Benchmarking Coordination Number Prediction Algorithms on Inorganic Crystal Structures

Hillary Pan,[†,¶] Alex M. Ganose,[†,¶] Matthew Horton,[†,‡] Muratahan Aykol,[†] Kristin Persson,[†,‡] Nils E.R. Zimmermann,[*,†] and Anubhav Jain[*,†]

†*Lawrence Berkeley National Laboratory, Energy Technologies Area, 1 Cyclotron Road, Berkeley, CA 94720, United States*

‡*Department of Materials Science & Engineering, University of California, Berkeley, United States*

¶*Equal contribution*

E-mail: nils.e.r.zimmermann@gmail.com; ajain@lbl.gov

Abstract

Coordination numbers and geometries form a theoretical framework for understanding and predicting materials properties. Algorithms to determine coordination numbers automatically are increasingly used for machine learning and automatic structural analysis. In this work, we introduce MaterialsCoord, a benchmark suite containing 56 experimentally-derived crystal structures (spanning elements, binaries, and ternary compounds) and their corresponding coordination environments as described in the research literature. We also describe CrystalNN, a novel algorithm for determining near neighbors. We compare CrystalNN against 7 existing near-neighbor algorithms on the MaterialsCoord benchmark, finding CrystalNN to perform similarly to several well-established algorithms. For each algorithm, we also assess computational demand and sensitivity towards small perturbations that mimic thermal motion. Finally, we

investigate the similarity between bonding algorithms when applied to the Materials Project database. We expect that this work will aid the development of coordination prediction algorithms as well as improve structural descriptors for machine learning and other applications.

Introduction

Coordination numbers and geometries (*e.g.*, tetrahedral, octahedral, trigonal planar) play a fundamental role in describing materials and dictating their properties. Some well-known examples throughout materials science include: (i) the local coordination of a site can predict the type of orbital interactions and crystal field splitting; (ii) the feasibility of hypothetical zeolites for catalysis, gas separation, or ion-exchange[1] is frequently assessed by the distortion of the tetrahedral SiO_4 building blocks;[2,3] (iii) in battery materials, diffusion path topologies can be classified using the coordination geometries of the diffusing ions;[4,5] (iv) the relative arrangement of octahedral Pb-halide motifs significantly influences the electronic properties of hybrid organic-inorganic halide perovskites.[6]

The primary challenge is to determine which atoms in the crystal are connected or bonded to one another and which are not. Although the definition of what constitutes a bonding interaction can be debated, in practice, assigning neighbors and thus coordination numbers for most crystals is typically intuitive for an expert in the field. However, manually assigning coordination numbers on a larger scale, say for tens of thousands of atoms, is impractical and therefore requires an automated approach. Machine learning (ML) of materials properties, where descriptions of the coordination environments of atoms can be important, is increasingly becoming an essential tool in the materials discovery process[7–9] and has been enabled by the large amounts of data provided by materials databases.[10–12] Coordination numbers have been used to predict formation enthalpies,[13] examine magnetic materials,[14] and as the basis of crystal graphs in convolutional graph-based neural networks.[15,16] Automated coordination number determination has also allowed researchers to reassess conventional rules

about the crystal structures of materials.[17] Accordingly, an ongoing challenge in materials science has been the development of reliable methods for determining the coordination numbers of atoms in crystal structures.

Various coordination number definitions have already been proposed. These definitions are typically based on interatomic distances or geometric principles. The former includes those proposed by Brunner,[18] O'Keeffe and Brese[19], and Hoppe[20]. Brunner suggested a cut-off system, in which coordination is determined by considering the largest reciprocal gap in interatomic distances. Hoppe developed a coordination number definition based on structure, whereas O'Keeffe and Brese proposed that near-neighbor atoms be determined by sums of bond valences. O'Keeffe also developed another approach using geometric principles in which atoms that share a Voronoi polyhedral face are considered coordinated to each other.[21] More recent coordination number predictions are often modified versions of these definitions. For instance, the valence-ionic radius estimator (VIRE) approach[22] takes oxidation state estimations along with coordination number estimations from Voronoi tessellations[23] to predict coordination environments. Despite the plethora of available methods, a rigorous framework for evaluating the performance of coordination algorithms does not, to our knowledge, exist. Consequently, a universal tried-and-tested approach for determining atomic coordination has not been established.

In this work, we introduce a benchmarking framework, MaterialsCoord, to compare near-neighbor finding algorithms using a diverse dataset composed of experimentally-determined structures from the Inorganic Crystal Structure Database (ICSD).[24] The MaterialsCoord dataset relies on literature descriptions of coordination environments in these structures to assign coordination numbers. We evaluate a new approach, crystal-near-neighbor (CrystalNN), which uses Voronoi decomposition and solid angle weights to determine coordination environments. We compare CrystalNN against existing near-neighbor algorithms using the MaterialsCoord benchmark. Algorithms are evaluated on the basis of: i) ability to reproduce literature descriptions of coordination numbers across a diverse range of structures, ii) sen-

sitivity towards small perturbations introduced to each crystal structure, and iii) the time taken to perform the analysis. We quantify the similarity between bonding algorithms using Jaccard distance plots applied to the Materials Project database.[10] Software implementations for all near-neighbor finding algorithms are available in the pymatgen library.[22]

Methods

Near-Neighbor Finding Algorithms

We first describe the near-neighbor finding methods evaluated in this work, all of which are implemented in the `local_env` module of the pymatgen library.[22] The pymatgen class for each implementation is given in parentheses and is used as an identifier throughout this work. Algorithms are split into two groups: the first five algorithms discussed are distance-based approaches and the rest are based on or involve Voronoi decomposition. We use the abbreviation CN to denote coordination number (*i.e.*, the number of "near neighbors" expected to participate in some kind of bonding interaction) and NN to denote "near-neighbor" finding algorithm. For consistency, we use the default value of each tolerance parameter, δ, for each algorithm provided in pymatgen.[22] In Sections S1 and S2 of the Supporting Information, we also introduce and benchmark the ToposPro AutoCN algorithm and the modified Voronoi approach outlined by Isayev et al.[25]. We note that ToposPro is a proprietary method that cannot be easily automated and only runs on the Windows operating system. We have thus run a manual analysis over the benchmark set for reference, but do not find it suitable for automated analyses. We find its overall score to be competitive with the best algorithms studied in this work (overall score of 9.7, see Section S1 of the Supporting Information).

One important comment about the near neighbor methods discussed in this work is that in many cases the coordination is not reciprocal by default. That is, if site A is coordinated to site B, it is not guaranteed that site B will be coordinated to site A. Thus, in practice we consider A and B to be neighbors if either condition holds — *i.e.*, either A has B as a neighbor

or B has A as a neighbor. Further information on the symmetry of bonding behavior for various algorithms is provided in Section S3 and Figure S4 of the Supporting Information. Furthermore, we note that all algorithms discussed in this work assign coordination that does not alter the original symmetry of the structure.

Minimum Distance Method

The simplest algorithm evaluated in this work (MinimumDistanceNN) determines the coordination of a site, i, based on the distance, $d_i^{\min}$, to the closest nearest neighbor site. Other neighboring sites are considered bonded neighbors if they fall within a cut-off, d_i^{cut}, defined as

$$d_i^{\text{cut}} = (1 + \delta) d_i^{\min}, \tag{1}$$

where δ is a (relative) tolerance parameter. This tolerance parameter was previously optimized by Zimmermann et al.[26] for detecting various coordination motifs in a database of 1,025 test structures; we use the suggested value of 0.1 for this parameter.

Emulation of Jmol's autoBond Algorithm

In Jmol,[27] a free, open-source software for visualizing molecules, bonds can be automatically detected using the autoBond algorithm. In this work, we use an emulation of Jmol's algorithm (JmolNN) implemented in pymatgen.[22] Atoms are considered bonded if the distance between them, d_{ij}, is such that

$$d_{ij} \leq r_i + r_j + \delta, \tag{2}$$

where r_i is the elemental radius of the atom at site i, r_j is the elemental radius of the atom at site j, and δ is a tolerance parameter fixed at 0.45 Å. A list of the elemental radii used is detailed elsewhere[28] and is included as part of pymatgen.[22] We note that this algorithm does not take into account oxidation states.

Brunner's Largest Reciprocal Gap Method

Three versions of Brunner's method[18] (BrunnerNN_reciprocal, BrunnerNN_real, and BrunnerNN_relative) are implemented in pymatgen.[22] Brunner's method of largest reciprocal gap (BrunnerNN_reciprocal), however, predicts coordination environments significantly better than the other two algorithms. We thus report the results of BrunnerNN_reciprocal in the main text and refer to this algorithm as BrunnerNN. Coordination number predictions using the other two Brunner algorithms are reported in Section S4 and Figure S5 of the Supporting Information.

Brunner's method[18] (BrunnerNN) chooses the distance cut-off by considering the largest reciprocal gap in interatomic distances from a central site. The equation

$$j^{\max} = \arg\max_{j} \left\{ \frac{1}{d_{ij}} - \frac{1}{d_{i(j+1)}} : j = 1 \dots n \right\} \tag{3}$$

is used to determine the largest reciprocal gap, where d_{ij} and $d_{i(j+1)}$ are the interatomic distances between a central site, i, and the j^{th} and $(j+1)^{\text{th}}$ neighboring sites, ordered in increasing distance from the central site. The distance cut-off for determining coordination is then given by

$$d_i^{\text{cut}} = d_{ij}^{\max} + \delta, \tag{4}$$

where δ is a tolerance parameter set to 0.0001 Å for numerical stability of the procedure.

O'Keeffe's Bond Valence Method

The minimum O'Keeffe algorithm (MinimumOKeeffeNN) determines atomic coordination based on a *minimum relative distance* approach. Here, the relative distance between two atoms, d_{ij}^{rel}, is given by

$$d_{ij}^{\text{rel}} = \frac{d_{ij}}{d_{ij}^{\text{O'Keeffe}}}, \tag{5}$$

where d_{ij} is the interatomic distance between sites i and j, $d_{ij}^{\mathrm{O'Keeffe}}$ is the bond valence parameter,[19] an ideal bond length defined as

$$d_{ij}^{\mathrm{O'Keeffe}} = r_i^{\mathrm{emp}} + r_j^{\mathrm{emp}} - \frac{r_i^{\mathrm{emp}} r_j^{\mathrm{emp}} (\sqrt{c_i} - \sqrt{c_j})^2}{c_i r_i^{\mathrm{emp}} + c_j r_j^{\mathrm{emp}}} \tag{6}$$

where r^{emp} is an empirical "size" parameter[19] based on the atomic radii and c is the electronegativity calculated using the Allred-Rochow scale.[29] Two atoms are considered bonded if

$$d_{ij}^{\mathrm{rel}} \leq (1 + \delta) \times \min\{d_{ij}^{\mathrm{rel}} : j = 1 \ldots n\}, \tag{7}$$

where δ is a tolerance parameter set to 0.1.

