


Cite this: *RSC Adv.*, 2020, **10**, 6063

Received 24th September 2019  
Accepted 20th January 2020

DOI: 10.1039/c9ra07755c  
rsc.li/rsc-advances

## Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity†

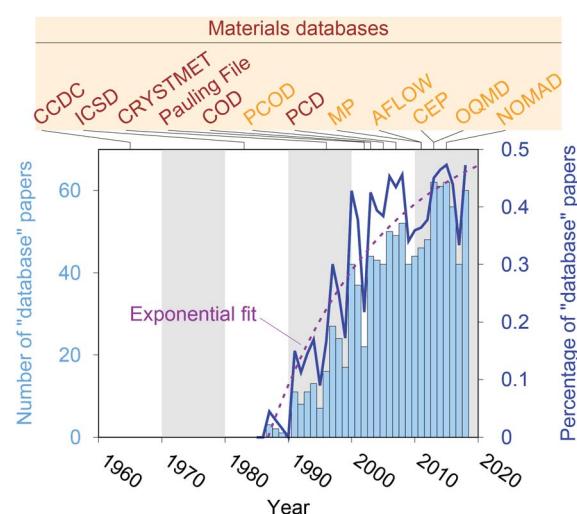
Nils E. R. Zimmermann  \*<sup>a</sup> and Anubhav Jain<sup>b</sup>

Structure characterization and classification is frequently based on local environment information of all or selected atomic sites in the crystal structure. Therefore, reliable and robust procedures to find coordinated neighbors and to evaluate the resulting coordination pattern (e.g., tetrahedral, square planar) are critically important for both traditional and machine learning approaches that aim to exploit site or structure information for predicting materials properties. Here, we introduce new local structure order parameters (LoStOPs) that are specifically designed to rapidly detect highly symmetric local coordination environments (e.g., Platonic solids such as a tetrahedron or an octahedron) as well as less symmetric ones (e.g., Johnson solids such as a square pyramid). Furthermore, we introduce a Monte Carlo optimization approach to ensure that the different LoStOPs are comparable with each other. We then apply the new local environment descriptors to define site and structure fingerprints and to measure similarity between 61 known coordination environments and 40 commonly studied crystal structures, respectively. After extensive testing and optimization, we determine the most accurate structure similarity assessment procedure to compute all 2.45 billion structure similarities between each pair of the  $\approx 70\,000$  materials that are currently present in the Materials Project database.

## 1 Introduction

Crystal structure databases<sup>1–16</sup> play an increasingly important role in materials science, chemistry, and related fields. Publication statistics gathered from Web of Science<sup>17</sup> on November 23, 2019, indicate that this trend started in the early 1990s (Fig. 1) and that the underlying potential is still not exhausted. The steady increase is (most likely) linked to continuously increasing computing power and memory storage, and it has fostered the creation of many different crystallographic databases that catalog existing materials such as the Cambridge Crystallographic Data Centre (CCDC) in 1965,<sup>1</sup> the Inorganic Crystal Structure Database (ICSD) in 1983,<sup>2,3</sup> CRYSTMET in 1993,<sup>4,5</sup> Pauling File in 2002,<sup>6,7</sup> the Crystallography Open Database (COD) in 2003<sup>8</sup> (previously known as “The American Mineralogist crystal structure database”<sup>9</sup>), and the Pearson’s Crystal Data (PCD) in 2007.<sup>10</sup> Furthermore, computational databases, which mainly use crystallographic databases as a source, and fully hypothetical structure databases are currently being created more and more: the Predicted

Crystallography Open Database (PCOD) appeared in 2005, the Materials Project (MP) database<sup>11,12</sup> and the Automatic FLOW for Materials Discovery database (AFLOW) in 2011,<sup>13</sup> the



**Fig. 1** Publication statistics for research articles that include “database” as keyword in the topic field within the Web of Science™ Core Collection for category “Chemistry, Inorganic & Nuclear.” Data was retrieved on November 23, 2019. Furthermore, we highlight the inception years of established materials databases.

<sup>a</sup>Energy Technology Area, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. E-mail: nils.e.r.zimmermann@gmail.com; Tel: +49 177 9077 532

<sup>b</sup>Energy Technology Area, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c9ra07755c

Harvard Clean Energy Project (CEP)<sup>14</sup> and the Open Quantum Materials Database (OQMD) in 2013,<sup>15</sup> as well as the Novel Materials Discovery Laboratory (NOMAD) in 2015.<sup>16</sup>

As computational resources still continue to grow<sup>18</sup> and to become more omnipresent and accessible, the computational chemistry, physics, and materials science communities have focused their efforts more and more on automation tools for materials database analysis and on employing statistical and machine learning (SML)<sup>19,20</sup> to help expedite materials discoveries and chemical innovations.<sup>21</sup> This includes, for example, predicting properties (e.g., formation energies, crystal structure dimensionalities, phase diagrams, band gaps, elastic moduli, ionic conductivity) of diverse materials from classes and families such as AX binary compounds,<sup>22</sup> M<sub>2</sub>AX ternary phases,<sup>23</sup> delafossite and related layered phases of composition ABX<sub>2</sub>,<sup>24</sup> conventional<sup>25</sup> and double perovskite halides (or elpasolites),<sup>26</sup> zeolites<sup>27,28</sup> and other silicates,<sup>29</sup> and other inorganic materials<sup>30–36</sup> as well as polymers;<sup>37</sup> indicating possible synthesis approaches by screening and predicting synthesis parameters and reactions of inorganic materials,<sup>38,39</sup> metal–organic frameworks,<sup>40</sup> and organic molecules;<sup>41,42</sup> generating interatomic potentials;<sup>43–46</sup> and expediting *ab initio*<sup>47–50</sup> calculations.<sup>51–55</sup>

Another important scientific problem is, in this context, the classification and categorization of entire crystal structures and the assessment of similarity between two materials (Fig. 2A).<sup>56–58</sup> Conventionally, crystal structures are characterized by their chemistry, crystal system, and space group. Other schemes employ the coordination number and pattern of the constituting atomic sites.<sup>59</sup> Because of the plethora of ways to classify structures, defining and automatically finding prototype structures is currently a very active research area.<sup>60–62</sup> In particular, the usage of coordination number and pattern has culminated in a larger current effort of the community to leverage fingerprinting.<sup>58,63–66</sup> This is the process of combining crucial information about the structure and/or its constituting local environments around each atom into a vector that represents the structure as a whole and includes, for example: a two-dimensional fingerprint based on simulated diffraction patterns;<sup>57</sup> the Coulomb matrix;<sup>67</sup> a many-body tensor representation;<sup>68</sup> deep tensor neural networks;<sup>45</sup> Voronoi tessellation;<sup>69</sup> radial distribution functions with<sup>70</sup> and without<sup>71</sup> incorporating partial atomic charges; and local environment-based crystal fingerprints.<sup>72</sup>

Crystal structure classification approaches that are specifically based on coordination information of the constituent atomic sites have the advantage that, apart from globally classifying the structure, they carry easily interpretable local information. This facilitates to ensure causality between the descriptor–property relation and the underlying mechanism,<sup>73</sup> when using the corresponding local coordination descriptors for design rules or machine learning applications. A critically important ingredient is then the effectiveness of and comparability across coordination site descriptors (Fig. 2B) and the resulting site fingerprints that characterize the coordination environment.

To address this we introduce here (i) a new neighbor finding method, (ii) several new local structure order parameters

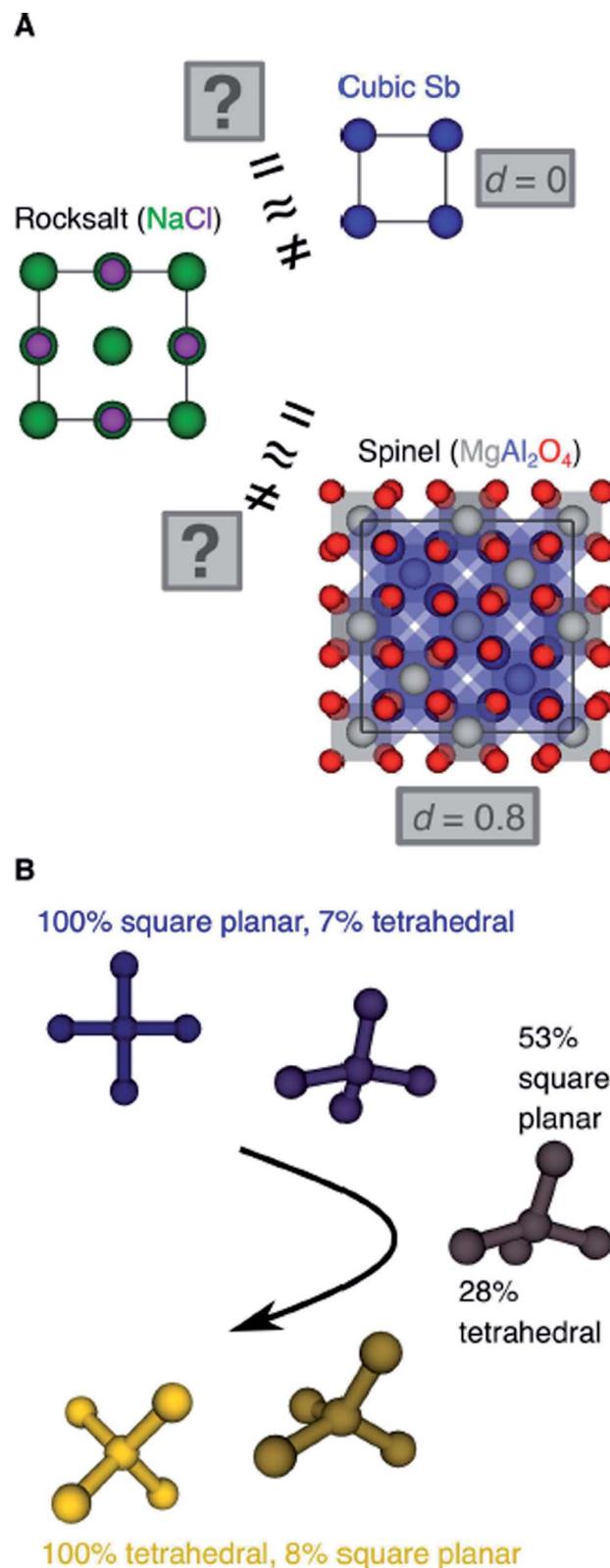


Fig. 2 (A) Is rocksalt structurally (more) similar to cubic Sb or to MgAl<sub>2</sub>O<sub>4</sub>–spinel? (B) Can we create descriptors which measure the degrees of resembling perfect coordination environments in a comparable way?

(LoStOPs), which are optimized with a tailor-made Monte Carlo procedure, as well as (iii) new site fingerprint and (iv) new structure fingerprints (Section 2), which are freely available through the python<sup>74</sup> packages pymatgen<sup>75,76</sup> and matminer.<sup>65,77</sup> We extensively test and optimize the new tools on known coordination environments and commonly investigated crystalline prototype materials with the main purpose of assessing site and structure similarity (Section 3).

## 2 Methodology

In this section, we introduce our methods for finding neighbors of a given central site in a crystal structure,<sup>78–80</sup> for performing pattern matching on the resulting coordination environment,<sup>81,82</sup> and for using that information to generate fingerprints that aim to characterize coordination environments and, ultimately, entire crystal structures on the basis of the coordination descriptors. Furthermore, we optimize our novel coordination environment descriptors [local structural order parameters (LoStOPs)] to improve inter-motif comparability,<sup>83</sup> and we describe the similarity measures that we test for comparing site and structure fingerprints. All of the methods and algorithms listed here are freely available through the local\_env module in pymatgen<sup>75,76</sup> and the featurizers module in matminer.<sup>65,77</sup>

### 2.1 Neighbor finding approaches

Two general and very popular neighbor finding approaches are employed in this work, which use (1) distance<sup>80</sup> and (2) topology-based,<sup>78,79</sup> information, respectively, to decide which neighbors are included in the near neighbor list. In both cases, an initial tentative neighbor list is constructed with a large hard cutoff radius (typically: 7–10 Ångström), which can, however, be dynamically increased, as we explain below. The methods are therefore approaches how to prune this initial (long) neighbor list.

The first approach, which we call “minimum distance” neighbor finding (MDNF), consists of 3 basic steps (Fig. 3A):

- Find neighbor  $k$  that has the smallest distance,  $d_{\min} = d_k = \min\{d_j\}$ , to (central) site  $i$ , given all distances,  $\{d_j\}$ , from the tentative neighbor list.

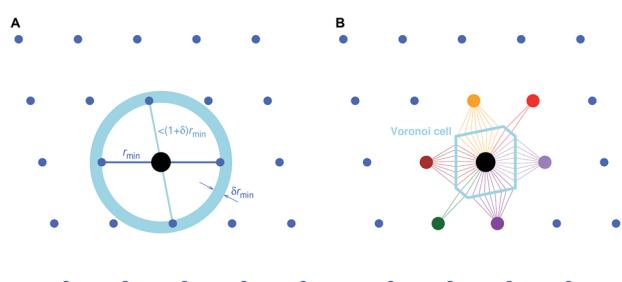


Fig. 3 Basic neighbor finding approaches used and investigated in this work: (A) minimum distance<sup>80</sup> and (B) Voronoi decomposition-based<sup>78,79</sup> near neighbor finding.

- Divide all distances,  $\{d_j\}$ , by  $d_{\min}$ , thus, yielding fractional distances,  $\{\tilde{d}_j\} = \{d_j/d_{\min}\}$ .

- Include neighbors that are at most 1.1 fractional distances away from the central site<sup>80</sup> (i.e.,  $\tilde{d} \leq 1.1$ ).

We choose the label “minimum distance” over “relative distance” for this neighbor finding approach in order to avoid confusion with similar methods<sup>80</sup> that use bond lengths<sup>84</sup> and/or atomic/ionic radii<sup>85</sup> to compute dimensionless (relative) neighbor distances.

