys. Chem.* 91, 4950–4963. doi:10.1021/j100303a014

Hoppe, R. (1979). Effective coordination numbers (ECoN) and mean fictive ionic radii (MEFIR). *Z. Kristallogr.* 150, 23–52. doi:10.1524/zkri.1979.150.1-4.23

Hosseini, S. M. (2008). Structural, electronic and optical properties of spinel $MgAl_2O_4$ oxide. *Phys. Stat. Sol. B* 245, 2800–2807. doi:10.1002/pssb.200844142

Igarashi, M., Khantha, M., and Vitek, V. (1991). *N*-body interatomic potentials for hexagonal close-packed metals. *Philos. Mag. B* 63, 603–627. doi:10.1080/13642819108225975

Isayev, O., Fourches, D., Muratov, E. N., Oses, C., Rasch, K., Tropsha, A., et al. (2015). Materials cartography: representing and mining materials space using structural and electronic fingerprints. *Chem. Mater.* 27, 735–743. doi:10.1021/cm503507h

Jain, A., Hautier, G., Ong, S. P., and Persson, K. (2016). New opportunities for materials informatics: resources and data mining techniques for uncovering hidden relationships. *J. Mater. Res.* 31, 977–994. doi:10.1557/jmr.2016.80

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., et al. (2013). Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* 1, 011002. doi:10.1063/1.4812323

Jungblut, S., Singraber, A., and Dellago, C. (2013). Optimising reaction coordinates for crystallisation by tuning the crystallinity definition. *Mol. Phys.* 111, 3527–3533. doi:10.1080/00268976.2013.832820

Kohn, W., and Sham, L. J. (1965). Self-consistent equations including exchange and correlation effects. *Phys. Rev. A* 140, A1133–A1138. doi:10.1103/PhysRev.140.A1133

Kresse, G., and Hafner, J. (1993). *Ab initio* molecular dynamics for liquid metals. *Phys. Rev. B* 47, 558–561. doi:10.1103/PhysRevB.47.558

Li, X., Xu, Y., and Wang, C. (2009). Suppression of Jahn–Teller distortion of spinel $LiMn_2O_4$ cathode. *J. Alloys Compd.* 479, 310–313. doi:10.1016/j.jallcom.2008.12.081

Li, Y., Yu, J., and Xu, R. (2013). Criteria for zeolite frameworks realizable for target synthesis. *Angew. Chem. Int. Ed.* 52, 1673–1677. doi:10.1002/anie.201206340

Limmer, D. T., and Chandler, D. (2013). The putative liquid-liquid transition is a liquid-solid transition in atomistic models of water. II. *J. Chem. Phys.* 138, 214504. doi:10.1063/1.4807479

Mazur, M., Wheatley, P. S., Navarro, M., Roth, W. J., Položij, M., Mayoral, A., et al. (2015). Synthesis of 'unfeasible' zeolites. *Nat. Chem.* 8, 58–62. doi:10.1038/NCHEM.2374

Mickel, W., Kapfer, S. C., Schröder-Turk, G. E., and Mecke, K. (2013). Shortcomings of the bond orientational order parameters for the analysis of disordered particulate matter. *J. Chem. Phys.* 138, 044501. doi:10.1063/1.4774084

Mills, G., Jónsson, H., and Schenter, G. K. (1995). Reversible work transition state theory: application to dissociative adsorption of hydrogen. *Surf. Sci.* 324, 305–337. doi:10.1016/0039-6028(94)00731-4

Morris, J. W. Jr. (2007). *A Survey of Materials Science I. Structure*. Berkeley, USA: University of California.

Navrotsky, A., and Kleppa, O. J. (1967). Thermodynamics of cation distributions in simple spinels. *J. Inorg. Nucl. Chem.* 29, 2701–2714. doi:10.1016/0022-1902(67)80008-3

O'Keeffe, M. (1979). Proposed rigorous definition of coordination number. *Acta Cryst. A* 35, 772–775. doi:10.1107/S0567739479001765

O'Keeffe, M., and Brese, N. E. (1991). Atom sizes and bond lengths in molecules and crystals. *J. Am. Chem. Soc.* 113, 3226–3229. doi:10.1021/ja00009a002

Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., et al. (2013). Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* 68, 314–319. doi:10.1016/j.commatsci.2012.10.028

Pereira, L. M. C., Wahl, U., Decoster, S., Correia, J. G., Amorim, L. M., da Silva, M. R., et al. (2012). Stability and diffusion of interstitial and substitutional Mn in GaAs of different doping types. *Phys. Rev. B* 86, 125206. doi:10.1103/PhysRevB.86.125206

Pereira, L. M. C., Wahl, U., Decoster, S., Correia, J. G., da Silva, M. R., Vantomme, A., et al. (2011). Direct identification of interstitial Mn in heavily *p*-type doped GaAs and evidence of its high thermal stability. *Appl. Phys. Lett.* 98, 201905. doi:10.1063/1.3592568

Peters, B. (2009). Competing nucleation pathways in a mixture of oppositely charged colloids: out-of-equilibrium nucleation revisited. *J. Chem. Phys.* 131, 244103. doi:10.1063/1.3271024

Rong, Z., Malik, R., Canepa, P., Gautam, G. S., Liu, M., Jain, A., et al. (2015). Materials design rules for multivalent ion mobility in intercalation structures. *Chem. Mater.* 27, 6016–6021. doi:10.1021/acs.chemmater.5b02342

Rosato, V., Guillope, M., and Legrand, B. (1989). Thermodynamical and structural properties of f.c.c. transition metals using a simple tight-binding model. *Philos. Mag. A* 59, 321–336. doi:10.1080/01418618908205062

Sands, D. E. (1993). *Introduction to Crystallography*. Mineola, USA: Dover Publications.