Hoppe's Method of Effective Coordination Numbers

The effective coordination number algorithm (EconNN) calculates coordination numbers using Hoppe's effective coordination number formula.[20] In this method, a weighted average bond length, ${}^0d^{\mathrm{avg}}$, is obtained according to

$${}^0d^{\mathrm{avg}} = \frac{\sum_j d_{ij} \exp\left[1 - \left(\frac{d_{ij}}{d_i^{\mathrm{min}}}\right)^6\right]}{\sum_j \exp\left[1 - \left(\frac{d_{ij}}{d_i^{\mathrm{min}}}\right)^6\right]} \tag{8}$$

where d_{ij} is the distance between site i and neighboring site j, and d_i^{min} is the distance from site i to its closest neighbor. To avoid small bond distances biasing the weighted average, an iterative procedure is employed in which ${}^nd^{\mathrm{avg}}$ is calculated according to

$${}^nd^{\mathrm{avg}} = \frac{\sum_j d_{ij} \exp\left[1 - \left(\frac{d_{ij}}{{}^{n-1}d^{\mathrm{avg}}}\right)^6\right]}{\sum_j \exp\left[1 - \left(\frac{d_{ij}}{{}^{n-1}d^{\mathrm{avg}}}\right)^6\right]}. \tag{9}$$

Starting with $n = 1$, d^{avg} is calculated until ${}^{n}d^{\text{avg}} - {}^{n-1}d^{\text{avg}} \leq 0.001$ Å. This procedure always converges, with the final value independent of ${}^{0}d^{\text{avg}}$. Two atoms are considered bonded if

$$\delta \leq \exp\left[1 - \left(\frac{d_{ij}}{{}^{n-1}d^{\text{avg}}}\right)^{6}\right], \tag{10}$$

where δ is a tolerance parameter set to 0.5. We investigate the impact of the tolerance parameter in Section S5 and Figure S5 of the Supporting Information and find that the results are largely insensitive for values from 0.1 - 0.8.

O'Keeffe's Method of Voronoi Coordination

O'Keeffe's method of Voronoi coordination (VoronoiNN) uses geometric principles to determine an atom's coordination.[21] The crystal structure is first partitioned using Voronoi decomposition of the atomic sites (Figure 1a, b). From this, an atom's "domain" is defined by a polyhedron, with faces determined by an equidistant border between the atom and a neighboring site.[30] Sites that share a face with the central atom are considered either direct or indirect neighbors. To distinguish between the two, atoms are weighted by the solid angle subtended by the polyhedral face. Since indirect neighbors usually subtend smaller angles, only neighboring atoms with weights within a specified tolerance of the largest weight are considered coordinated to the central atom. In this work, atoms are considered bonded if the weights are within 50% of the largest weight for that site. This tolerance was found to be close to optimal for the MaterialsCoord benchmark and was chosen for simplicity and to avoid overfitting to the materials included in the dataset (see Section S6 and Figure S7 of the Supporting Information).

Valence Ionic Radius Evaluator Method

The minimum valence ionic-radius evaluator (VIRE) method[22] for determining coordination (MinimumVIRENN) uses a similar "minimum relative distance" approach as the minimum

O'Keeffe algorithm. The relative distance between two atoms is given by

$$d_{ij}^{\text{rel}} = \frac{d_{ij}}{d_{ij}^{\text{VIRE}}}, \tag{11}$$

where $\frac{d_{ij}}{d_{ij}^{\text{VIRE}}}$ is the ideal bond length, calculated according to

$$d_{ij}^{\text{VIRE}} = r_i^{\text{Shannon}} + r_j^{\text{Shannon}}, \tag{12}$$

in which r_i^{Shannon} is the Shannon crystal radius for site i, computed using the VIRE method implemented in pymatgen.[22] In the VIRE approach, the valence of a site is first calculated using O'Keeffe's bond valence sum method.[19] Next, an initial guess for the coordination is obtained from O'Keeffe's method of Voronoi coordination (VoronoiNN). The element type, oxidation state, and coordination number are then used to look up the associated radius in tabulated Shannon crystal radii data.[31] Where information on ionic radii is lacking, for example in structures without oxidation states or for species without associated Shannon radii, the atomic radius is used.[28] Finally, two atoms are considered bonded if

$$d_{ij}^{\text{rel}} \leq (1+\delta) \times \min\{d_{ij}^{\text{rel}} : j = 1 \dots n\}, \tag{13}$$

where δ is a tolerance parameter set to 0.1. The MinimumVIRENN algorithm is not self-consistent; coordination numbers are determined once using the VoronoiNN method to aid in determining the associated Shannon radii. Coordination numbers are not recalculated once the Shannon radii have been determined.

Crystal Near-neighbor algorithm

The crystal near-neighbor method (CrystalNN) is an algorithm we recently introduced[32] that uses Voronoi decomposition[23] to determine the probability of various coordination environments and selects the one with highest probability. The first step of this approach is to

determine a set of weights, w_{ij}, that correspond to the likelihood of a central atom i being a neighbor to surrounding atoms j. This weight has multiple components.

A first component of the weight w_{ij} is based on the Voronoi construction, which we call w^{Vor}. In the simplest case, w^{Vor} can be set to the solid angle of the neighbor atom, w^{sa}. However, we note that for porous structures, the solid angle weight can be quite high even for distant atoms; thus, by default we scale this quantity by the ratio of the solid angle to the Voronoi facet area, thereby penalizing distant atoms, with $w^{\mathrm{Vor}} = w^{\mathrm{sa}^2}/w^{\mathrm{fa}}$.

A second component of the neighbor weights, w_{ij}^{dc}, more directly penalizes atoms that are too far from the central atom, according to

$$
\begin{aligned}
w_{ij}^{dc} &= \begin{cases} 1, & d_{ij} \leq d_{\mathrm{low}}^{\mathrm{cut}} \\ \sqrt{\cos\frac{\pi(d_{ij}-d_{\mathrm{low}}^{\mathrm{cut}})}{d_{\mathrm{high}}^{\mathrm{cut}}-d_{\mathrm{low}}^{\mathrm{cut}}}+1}, & d_{ij} < d_{\mathrm{high}}^{\mathrm{cut}} \\ 0, & d_{ij} \geq d_{\mathrm{high}}^{\mathrm{cut}} \end{cases}, \\
d_{\mathrm{low}}^{\mathrm{cut}} &= r_i + r_j + \delta_{\mathrm{low}}^{\mathrm{cut}}, \\
d_{\mathrm{high}}^{\mathrm{cut}} &= r_i + r_j + \delta_{\mathrm{high}}^{\mathrm{cut}},
\end{aligned}
\tag{14}
$$

where d_{ij} is the distance between site i and neighboring site j, r_i is the radius of the species at site i, and $\delta_{\mathrm{low}}^{\mathrm{cut}}$ and $\delta_{\mathrm{high}}^{\mathrm{cut}}$ are the low and high distance cut-offs, set to 0.5 Å and 1 Å, respectively. Essentially, this function gradually starts penalizing atoms that are greater in distance than $d_{\mathrm{low}}^{\mathrm{cut}}$, and explicitly excludes neighbors that are further than $d_{\mathrm{high}}^{\mathrm{cut}}$. We note that the type of radius that is used depends on what information is available about the structure, and is (in order of decreasing preference): the ionic radius (if the oxidation state is known and an ionic radius is available), an averaged cation / anion radius (if an ionic radius is not tabulated for that species), a covalent radius, and finally an atomic radius (if a covalent radius is not available). CrystalNN will use a mixture of radii types in cases where higher-preference radii information is available for some sites but not others.. In this work we remove all oxidation states from the test structures, so only the covalent or atomic radii

are used by CrystalNN.

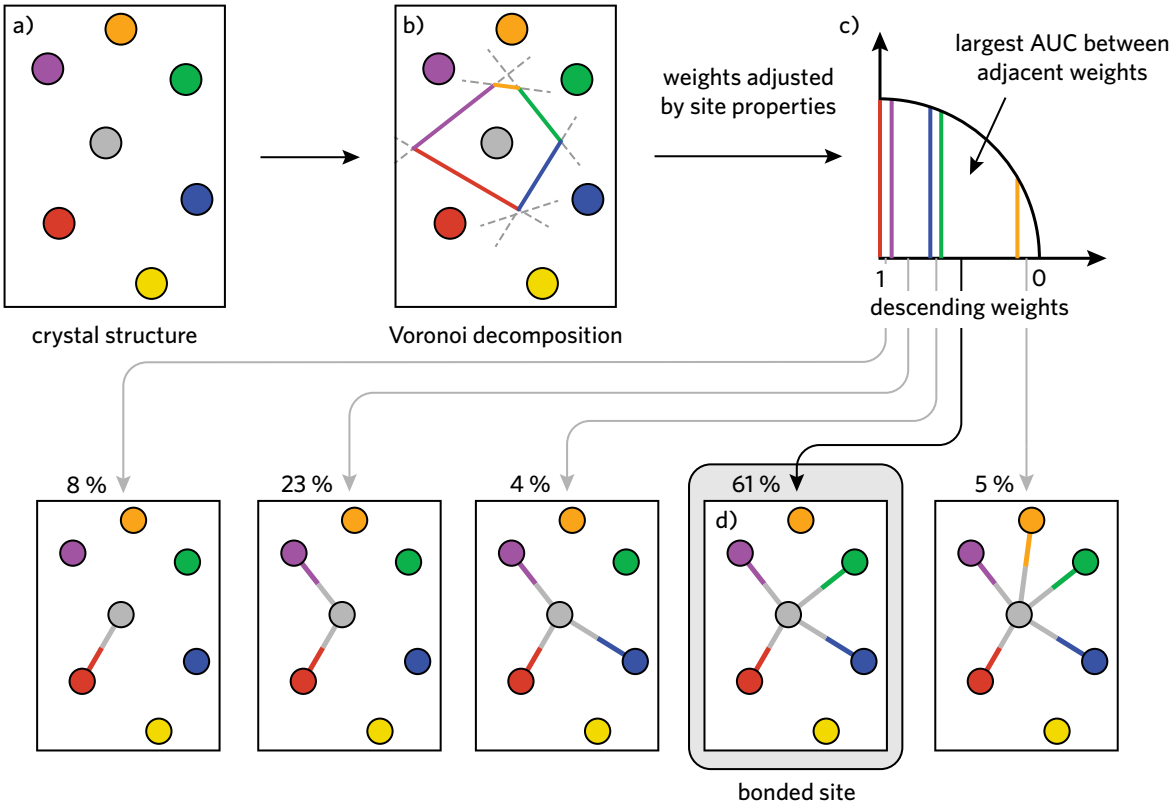

Figure 1: Schematic of CrystalNN bonding algorithm for determining the coordination of a site. A crystal structure (a) is partitioned using Voronoi decomposition (b). Only the Voronoi polyhedron for the gray central site is illustrated for clarity. The Voronoi polyhedral faces are formed by equidistant borders between the central site and its neighbors. The solid angle weights defined by the Voronoi polyhedron are rescaled based on site properties, such as electronegativity differences and distance cut-offs. The weights are normalized and projected onto a quadrant of a circle in descending order (c). The relative probability of a certain coordination number is defined by the area under the curve (AUC) between adjacent weights. If a single coordination number is desired, the environment with the highest probability is used. In this example, the largest area is between the green and orange weights and so the weight associated with the green line is set as the minimum weight cut-off. All sites with weights larger than this cut-off are considered bonded to the central site, as shown in (d).