The second approach (VNF) uses Voronoi decomposition<sup>78,79</sup> to identify neighbors from the tentative list by employing the solid angle as a weighting measure.<sup>86</sup> In this case, we search for the largest solid angle among all tentative neighbors, divide all solid angles by this maximum angle, and use a threshold (typically, 0.05) to cut out all neighbors that have a fractional solid angle that is smaller than the threshold.

A third neighbor-finding scheme, which we label “CrystalNN” (CNN), assigns probabilities to multiple coordination environments (Fig. 4). The algorithm begins similarly to the Voronoi strategy in which each neighbor is assigned a weight based on solid angle and facet area. Next, all weights are normalized such that the maximum weight is 1. All distinct values of normalized weight (multiple neighbor sites can belong to the same weight) are arranged from 0 to 1. For each distinct weight, the probability of all sites with that weight or greater being considered neighbors is proportional to the integral of the area under a curve (in our case, a semicircle that has a value of zero at zero weight and a value of 1 at a weight of 1) starting from 0 and ending in that weight. Thus, neighbors with higher weights are given a larger “section of the pie” in terms of likelihood to be part of the coordination environment. Also note that this method computes coordination likelihoods,  $w_{CN=i}$ , on the basis of these neighbor coordination probabilities, which quantify the probability that a given coordination environment

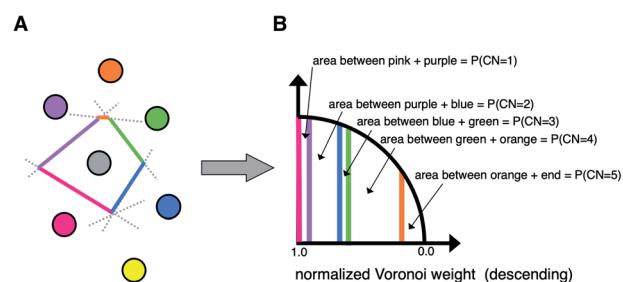


Fig. 4 Schematic of the novel neighbor-finding approach, “CrystalNN”. (A) An example neighbor environment; the Voronoi weights of each neighboring site can be considered proportional to the length of its corresponding colored lined segment. (B) Demonstrating the calculation of coordination probabilities from the normalized Voronoi weights. For example, the section of the semicircle between the pink and purple segments indicates the small probability that only the strongest neighbor (pink) should be considered a neighbor. The area between the purple and blue segment indicates the larger probability that the system should be 2-fold coordinated (the two strongest neighbors, pink and purple). The highest coordination probability for this particular system would be 4-fold coordinated, corresponding to pink, purple, blue and green sites being neighbors.



Table 1 New and existing local structure order parameters (LoStOPs)

Coordination environment (CE)	LoStOP	Coordination number (CN)	Reference
Single bond	$q_{\text{sgl\_bd}}$	1	This work
Bent bonds	$q_{\text{bent}}$	2	This work
Trigonal planar	$q_{\text{tri\_plan}}$	3	This work
T-shape	$q_T$	3	This work
Square planar	$q_{\text{sq\_plan}}$	4	This work
Square non-coplanar	$q_{\text{sq}}$	4	This work
Tetrahedral	$q_{\text{tet}}$	4	82
See-saw <sup>a</sup>	$q_{\text{see\_saw\_rect}}$	4	This work
Trigonal pyramid	$q_{\text{tri\_pyr}}$	4	This work
Pentagonal planar	$q_{\text{pent\_plan}}$	5	This work
Square pyramid	$q_{\text{sq\_pyr}}$	5	This work
Trigonal bipyramidal	$q_{\text{tri\_bipyr}}$	5	This work
Hexagonal planar	$q_{\text{hex\_plan}}$	6	This work
Pentagonal pyramid	$q_{\text{pent\_pyr}}$	6	This work
Octahedral	$q_{\text{oct}}$	6	82
Hexagonal pyramid	$q_{\text{hex\_pyr}}$	7	This work
Pentagonal bipyramidal	$q_{\text{pent\_bipyr}}$	7	This work
Hexagonal bipyramidal	$q_{\text{hex\_bipyr}}$	8	This work
BCC	$q_{\text{bcc}}$	8	81
Cuboctahedron	$q_{\text{cuboct}}$	12	This work

<sup>a</sup> The target see-saw motif has a 90° angle instead of the typical 120° as explained in the text.

should be considered  $i$ -fold coordinated. In an upcoming paper, we investigate the performance of various existing methods<sup>79,80,84,85,87,88</sup> and CNN to predict coordination numbers in inorganic materials, where our novel method performs particularly well.

## 2.2 Local structure order parameters

We use local structure order parameters (LoStOPs) such as those introduced by Peters<sup>81</sup> ( $q_{\text{bcc}}$ ) and by Zimmermann *et al.*<sup>80,82</sup> ( $q_{\text{tet}}$  and  $q_{\text{oct}}$ ) to determine the degree to which the angles in a given observed coordination environment agree with those in the perfect target environment. The LoStOPs are being increasingly exploited for rating the feasibility<sup>89,90</sup> of zeolites,<sup>27,28</sup> finding interstitials and evaluating diffusional pathways in materials,<sup>80</sup> hierarchically visualizing structural similarity,<sup>91</sup> and predicting the magnetic ordering of inorganic materials.<sup>92</sup>

In this work, we vastly extend the existing library from 3 to 20 by introducing new LoStOPs (Table 1). The new LoStOPs permit the detection of both highly symmetric motifs (*e.g.*, Platonic solids<sup>93</sup> such as a tetrahedron or an octahedron) as well as less symmetric motifs (*e.g.*, Johnson solids<sup>94</sup> such as a square pyramid). This is possible because we change our ansatz slightly while keeping the key ideas.

We still test whether a given coordination environment (*e.g.*, the blue T-motif in Fig. 5) resembles a perfect target motif (gray “template” in Fig. 5) by using the neighbors of the central site to locally set up spherical coordinate systems.<sup>81,82</sup> And, we still check all permutations explicitly; that is, we use each neighbor as a tentative North pole for creating the coordinate system (marked with a dark-blue “N” in each configuration) and each remaining neighbor as a prime meridian. We also stay with the general strategy of using cosine functions and Gaussian kernels

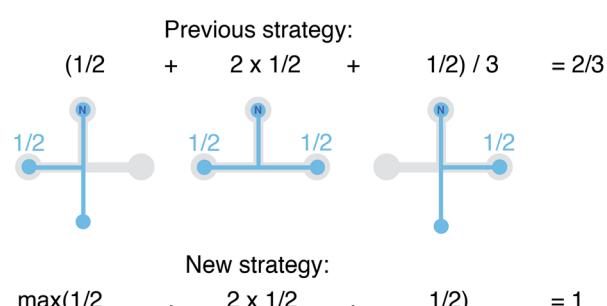


Fig. 5 Visual depiction of change in the computation of local structure order parameters: going from the previous strategy of averaging (top) to the new strategy of finding the maximal resemblance, given all possible rotations.

[without normalization constant  $1/(\sqrt{2\pi}\sigma)$ ] to penalize positions of the remaining neighbors that are not at expected positions. The crucial difference to our earlier motif resemblance metrics<sup>80–82</sup> is that we do not average over all permutations anymore. Instead, we use the highest motif resemblance,

$$q_m = \max(\{q_{m,j}\}) \quad (1)$$

given all the individual resemblance values,  $q_{m,j}$ , each of which is obtained with one single neighbor  $j$  as the North pole for resemblance evaluation to motif type  $m$  around central site  $i$  (Fig. 5). For example, the LoStOP for the T-shaped coordination environment,  $q_T$ , is given by:

$$q_T = \max_{\substack{j \in N_{\text{nn}} \\ k \neq j}} \left\{ \sum_{l \neq j}^{|N_{\text{nn}}|} \exp \left[ -\frac{(\theta_{ijl} - 90^\circ)^2}{2\Delta\theta^2} \right] \cos^2 \varphi_{ijkl} \right\}, \quad (2)$$



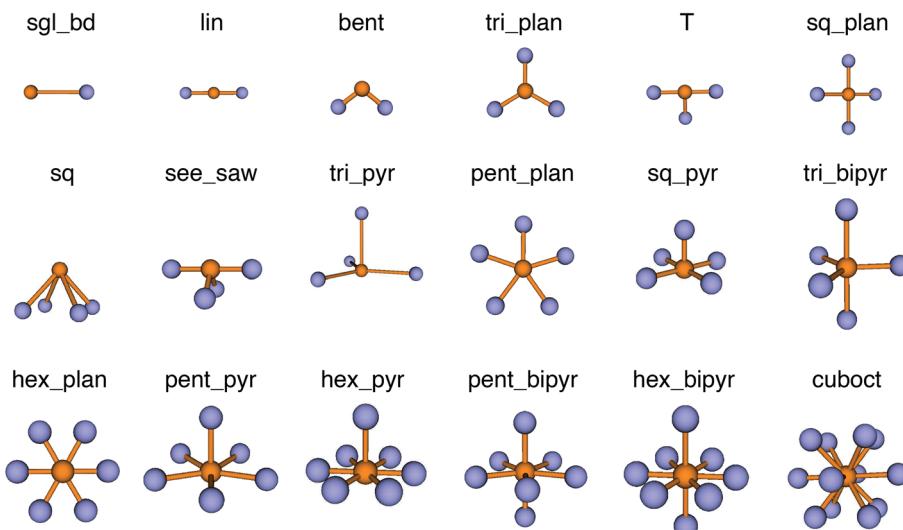


Fig. 6 Coordination motifs for which we introduce new local structure order parameters (LoStOPs) along with their labels that we use in this work.

where  $N_{nn}$  is the number of those near neighbors that are to be considered for LoStOP calculation,  $\theta_{ijl}$  is the polar angle of neighbor  $l$  (*i.e.*, the angle between North pole neighbor  $j$ , central site  $i$ , and neighbor  $l$ ),  $\Delta\theta$  is the parameter penalizing positions of  $l$  that are not forming a perfect  $90^\circ$  angle with the bond between  $i$  and  $j$ , and neighbor  $k$  is used to construct the prime meridian; hence,  $\varphi_{ijkl}$  is zero for  $l = k$ . Note that we also redefine the tetrahedral and octahedral LoStOPs using the new ansatz regarding the permutations.

The new LoStOPs allow detection of even more motifs than suggested by Table 1 (*cf.*, Fig. 6) because the linear/bent LoStOP can be used for various bent angles, the trigonal bipyramidal LoStOP can be used for regular see-saw motifs, and we can use  $q_{\text{tet}}$  to detect triangular non-coplanar arrangements with tetrahedral angles. Later, when we define site fingerprints, we define multiple instances of the bent LoStOP with different target angles. Note that the new LoStOPs, just as the originals ones, are invariant to translational and rotational operations, which represents an important prerequisite to be used as an element of a numerical materials fingerprint.<sup>63</sup> Furthermore, the new LoStOPs still smoothly vary between 0 and 1, and a value of 1 flags, as usual, perfect resemblance with the underlying target motif, whereas 0 indicates no resemblance or match.

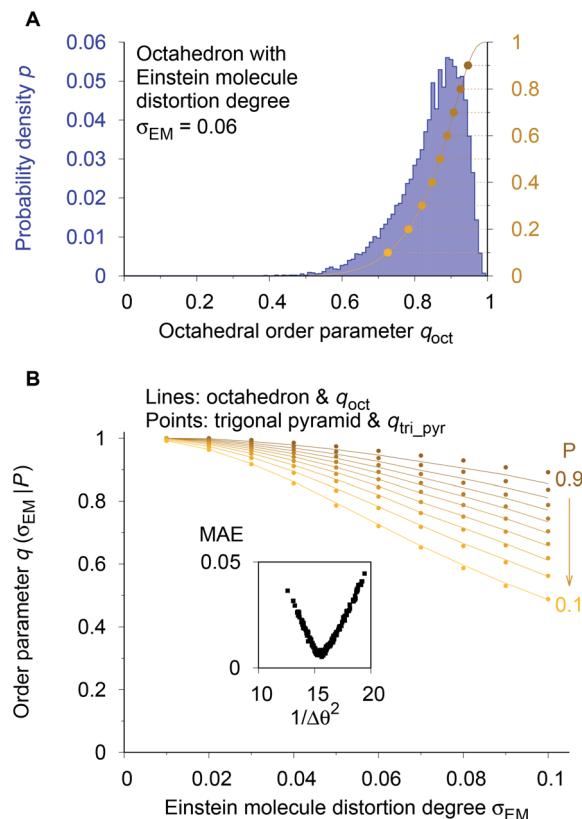
**2.2.1 Optimization.** Any LoStOP computation requires the *a priori* choice of certain calculation parameters such as the penalty for the neighbor that is closest to the South pole position in an octahedron-like environment to not exactly form a straight line with the (tentative) North pole neighbor and the central site. Our goal for setting these parameters was to maximize the comparability between different motifs (*e.g.*, degree of resembling a perfect tetrahedron equals degree of a perfect square pyramid). Consequently, we had to make a decision about which distortion in one motif (*e.g.*, moving one of the non-North pole neighbors in a tetrahedron slightly towards the South pole position) should be compared to

another distortion in a different motif (*e.g.*, moving two neighbors in the basal plane of the square pyramid closer together). Identifying all possible elemental distortions would have been unfeasible, notwithstanding intermediate distortion pathways (*cf.*, Fig. 4 in ref. 80). Also, it would have been somewhat arbitrary to weight the different elemental and intermediate distortions for their occurrence or relevance: should they contribute equally or are some less likely or less important than others?

To circumvent these issues we decided to employ a numerical procedure that leverages our recently introduced Einstein crystal test rig procedure.<sup>80</sup> This order parameter testing framework assigns Gaussian-distributed random perturbations to all sites in the initially perfect coordination environment using the polar form of Box–Muller transforms<sup>95</sup> as implemented in numpy.<sup>96</sup> The resulting spatial distributions of the atoms around their perfect motif positions resemble those in an Einstein crystal or molecule.<sup>97–99</sup> The atom displacement distribution width,  $\sigma_{\text{EM}}$ , is an input parameter and provides a well-controllable way to *a priori* define the average distortion degree of the entire coordination environment. Thus, it also provides a definition for comparing distortion states between different motifs. Furthermore, because we aim at the assessment of real materials and their coordination patterns and because those can be subject to thermal fluctuations, we, hence, use a physically meaningful reference model.