Santiso, E. E., and Trout, B. L. (2011). A general set of order parameters for molecular crystals. *J. Chem. Phys.* 134, 064109. doi:10.1063/1.3548889

Seletskaia, T., Osetsky, Y., Stoller, R. E., and Stocks, G. M. (2005). Magnetic interactions influence the properties of helium defects in iron. *Phys. Rev. Lett.* 94, 046403. doi:10.1103/PhysRevLett.94.046403

Shannon, R. D. (1976). Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Cryst.* A32, 751–767. doi:10.1107/S0567739476001551

Shetty, R., Escobedo, F. A., Choudhary, D., and Clancy, P. (2002). A novel algorithm for characterization of order in materials. *J. Chem. Phys.* 117, 4000–4009. doi:10.1063/1.1494986

Silva, D. J., Wahl, U., Correia, J. G., Pereira, L. M. C., Amorim, L. M., da Silva, M. R., et al. (2014). Lattice location and thermal stability of implanted nickel in silicon studied by on-line emission channeling. *J. Appl. Phys.* 115, 023504. doi:10.1063/1.4861142

Silva, M. R., Wahl, U., Correia, J. G., Amorim, L. M., and Pereira, L. M. C. (2013). A versatile apparatus for on-line emission channeling experiments. *Rev. Sci. Instrum.* 84, 073506. doi:10.1063/1.4813266

Sivia, D. S. (2012). *Data Anlaysis – A Bayesian Tutorial*, 2 Edn. Oxford, UK: Oxford Science Publishing.

Sprik, M. (1998). Coordination numbers as reaction coordinates in constrained molecular dynamics. *Faraday Discuss.* 110, 437–445. doi:10.1039/a801517a

Steinhardt, P. J., Nelson, D. R., and Ronchetti, M. (1983). Bond-orientational order in liquids and glasses. *Phys. Rev. B* 28, 784–805. doi:10.1103/PhysRevB.28.784

Tao, Y., Zou, W., Jia, J., Li, W., and Cremer, D. (2017). Different ways of hydrogen bonding in water – why does warm water freeze faster than cold water? *J. Chem. Theory Comput.* 13, 55–76. doi:10.1021/acs.jctc.6b00735

ten Wolde, P. R., Ruiz-Montero, M. J., and Frenkel, D. (1996). Numerical calculation of the rate of crystal nucleation in a Lennard–Jones system at moderate undercooling. *J. Chem. Phys.* 104, 9932–9947. doi:10.1063/1.471721

Tropsha, A., Gramatica, P., and Gombar, V. K. (2003). The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* 22, 69–77. doi:10.1002/qsar.200390007

van der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* 13, 22–30. doi:10.1109/MCSE.2011.37

Voronoi, G. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Sur quelques propriétés des formes quadratiques positives parfaites. *J. Reine Angew. Math.* 133, 97–178.

Wagner, N., and Rondinelli, J. M. (2016). Theory-guided machine learning in materials science. *Front. Mater.* 3:28. doi:10.3389/fmats.2016.00028

Wang, Y., Richards, W. D., Ong, S. P., Miara, L. J., Kim, J. C., Mo, Y., et al. (2015). Design principles for solid-state lithium superionic conductors. *Nat. Mater.* 14, 1026–1031. doi:10.1038/NMAT4369

Ward, L., and Wolverton, C. (2017). Atomistic calculations and materials informatics: a review. *Curr. Opin. Solid State Mater. Sci.* 21, 167–176. doi:10.1016/j.cossms.2016.07.002

Werner, A. (1912). Zur Kenntnis des asymmetrischen Kobaltatoms. *V. Ber. Deu. Chem. Gesell.* 45, 121–130. doi:10.1002/cber.19120450116

Wu, H., Lorenson, A., Anderson, B., Witteman, L., Wu, H., Meredig, B., et al. (2017). Robust fcc solute diffusion predictions from ab-initio machine learning methods. *Comput. Mater. Sci.* 134, 160–165. doi:10.1016/j.commatsci.2017.03.052

Yu, Q., Qi, L., Tsuru, T., Traylor, R., Rugg, D., Morris, J. W. Jr., et al. (2015). Origin of dramatic oxygen solute strengthening effect in titanium. *Science* 347, 635–639. doi:10.1126/science.1260485

Zimmermann, N. E. R., and Haranczyk, M. (2016). History and utility of zeolite framework-type discovery from a data-science perspective. *Cryst. Growth Des.* 16, 3043–3048. doi:10.1021/acs.cgd.6b00272

Zimmermann, N. E. R., Vorselaars, B., Quigley, D., and Peters, B. (2015). Nucleation of NaCl from aqueous solution: critical sizes, ion-attachment kinetics, and rates. *J. Am. Chem. Soc.* 137, 13352–13361. doi:10.1021/jacs.5b08098

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, JJ, and handling Editor declared their shared affiliation.