Finally, atoms that have greater electronegativity difference from the central atom are weighted higher according to

$$w_{ij}^{\text{en}} = 1 + \delta^{\text{en}} \frac{\sqrt{|\chi_i - \chi_j|}}{3.3}, \tag{15}$$

where χ_i is the Pauling electronegativity of site i, and δ^{en} is a parameter that controls the preference for neighbors with higher electronegativity differences, set to the default value of 3. The normalization factor of 3.3 on the denominator is chosen as it is the largest electronegativity difference possible between any two elements. The final normalized weighting is calculated as

$$w_{ij} = \frac{w_{ij}^{\text{Vor}} \times w_{ij}^{\text{dc}} \times w_{ij}^{\text{en}}}{\max(w)}. \tag{16}$$

We have evaluated the importance of each weight by disabling individual features of the algorithm and investigating the resulting performance on the benchmark, with the results provided in Section S13 and Figure S21 of the Supporting Information.

In the CrystalNN approach, the coordination number of a site is determined by i) projecting the normalized weights onto a quadrant of a unit circle, ordered from largest to smallest weight, ii) calculating the area under the circle between adjacent weights to obtain coordination probabilities, and iii) choosing the coordination number with the largest probability. This procedure is illustrated in Figure 1. The end result is that one can either obtain a probabilistic assessment of different coordination scenarios or take the maximum likelihood scenario and obtain a single coordination environment (as is done in this work).

Benchmarking Framework

To compare the predictive ability of NN algorithms to reproduce literature-reported coordination numbers, we have developed a package called MaterialsCoord.[33] Using this package, a NN algorithm can be tested against a database of reported coordination environments, built from a literature search of prototypical crystal structures from the ICSD.[12] The data set contains 56 structures, broken down into 16 elementary, 11 binary, and 29 ternary compounds. The MaterialsCoord benchmark includes a wide variety of material types covering metallic and intermetallic compounds, semiconductors, and insulators. All structures are stable at ambient temperatures and pressures. Coordination numbers for these structures

are tabulated in the MaterialsCoord GitHub repository.[33] We stress that coordination numbers are fundamentally subjective quantities and are not an intrinsic or measurable property of a structure. Accordingly, MaterialsCoord is only so useful as to identify the bonding algorithms that agree with a human interpretation of coordination. In many cases, the assigned coordination numbers are well justified. For instance, structures that have basic coordination geometries (*e.g.*, tetrahedral and octahedral coordination), in which further neighboring atoms are clearly not within first neighbor shells, have robust coordination numbers. The coordination numbers of more complex structures with highly asymmetrical bonding, such as oxides or intermetallics, are more difficult to assign consistently; in several cases, many bonding descriptions for the same structure can be found in the literature. We rely on literature-reported data and descriptions for each structure in the data set and cite accordingly. Structures with basic arrangements (*e.g.*, fcc, bcc, and hcp) and well-versed coordination environments are not given a specific citation. Complex structures with ambiguous coordination environments are discussed further in the following sections.

For a given structure, each NN algorithm is assigned a score:

$$Z = \frac{\sum_{i=1}^{N_{\text{sites}}^{\text{unique}}} |\text{CN}_i^{\text{calc}} - \text{CN}_i^{\text{expected}}| N_i^{\text{degen}}}{N_{\text{sites}}} \tag{17}$$

where $N_{\text{sites}}^{\text{unique}}$ is the number of symmetrically distinct atomic sites, N_i^{degen} is the number of degenerate atomic sites, and N_{sites} is the total number of atomic sites in a structure's unit cell. For ionic compounds, we distinguish between cation and anion sites (*e.g.*, $N_{\text{sites}}^{\text{unique}}$ and N_{cations} for calculating Z_{cations}). The $\text{CN}_i^{\text{calc}}$ and $\text{CN}_i^{\text{expected}}$ are the calculated and expected coordination numbers of the i^{th} site. A score of zero indicates that the algorithm is in consensus with the coordination description in the literature. Values greater than zero indicate that there are inconsistencies between the literature and computed coordination number for a particular structure.

Several structures have multiple coordination interpretations corresponding to primary

and secondary bonding interactions. For example, in α-U, atoms are tetrahedrally coordinated to four neighbors, forming corrugated sheets held together by secondary covalent bonds to form the overall structure (accepted coordination numbers = 4 or 12).[34] For these cases, algorithms are penalized based on the smallest deviation from any of the possible coordination definitions. Using α-U as an example, if an algorithm were to predict the coordination as 11, the score would be 1.

We use the Einstein crystal test rig method[26] to determine how robust different neighbor-finding methods are towards small distortions in the crystal structures. The method mimics thermal vibrations and can thus assess the performance of different algorithms when analyzing partially relaxed structures and molecular dynamics simulations. The Einstein crystal test rig method is also useful as a framework to perform uncertainty quantification of the coordination number prediction methods in a more statistically rigorous way.

MaterialsCoord is provided as an open-source package.[33] The benchmark suite is implemented in Python 3.5+ and is designed to be easily extensible to both user defined structures and additional near-neighbor finding algorithms. Instructions on how to benchmark additional test structures and alternative near-neighbor finding algorithms are provided as tutorial notebooks in the MaterialsCoord GitHub repository. Further documentation on MaterialsCoord and a diagram of how the benchmarking scores are calculated can be found in Section S7 and Figure S8, respectively, of the Supporting Information.

Results

We compare how well the eight near-neighbor finding algorithms mentioned in Section 2 can reproduce literature descriptions by testing them on the MaterialsCoord dataset of 56 experimentally-determined prototypical structures from the ICSD.[24] This test set includes 16 elementary, 11 binary, and 29 ternary structures, of which many of the compounds are oxides. In addition to the mostly ceramic compounds discussed here, we also separately tested

intermetallic structures for which coordination can be even more ambiguous (see Section S8 and Figure S9 of the Supporting Information). The results of our benchmarking efforts are presented in the form of heatmaps, in which algorithms are assigned a score for each structure reflecting their ability to match literature-reported coordination numbers (lower scores indicate greater consensus with reported values). In our discussion, we focus on structures for which multiple algorithms deviate from the expected coordination environments.

Elemental Structures

The benchmarking scores for the 16 elemental structures in the MaterialsCoord data set are shown in Figure 2. The set includes "simple" structures, such as face-centered cubic (fcc) Cu, body-centered cubic (bcc) α-W, hexagonal close-packed (hcp) La and Mg, and diamond.[35] In addition, the set includes layered compounds (*e.g.*, α-As,[36] black P,[37] graphite, and Sm[38]) and several elements with complex, low-symmetry structures, such as α-Mn[39] and β-Mn.[40] The literature coordination environments for all elemental materials are provided in Section S9 and Figure S10 of the Supporting Information. In general, the algorithms obtain similar bonding descriptions for the elemental structures, with all matching literature-reported coordination descriptions in 80% or more of the structures. CrystalNN demonstrates the greatest consensus with the literature by reproducing the human determined coordination environments for all test structures. The threshold-based cutoff approaches (MinimumDistanceNN, MinimumOKeeffeNN, and MinimumVIRENN), EconNN, BrunnerNN, and VoronoiNN perform similarly, achieving scores between 4 and 10. JmolNN shows the greatest disagreements, dramatically over-predicting the coordination of Mg to achieve an overall score of 21.

All algorithms agree with the literature when predicting the coordination of basic structures (bcc, hcp, fcc, and diamond-like). Similar behavior is seen for β-Sn, which has a distorted octahedral geometry (CN = 6).[41] For Se, which is composed of parallel helical chains of Se atoms (CN = 2),[42] only VoronoiNN predicts a coordination that does not match the reported literature value. Furthermore, all algorithms agree with the literature

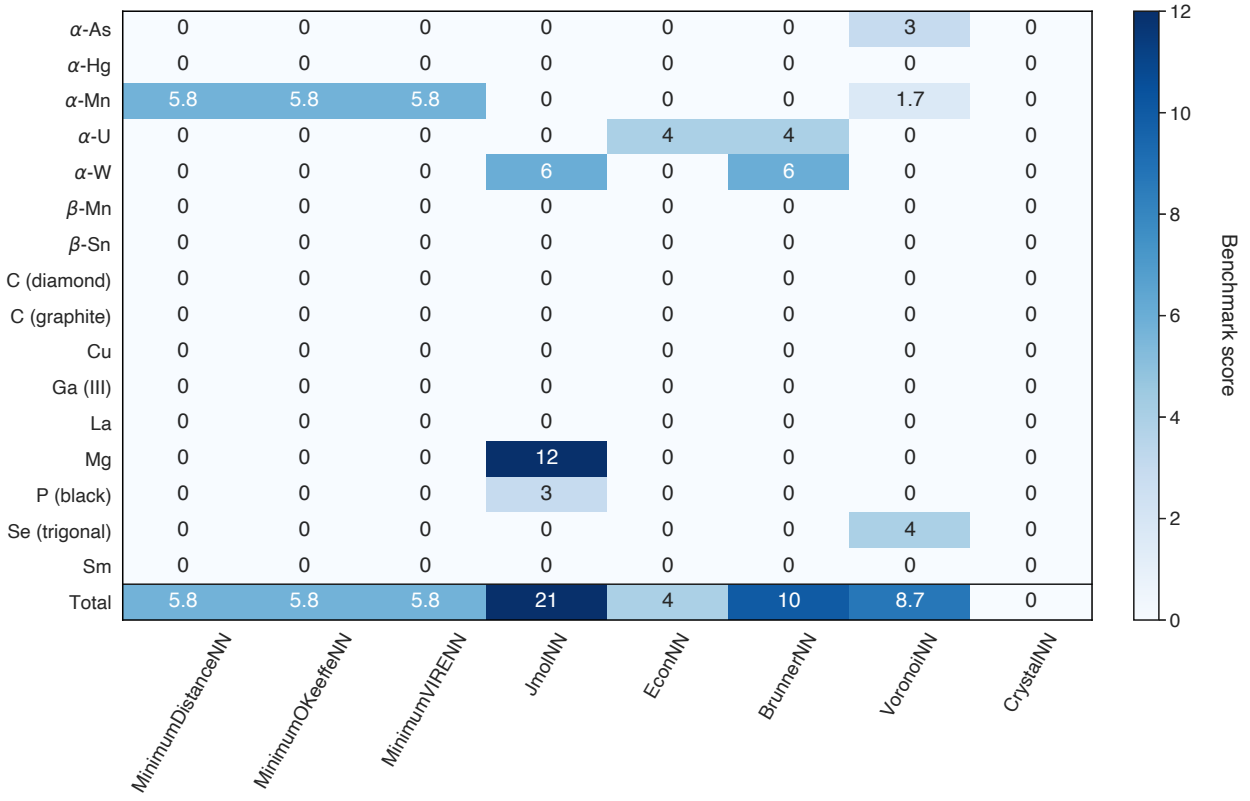

Figure 2: Ability of near neighbor algorithms (x-axis) to reproduce literature descriptions of cation coordination for the elemental structures in the MaterialsCoord benchmark suite (y-axis). Scores are color-coded, with darker colors indicating greater deviation from the literature coordination number. The total score for each algorithm is calculated as the sum of the scores across all structures.

for layered structures, albeit with a few exceptions in which additional interlayer bonds are predicted. In particular, the relative interlayer spacing appears to correlate with the difficulty of determining the coordination number. For example, all algorithms reproduce the literature description of graphite which possesses the largest interlayer spacing, with next-nearest neighbor distances 42 % larger than the nearest neighbor distance. In contrast, black P and α-As possess next-nearest neighbor spacings of 32 % and 20 % of the nearest neighbor distance with JmolNN and VoronoiNN obtaining inconsistent descriptions for each structure, respectively.