As a reference point, we use the Einstein molecule response behavior of the octahedral order parameter,  $q_{\text{oct}}$ . For varying Einstein molecule distortion degrees ( $\sigma_{\text{EM}} = 0.01 \cdots 0.1$  in 0.01 increments), we compute histograms of the order parameter by perturbing all atoms in an octahedral coordination motif 10 000 times. From each density distribution  $p(q_{\text{oct}}|\sigma_{\text{EM}})$  (blue line in Fig. 7A), we compute cumulative probability distributions  $P(q_{\text{oct}}|\sigma_{\text{EM}})$  (gold-brown line and points), which are smooth functions of  $q_{\text{oct}}$ . To reduce the number of data points we





**Fig. 7** Procedure for optimizing the calculation parameters to maximize inter-motif comparability. (A) The cumulative probability distributions of the octahedral LoSTOP as resulting from random displacements (Einstein crystal test rig; cf. ref. 80) are calculated for different Einstein molecule displacement degrees,  $\sigma_{\text{EM}}$ , to give the baselines. (B) The parameters involved in the calculation of any and all other LoSTOPs (here: trigonal pyramid) are optimized by minimizing the difference of curves obtained from plotting the LoSTOP given a specific cumulative probability value,  $q(\sigma_{\text{EM}}|P)$ , as it varies with Einstein molecule displacement degree,  $\sigma_{\text{EM}}$ . The inset shows the mean absolute error, MAE, vs. the (inverse and squared) parameter variation,  $1/\Delta\theta^2$  for the case of  $q_{\text{tri\_pyr}}$ .

consider only 9 values of the cumulative distributions in the following: the points at  $P = 0.1, 0.2, 0.3, \dots, 0.9$ . Below, we use and refer to these points in an inverse manner: the order parameter values as a function of the Einstein molecule distortion degree for a given cumulative probability value,  $q(\sigma_{\text{EM}}|P)$ .

Using the octahedral order parameter results as reference data, our optimization procedure for any remaining order parameter (e.g., trigonal pyramid,  $q_{\text{tri\_pyr}}$ ) is a basic Monte Carlo approach and consists of following steps:

(1) Create a set of new trial calculation parameters (e.g., for  $q_{\text{tri\_pyr}}$ , the penalty,  $\Delta\theta$ , for polar angles being different than their expected values of  $90^\circ$ ) by randomly perturbing the current parameters within a pre-defined maximum range.

(2) Compute the  $q(\sigma_{\text{EM}}|P)$  data (in Fig. 7B: gold-brown points) with the motif for which the OP is designed (here:  $q_{\text{tri\_pyr}}$  and a trigonal pyramid).

(3) Calculate the mean absolute error (MAE) between these data points and the octahedral reference at the same Einstein

molecule distortion degrees (*i.e.*, sum up the absolute difference between points and lines in Fig. 7B).

(4) Compare the new MAE,  $\Delta_{\text{new}}$ , to the MAE resulting from the previous parameter set,  $\Delta_{\text{old}}$ . If  $\exp[-k_{\text{MC}}(\Delta_{\text{new}} - \Delta_{\text{old}})]$  is larger than a random number drawn from a uniform distribution, then save the new trial parameter set; else, continue to use the previous parameter set.

(5) Go to step 1.

We usually perform 2000 MC trials and, in each separate trial, we compute the  $q(\sigma_{\text{EM}}|P)$  data by 1000 Einstein molecule perturbations. We carry out at least two optimization runs. In the first optimization, we used a small MC parameter,  $k_{\text{MC}} = 100$ , and a larger maximum parameter perturbation range, which allowed sampling of a larger parameter space. The second run proceeded with a larger MC parameter and a smaller maximum perturbation range, which forced the system to quickly go to the (nearest) MAE minimum. The first optimization served as a check whether or not there is another minimum that might even have a lower MAE than the one that is closest to the initial state during optimization. Finally, the inset of Fig. 7B indicates that the procedure is appropriate because we obtain a convex parameter space.

### 2.3 Site fingerprint

We define site fingerprints on the basis of the above described coordination likelihood and local structure order parameters. The site fingerprints are single-column vectors, the elements of which exhibit an ordered structure. Specifically, the different coordination features are arranged in blocks that represent well-defined coordination numbers so that a site fingerprint  $\mathbf{v}$  can be written as:

$$\mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{12}]^T, \quad (3)$$

where  $\mathbf{v}_i$  denote the (sub)fingerprint that contains all coordination features fulfilling the condition that the coordination number is  $i$ . For example,  $\mathbf{v}_4$  can carry the likelihood of being 4-fold coordinated, the square planar LoSTOP, and the tetrahedral LoSTOP, but it cannot carry the likelihood of being 3-fold coordinated or the bcc (8-fold coordinated) LoSTOP. While it might be necessary to increase the length to also include  $\mathbf{v}_{13}$  and  $\mathbf{v}_{14}$ , our tests have shown that there is little benefit of doing so because the vast majority of sites in crystal structures typically have a coordination number  $\leq 12$ . In the following, we describe specific details of three site fingerprint (CrystalNNFingerprint, OPSiteFingerprint, and ChemEnvFingerprint) as they are currently implemented in matminer<sup>65,77</sup> (version: 0.3.3). Note that there is a mutual strategy in computing the first two fingerprints: grouping the features according to their underlying coordination number, as outlined in eqn (3).

The CrystalNNFingerprint (CNN fingerprint) computes the fingerprint of a given site  $i$  as follows. First, we choose the coordination features to be included (e.g.,  $w_{\text{CN}=4}$ ,  $q_{\text{sq\_plan}}$ , and  $q_{\text{tet}}$ ). Second, the neighbors of site  $i$  are determined with the CrystalNN neighbor-finding algorithm. Then, we loop over all theoretically possible coordination numbers between 1 and the maximum of the set of coordination numbers that underly all



chosen features from step 1. For each coordination number  $j$ , we compute the coordination likelihood,  $w_{CN=j}$ . Subsequently, we compute each LoStOP (e.g.,  $q_{\text{tet}}$ ) that is to be considered for this coordination number (here:  $CN = 4$ ) using  $j$  neighbors (here: 4) with the highest coordination weight, we multiply the value with  $w_{CN=j}$  (here:  $q_{\text{tet}} \times w_{CN=4}$ ), and we then add the resulting value to the (growing) site fingerprint vector. If there is no feature type for a given coordination number  $j$ , a single zero entry is added for the entire  $j$ -block. Furthermore, note that the implementation of the CNN site fingerprint in matminer<sup>65,77</sup> features two convenient presets for rapid calculation setup: the “cn” preset only computes coordination likelihoods ( $w_{CN=j}$ ), whereas the “ops” preset adds all available LoStOP features in addition to the  $w_j$ ’s.

Similar to the CNN site fingerprint, the OPSiteFingerprint (OPS fingerprint) first requires the choice of the features to be included. The main differences to the CNN fingerprint are:

(1) The minimum distance neighbor finding method is used with the default fractional cutoff (1.1).

(2) A more elaborate binning scheme is used to determine whether neighbors belong to the same shell or not, which basically employs a bin width variation approach.

(3) Based on the variational neighbor finding results, multiple values for a given LoStOP are obtained, from which the most stable (*i.e.*, most frequently occurring) value is extracted *via* histogramming.

(4) Coordination weights ( $w_j$ ) are not used (*i.e.*, the fingerprint fully relies on the LoStOP features). However, an additional distance variation factor,  $f_d$ , can be chosen which is, for a motif with  $N_{\text{nn}}$  neighbors around central site  $i$  given by:

$$f_d = \left[ \sum_{j=1}^{N_{\text{nn}}} 1/d_{ij} \right]^2. \quad (4)$$

Lastly, the ChemEnvFingerprint (CE fingerprint) makes full use of the ChemEnv module in pymatgen,<sup>75,76</sup> which provides alternative tools for automatically identifying the coordination environments of atoms in materials.<sup>100</sup> Two presets are available (“simple” and “multi\_weights”), which directly relate to the corresponding ChemEnv strategies. The principal neighbor finding approach is Voronoi decomposition (VNN). The features that are computed with the CE fingerprint are the continuous symmetry measures<sup>101</sup> (CSM) between a given motif and all available ideal coordination environments supported by ChemEnv.<sup>100</sup> In particular, the considered environments are: (1) single neighbor (S:1), (2) linear (L:2), (3) angular (A:2), (4) trigonal plane (TL:3), (5) triangular non-coplanar (TY:3), (6) T-shape (TS:3), (7) tetrahedron (T:4), (8) square plane (S:4), (9) square non-coplanar (SY:4), (10) see-saw (SS:4), (11) pentagonal plane (PP:5), (12) square pyramid (S:5), (13) trigonal bipyramidal (T:5), (14) octahedron (O:6), (15) trigonal prism (T:6), (16) pentagonal pyramid (PP:6), (17) pentagonal bipyramidal (PB:7), (18) square-face capped trigonal prism (ST:7), (19) end-trigonal-face capped trigonal prism (ET:7), (20) face-capped octahedron (FO:7), (21) cube (C:8), (22) square antiprism (SA:8), (23) square-face bicapped trigonal prism (SBT:8), (24) triangular-face bicapped trigonal prism (TBT:8), (25)

dodecahedron with triangular faces (DD:8), (26) dodecahedron with triangular faces—p2345 plane normalized (DDPN:8), (27) hexagonal bipyramid (HB:8), (28) bicapped octahedron (opposed cap faces) (BO\_1:8), (29) bicapped octahedron (cap faces with one atom in common) (BO\_2:8), (30) bicapped octahedron (cap faces with one edge in common) (BO\_3:8), (31) triangular cupola (TC:9), (32) tricapped triangular prism (three square-face caps) (TT\_1:9), (33) tricapped triangular prism (two square-face caps and one triangular-face cap) (TT\_2:9), (34) tricapped triangular prism (one square-face cap and two triangular-face caps) (TT\_3:9), (35) heptagonal dipyramid (HD:9), (36) tridiminished icosahedron (TI:9), (37) square-face monocapped antiprism (SMA:9), (38) square-face capped square prism (SS:9), (39) tricapped octahedron (all 3 cap faces share one atom) (TO\_1:9), (40) tricapped octahedron (cap faces are aligned) (TO\_2:9), (41) tricapped octahedron (all 3 cap faces are sharing one edge of a face) (TO\_3:9), (42) pentagonal prism (PP:10), (43) pentagonal antiprism (PA:10), (44) square-face bicapped square antiprism (SBSA:10), (45) metabidiminished icosahedron (MI:10), (46) bicapped square prism (opposite faces) (BS\_1:10), (47) bicapped square prism (adjacent faces) (BS\_2:10), (48) trigonal-face bicapped square antiprism (TBSA:10), (49) pentagonal-face capped pentagonal antiprism (PCPA:11), (50) hendecahedron (H:11), (51) sphenoid hendecahedron (SH:11), (52) diminished icosahedron (DI:11), (53) icosahedron (I:12), (54) pentagonal-face bicapped pentagonal prism (PBP:12), (55) truncated tetrahedron (TT:12), (56) cuboctahedron (C:12), (57) anticuboctahedron (AC:12), (58) square cupola (SC:12), (59) hexagonal prism (HP:12), (60) hexagonal antiprism (HA:12), and (61) square-face capped hexagonal prism (SH:13).

## 2.4 Structure fingerprint

On the basis of the site fingerprints of all atoms in a crystal structure we can compute meaningful structure fingerprints (Fig. 8). Our approach consists of four steps:

(1) Choose a site fingerprint type.

(2) Choose the statistics to be computed [*e.g.*, only the mean or the mean, the standard deviation, the minimum, and the maximum].

(3) Compute the site fingerprint feature vector of each atom in a structure. Note that all feature vectors have the same length and that each element in one site fingerprint is of the same type (*e.g.*, tetrahedral LoStOP value) as the element at the same location in a fingerprint of another site.

(4) Calculate statistics across all values of a given feature vector element type (*e.g.*, tetrahedral LoStOP value) and arrange them to a new fingerprint that is representative of the coordination patterns in the entire structure.

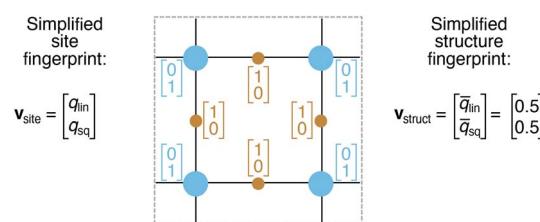


Fig. 8 Illustration of site and structure fingerprints.

In analogy to the site fingerprint, the structure fingerprints are arranged in an ascending order with respect to the coordination number underlying the different feature vector elements (e.g., first come the statistics values from feature types relating to  $\text{CN} = 1$ , then  $\text{CN} = 2$ , ...). For example, if we consider a generic site fingerprint that is only based on site features  $w_{\text{CN}=1}$ ,  $q_{\text{sgl\_bd}}|\text{CN} = 1$ ,  $w_{\text{CN}=2}$ , and  $q_L|\text{CN} = 2$  and if we, furthermore, only consider the mean,  $\bar{x}_i$ , and the standard deviation,  $\sigma_{x_i}$ , as statistics types to be computed, we obtain:

$$\mathbf{v} = [\bar{w}_{\text{CN}=1}, \sigma_{w_{\text{CN}=1}}, \bar{q}_{\text{sgl\_bd}}|\text{CN} = 1, \sigma_{q_{\text{sgl\_bd}}|\text{CN}=1}, \bar{w}_{\text{CN}=2}, \sigma_{w_{\text{CN}=2}}, \bar{q}_L|\text{CN} = 2, \sigma_{q_L|\text{CN}=2}]^T \quad (5)$$

## 2.5 Similarity measures

We consider three common similarity measures for comparing two fingerprint vectors: the Euclidean distance, the dot product, and the cosine similarity.