Of all the elemental structures, α-Mn exhibits the greatest divergence in bonding descriptions, with half of the algorithms obtaining coordination environments at variance to

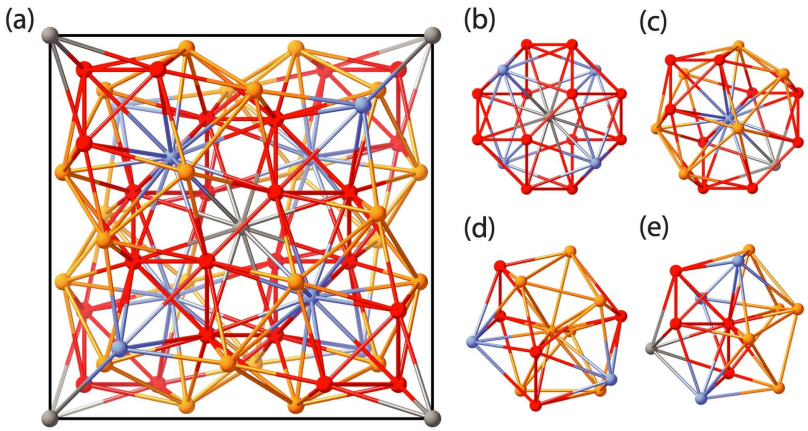

Figure 3: (a) α-Mn unit cell consisting of Mn′ (gray), Mn″ (light blue), Mn‴ (orange), and Mn⁗ (red) atomic sites and their respective coordination environments (b-e).[39]

the literature. This can be ascribed to the presence of mixed coordination environments: α-Mn contains 58 atoms, comprised of 2 Mn′ sites (CN = 16), 8 Mn″ sites (CN = 16), 24 Mn‴ sites (CN = 13), and 24 Mn⁗ sites (CN = 12), as illustrated in Figure 3.[39] This structure is the only elemental compound for which the threshold-based cutoff approaches (MinimumDistanceNN, MinimumOKeeffeNN, and MinimumVIRENN) deviate from the literature. All three algorithms under-predict the coordination identically; they assign a CN of 10 (instead of 16) to Mn″, a CN of 4 (instead of 13) to both Mn‴ and a CN of 9 (instead of 12) to Mn⁗. VoronoiNN also under-predicts the coordination of α-Mn, but to a lesser extent, assigning the coordination of Mn″ as 10, Mn‴ as 12, and Mn⁗ as 11.

Binary Structures

The benchmarking scores for the 11 binary structures in the MaterialsCoord data set are illustrated in Figure 4. For ionic compounds, we abbreviate coordination using the nomenclature A:X, where A and X are the coordination numbers of the cations and anions, respectively (*e.g.*, NaCl has 6:6 coordination). We follow bonding literature convention and focus our analysis on cation coordination in the main text — the results for anions show qualitatively the same trends and are provided in Section S10 and Figure S12 of the Supporting Informa-

tion. The set includes common simple binary solids, including rock-salt[43,44] (6:6), CsCl[43,44] (8:8), sphalerite[44] (4:4), wurtzite[43] (4:4), rutile[43,45] (6:3), and corundum[46] (6:4). In addition, we include γ-brass (Cu_5Zn_8), a more complicated structure with metallic bonding.[47] The literature coordination environments for all binary materials are provided in Section S9 and Figure S11 of the Supporting Information.

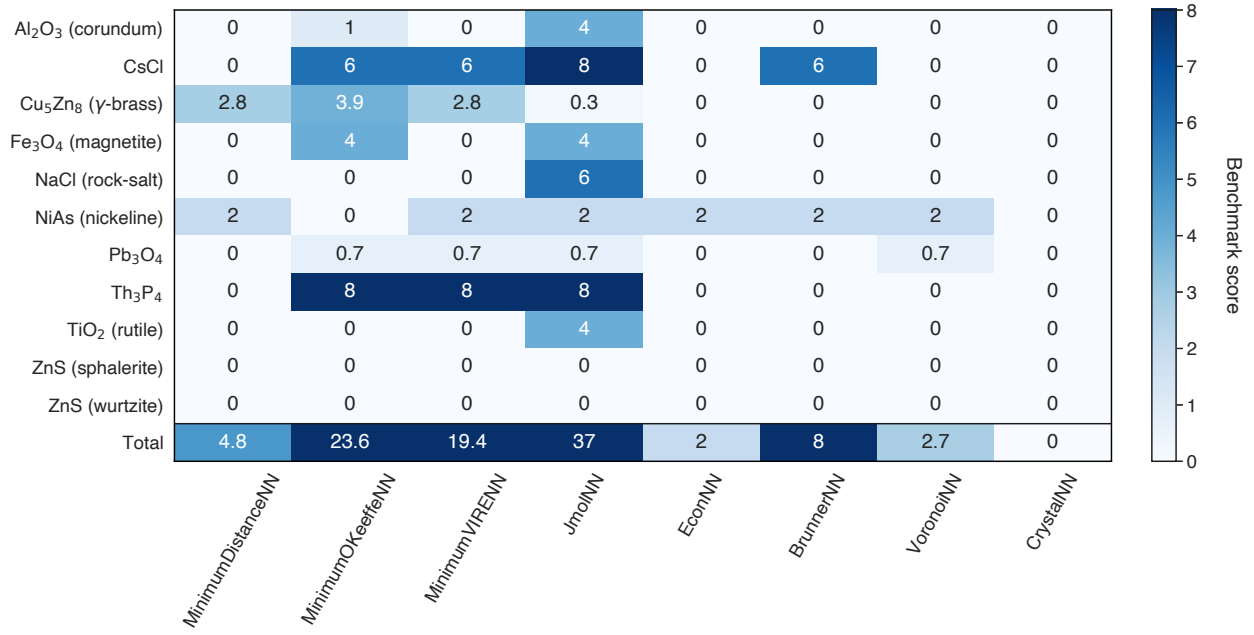

Figure 4: Ability of near neighbor algorithms (x-axis) to reproduce literature descriptions of cation coordination for the binary structures in the MaterialsCoord benchmark suite (y-axis). Scores are color-coded, with darker colors indicating greater deviation from the literature coordination number. The total score for each algorithm is calculated as the sum of the scores across all structures.

For the binary compounds, only CrystalNN matches the literature coordination in all cases. EconNN, VoronoiNN, MinimumDistanceNN, and BrunnerNN also obtain similar predictions, achieving scores of 2, 3, 5 and 8, respectively. The largest deviation is exhibited by JmolNN, MinimumOKeeffeNN, and MinimumVIRENN, which only match the literature coordination for 2 (score of 37.0), 5 (score of 23.6), and 6 (score of 19.4) structures, respectively out of 11 total structures.

Of the simple binary structures, CsCl (Figure 5) appears particularly challenging, with

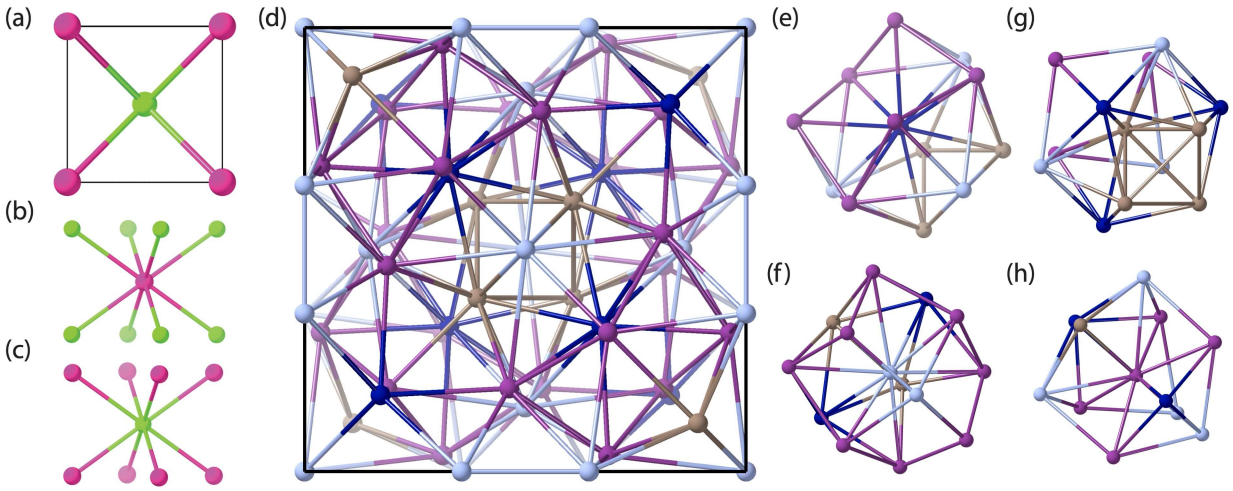

Figure 5: (a) CsCl unit cell and corresponding coordination environments (b-c) of Cs (pink) and Cl (green) atomic sites.[43,44] (d) γ-brass unit cell consisting of Cu′ (dark blue), Cu″ (blue), Zn′ (brown), and Zn″ (purple) atomic sites and their respective coordination environments (e-h).[47]

only MinimumDistanceNN, VoronoiNN, EconNN, and CrystalNN matching literature-reported coordination values.[43] The disagreements of the other algorithms can be attributed to several factors. The relatively large distance between Cs and its nearest neighbor Cl atoms (3.6 Å — larger than any other anion-cation near neighbor distance in the dataset) causes JmolNN, which employs radii tables, to entirely miss the Cs–Cl bonds. In addition, several algorithms predict bonding between adjacent Cs atoms, despite the large distance (4.1 Å) separating these sites (MinimumOKeeffeNN, MinimumVIRENN, and BrunnerNN). Since the MinimumOKeeffeNN approach explicitly accounts for electronegativity differences this behaviour is especially surprising.

The metallically bonded γ-brass shown in Figure 5 also proved difficult, with half of the algorithms predicting coordinations that deviate from the literature description.[47] In most cases, disagreements originate from the Cu″ site, which is bonded in a distorted icosahedra coordination geometry to 10 Zn and 3 Cu atoms (CN = 13). Perhaps due to their reliance on distance cut-offs, MinimumDistanceNN, MinimumOKeeffeNN, and MinimumVIRENN miss the coordination between Cu′ and five of the neighboring Zn atoms.

Most of the algorithms suffer from some degree of erroneous cation–cation bonding. An egregious example is nickeline (NiAs), in which Ni is bonded in an octahedral configuration to 6 As atoms.[48] All algorithms assign the expected Ni–As bonds but most — except CrystalNN and MinimumOKeeffe — also predict bonding between Ni and two Ni neighbors. A similar effect is entirely responsible for the high scores for corundum (Al_2O_3),[46] magnetite (Fe_3O_4),[49] Th_3P_4,[50] rutile (TiO_2),[45] and Pb_3O_4.[51] To assess this effect further, we have calculated the MaterialsCoord scores when coordination is restricted to sites of opposing charge, *i.e.*, only considering cation to anion bonding (see Section S11 and Figures S14 and S15 of the Supporting Information). This constraint significantly improves the agreement of the algorithms against the literature bonding descriptions, with the scores of EconNN, BrunnerNN, and VoronoiNN reducing to zero and the scores of MinimumOKeeffeNN, MinimumVIRENN, and JmolNN more than halved.

Ternary Structures

We next report the benchmarking results for the 29 ternary compounds in the MaterialsCoord data set. The structures comprise oxides and fluorides with ABX_3, ABX_4, and A_2BX_4 stoichiometries, where A and B are cations. In our data set, A is typically larger and heavier than B, and X is either O or F. The performance of all NN algorithms for predicting cation coordination numbers is illustrated in Figure 6 — the results for anions show qualitatively the same trends and are provided in Section S10 and Figure S13 of the Supporting Information. Compared to the elemental and binary structures, the ternary compounds produce greater deviations against human interpretations of bonding for most algorithms. The greatest consensus is exhibited by VoronoiNN (score of 2), CrystalNN (4.8), EconNN (7), and BrunnerNN (10.7) which agree with the literature description in over 90% of structures. Interestingly, MinimumDistanceNN (15), and MinimumVIRENN (19), show almost exactly the same scores for each structure in the test set. MinimumOKeeffeNN (124) and JMolNN

(89) achieve the highest scores and only identify the expected coordination in 31%, and 14% of the structures.

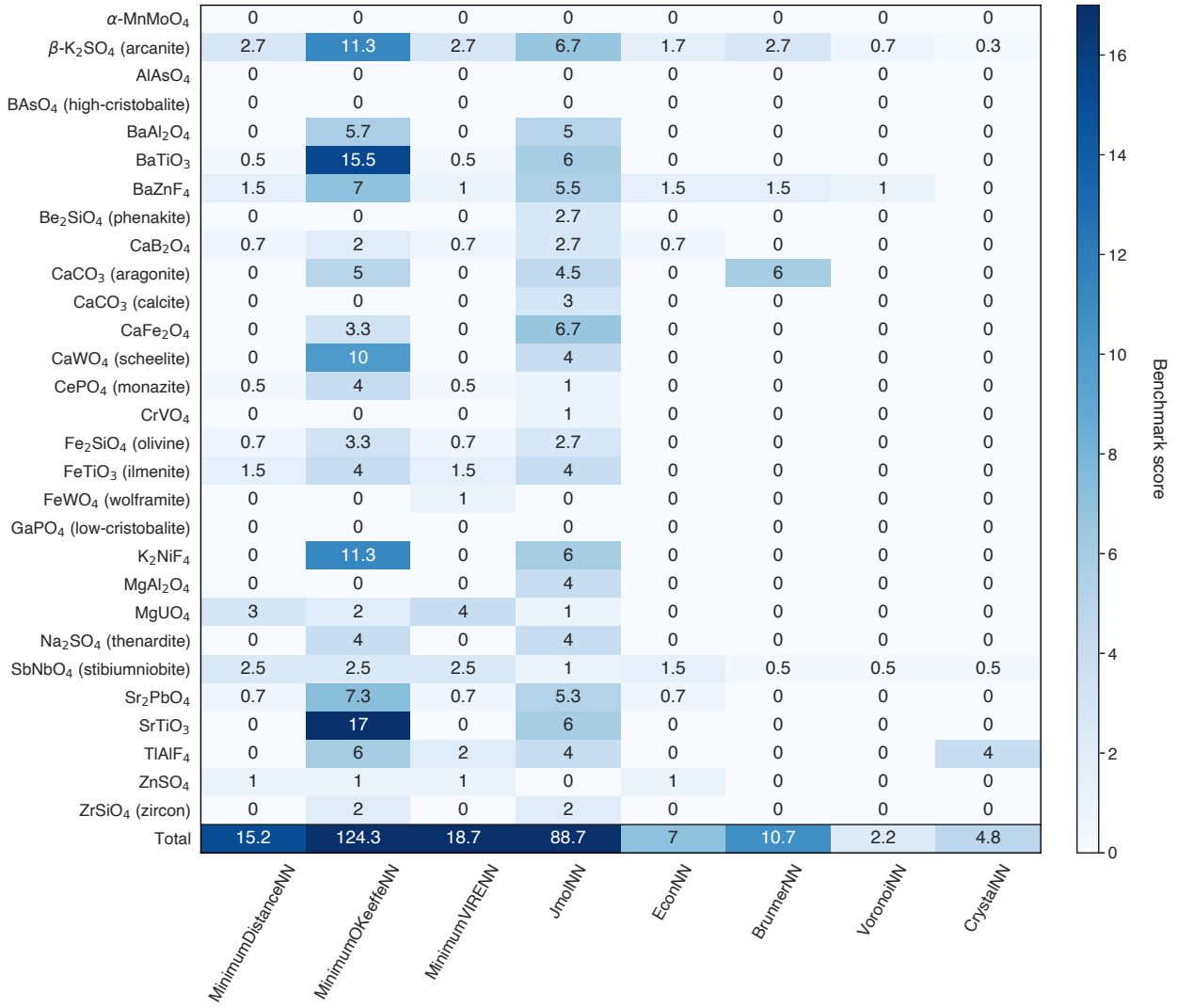

Figure 6: Ability of near neighbor algorithms (x-axis) to reproduce literature descriptions of cation coordination for the ternary structures in the MaterialsCoord benchmark suite (y-axis). Scores are color-coded, with darker colors indicating greater deviation from the literature coordination number. The total score for each algorithm is calculated as the sum of the scores across all structures.

All algorithms reproduce the literature coordination for several structures including zeolite-like materials ($AlAsO_4$, $GaPO_4$, and $BAsO_4$) — in which A- and B-site cations are tetrahedrally coordinated to O atoms[52–54] — and $MgAl_2O_4$ and $MnMoO_4$ — which have

octahedral A-site cations bound to tetrahedral B-site cations.[55,56] Noticeably, all algorithms match the coordination of tetrahedral- and trigonal planar-coordinated B-site cations. B sites with larger coordination numbers, however, often show greater deviations. For example, MinimumVIRENN underestimates the octahedral coordination environment of W (CN = 6) in $FeWO_4$ as being 4-coordinated.[55] The same effect is observed for octahedrally coordinated Tl in $TlAlF_4$[57] where most algorithms (MinimumOKeeffeNN, MinimumVIRENN, JmolNN, and CrystalNN), underestimate the coordination number.

The coordination environments for β-K_2SO_4 and $SbNbO_4$ show large variation from the literature for all algorithms. In the β-K_2SO_4 structure, units of tetrahedrally coordinated SO_4 are bonded to two unique K sites.[58] K′ is bonded to 11 O atoms, whereas K″ is bonded to 9 atoms (Figure 7). We note that β-K_2SO_4 is a highly complex structure for which reproducing the literature description of bonding may be difficult even for experienced researchers. All algorithms match the expected coordination of the SO_4 unit but exhibit inconsistencies with the K sites. The trend across algorithms is to underestimate the coordination. MinimumDistanceNN, MinimumOKeeffeNN, MinimumVIRENN, JmolNN, CrystalNN, and VoronoiNN all underestimate the coordination of at least one of the K sites. JmolNN exhibits the largest disagreement, assigning a coordination of 0 to both sites. In contrast, BrunnerNN and EconNN both overbind by assigning additional bonding to neighboring S sites and K sites. $SbNbO_4$ comprises layers of distorted NbO_6 octahedra (CN = 6) connected by layers of trigonal prismatic SbO_6 (CN = 6).[59] Again, in most cases, the coordination of the cations Nb and Sb is underestimated. MinimumDistanceNN, MinimumOKeeffeNN, and MinimumVIRENN under-predict the coordination number of both cations as either 3 or 4. EconNN, BrunnerNN, VoronoiNN, and CrystalNN match the coordination of Nb but determine Sb to be 3- or 5-coordinated rather than 6. JmolNN behaves similarly but further bonds Sb to two neighboring Sb sites.

Any trends between algorithms are generally difficult to determine due to the large variation in coordination predictions. However, the threshold distance-based algorithms

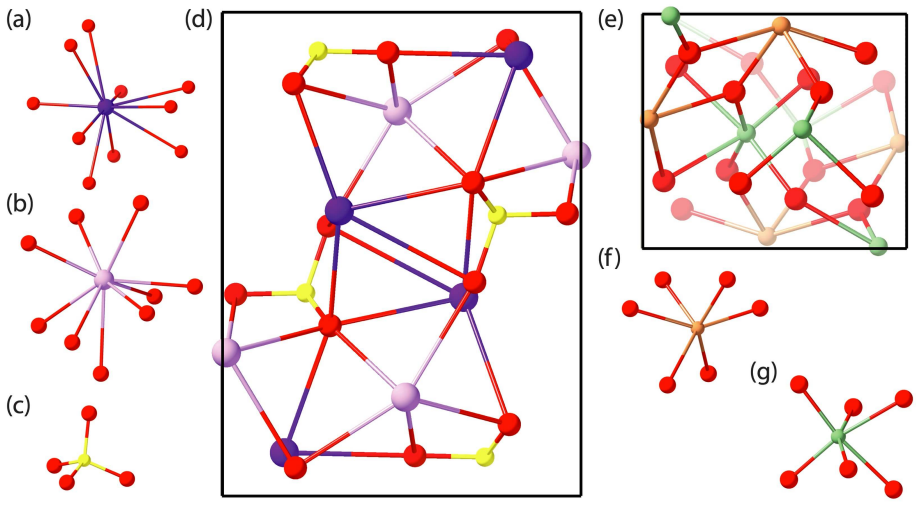

Figure 7: (a-c) Coordination environments of K′ (dark purple), K″ (light purple), and S (yellow) to O atoms (red), respectively, in the β-K_2SO_4 structure (d). (e) $SbNbO_4$ unit cell and coordination environments of Sb (orange) and Nb (green) to O (red) atomic sites (f-g).[55]

(MinimumDistanceNN, MinimumOKeeffeNN, and MinimumVIRENN) often show similar behavior. When these algorithms diverge from the literature description they almost always under-bind, *i.e.,* predict lower coordination numbers than the reference. For instance, all predict the edge-sharing octahedral MgO_6 and UO_6 units in $MgUO_4$ to be 2-coordinated rather than 6.[60] The under-predicted coordination environments are due to the differences in interatomic distances between the cation and the oxygens: each Mg is coordinated to 2 O atoms at 1.98 Å and 4 O atoms further away at 2.19 Å. Likewise, the U atoms are coordinated to 2 O atoms at 1.92 Å and 4 O atoms further away at 2.21 Å. A similar behavior is seen in $ZnSO_4$, comprising edge-sharing ZnO_6 octahedral chains linked by edge-sharing SO_4 tetrahedra.[61,62] In each case, these algorithms under predict the 6-coordinate Zn cations as 4-coordinate. For thirteen of the twenty-nine structures, JmolNN predicts A-site cations to be uncoordinated. This behavior persists over a range of structural prototypes including scheelite ($CaWO_4$),[63] stuffed tridymite-type $BaAl_2O_4$,[64] phenakite (Be_2SiO_4),[65] and perovskite structured $SrTiO_3$ and $BaTiO_3$.[55] Interestingly, while JmolNN often diverges from the literature on the A-site coordination, it matches the coordination of all B-sites.