The Euclidean distance,  $d$ , is defined as the  $L^2$  norm of the fingerprint difference vector,  $\mathbf{v}_i - \mathbf{v}_j$ :

$$d = \|\mathbf{v}_i - \mathbf{v}_j\| \quad (6)$$

$$= \sqrt{\sum_{k=1}^{N_{\text{dim}}} (v_i^k - v_j^k)^2}, \quad (7)$$

where  $\mathbf{v}_i$  denotes fingerprint vector  $i$  and  $N_{\text{dim}}$  the size (or, number of elements) of the vector. The distance can assume values between 0, indicating highest possible similarity, and  $\sqrt{N_{\text{dim}}}$  for lowest theoretically possible similarity, the latter for the condition that the vector elements are constrained between 0 and 1. The  $L^2$  norm is a frequently used (dis)similarity measure (cf., ref. 73 and references therein).

The dot product,  $s_{\text{dot}}$ , is defined by:

$$s_{\text{dot}} = \mathbf{v}_i \cdot \mathbf{v}_j \quad (8)$$

$$= \sum_{k=1}^{N_{\text{dim}}} v_i^k \cdot v_j^k, \quad (9)$$

whereas the cosine similarity,  $s_{\text{cos}}$ , is given by:

$$s_{\text{cos}} = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \cdot \|\mathbf{v}_j\|}. \quad (10)$$

Conceptually, the cosine similarity is especially attractive because we operate in positive space (i.e., vector elements are between 0 and 1) and the similarity measure itself will thus also assume values between 0 and 1. However, note that there should be straightforward ways to convert the distance from a non-normalized dissimilarity measure to a normalized similarity measure; for example, *via*:

$$s_{\text{dist}} = \exp(-d). \quad (11)$$

Finally, we measure the dissimilarity between two probability density distributions  $p_1$  and  $p_2$  *via* the overlapping coefficient (OVL),<sup>102</sup> which is defined by the following integral:

$$\text{OVL} = \int_{\min(X)}^{\max(X)} \min[p_1(X), p_2(X)] \text{d}X, \quad (12)$$

where  $X$  denotes the variable or observable over which the two distributions are defined.

## 3 Results & discussion

In this section, we present the benchmarking results for (i) using site fingerprints that are based on coordination number likelihood and the local structure order parameters (LoStOPs) to distinguish different local coordination environments and (ii) using structure fingerprints that are, in turn, based on site fingerprints to distinguish different prototype crystal structures.

### 3.1 Site fingerprints

We systematically test the performance of the here introduced site descriptors [local structure order parameters (LoStOPs)] in conjunction with the concept of a site fingerprint to help distinguishing different local coordination environments by employing 61 standard coordination motifs.<sup>103,104</sup> The motifs are available *via* the ChemEnv module<sup>100</sup> in pymatgen.<sup>75,76</sup> In order to focus on the LoStOPs performance we neglect the coordination number likelihood, and we decouple the pattern matching part (i.e., LoStOP calculation) from the neighbor finding part. The latter is possible because the separate motifs are well-defined by identification of a central site and the neighboring sites as provided by the ChemEnv module.<sup>100</sup> The (simplified) site fingerprint that we use to distinguish unique coordination environments is a vector,  $\mathbf{v}^{\text{OP}}$ , with 37 components. The vector elements,  $v_i^{\text{OP}}$ , are zero if the (observed) motif coordination number is not equal to the target coordination number, whereas the vector element is  $q_j$  if the coordination numbers are equal:

$$v_i^{\text{OP}} = \begin{cases} 0, & \text{if } \text{CN} \neq \text{CN}^{\text{target}} \\ q_j, & \text{if } \text{CN} = \text{CN}^{\text{target}} \end{cases} \quad (13)$$

Since we only have a single motif-specific order parameter for coordination numbers beyond 8, we additionally evaluate the bond-orientational order parameters<sup>105</sup>  $q_2$ ,  $q_4$ , and  $q_6$  for coordination numbers from 8 to 13 because those OPs are also helpful in discerning different structural motifs.<sup>80,106,107</sup> Thus, the resulting vector is:

$$\begin{aligned} \mathbf{v}^{\text{OP}} = [ & q_{\text{sgl\_bd}}|\text{CN} = 1, q_{\text{bent}(90^\circ)}|\text{CN} = 2, \\ & q_{\text{bent}(104.45^\circ)}|\text{CN} = 2, q_{\text{bent}(120^\circ)}|\text{CN} = 2, \\ & q_{\text{bent}(150^\circ)}|\text{CN} = 2, q_{\text{bent}(180^\circ)}|\text{CN} = 2, \\ & q_{\text{tri\_plan}}|\text{CN} = 3, q_{\text{tet}}|\text{CN} = 3, q_{\text{T}}|\text{CN} = 3, \\ & q_{\text{sq\_plan}}|\text{CN} = 4, q_{\text{tet}}|\text{CN} = 4, q_{\text{see\_saw\_rect}}|\text{CN} = 4, \\ & q_{\text{tri\_bipyrl}}|\text{CN} = 4, q_{\text{tri\_pyrl}}|\text{CN} = 4, q_{\text{pent\_plan}}|\text{CN} = 5, \\ & q_{\text{sq\_pyrl}}|\text{CN} = 5, q_{\text{tri\_bipyrl}}|\text{CN} = 5, q_{\text{hex\_plan}}|\text{CN} = 6, \\ & q_{\text{oct}}|\text{CN} = 6, q_{\text{pent\_pyrl}}|\text{CN} = 6, q_{\text{hex\_pyrl}}|\text{CN} = 7, \\ & q_{\text{pent\_bipyrl}}|\text{CN} = 7, q_{\text{bcc}}|\text{CN} = 8, q_{\text{hex\_bipyrl}}|\text{CN} = 8, \\ & q_2|\text{CN} = 9, q_4|\text{CN} = 9, q_6|\text{CN} = 9, q_2|\text{CN} = 10, \\ & q_4|\text{CN} = 10, q_6|\text{CN} = 10, q_2|\text{CN} = 11, q_4|\text{CN} = 11, \\ & q_6|\text{CN} = 11, q_{\text{cuboctl}}|\text{CN} = 12, q_2|\text{CN} = 12, \\ & q_4|\text{CN} = 12, q_6|\text{CN} = 12 ]^T. \end{aligned} \quad (14)$$



where  $q_{\text{bent}}(\alpha)$  refers to the bent LoStOP with a target angle of  $\alpha$ ,  $q_{\text{tet}}|\text{CN} = 3$ , aims at identifying trigonal non-coplanar environments, and  $q_{\text{tri\_bipy}}|\text{CN} = 4$  should work together with  $q_{\text{tri\_pyr}}|\text{CN} = 4$  to identify regular see-saw motifs (angle: 120°).

In Fig. 9, we compare results obtained using four site fingerprint similarity metrics: the square root of the dot product,  $\sqrt{s_{\text{dot}}}$ , (top left), the modified (top right) and conventional (bottom left) cosine similarity,  $s_{\text{cos}}^*$  and  $s_{\text{cos}}$ , respectively, and the distance (or  $L^2$  norm),  $d$ , (bottom right). By construction of the simplified site fingerprint  $\mathbf{v}^{\text{OP}}$ , the square root of the dot product and the cosine similarities are zero (blue color) for any two site fingerprints obtained from coordination environments that have different coordination numbers [e.g., tetrahedron ( $\text{CN} = 4$ ) vs. octahedron ( $\text{CN} = 6$ )] because those are strictly orthogonal. The conventional cosine similarity is, however, in many cases close to unity (yellow color) when we consider coordination environments which have the same coordination number (cf., yellow/gray triangles along the diagonal in bottom left panel).

To understand this potentially unexpected behavior let us consider some concrete examples: the 6-fold coordinated motifs octahedron, pentagonal pyramid, and trigonal prism. The three non-zero elements of the fingerprints are  $q_{\text{hex\_plan}}|\text{CN} = 6$ ,  $q_{\text{oct}}|\text{CN} = 6$ , as well as  $q_{\text{pent\_pyr}}|\text{CN} = 6$ . Removing the zero entries, the resulting motif fingerprints are:

$$\mathbf{v}_{\text{oct}}^{\text{OP}} = [0.20, 1.00, 0.50]^T,$$

$$\mathbf{v}_{\text{pent\_pyr}}^{\text{OP}} = [0.10, 0.50, 1.00]^T, \text{ and}$$

$$\mathbf{v}_{\text{trig\_prism}}^{\text{OP}} = [0.03, 0.28, 0.48]^T.$$

Thus, the dot products are 1.02, 0.53, and 0.62 for motif pairs oct-pent\_pyr, oct-trig\_prism, and trig\_prism-pent\_pyr, respectively, whereas the corresponding cosine similarities are 0.8, 0.83, and 0.997. Since the maximum dot product is around 3.2 given our motif test set, all three dot products are small compared to this upper bound. On the contrary, the cosine similarities are all close to the inherent upper bound of 1. Evidently, the normalization term  $\|\mathbf{v}_i^{\text{OP}}\| \times \|\mathbf{v}_j^{\text{OP}}\|$  changes the dot products in an undesirable manner, which is true in particular for the last pair (trig\_prism-pent\_pyr). The problem is that the “length” of the fingerprint carries actually important information. The “angle” or alignment of the fingerprints alone is, thus, not sufficient information for similarity purposes (it is however crucially important and works well if we consider coordination environments with different CN, as we have seen). To clarify this point further consider the pair pentagonal pyramid-trigonal prism. The fingerprint of the first motif is approximately the same as the fingerprint of the second motif multiplied by 2:  $\mathbf{v}_{\text{pent\_pyr}}^{\text{OP}} \approx 2 \times \mathbf{v}_{\text{trig\_prism}}^{\text{OP}}$ . Although the two fingerprints are parallel, they have very different interpretations. The first fingerprint,  $\mathbf{v}_{\text{pent\_pyr}}^{\text{OP}}$ , has an order parameter that equals 1 and, thus, flags a perfect motif, whereas all elements of the second fingerprint,  $\mathbf{v}_{\text{trig\_prism}}^{\text{OP}}$  are markedly smaller than 1 (maximum:  $\approx 0.5$ ), thus, signifying no perfect motif match

(given our set of LoStOPs). Hence, the cosine similarity can be a misleading similarity metric, especially if we consider same-coordination number motifs. Before moving on, we like to highlight at this point the nice symmetric behavior of  $q_{\text{pent\_pyr}}$  and  $q_{\text{oct}}$ :  $q_{\text{pent\_pyr}}$  gives a value of 0.5 for a perfect octahedron and,  $q_{\text{oct}}$ , in turn, gives 0.5 for a perfect pentagonal pyramid.

Omitting the normalization term of the cosine similarity but taking the square root of the dot product to preserve the units, we obtain a similarity metric that can alleviate the problem with putatively similar same-coordination number environments (cf., top left panel in Fig. 9), as can replacing the typical normalization constant  $\|\mathbf{v}_i^{\text{OP}}\| \cdot \|\mathbf{v}_j^{\text{OP}}\|$  with  $\max^2(\|\mathbf{v}_i^{\text{OP}}\|, \|\mathbf{v}_j^{\text{OP}}\|)$  to some extent (top right panel). So, choosing the square root of the dot product instead of the cosine similarity has here the advantage that the vast majority of coordination environments are identified as dissimilar, as expected. On the other hand, we lose the nice behavior of the similarity metric to be constrained between zero and unity. The majority of square root of the dot products for same-coordination number motifs are  $\leq 1$ . Thus,  $\sqrt{s_{\text{dot}}} = 1$  seems to be a reasonable *ad-hoc* threshold value for distinguishing any coordination environments by means of the here introduced site fingerprints and the square root of the dot product similarity metric.

Finally, the distance dissimilarity metric (bottom right panel), for which we expected high values for the majority of cases (i.e., also blue color because the range was inverted), gives a rather diffuse picture. A subtle trend might be seen because high ( $\approx 2$ : blue) and moderate ( $\approx 1$ : gray) values are predominant along the axes, whereas small distances ( $\approx 0$ : yellow) accumulate in the center of the graph. But the overall results strongly suggest that the square root of the dot product is the best similarity metric for coordination environment comparison using the here introduced order parameter-based site fingerprints.

The detailed analysis of the site fingerprints underscores that the LoStOPs can reliably distinguish between motifs sharing the same coordination number (e.g., tetrahedron vs. square planar). This capability has already been exploited to develop a computational tool that can automatically detect sites in crystal structures (e.g., “tet” and “oct”) and translate the information into human readable text (e.g., “This site is a tetrahedral site”),<sup>108</sup> thus, demonstrating the unique power of the here introduced tools and concepts.