As observed in the binary structures, many algorithms assign unexpected bonding from cations to other cations. This behavior is highlighted by aragonite structured $CaCO_3$, containing Ca cations bonded to 9 oxygen atoms.[55] Both BrunnerNN and MinimumOKeeffe assign additional bonds from Ca to neighboring C and Ca sites. We investigate this effect further by calculating the MaterialsCoord benchmark scores with coordination limited to sites of opposing charge (see Section S11 and Figures S16 and S17 of the Supporting Information). This constraint improves the agreement with the literature for many algorithms. In particular, the scores for MinimumOKeeffeNN (124), JmolNN (89) and BrunnerNN (11) are dramatically reduced to 40, 66, and 2, respectively. In contrast, the scores of MinimumDistanceNN, VoronoiNN, and CrystalNN remain unaffected, indicating that these algorithms do not assign any cation–cation bonds in the ternary structure test set.

Analysis of coordination trends

Figure 8 illustrates the tendency for algorithms to either under- or over-predict coordination numbers. Here, the deviation in coordination prediction ($CN_{calc} - CN_{expected}$) of every site across all structures is plotted as a histogram for each coordination algorithm. Only prediction differences are included, with the area of the distribution being proportional to the number of diverging predictions. Distributions with greater area above zero signify over-coordination, whereas larger negative areas indicate under-coordination. A theoretical algorithm that can reproduce all literature descriptions would have no area at all.

Most algorithms tend to underpredict coordination numbers, as indicated by the tails of the distributions which are generally negative. In particular, MinimumDistanceNN, EconNN, VoronoiNN, and CrystalNN show very little positive area. Accordingly, the ability of these methods to reproduce literature descriptions might be improved by adjusting their tolerance parameters (δ) to yield more balanced prediction differences. The considerable disagreements of MinimumOKeeffe and JmolNN against literature descriptions is reflected in the large area of their distributions. These algorithms are the only methods which frequently over-

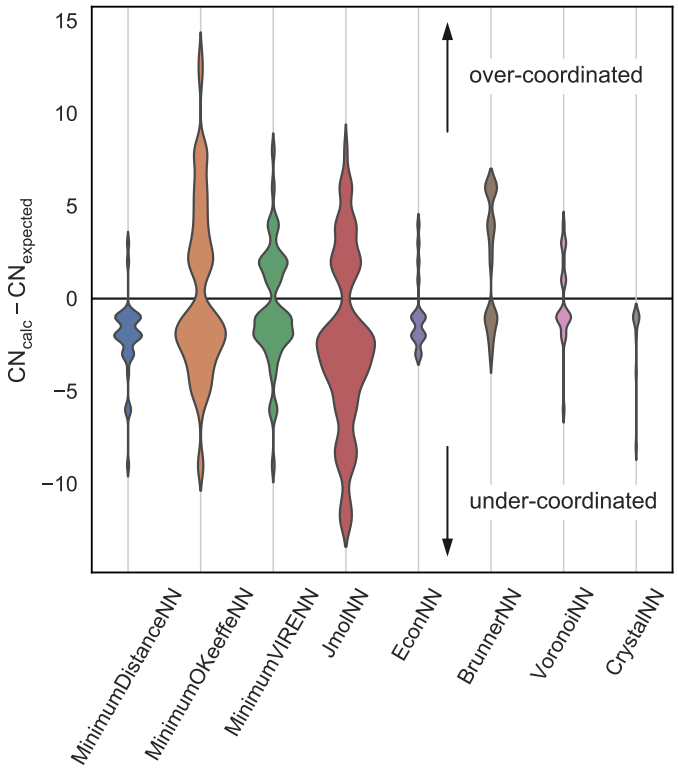

Figure 8: Tendency of NN algorithms (x-axis) to either under- or over-predict coordination numbers (y-axis). Distributions that tend towards positive values indicate over-coordination whereas negative values indicate under-coordination. Only finite prediction differences are included, with the area of the distribution being proportional to the number of diverging predictions. Thus, the density at zero error (which dominates the data) is not plotted. A theoretical algorithm that can reproduce all literature descriptions would have no area at all.

coordinate, with prediction differences reaching 14 for some sites. To further analyze the behaviour of the algorithms, we break down the data set into elemental, binary, and ternary compounds and report their coordination trends in Section S12 and Figures S18, S19, and S20 of the Supporting Information.

Perturbation of Crystal Structures

We next discuss our benchmarking results for structures containing atomic perturbations introduced using the Einstein crystal test rig method.[26] Coordination analysis of perturbed structures has already found use tracking the local coordination of sites in molecular dynam-

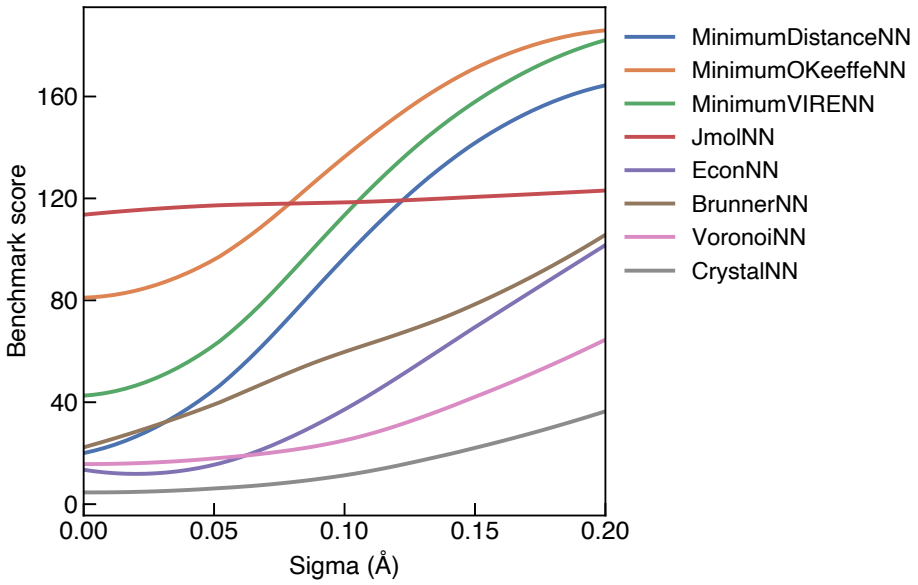

Figure 9: Robustness of coordination prediction to random atomic displacements. CrystalNN, VoronoiNN, and JmolNN show greater stability against perturbation than other methods. Displacements introduced according to the Einstein crystal test rig approach.

ics simulations.[66] Furthermore, by assessing the stability of coordination predictions against small perturbations, the robustness of coordination algorithms can be determined. It is important to note that this analysis assumes the coordination number should remain constant when perturbations are introduced. For small atomic displacements (less than $\sim$0.1 Å), this assumption is reasonable. For larger displacements the true coordination number is not well defined and it is unclear whether the coordination number should remain the same. It may be that at such large perturbation values, the coordination number can significantly vary from that of the ideal crystal structure. Regardless, the performance of the algorithms against large displacements can still be instructive.

The results of the perturbation analysis are illustrated in Figure 9. Most algorithms follow a similar trend, in which the prediction differences increase with increasing perturbation distance. Within small perturbation values (< 0.05 Å), the sensitivity of most algorithms seems reasonable as there is not much change in the benchmark scores. Slightly larger perturbations between $0.05 < \sigma < 0.15$ Å, however, results in higher sensitivity to perturbation,

particularly for MinimumDistanceNN, MinimumOKeeffeNN, MinimumVIRENN, and BrunnerNN algorithms. Benchmarking scores for these algorithms rise rapidly in this region. In contrast, CrystalNN, VoronoiNN, and JmolNN are considerably more robust to atomic displacements, with scores that vary little up to displacements of 0.1 Å. JmolNN in particular is extremely insensitive to atomic displacements, showing minimal change in benchmark score even with 0.2 Å perturbations. This is likely because it employs absolute cutoffs that do not depend on the relative distances or weights between sites. In Section S14 and Figure S22 of the Supporting Information, we investigate the performance of the near neighbor algorithms on structures containing point defects and find that all algorithms are relatively tolerant to atomic vacancies.