### 3.2 Structure fingerprints

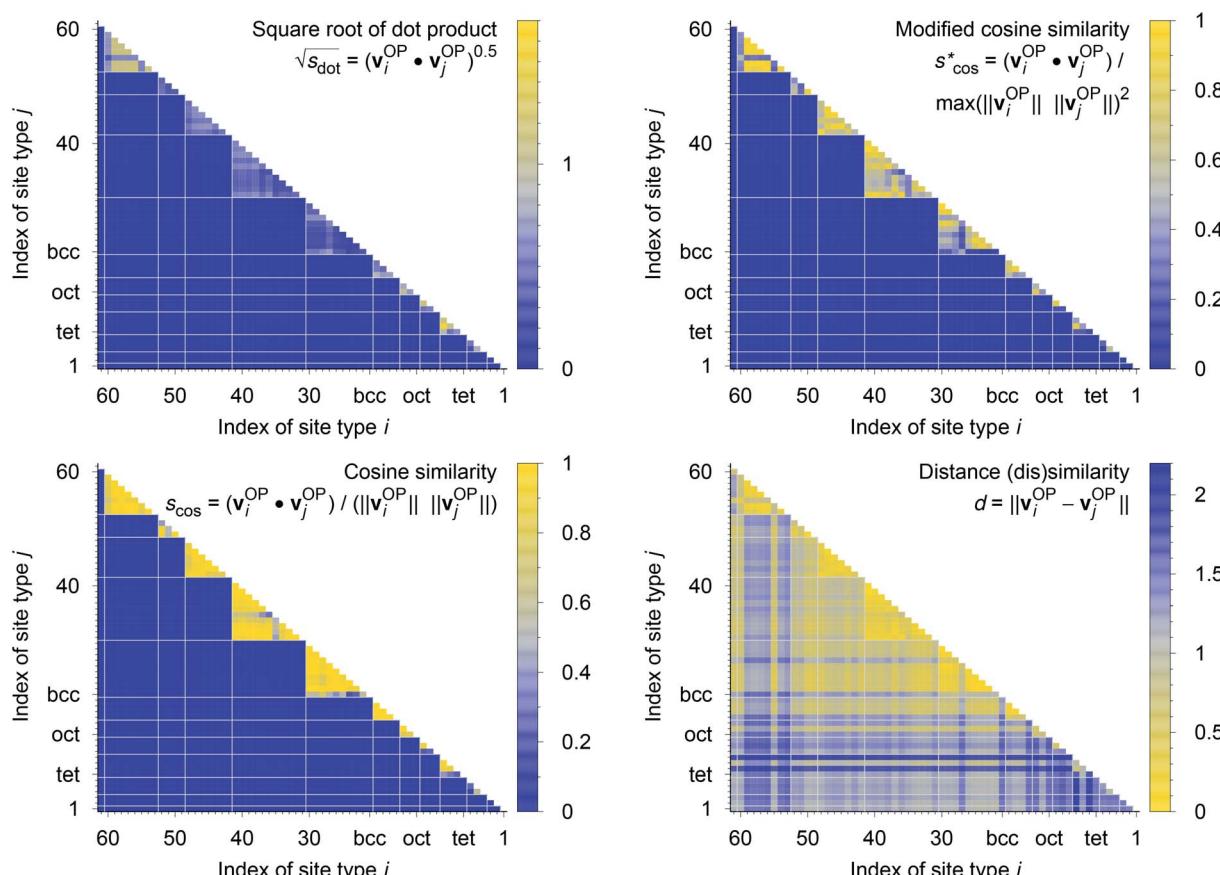
The order parameter-based site fingerprints work well for distinguishing many different isolated coordination environments, especially if the square root of the dot product is used as a similarity metric. In this section, we aim to leverage this newly discovered powerful capability to accomplish a more complex, yet exceptionally important, task in materials science: automatically distinguishing and quantitatively comparing different (prototype) crystal structures using the earlier introduced concept of structure fingerprints.

In order to thoroughly test our structure fingerprints we have constructed a benchmark test set consisting of 40 groups of

relevant prototype structures (Fig. 10) ranging from metals over semiconductors to insulators. Some of the groups stem from one of our previous papers,<sup>80</sup> but we have systematically and significantly increased the number of groups using a canonical reference: “The Major Ternary Structure Families” by Muller and Roy.<sup>59</sup> We included the main structure of any structure family given by ref. 59. For example, we included regular spinel, but we did not include spinels with Jahn–Teller distortions such as hausmannite. In the future, we will further increase the test set. Given a structure prototype described in ref. 59, we proceed by identifying the respective structure in the Materials Project<sup>11</sup> database—typically *via* the chemical formula and the space group. Subsequently, we use pymatgen’s structure matcher in

conjunction with a framework comparator for finding more structures of the same (framework) prototype but with different chemistry, as we did previously.<sup>80</sup> This procedure yields 6528 structures, each of which belongs to one of the 40 prototype groups and all together amounting to a test set that comprises almost 10% of the entire MP database<sup>11</sup> (currently:<sup>12</sup> 69 640 inorganic structures).

We consider the cosine similarity and the distance dissimilarity metrics for testing the performance of distinguishing prototype structure groups on the basis of our structure fingerprints. Furthermore, we test 3 different site fingerprint definitions with different options and presets, resulting in 8 distinctly different site fingerprint types, and we investigate 7



1: Single neighbor. 2: Linear. 3: Angular. 4: Trigonal plane. 5: Triangular non-coplanar. 6: T-shaped. 7: Tetrahedron. 8: Square plane. 9: Square non-coplanar. 10: See-saw. 11: Pentagonal plane. 12: Square pyramid. 13: Trigonal bipyramidal. 14: Octahedron. 15: Trigonal prism. 16: Pentagonal pyramid. 17: Pentagonal bipyramidal. 18: Square-face capped trigonal prism. 19: End-trigonal-face capped trigonal prism. 20: Face-capped octahedron. 21: Cube. 22: Square antiprism. 23: Square-face bicapped trigonal prism. 24: Triangular-face bicapped trigonal prism. 25: Dodecahedron with triangular faces. 26: Dodecahedron with triangular faces - p2345 plane normalized. 27: Hexagonal bipyramid. 28: Bicapped octahedron (opposed cap faces). 29: Bicapped octahedron (cap faces with one atom in common). 30: Bicapped octahedron (cap faces with one edge in common). 31: Triangular cupola. 32: Tricapped triangular prism (three square-face caps). 33: Tricapped triangular prism (two square-face caps and one triangular-face cap). 34: Tricapped triangular prism (one square-face cap and two triangular-face caps). 35: Heptagonal dipyramid. 36: Tridiminished icosahedron. 37: Square-face monocapped antiprism. 38: Square-face capped square prism. 39: Tricapped octahedron (all 3 cap faces share one atom). 40: Tricapped octahedron (cap faces are aligned). 41: Tricapped octahedron (all 3 cap faces are sharing one edge of a face). 42: Pentagonal prism. 43: Pentagonal antiprism. 44: Square-face bicapped square antiprism. 45: Metabidiminished icosahedron. 46: Bicapped square prism (opposite faces). 47: Bicapped square prism (adjacent faces). 48: Trigonal-face bicapped square antiprism. 49: Pentagonal-face capped pentagonal antiprism. 50: Hendecahedron. 51: Sphenoid hendecahedron. 52: Diminished icosahedron. 53: Icosahedron. 54: Pentagonal-face bicapped pentagonal prism. 55: Truncated tetrahedron. 56: Cuboctahedron. 57: Anticuboctahedron. 58: Square cupola. 59: Hexagonal prism. 60: Hexagonal antiprism. 61: Square-face capped hexagonal prism.

**Fig. 9** (Dis)similarity between 61 standard coordination environments based on a simplified site fingerprint,  $v^{\text{OP}}$ , that is defined by local structure order parameters only; top left: square root of the dot product,  $\sqrt{s_{\text{dot}}}$ ; top right: modified cosine similarity,  $s_{\text{cos}}^*$ ; bottom left: cosine similarity,  $s_{\text{cos}}$ ; bottom right: distance dissimilarity,  $d$ . The names of the coordination environments are given beneath the plots along with the respective indices used for plotting purposes. For clarity, the white grid lines separate coordination environments with different coordination numbers.

different combinations of tentatively relevant statistics types: the mean of each coordination feature across all sites; mean and maximum; mean and minimum; mean and standard

deviation; mean, standard deviation, and minimum; mean, standard deviation, and maximum; as well as mean, standard deviation, minimum, and maximum. In summary, we thus test

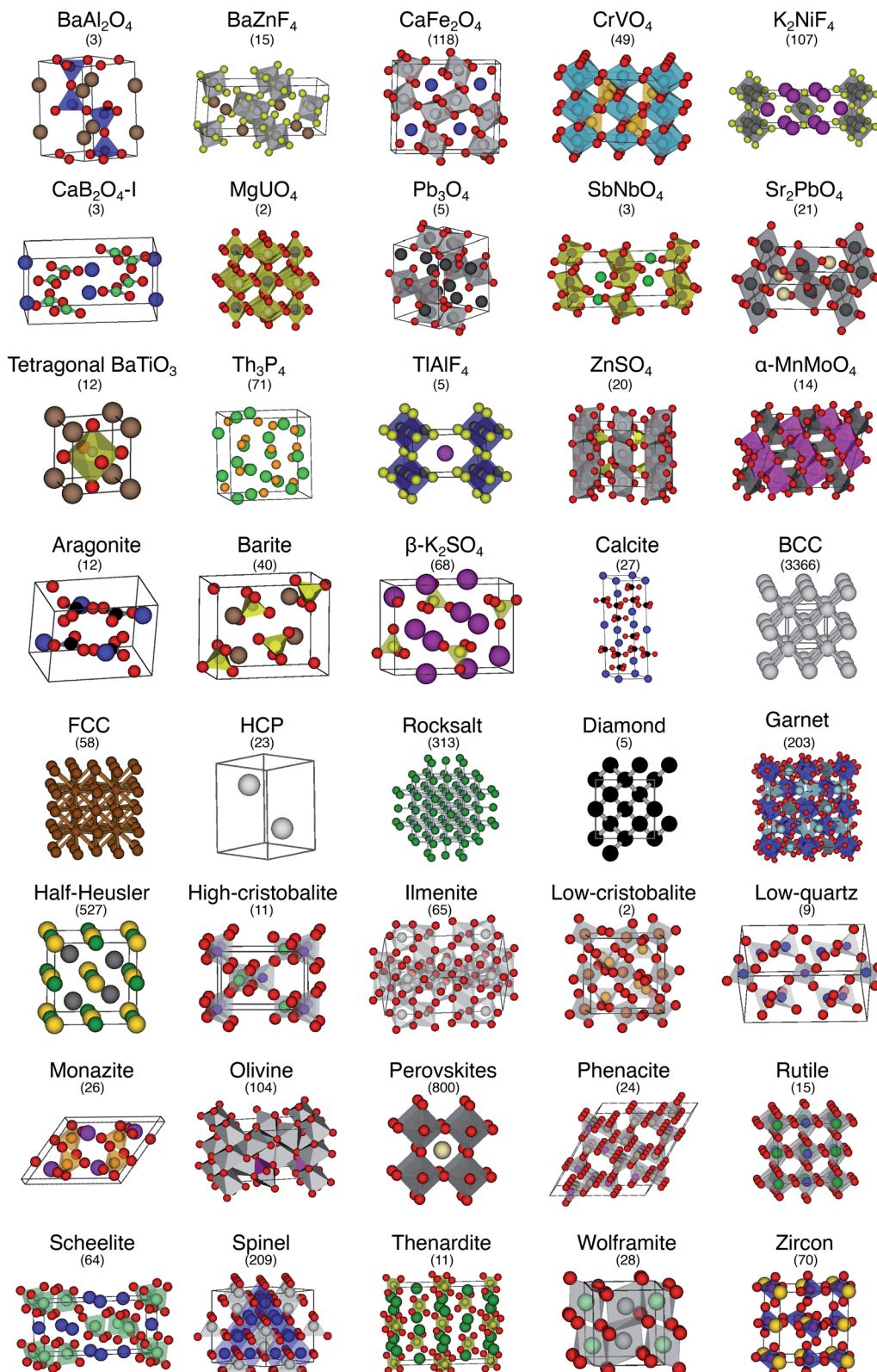


Fig. 10 The 40 prototype structures considered in this work for testing the performance of structure similarity assessment with our novel structure fingerprints. In parentheses, we provide the number of materials that we found in the MP database<sup>11,12</sup> for a given prototype group.

$2 \times 8 \times 7 = 112$  different ways of assessing structure similarity for each of the 40 structure groups separately and for all groups together. Note that each structure group is weighted equally; that is, we remove the bias of a group that has many structure members (e.g., bcc + CsCl + Heusler: 3366) in comparison to a group that only features a few structures (low-cristobalite: 2).

The best performing combination (OVL: 1.9%) is the CrystallNNFingerprint with “ops” preset and all additional flags turned off, mean and maximum as statistics types, as well as distance as comparison metric (Fig. 11). The mean as statistics type alone already provides excellent results (OVL: 2.6%), which can be regarded as a reflection of Pauling’s 5th rule, the Rule of Parsimony:<sup>109</sup> “the number of essentially different kinds of constituents in a crystal tends to be small.” Furthermore, it is interesting that the “cn” preset variant, which uses only coordination number likelihoods,  $w_{CN=j}$ , and, thus, no order parameters at all, performs similarly well (OVL: 1.8%) while having  $183/72 = 2.5$  times less features in the structure fingerprint. Concepts like the Bayesian<sup>110</sup> or Akaike information criterion<sup>111</sup> that aim at guiding model selection might suggest here that we added too many parameters to our model. However, we stress that the local structure order parameter features are not simply “random” features that we test in a “black box” approach. The features have clear and, most importantly, desirable interpretations or capabilities: they can reliably identify coordination environments, as the previous section unequivocally highlighted. Since the structure fingerprint results underscore that the most basic coordination information suffices to reliably distinguish different structures, this hints at a different problem.

The issue seems to be that our test set might not be diverse enough in order to make the LoStOPs more critical components of the structure fingerprints. This can be understood when recalling that occurrence statistics of different coordination motifs are not uniform, but there are certain motifs that are observed disproportionately more frequently. In particular, tetrahedra and octahedra are the most frequently occurring coordination motifs for anions. For example, we found in the structures that we used from ref. 59 that 23 structures have (anion) octahedral sites, 21 structures with tetrahedral sites, as well as 3 with trigonal planar and with 2 trigonal non-coplanar sites. Furthermore, almost all of these tetrahedral and octahedral sites are nearly perfect Platonic polyhedra. Consequently, the issue or bias that there are only a few important perfect motifs and that the vast majority of coordination environments considered in the previous section do not occur frequently enough is a natural external condition in the present context. Nonetheless, our results indicate that the key to a reliable coordination environment-based structure similarity assessment lies in the accuracy of the neighbor finding method.

The full collection of results for structure group (dis)similarity is provided in the ESI.† In Fig. 12, we present the data using the setup that gives the smallest overall overlapping coefficient. The blue lines indicate the distributions obtained from calculating distances between fingerprints from structure of the same prototype group, whereas the orange lines are the

distribution of structure fingerprint distances to members of other prototype groups. The results for each prototype group separately underline that our approach using the overlapping coefficient is a valuable choice because, in most cases, there is a nice consistent separation between the blue area on the left and orange area in the center and on the right. Very few groups exhibit slightly misleading results such as low-cristobalite (right column, 6th panel from the top). Despite the very low OVL (0.5%) observed for this group, there are a lot of unlike structures having a smaller distance to a member of the low-cristobalite group than the peak location representing the (little) distance variation within the low-cristobalite prototype group itself.

We emphasize that our new structure similarity quantification procedure has a very important advantage to conventional structure matching or similarity methods. It does not only provide a global similarity value (e.g., RMSD). The separate elements of the fingerprints and, thus, difference vectors also carry local coordination information in a very compact way, which can be readily used in machine learning approaches and in formulating design rules. For example, consider diamond (C, mp-66), rocksalt (NaCl, mp-22862), and  $\alpha$ -iron (Fe, mp-13) as well as a Heusler (Cu<sub>2</sub>MnSn, mp-22221), a spinel (MgAlO<sub>3</sub>, mp-3536), and a perovskite material (CaTiO<sub>3</sub>, mp-5827). Ignoring zero entries, the values of the best performing fingerprint for similarity assessment (“CrystallNN\*” with “ops” preset and mean and maximum values) are provided in Table 2.

Diamond’s leading fingerprint elements are the mean 4-fold coordination likelihood and the mean tetrahedral LoStOP. Because diamond<sup>112</sup> consists exclusively of tetrahedrally coordinated carbon atoms both LoStOP entries are 1, whereas all other 4-fold coordinated LoStOPs are much smaller ( $\leq 0.25$ ). Similarly, rocksalt<sup>113</sup> has only octahedrally coordinated Na<sup>+</sup> and Cl<sup>-</sup> ions, for which reason the mean 6-fold coordination likelihood and the mean octahedral LoStOP is 1 and all other 6-fold coordination descriptors considerably smaller.

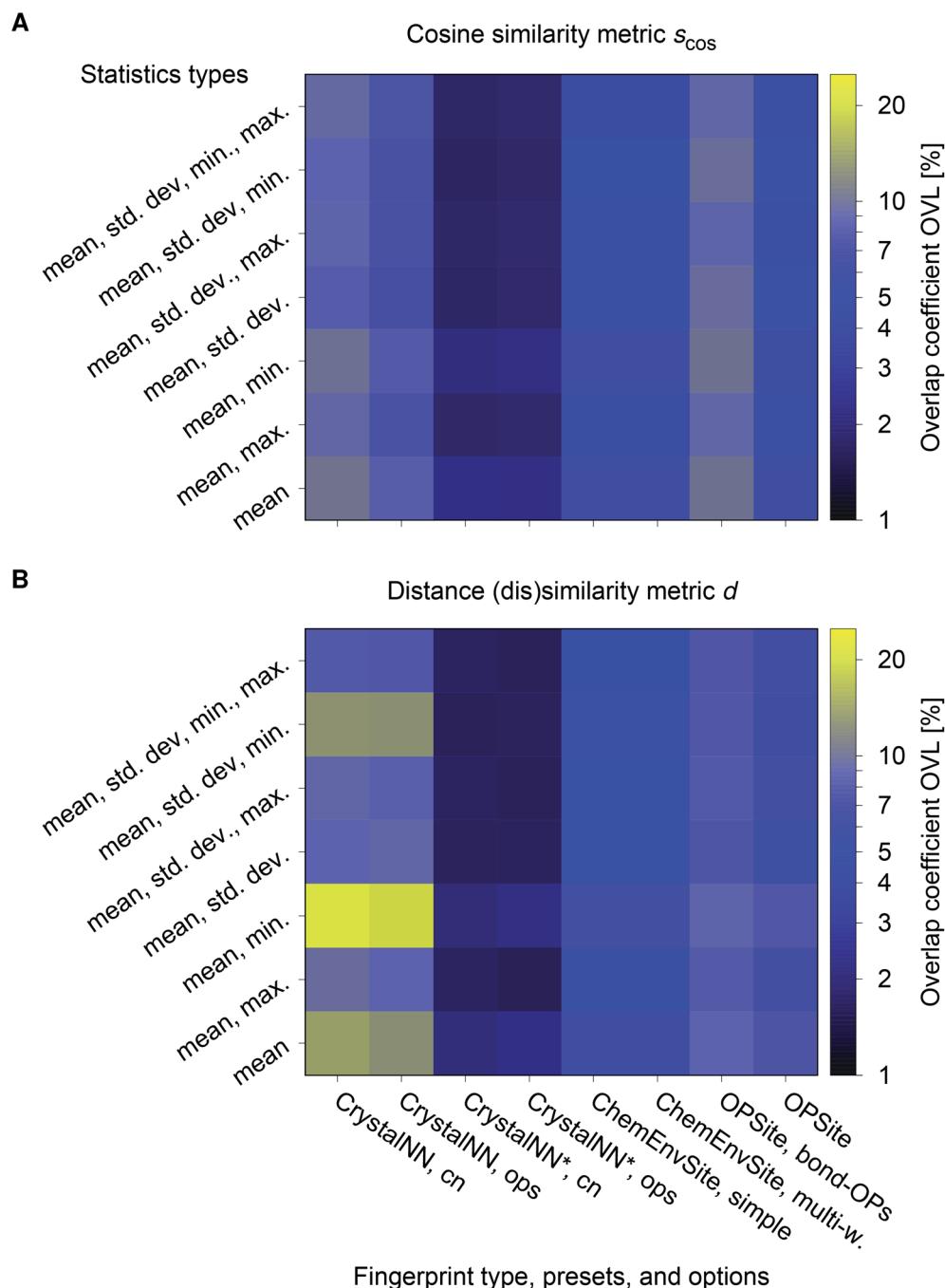
All sites in  $\alpha$ -iron<sup>114</sup> and the full Heusler material<sup>115</sup> Cu<sub>2</sub>MnSn are expected to be coordinated 8-fold in a body centered cubic manner. Among the leading fingerprint elements, we find in fact  $\bar{w}_{CN=8}$  and  $\bar{q}_{bcc}|CN = 8$ , but their deviation from 1 (0.58 and 0.54, respectively) and a slightly higher maximum value together with  $\max(w_{CN=14})$  flag a special case. The second nearest neighbors in BCC structures are quite close to a given site in comparison to nearest neighbors,<sup>80</sup> and the structure produces rather large Voronoi facets for those second shell atoms.<sup>107</sup> Hence, it is not unreasonable to consider the second nearest neighbors as being directly coordinated to a given site—only with a smaller coordination likelihood. Also note that, in the case of  $\alpha$ -iron, the 6 second nearest neighbors exhibit an as small contribution as 0.21 to the 14-fold coordination likelihood if the contribution of the 8 nearest neighbors can be set to  $w_{CN=8} = 0.58$ .

The MgAlO<sub>3</sub> spinel structure possesses perfect MgO<sub>4</sub> tetrahedra, which share corners with slightly distorted AlO<sub>6</sub> octahedra.<sup>59</sup> Since the oxygen atoms are 4-fold coordinated just as the Mg ions,<sup>59</sup> 80% of the sites are 4-fold coordinated, which explains the high value of  $\bar{w}_{CN=4} = 0.7$  (all Mg and O site  $w_{CN=4}$



$\geq 0.98$ ). The fact that the mean tetrahedrality is markedly larger than 1/5 indicates that there are more sites that have a certain tetrahedral character than only Mg sites with  $q_{\text{tet}} = 1$ : the oxygen sites with  $q_{\text{tet}} \approx 0.5$ . However, several of the maximum values of the other 4-fold coordination motif LoStOPs are also around 0.5,

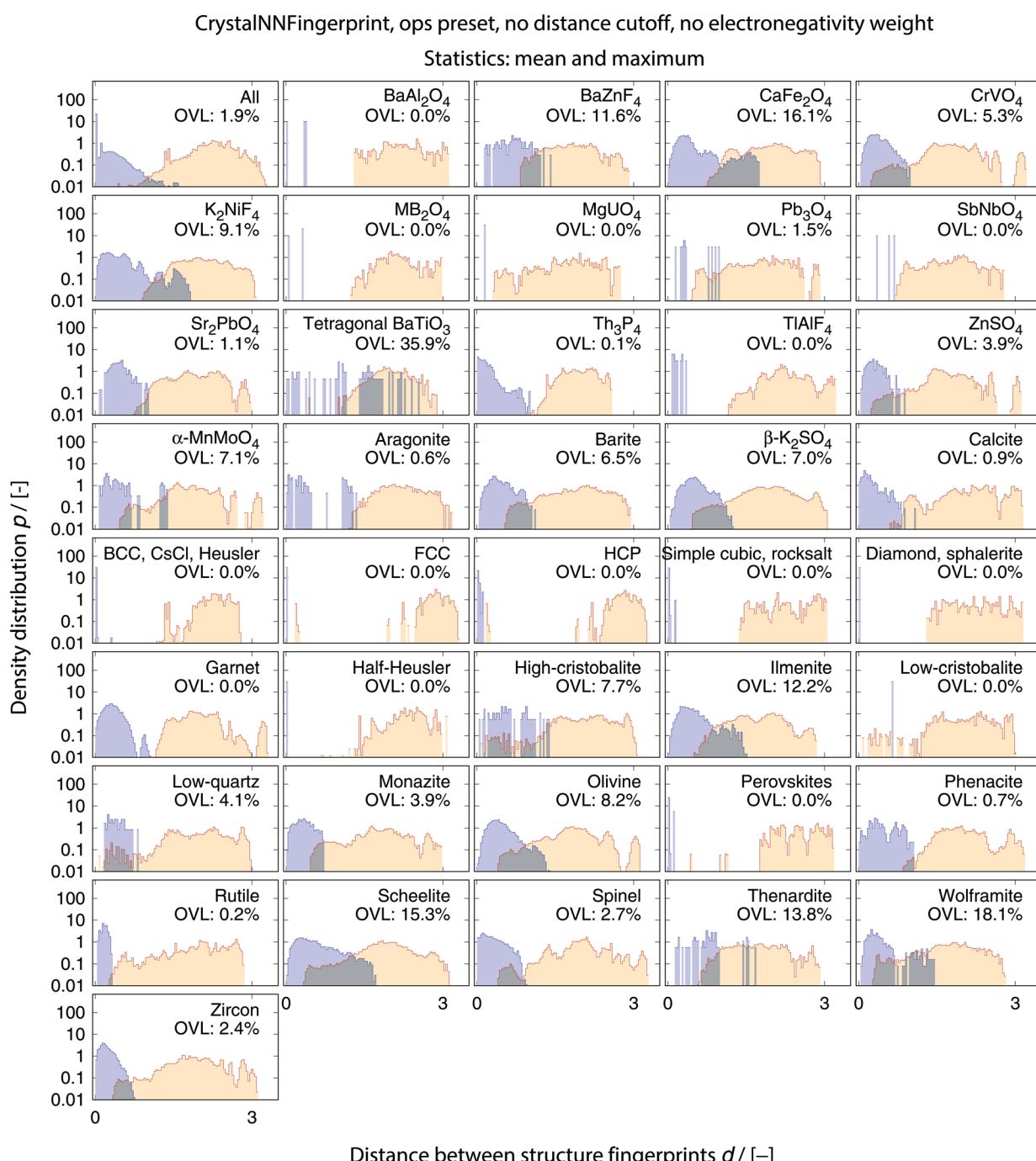
thus, underlining that the oxygen atoms have no distinctly identifiable coordination motif. The distorted  $\text{AlO}_6$  octahedra are clearly represented by  $\max(w_{\text{CN}=6}) = 1$  and  $\max(q_{\text{oct}}|\text{CN} = 6) = 0.83$ .



**Fig. 11** Overlapping coefficient, OVL, as a function of (i) the underlying site fingerprint and (ii) the statistics run across all site fingerprints in a structure, both together centrally define a given structure fingerprint, and (iii) as obtained by using the cosine similarity,  $s_{\text{cos}}$ , (A) and the distance (dis)similarity metric,  $d$ , (B) for comparing two structure, respectively. Note that "CrystalNN, cn" and "CrystalNN, ops" refer to the CrystalNNFingerprint in matminer using preset "cn" and "ops," respectively, an additional asterisk indicates that the two available fingerprint options are set to none, "ChemEnvSite, simple" and "ChemEnvSite, multi-weight" refer to the ChemEnvSiteFingerprint in matminer using the "simple" and "multi-weight" preset, respectively, "OPSite" and "OPSite, bond-OPs" refer to the OPSiteFingerprint with the default OP set and using bond-orientational order parameters<sup>105</sup> for all coordination numbers tested (i.e., CN = 1, ..., 12), and "std. dev.", "min.", as well as "max." refer to standard deviation, minimum, and maximum, respectively.

Ultimately, ideal perovskite materials such as  $\text{CaTiO}_3$  have Ca atoms that are 12-fold coordinated by oxygen atoms in a cuboctahedron fashion, perfect  $\text{TiO}_6$  octahedra, and oxygen atoms that are considered 6-fold coordinated (2 Ti atoms are nearest neighbors forming linear bonds, but the 4 next nearest Ca neighbors are also considered coordinated to oxygen

atoms).<sup>59</sup> All these features are nicely reproduced by our structure fingerprint. 60% of the sites (oxygen atoms) can to some degree ( $w_{\text{CN}=2} \approx 0.5$ ) be considered 2-fold coordinated ( $0.6 \times w_{\text{CN}=2} \approx \bar{w}_{\text{CN}=2} = 0.31$ ). And, we can infer the linear bond character of these sites by comparing  $\max(w_{\text{CN}=2})$  and  $\max(q_{\text{lin}}|\text{CN} = 2)$  values. The perfect  $\text{TiO}_6$  octahedron results in



**Fig. 12** Distributions of distances between structure fingerprints,  $\|\mathbf{v}_{\text{struct},i} - \mathbf{v}_{\text{struct},j}\|$ , for measuring dissimilarities between structures that belong to the same prototype group (blue) and between structures that belong to different prototype groups (orange) for the optimal fingerprint settings, etc. The overlapping coefficients (OVLs), which quantify the overlap between the two distributions in each panel, are also provided. The top left panel represents results averaged over data from all prototype structure groups, whereas the remaining panels display the results for a specific target prototype structure group.