Jaccard Distance Quantification

It is interesting to understand whether two algorithms show similar behavior despite different scientific justifications on a large scale. Although two algorithms can be compared based on their benchmark scores, this approach does not provide a reliable indication of similarity. For example, the same coordination number can be achieved through completely different bonding. To robustly compare the behavior of coordination algorithms we therefore employ the Jaccard distance, which is a measure of dissimilarity between two sets.[67] Here, we only consider the set of bonds present in the primitive crystallographic cell. Each bond is characterized by: i) the origin atom, ii) the destination atom, and iii) the periodic image of the destination atom (keeping the origin atom in the origin image by convention). The Jaccard distance between two algorithms on a specific structure is calculated as

$$J_{\text{dist}} = 1 - \frac{|\text{Bonds}_A \cap \text{Bonds}_B|}{|\text{Bonds}_A| + |\text{Bonds}_B| - |\text{Bonds}_A \cap \text{Bonds}_B|} \tag{18}$$

where Bonds_A and Bonds_B are the sets of bonds determined by algorithm A and B, respectively. The Jaccard distance is 0 if two algorithms behave identically and 1 if they do not

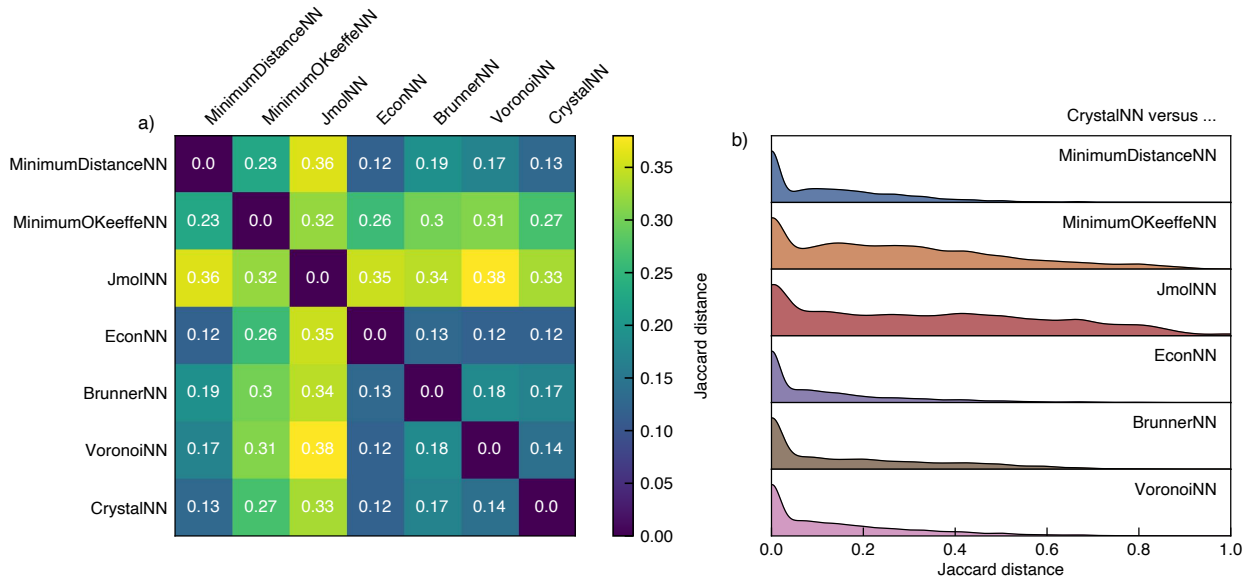

Figure 10: a) Heatmap of average Jaccard distances illustrating the similarity in bonding behavior between two algorithms. The Jaccard distance is 0 if two algorithms behave identically and 1 if they do not share a single bond in common. b) Histogram of Jaccard Distances for CrystalNN against other NN algorithms, calculated across all experimental compounds in the Materials Project database.

share a single bond in common.

The Jaccard distance algorithm was implemented in pymatgen[22] for the purposes of this analysis. We calculate the Jaccard distance for all structures in the Materials Project database that have been characterized experimentally — 42,500 compounds at the time of writing. We note that these structures are calculated with density functional theory[68] using the Perdew-Burke-Ernzerhof[69] parameterization of the generalized gradient approximation (GGA)[70] along with (for most transition metal oxides) Hubbard +U corrections.[71] Notably, lattice parameters may be slightly overestimated in general compared to experimental values.[72] Due to inclusion of organic crystals in this dataset, we exclude MinimumVIRENN from our analysis as it is specifically formulated for ionic materials. For each structure in the dataset the Jaccard distance was calculated between every pair of algorithms. Finally, the pairwise Jaccard distances were averaged across all structures to give a single distance metric characterizing the similarity between two algorithms.

The averaged Jaccard distance results are illustrated in Figure 10a. BrunnerNN, MinimumDistanceNN, CrystalNN, VoronoiNN, and EconNN exhibit similar bonding behavior, with average Jaccard distances between 0.12 and 0.19. The smallest Jaccard distance (0.12) is found between CrystalNN and EconNN. Surprisingly, MinimumDistanceNN and CrystalNN also share a small Jaccard distance (0.13) despite their vastly different underlying formulation. The largest Jaccard distance is between JmolNN and VoronoiNN (0.38). In general, EconNN and MinimumOKeeffeNN exhibit different bonding behavior from all other algorithms. Furthermore, the methods themselves share a high Jaccard distance (0.32) indicating they often assign different bonding. As these algorithms exhibit the largest scores on the MaterialsCoord benchmark, this indicates the algorithms often diverge from literature bonding descriptions but in different ways. For CrystalNN, we report the distribution of Jaccard distances (non-averaged) across all materials (Figure 10b). This analysis further illustrates CrystalNN's similarity to BrunnerNN, MinimumDistanceNN, VoronoiNN, and EconNN while highlighting its contrasting behavior to MinimumOKeeffeNN and JmolNN.

Timing analysis

A common practical use for coordination prediction algorithms is providing local environment information in machine learning studies or in large database analyses. Often machine learning models are trained on tens or hundreds of thousands of materials simultaneously. Accordingly, the computational demand of the prediction algorithm should be minimized. To assess the tradeoff between speed and ability to reproduce literature-reported coordination numbers of the near neighbor algorithms we calculate the time taken to run each algorithm on all elemental, binary, and ternary materials in the MaterialsCoord benchmark. We note that the implementation of a particular algorithm might be slow even if the method could be much faster. For example, in principle the timing of VoronoiNN should be approximately equal to that of CrystalNN, but the implementation of CrystalNN in pymatgen employs an intelligent cut-off scheme for Voronoi construction that reduces runtime. Accordingly, our

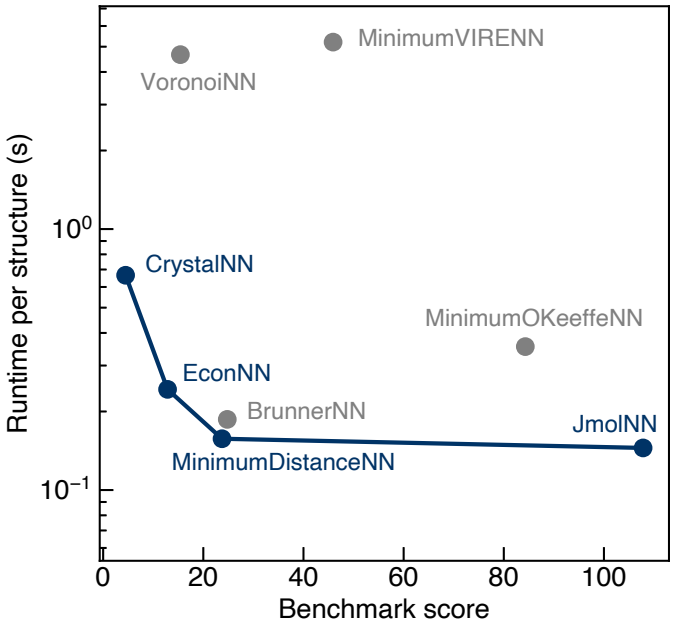

Figure 11: Tradeoff between computational speed and ability to reproduce literature-reported coordination numbers of near neighbor methods on the MaterialsCoord benchmark. The algorithms highlighted in blue form the Pareto frontier. These are the most-optimal methods that are not dominated in both score and runtime by any other method.

results provide an indication of the computational demand of the algorithms as implemented in pymatgen but that might be subject to further optimization. Regardless, our results may still pragmatically guide materials scientists in their choice of neighbor-finding algorithm when constrained by computational resources.

The tradeoff between speed and ability to reproduce literature-reported coordination numbers of the near neighbor algorithms is illustrated in Figure 11. The Pareto frontier of most optimal algorithms is highlighted in blue. These algorithms are not dominated in both score and runtime simultaneously by any other method. While CrystalNN obtains the lowest benchmark score, it is the third most computationally expensive method in terms of runtime (0.66 s per structure). Accordingly, the reduced computational demand of EconNN (0.24 s per structure) or MinimumDistanceNN (0.16 s per structure) may be a more attractive option when analyzing large computational datasets or long molecular dynamics simulations. However, because the computational cost of CrystalNN falls within the same order of magni-

tude as other approaches, we expect its ability to reproduce literature-reported coordination numbers will make it a viable option for all but the most demanding computational applications.

Discussion

The MaterialsCoord benchmark is, to our knowledge, the first tool for the quantitative assessment of near neighbor finding methods. The primary use of MaterialsCoord is to identify the algorithms which show the greatest consensus with human interpretations of coordination. CrystalNN shows the greatest agreement with literature-reported coordination numbers, with a total score of 5 across all structures — including cation and anion sites. EconNN, VoronoiNN, MinimumDistanceNN, and BrunnerNN also perform similarly, with overall scores of 13, 15, 24, and 25 respectively. The remaining algorithms, MinimumVIRENN, MinimumOKeeffeNN, and JMolNN, show greater deviations achieving scores of 46, 84, and 108, respectively. Along with its ability to predict literature-reported coordination numbers, CrystalNN is also one of the more robust algorithms against structures with small atomic perturbations. The ability to reproduce human interpretations of bonding, combined with relatively high speed, robustness to small changes in bond length and built-in avoidance of cation–cation bonding make CrystalNN a viable new option for use in a variety of applications.

Nevertheless, there will be situations in which to prefer other algorithms. For applications in which speed or simplicity is paramount, MinimumDistanceNN performs relatively well on the MaterialsCoord benchmark and its results agree with CrystalNN to a high degree (the two algorithms have a relatively small Jaccard distance). However, a weakness of this algorithm is that small perturbations to atomic distances can potentially result in different coordination assignments, which may be problematic for applications such as constructing graph neural networks or analyzing molecular dynamics trajectories. The EconNN represents

a relatively good compromise between speed, ability to reproduce coordination numbers from the literature, and robustness to atomic displacement. This method is also relatively insensitive to its single tolerance parameter (see Supporting Information), and thus one does not need to worry about overparameterization. Finally, although we find that CrystalNN generally outperforms a simpler Voronoi procedure on the MaterialsCoord benchmark, the Voronoi algorithm is conceptually simpler and results are also relatively insensitive to the choice of solid angle tolerance parameter in the range of 0.3 – 0.6. Furthermore, the speed of this algorithm should be able to match that of CrystalNN with further code optimization.

Most near-neighbor methods evaluated in this work assign bonds between sites of like charge — i.e., cation to cation or anion to anion bonds. One route to improving the ability of these algorithms to match literature coordination numbers would be to manually restrict bonding to sites of opposing charge. Unfortunately, this approach is complicated by several factors. Primarily, the oxidation states of the atomic sites may not be known in advance. In addition, this route will fail for strongly covalent materials — such as organic molecules — where formal oxidation states are not well defined and not necessarily reflective of bonding.

Improvements to coordination prediction has benefits in a broad range of applications. For example, databases such as the Materials Project rely on neighbor algorithms for text-descriptions (robocrystallographer)[73] and structural similarity analysis.[32] MaterialsCoord may assist in the development of novel coordination algorithms. In particular, analysis of over-coordination vs. under-coordination can be applied to understand how algorithms fail in order to produce more balanced predictions. Although the algorithms investigated here rely solely on crystal structure as input, MaterialsCoord may also be used to assess more advanced methods such as those that rely on charge densities from first-principles calculations.