$\max(w_{CN=6}) = \max(q_{oct}|CN = 6) = 1$ , whereas the corresponding mean values of  $\approx 0.5$  indicate that more sites than the Ti atoms possess distinct octahedral coordination geometry: the oxygen atoms. Finally,  $\bar{w}_{CN=12} = 0.2$  and  $\max(w_{CN=12}) = 1$  clearly highlight the 20% of sites (Ca) that are perfectly 12-fold coordinated, and  $\bar{q}_{cuboct}|CN = 12 = 0.2$  and  $\max(q_{cuboct}|CN = 12) = 1$  that the coordination geometry is a perfect cuboctahedron.

### 3.3 Discussion

Two aspects of our work require additional discussion: our novel neighbor-finding algorithm “CrystallNN” (CNN) and the metric used to assess similarity—especially between structures.

**3.3.1 Novel neighbor-finding method.** The structure similarity analysis suggests that CNN is a valuable new way of finding neighbors and computing coordination numbers. It yields better structure similarity assessments than the minimum-distance near-neighbor finding method used in the OPSite structure fingerprint and the conventional Voronoi method used by the ChemEnvSite fingerprint (Fig. 11). Here, it does not matter whether or not to use our new order parameters for structure similarity assessment, given our structure test set. Essentially, the usage of CNN-based coordination likelihoods lends the CrystallNN structure fingerprint its similarity assessment power. Note also that we are currently investigating the performance of CNN for predicting coordination numbers in

greater detail and that CNN frequently outperforms conventional approaches in the upcoming work.

We introduced, to the best of our knowledge, a new element to the neighbor-finding issue: the coordination likelihood computation using the normalized Voronoi weights in conjunction with a semicircle. We invented this approach in an adaptive data analysis<sup>116</sup> fashion. Given a small number of pathological cases, we optimized our algorithm *via* trial-and-error. We stress here however that we did not use the entire benchmark test set for structure similarity assessment (Fig. 10). Therefore, the algorithm development procedure resembled a model selection process, where we took care about not using the same test set for testing.

The new element of our method improved structure similarity assessment. However, we have to underscore that the approach might appear purely algorithmic in nature. It lacks any theory or physical model justifying, for example, the specific use of the semicircle. This can be interpreted as a conceptual disadvantage over more direct or intuitive approaches such as the minimum distance neighbor-finding method.

**3.3.2 Similarity metric issue.** The results of the site fingerprints clearly indicate that dot product-based similarity metrics are advantageous for more stringently distinguishing dissimilar coordination environments. The situation is more nuanced in the case of the structure fingerprints because the overlapping coefficient (OVL) suggests that the distance (dis)

Table 2 Examples of structure fingerprints

Structure	Leading fingerprint elements
Diamond <sup>a</sup>	$\bar{w}_{CN=4} = 1$ $\bar{q}_{see\_saw} CN = 4 = 0.23$
Rocksalt <sup>a</sup>	$\bar{w}_{CN=6} = 1$ $\bar{q}_{hex\_plan} CN = 6 = 0.2$
$\alpha$ -iron <sup>a</sup>	$\bar{w}_{CN=8} = 0.58$ $\bar{w}_{CN=14} = 0.42$
Heusler (Cu <sub>2</sub> MnSn)	$\bar{w}_{CN=4} = 0.09$ $\max(q_{tet} CN = 4) = 0.18$ $\bar{q}_{tri\_pyr} CN = 4 = 0.02$ $\max(q_{sq\_plan} CN = 4) = 0.01$ $\bar{q}_{bcc} CN = 8 = 0.54$ $\max(q_{hex\_bipy} CN = 8) = 0.26$
Spinel (MgAgO <sub>3</sub> )	$\bar{w}_{CN=1} = 0.01$ $\max(w_{CN=4}) = 1$ $\bar{q}_{tri\_pyr} CN = 4 = 0.34$ $\max(q_{see\_saw} CN = 4) = 0.52$ $\bar{q}_{sq\_plan} CN = 4 = 0.17$ $\max(w_{CN=6}) = 1$ $\bar{q}_{pent\_pyr} CN = 6 = 0.13$ $\max(q_{hex\_plan} CN = 6) = 0.2$
Perovskite (CaTiO <sub>3</sub> )	$\bar{w}_{CN=2} = 0.31$ $\max(q_{lin} CN = 2) = 0.52$ $\bar{w}_{CN=6} = 0.49$ $\max(q_{oct} CN = 6) = 1$ $\bar{q}_{hex\_plan} CN = 6 = 0.1$ $\max(w_{CN=12}) = 1$ $\bar{q}_6 CN = 12 = 0.12$ $\max(q_4 CN = 12) = 0.19$
	$\bar{q}_{tet} CN = 4 = 1$ $\bar{q}_{sq\_plan} CN = 4 = 0.08$ $\bar{q}_{oct} CN = 6 = 1$  $\bar{q}_{bcc} CN = 8 = 0.58$  $\max(w_{CN=4}) = 0.18$ $\bar{q}_{see\_saw} CN = 4 = 0.02$ $\max(q_{tri\_pyr} CN = 4) = 0.04$ $\bar{w}_{CN=8} = 0.54$ $\max(q_{bcc} CN = 8) = 0.6$ $\bar{w}_{CN=14} = 0.37$ $\max(w_{CN=1}) = 0.02$ $\bar{q}_{tet} CN = 4 = 0.43$ $\max(q_{tri\_pyr} CN = 4) = 0.53$ $\bar{q}_{rect\_see\_saw} CN = 4 = 0.23$ $\max(q_{sq\_plan} CN = 4) = 0.28$ $\bar{q}_{oct} CN = 6 = 0.24$ $\max(q_{pent\_pyr} CN = 6) = 0.44$ $\max(w_{CN=8}) = 0.01$ $\max(w_{CN=2}) = 0.52$ $\bar{q}_{150^\circ} CN = 2 = 0.04$ $\max(w_{CN=6}) = 1$ $\bar{q}_{pent\_pyr} CN = 6 = 0.24$ $\max(q_{hex\_plan} CN = 6) = 0.2$ $\bar{q}_{cuboct} CN = 12 = 0.2$ $\max(q_6 CN = 12) = 0.58$  $\bar{q}_{tri\_pyr} CN = 4 = 0.25$ $\bar{q}_{see\_saw\_rect} CN = 4 = 0.01$ $\bar{q}_{pent\_pyr} CN = 6 = 0.5$  $\bar{q}_{hex\_bipy} CN = 8 = 0.25$  $\bar{q}_{tet} CN = 4 = 0.09$ $\max(q_{see\_saw} CN = 4) = 0.04$ $\bar{q}_{sq\_plan} CN = 4 = 0.01$ $\max(w_{CN=8}) = 0.6$ $\bar{q}_{hex\_bipy} CN = 8 = 0.24$ $\max(w_{CN=14}) = 0.6$ $\bar{w}_{CN=4} = 0.7$ $\max(q_{tet} CN = 4) = 1$ $\bar{q}_{see\_saw} CN = 4 = 0.33$ $\max(q_{rect\_see\_saw} CN = 4) = 0.4$ $\bar{w}_{CN=6} = 0.29$ $\max(q_{oct} CN = 6) = 0.82$ $\bar{q}_{hex\_plan} CN = 6 = 0.06$ $\max(q_{bcc} CN = 8) = 0.01$ $\bar{q}_{lin} CN = 2 = 0.31$ $\max(q_{150^\circ} CN = 2) = 0.06$ $\bar{q}_{oct} CN = 6 = 0.49$ $\max(q_{pent\_pyr} CN = 6) = 0.5$ $\bar{w}_{CN=12} = 0.2$ $\max(q_{cuboct} CN = 12) = 1$ $\bar{q}_4 CN = 12 = 0.04$

<sup>a</sup> We omit the maximum values of the fingerprint for sake of brevity because those values were identical to the mean values.



similarity metric gives the best performance. However, this outcome should be viewed cautiously.

On an average, the cosine similarity yields an OVL that is  $\approx 0.01$  smaller (or  $\approx 11\%$ ) than the distance (dis)similarity. Using only the CNN data ("CrystalNN\*", ops" and "CrystalNN\*", cn" in Fig. 11), the trend is opposite ( $\approx -0.001$ ,  $\approx -7\%$ ) but considerably weaker. Hence, the (dis)similarity metric does not significantly impact the outcome of the best performing group of fingerprints that we have considered. This is confirmed by comparing results of individual material prototype groups of the best combination of fingerprint and (dis)similarity metric (Fig. 12) with the same fingerprint but cosine similarity metric (ESI: Fig. 26†). The OVLs are usually almost identical. There are in fact only two cases (tetragonal BaTiO<sub>3</sub> and ilmenite) for which the OVLs differ more evidently ( $\approx -5\%$  and  $\approx +4\%$ , respectively). However, the corresponding OVLs are also exceptionally large (36% and 12%, respectively) so that we conclude that even these few extreme cases represent only moderate discrepancies between the two (dis)similarity metrics. Moreover, distribution features such as fragmentation (tetragonal BaTiO<sub>3</sub>) and second similarity peaks (aragonite) are similarly clearly observable with both (dis)similarity metrics.

Because normalization is very desirable in most machine learning approaches,<sup>117</sup> we conclude that the cosine similarity metric is—at the very least—an often unjustifiably ignored metric. We can solidly recommend to use the cosine similarity as an *ad-hoc* assessment metric for computational materials science applications.

## 4 Conclusions

Despite the facts that coordination chemistry was already established more than a century ago<sup>118</sup> and Brunner, more than 4 decades ago, still realized that "the term coordination ha[d] no satisfying definition,"<sup>87</sup> determination and evaluation of coordinated neighbors in crystal structures is even today continuing to be a nontrivial scientific task—let alone automated versions of those processes. To aid in filling this gap, we have, in this work, introduced a new neighbor finding method and several new local structure order parameters (LoStOPs). We have used a Monte Carlo framework to maximize comparability between different LoStOPs. Subsequently, we have employed the new descriptors to define feature vectors that are characteristic of known coordination environments (site fingerprints and similarities) and that are capable of distinguishing commonly investigated crystalline prototype materials (structure fingerprints and similarities). In-depth testing has enabled us to give recommendations which type of fingerprint should be combined with which similarity metric in order to most reliably categorize site environments and crystal structures. We actively utilize our novel capabilities on the Materials Project<sup>11</sup> website<sup>12</sup> to assess the crystal structure similarity between different materials. To this end, we have computed all  $\approx 2.45$  billion structure similarity distances between each pair of the  $\approx 70\,000$  materials in the MP database. The structure similarity data greatly facilitate browsing of the website and, thus, exploration

of the MP materials database because it adds an intuitive connection mechanism.

We hope that our novel coordination descriptors will be helpful in the context of inverse design approaches,<sup>119,120</sup> where, specifically, the distinction between perfect and slightly distorted coordination environments<sup>121</sup> or building blocks<sup>27,28</sup> is decisive for the target property. Finally, we believe that the here introduced methods and concepts, which are freely available through the python<sup>74</sup> packages pymatgen<sup>75,76</sup> and matminer,<sup>65,77</sup> will greatly facilitate future data-driven and machine learning studies at the interface between materials science, chemistry, and engineering, for example, by producing standardized metadata.<sup>122</sup>

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

This work was intellectually led by the U. S. Department of Energy (DOE) Basic Energy Sciences (BES) program—the Materials Project—under Grant No. KC23MP. We thank Donny Winston and Patrick Huck for support to generate and integrate data on structure similarities on the Materials Project website. Furthermore, we thank Özlem Çavuşoğlu for the suggestion to take the square root of the dot product to make units consistent. 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.

## Notes and references

- 1 F. H. Allen, S. Bellard, M. D. Brice, B. A. Cartwright, A. Doubleday, H. Higgs, T. Hummelink, B. G. Hummelink-Peters, O. Kennard, W. D. S. Motherwell, J. R. Rodgers and D. G. Watson, *Acta Crystallogr., Sect. B: Struct. Sci.*, 1979, **35**, 2331–2339.
- 2 G. Bergerhoff, R. Hundt, R. Sievers and I. D. Brown, *J. Chem. Inf. Comput. Sci.*, 1983, **23**, 66–69.
- 3 A. Belsky, M. Hellenbrandt, V. L. Karen and P. Luksch, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2002, **58**, 364–369.
- 4 G. H. Wood, J. R. Rodgers and S. R. Gough, *J. Chem. Inf. Comput. Sci.*, 1993, **33**, 31–35.
- 5 P. S. White, J. R. Rodgers and Y. Le Page, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2002, **58**, 343–348.
- 6 P. Villars, M. Berndt, K. Brandenburg, K. Cenzual, J. Daams, F. Hulliger, T. Massalski, H. Okamoto, K. Osaki, A. Prince, H. Putz and S. Iwata, *PAULING FILE, Binaries Edition*, ASM International, Materials Park, Ohio, U.S.A., 1st edn, 2002.
- 7 P. Villars, M. Berndt, K. Brandenburg, K. Cenzual, J. Daams, F. Hulliger, T. Massalski, H. Okamoto, K. Osaki, A. Prince, H. Putz and S. Iwata, *J. Alloys Compd.*, 2004, **367**, 293–297.

8 S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck and A. Le Bail, *J. Appl. Crystallogr.*, 2009, **42**, 726–729.

9 R. T. Downs and M. Hall-Wallace, *Am. Mineral.*, 2003, **88**, 247–250.

10 K. Villars and P. Cenzual, *Pearson's Crystal Data: Crystal Structure Database for Inorganic Compounds*, ASM International, Materials Park, Ohio, U.S.A., 1st edn, 2017.

11 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.

12 Materials Project website, <http://materialsproject.org>.

13 S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo and O. Levy, *Comput. Mater. Sci.*, 2012, **58**, 227–235.

14 J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.

15 J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, *J. Miner. Met. Mater. Soc.*, 2013, **65**, 1501–1509.

16 C. Draxl and M. Scheffler, *MRS Bull.*, 2018, **43**, 676–682.

17 Web of Science, <https://webofknowledge.com>.

18 T. Heine, *Front. Mater.*, 2014, **1**, 7.

19 N. Wagner and J. M. Rondinelli, *Front. Mater.*, 2016, **3**, 28.

20 B. R. Goldsmith, M. Boley, J. Vreeken, M. Scheffler and L. M. Ghiringhelli, *New J. Phys.*, 2017, **19**, 013031.

21 A. Jain, G. Hautier, S. P. Ong and K. Persson, *J. Mater. Res.*, 2016, **31**, 977–994.

22 J. Lee, A. Seko, K. Shitara, K. Nakayama and I. Tanaka, *Phys. Rev. B*, 2016, **93**, 115104.

23 T. Lookman, P. V. Balachandran, D. Xue, J. Hogden and J. Theiler, *Curr. Opin. Solid State Mater. Sci.*, 2017, **21**, 121–128.

24 J. Shi, T. F. T. Cerqueira, W. Cui, F. Nogueira, S. Botti and M. A. L. Marques, *Sci. Rep.*, 2017, **7**, 43179.

25 T. Krishnamoorthy, H. Ding, C. Yan, W. L. Leong, T. Baikie, Z. Zhang, M. Sherburne, S. Li, M. Asta, N. Mathews and S. G. Mhaisalkar, *J. Mater. Chem. A*, 2015, **3**, 23829–23832.

26 G. Pilania, J. E. Gubernatis and T. Lookman, *Comput. Mater. Sci.*, 2017, **129**, 156–163.

27 N. E. R. Zimmermann and M. Haranczyk, *Cryst. Growth Des.*, 2016, **16**, 3043–3048.

28 J. L. Salcedo Perez, M. Haranczyk and N. E. R. Zimmermann, *Z. Kristallogr. - Cryst. Mater.*, 2019, **234**, 437–450.

29 M. A. Shandiz and R. Gauvin, *Comput. Mater. Sci.*, 2016, **117**, 270–278.

30 F. Faber, A. Lindmaa, O. A. von Lilienfeld and R. Armiento, *Int. J. Quantum Chem.*, 2015, **115**, 1094–1101.

31 L. Ward, A. Agrawal, A. Choudhary and C. A. Wolverton, *npj Comput. Mater.*, 2016, **2**, 16028.

32 A. D. Sendek, Q. Yang, E. D. Cubuk, K.-A. N. Duerloo, Y. Cuic and E. J. Reed, *Energy Environ. Sci.*, 2017, **10**, 306–320.

33 M. Ashton, J. Paul, S. B. Sinnott and R. G. Hennig, *Phys. Rev. Lett.*, 2017, **118**, 106101.

34 G. Cheon, K.-A. N. Duerloo, A. D. Sendek, C. Porter, Y. Chen and E. J. Reed, *Nano Lett.*, 2017, **17**, 1915–1923.

35 C. Y. Lau, M. T. Dunstan, W. Hu, C. P. Grey and S. A. Scott, *Energy Environ. Sci.*, 2017, **10**, 818.

36 K. Choudhary, I. Kalish, R. Beams and F. Tavazza, *Sci. Rep.*, 2017, **7**, 5179.

37 C. Kim, A. Chandrasekaran, T. D. Huan, D. Das and R. Ramprasad, *J. Phys. Chem. C*, 2018, **122**, 17575–17585.

38 E. Kim, K. Huang, S. Jegelka and E. Olivetti, *npj Comput. Mater.*, 2017, **3**, 53.

39 E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder and E. Olivetti, *Chem. Mater.*, 2017, **29**, 9436–9444.

40 S. M. Moosavi, A. Chidambaram, L. Talirz, M. Haranczyk, K. C. Stylianou and B. Smit, *Nat. Commun.*, 2019, **10**, 539.

41 M. H. S. Segler and M. P. Waller, *Chem.-Eur. J.*, 2017, **23**, 5966–5971.

42 B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, J. Ho, S. Sloane, P. Wender and V. Pande, *ACS Cent. Sci.*, 2017, **3**, 1103–1113.

43 J. Behler, *J. Chem. Phys.*, 2016, **145**, 170901.

44 V. Botu, R. Batra, J. Chapman and R. Ramprasad, *J. Phys. Chem. C*, 2016, **121**, 511–522.

45 K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller and A. Tkatchenko, *Nat. Commun.*, 2017, **8**, 13890.

46 T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen and R. Ramprasad, *npj Comput. Mater.*, 2017, **3**, 37.

47 P. Hohenberg and W. Kohn, *Phys. Rev. B: Solid State*, 1964, **136**, B864–B871.

48 W. Kohn and L. J. Sham, *Phys. Rev. A*, 1965, **140**, A1133–A1138.

49 G. Kresse and J. Hafner, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1993, **47**, 558–561.

50 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

51 F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke and K.-R. Müller, *Nat. Commun.*, 2017, **8**, 872.

52 A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf and M. Ceriotti, *ACS Cent. Sci.*, 2018, **5**, 57–64.

53 N. E. R. Zimmermann, D. C. Hannah, Z. Rong, M. Liu, G. Ceder, M. Haranczyk and K. A. Persson, *J. Phys. Chem. Lett.*, 2018, **9**, 628–634.

54 G. P. P. Pun, R. Batra, R. Ramprasad and Y. Mishin, *Nat. Commun.*, 2019, **10**, 2339.

55 A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen and R. Ramprasad, *npj Comput. Mater.*, 2019, **5**, 22.

56 A. O. Oliynyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig and A. Mar, *Chem. Mater.*, 2016, **28**, 7324–7331.

57 A. Ziletti, D. Kumar, M. Scheffler and L. M. Ghiringhelli, *Nat. Commun.*, 2018, **9**, 2775.

58 R. Batra, H. D. Tran, C. Kim, J. Chapman, L. Chen, A. Chandrasekaran and R. Ramprasad, *J. Phys. Chem. C*, 2019, **123**, 15859–15866.

59 O. Muller and R. Roy, *The Major Ternary Structural Families*, Springer-Verlag, Berlin, Germany, 1st edn, 1974.

60 M. J. Mehl, D. Hicks, C. Toher, O. Levy, R. M. Hanson, G. Hart and S. Curtarolo, *Comput. Mater. Sci.*, 2017, **136**, S1–S28.

61 L. Bindi and J. Plášil, *Z. Kristallogr.*, 2018, **233**, 153–154.

62 D. Hicks, M. J. Mehl, E. Gossett, C. Toher, O. Levy, R. M. Hanson, G. Hart and S. Curtarolo, *Comput. Mater. Sci.*, 2019, **161**, S1–S1011.

63 T. Mueller, A. G. Kusne and R. Ramprasad, in *Machine learning in materials science: recent progress and emerging applications*, John Wiley & Sons, Inc., 2016, pp. 186–273.

64 R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi and C. Kim, *npj Comput. Mater.*, 2017, **3**, 54.

65 L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster and A. Jain, *Comput. Mater. Sci.*, 2018, **152**, 60–69.

66 A. Grisafi, D. M. Wilkins, M. J. Willatt and M. Ceriotti, arXiv:1904.01623, 2019.

67 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.

68 H. Huo and M. Rupp, arXiv:1704.06439, 1–5, 2017.

69 L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrarwal, A. Choudhary and C. Wolverton, *Phys. Rev. B*, 2017, **96**, 024104.

70 E. L. Willighagen, R. Wehrens, P. Verwer, R. De Gelder and L. M. C. Buydens, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2005, **61**, 29–36.

71 M. Valle and A. R. Oganov, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 2010, **66**, 507–517.

72 L. Zhu, M. Amsler, T. Fuhrer, B. Schaefer, S. Faraji, S. Rostami, S. A. Ghasemi, A. Sadeghi, M. Grauzintyte, C. Wolverton and S. Goedecker, *J. Chem. Phys.*, 2016, 034203.

73 L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl and M. Scheffler, *Phys. Rev. Lett.*, 2015, **114**, 105503.

74 G. Van Rossum, *USENIX annual technical conference*, 2007, p. 36.

75 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Comput. Mater. Sci.*, 2013, **68**, 314–319.

76 pymatgen's GitHub repository, <https://github.com/materialsproject/pymatgen>.

77 matminer's GitHub repository, <https://github.com/hackingmaterials/matminer>.

78 G. L. Dirichlet, *J. Reine Angew. Math.*, 1850, **40**, 209–227.

79 G. Voronoi, *J. Reine Angew. Math.*, 1908, **133**, 97–178.

80 N. E. R. Zimmermann, M. K. Horton, A. Jain and M. Haranczyk, *Front. Mater.*, 2017, **4**, 34.

81 B. Peters, *J. Chem. Phys.*, 2009, **131**, 244103.

82 N. E. R. Zimmermann, B. Vorselaars, D. Quigley and B. Peters, *J. Am. Chem. Soc.*, 2015, **137**, 13352–13361.

83 J. Cumby and J. P. Attfield, *Nat. Commun.*, 2019, **8**, 14235.

84 M. O'Keeffe and N. E. Brese, *J. Am. Chem. Soc.*, 1991, **113**, 3226–3229.

85 R. Hoppe, *Z. Kristallogr.*, 1979, **150**, 23–52.

86 M. O'Keeffe, *Acta Crystallogr., Sect. A: Cryst. Phys., Diffraction Theory. Gen. Crystallogr.*, 1979, **35**, 772–775.

87 G. O. Brunner, *Acta Crystallogr., Sect. A: Cryst. Phys., Diffraction Theory. Gen. Crystallogr.*, 1977, **33**, 226–227.

88 Jmol: an open-source Java viewer for chemical structures in 3D, <http://www.jmol.org/>.

89 Y. Li, J. Yu and R. Xu, *Angew. Chem., Int. Ed.*, 2013, **52**, 1673–1677.

90 M. Mazur, P. S. Wheatley, M. Navarro, W. J. Roth, M. Položij, A. Mayoral, P. Eliášová, P. Nachtigall, J. Čejka and R. E. Morris, *Nat. Chem.*, 2015, **8**, 58–62.

91 T. Xie and J. C. Grossman, *J. Chem. Phys.*, 2018, **149**, 174111.

92 M. K. Horton, J. H. Montoya, M. Liu and K. A. Persson, *npj Comput. Mater.*, 2019, **5**, 64.

93 M. Atiyah and P. Sutcliffe, *Milan J. Math.*, 2003, **1**, 33–58.

94 N. W. Johnson, *Can. J. Math.*, 1966, **18**, 169–200.

95 G. E. P. Box and M. E. Muller, *Ann. Math. Stat.*, 1958, **29**, 610–611.

96 S. van der Walt, S. C. Colbert and G. Varoquaux, *Comput. Sci. Eng.*, 2011, **13**, 22–30.

97 A. Einstein, *Ann. Phys.*, 1906, **22**, 180–190.

98 D. Frenkel and A. J. C. Ladd, *J. Chem. Phys.*, 1984, **81**, 3188–3193.

99 J. L. Aragones, E. Sanz and C. Vega, *J. Chem. Phys.*, 2012, **136**, 244508.

100 D. Waroquiers, X. Gonze, G.-M. Rignanese, C. Welker-Nieuwoudt, F. Rosowski, M. Göbel, S. Schenk, P. Degelmann, R. André, R. Glaum and G. Hautier, *Chem. Mater.*, 2017, **29**, 8346–8360.

101 M. Pinsky and D. Avnir, *Inorg. Chem.*, 1998, **37**, 5575–5582.

102 H. F. Inman and E. L. Bradley Jr, *Commun. Stat. Theor. Methods*, 1989, **18**, 3851–3874.

103 J. Lima-de Faria, E. Hellner, F. Liebau, E. Makovicky and E. Parthé, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 1990, **46**, 1–11.

104 R. M. Hartshorn, E. Hey-Hawkins, R. Kalio and G. J. Leigh, *Pure Appl. Chem.*, 2007, **79**, 1779–1799.

105 P. J. Steinhardt, D. R. Nelson and M. Ronchetti, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1983, **28**, 784–805.

106 G. J. Ackland and A. P. Jones, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2006, **73**, 054104.

107 W. Mickel, S. C. Kapfer, G. E. Schröder-Turk and K. Mecke, *J. Chem. Phys.*, 2013, **138**, 044501.

108 A. M. Ganose and A. Jain, *MRS Commun.*, 2019, **9**, 874–881.

109 L. Pauling, *J. Am. Chem. Soc.*, 1929, **51**, 1010–1026.

110 G. Schwarz, *Ann. Stat.*, 1978, **6**, 461–464.

111 H. Akaike, *IEEE Trans. Autom. Control*, 1974, **19**, 716–723.

112 W. H. Hull and W. L. Bragg, *Proc. R. Soc. London, Ser. A*, 1913, **33**, 277.

113 W. H. Bragg and W. L. Bragg, *Proc. R. Soc. London, Ser. A*, 1913, **88**, 428.

114 A. W. Hull, *Phys. Rev.*, 1917, **10**, 661–696.

115 A. J. Bradley and J. W. Rodgers, *Proc. R. Soc. London, Ser. A*, 1934, **144**, 340–359.

116 C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold and A. Roth, *Science*, 2015, **349**, 636–638.

117 S. Raschka, *Python Machine Learning*, Packt Publishing, Birmingham, UK, 1st edn, 2016.

118 A. Werner, *Ber. Dtsch. Chem. Ges.*, 1912, **45**, 121–130.

119 K. Huwig, C. Fan and M. Springborg, *J. Chem. Phys.*, 2017, **147**, 234105.

120 B. Kim, S. Lee and J. Kim, *Sci. Adv.*, 2020, **6**, eaax9324.

121 W. S. Drisdell, L. Leppert, C. M. Sutter-Fella, Y. Liang, Y. Li, Q. P. Ngo, S. Wan, L. F. Gul, T. Kroll, D. Sokaras, A. Javey, J. Yano, J. B. Neaton, F. M. Toma, D. Prendergast and I. D. Sharp, *ACS Energy Lett.*, 2017, **2**, 1183–1189.

122 L. Himanen, A. Geurts, A. S. Foster and P. Rinke, *Adv. Sci.*, 2019, **6**, 1900808.