Conclusion

We have introduced MaterialsCoord, an open-source benchmark suite for evaluating the agreement of near-neighbor finding algorithms with human interpretations of coordination. The benchmark contains 56 experimentally determined prototype structures from the ICSD, comprising a diverse test set of elemental, binary, and ternary compounds. We introduce CrystalNN, a novel algorithm for determining near neighbors and benchmark it against other existing near neighbor finding methods on MaterialsCoord. We demonstrate CrystalNN to be a viable coordination number prediction algorithm, able to compete with other well-established methods such as MinimumDistanceNN, VoronoiNN, and EconNN. We reveal that CrystalNN is relatively fast and is particularly tolerant to structures with small perturbations (*e.g.*, those mimicking thermal motion). Accordingly, CrystalNN is a viable option for many near neighbor finding applications. We believe that this work will aid the development of coordination prediction algorithms as well as improve structural descriptors for machine learning.

Supporting Information

Additional analysis on the symmetry of near neighbor algorithms, comparing variations of BrunnerNN algorithms, comparing VoronoiNN and EconNN tolerance parameters, benchmarking intermetallic structure types, anion coordination environments, cation-anion bonding effects, overbinding versus underbinding in separated elemental, binary, and ternary structure test sets, CrystalNN algorithm feature sensitivity, and vacancy structures as well as documentation on MaterialCoord and figures of coordination environments for elemental and binary structures reported in the main text.

Acknowledgements

This work was funded and intellectually led by the U.S. Department of Energy (DOE) Basic Energy Sciences (BES) program — the Materials Project — under Grant No. KC23MP. 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. We thank Alireza Faghaninia for many fruitful discussions and Julia Dshemuchadse for lending her expertise on intermetallics.

References

(1) Deem, M. W.; Pophale, R.; Cheeseman, P. A.; Earl, D. J. Computational Discovery of New Zeolite-Like Materials. *J. Phys. Chem. C* **2009**, *113*, 21353–21360, DOI: 10.1021/jp906984z.

(2) Li, Y.; Yu, J.; Xu, R. Criteria for Zeolite Frameworks Realizable for Target Synthesis. *Angew. Chem. Int. Ed.* **2013**, *52*, 1673–1677, DOI: 10.1002/anie.201206340.

(3) Salcedo Perez, J. L.; Haranczyk, M.; Zimmermann, N. E. R. High-Throughput Assessment of Hypothetical Zeolite Materials for Their Synthesizeability and Industrial Deployability. *Z. Kristallogr. Cryst. Mater.* **2019**, *234*, 437–450, DOI: 10.1515/zkri-2018-2155.

(4) Rong, Z.; Malik, R.; Canepa, P.; Sai Gautam, G.; Liu, M.; Jain, A.; Persson, K.; Ceder, G. Materials Design Rules for Multivalent Ion Mobility in Intercalation Structures. *Chem. Mater.* **2015**, *27*, 6016–6021, DOI: 10.1021/acs.chemmater.5b02342.

(5) He, B.; Ye, A.; Chi, S.; Mi, P.; Ran, Y.; Zhang, L.; Zou, X.; Pu, B.; Zhao, Q.; Zou, Z., et al. CAVD, towards better characterization of void space for ionic transport analysis. *Sci. Data* **2020**, *7*, 1–13, DOI: 10.1038/s41597-020-0491-x.

(6) Drisdell, W. S. et al. Determining Atomic-Scale Structure and Composition of Organo-Lead Halide Perovskites by Combining High-Resolution X-Ray Absorption Spectroscopy and First-Principles Calculations. *ACS Energy Lett.* **2017**, *2*, 1183–1189, DOI: 10.1021/acsenergylett.7b00182.

(7) Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials discovery and design using machine learning. *J Mater.* **2017**, *3*, 159–177, DOI: 10.1016/j.jmat.2017.08.002.

(8) Ward, L. et al. Matminer: An Open Source Toolkit for Materials Data Mining. *Comput. Mater. Sci.* **2018**, *152*, 60–69, DOI: 10.1016/j.commatsci.2018.05.018.

(9) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555, DOI: 10.1038/s41586-018-0337-2.

(10) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002, DOI: 10.1063/1.4812323.

(11) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65*, 1501–1509, DOI: 10.1007/s11837-013-0755-4.

(12) Bergerhoff, G.; Brown, I. D. *Crystallographic Databases*; International Union of Crystallography, Chester, 1987; pp 77–95.

(13) Ward, L.; Liu, R.; Krishna, A.; Hedge, V. I.; Agrawal, A.; Choudhary, A.; Wolverton, C. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* **2017**, *96*, 024104, DOI: 10.1103/PhysRevB.96.024104.

(14) Pham, T. L.; Kino, H.; Terakura, K.; Miyake, T.; Tsuda, K.; Takigawa, I.; Dam, H. C. Machine learning reveals orbital interaction in materials. *Sci. Technol. Adv. Mater.* **2017**, *18*, 756–765, DOI: 10.1080/14686996.2017.1378060.

(15) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Materials Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301, DOI: 10.1103/PhysRevLett.120.145301.

(16) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31*, 3564–3572, DOI: 10.1021/acs.chemmater.9b01294.

(17) George, J.; Waroquiers, D.; Stefano, D. D.; Petretto, G.; Rignanese, G.; Hautier, G. The Limited Predictive Power of the Pauling Rules. *Angew. Chem.* **2020**, *59*, 7569–7575, DOI: 10.1002/anie.202000829.

(18) Brunner, G. O. A definition of coordination and its relevance in the structure types AlB_2 and NiAs. *Acta Crystallogr. A* **1977**, *33*, 226–227, DOI: 10.1107/S0567739477000461.

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

(20) Hoppe, R. Effective coordination numbers (ECoN) and mean fictive ionic radii (MEFIR). *Z. Kristallogr. Cryst. Mater.* **1979**, *150*, 23–52, DOI: 10.1524/zkri.1979.150.14.23.

(21) O'Keeffe, M. A proposed rigorous definition of coordination number. *Acta Crystallogr. A* **1979**, *35*, 772–775, DOI: 10.1107/S0567739479001765.

(22) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A

robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314–319, DOI: 10.1016/j.commatsci.2012.10.028.

(23) Voronoi, G. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Premier mémoire. Sur quelques propriétés des formes quadratiques positives parfaites. *J. Reine Angew. Math* **1908**, *133*, 97–178, DOI: 10.1515/crll.1908.133.97.

(24) Hellenbrandt, M. The Inorganic Crystal Structure Database (ICSD)—Present and Future. *Crystallogr. Rev.* **2004**, *10*, 17–22, DOI: 10.1080/08893110410001664882.

(25) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **2017**, *8*, 1–12, DOI: 10.1038/ncomms15679.

(26) Zimmermann, N. E. R.; Horton, M. K.; Jain, A.; Haranczyk, M. Assessing Local Structure Motifs Using Order Parameters for Motif Recognition, Interstitial Identification, and Diffusion Path Characterization. *Front. Mater.* **2017**, *4*, 34, DOI: 10.3389/fmats.2017.00034.

(27) Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/.

(28) Default bond distances (when charge is unknown) from Jmol. Available at: https://github.com/materialsproject/pymatgen/blob/master/pymatgen/analysis/bonds_jmol_ob.yaml.

(29) Allred, A. L.; Rochow, E. G. A scale of electronegativity based on electrostatic force. *J. Inorg. Nucl. Chem.* **1958**, *5*, 264–268, DOI: 10.1016/0022-1902(58)80003-2.

(30) Frank, F. C.; Kasper, J. S. Complex alloy structures regarded as sphere pack-

ings. I. Definitions and basic principles. *Acta Crystallogr.* **1958**, *11*, 184–190, DOI: 10.1107/S0365110X58000487.

(31) Shannon, R. D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr. A* **1976**, *32*, 751–767, DOI: 10.1107/S0567739476001551.

(32) Zimmermann, N. E. R.; Jain, A. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC Adv.* **2020**, *10*, 6063–6081, DOI: 10.1039/C9RA07755C.

(33) MaterialsCoord. (2020). Available at: https://github.com/hackingmaterials/materials-coord.

(34) Lander, G. H.; Mueller, M. H. Neutron diffraction study of a α-uranium at lower temperatures. *Acta Crystallogr. B* **1970**, *26*, 129–136, DOI: 10.1107/S0567740870002066.

(35) Donohue, J. *The Structure of the Elements*; Robert E. Krieger Publishing Company, 1982.

(36) Schiferl, D.; Barrett, C. S. The crystal structure of arsenic at 4.2, 78 and 299K. *J. Appl. Crystallogr.* **1969**, *2*, 30–36, DOI: 10.1107/S0021889869006443.

(37) Brown, A.; Rundqvist, S. Refinement of the crystal structure of black phosphorus. *Acta Crystallogr. A* **1965**, *19*, 684–685, DOI: 10.1107/S0365110X65004140.

(38) Daane, A. H.; Rundle, R. E.; Smith, H. G.; Spedding, F. H. The crystal structure of samarium. *Acta Crystallogr. A* **1954**, *7*, 532–535, DOI: 10.1107/S0365110X54001818.

(39) Oberteuffer, J. A.; Ibers, J. A. A refinement of the atomic and thermal parameters of α-manganese from a single crystal. *Acta Crystallogr. B* **1970**, *26*, 1499–1504, DOI: 10.1107/S0567740870004399.

(40) Shoemaker, C. B.; Shoemaker, D. P.; Hopkins, T. E.; Yindepit, S. Refinement of the structure of β-manganese and of a related phase in the Mn–Ni–Si system. *Acta Crystallogr. B* **1978**, *34*, 3573–3576, DOI: 10.1107/S0567740878011620.

(41) Wołcyrz, M.; Kubiak, R.; Maciejewski, S. X-ray investigation of thermal expansion and atomic thermal vibrations of tin, indium, and their alloys. *physica status solidi (b)* **1981**, *107*, 245–253, DOI: 10.1002/pssb.2221070125.

(42) Lucovsky, G.; Mooradian, A.; Taylor, W.; Wright, G.; Keezer, R. Identification of the fundamental vibrational modes of trigonal, α-monoclinic and amorphous selenium. *Solid State Commun.* **1967**, *5*, 113–117, DOI: 10.1016/0038-1098(67)90006-3.

(43) Gruner, J. W. Crystal structure types. *Am. Mineral* **1929**, *14*, 173–187.

(44) Madelung, O. *Semiconductors: Data Handbook.*; Springer, 2013.

(45) Bursill, L. A. Structural Relationships between β-Gallia, Rutile, Hollandite, Psilomelane, Ramsdellite and Gallium Titanate Type Structures. *Acta Crystallogr. B* **1979**, *35*, 530–538, DOI: 10.1107/S0567740879004088.

(46) Pauling, L.; Hendricks, S. B. The crystal structures of hematite and corundum. *J. Am. Chem. Soc.* **1925**, *47*, 781–790, DOI: 10.1021/ja01680a027.

(47) Brandon, J. K.; Brizard, R. Y.; Chieh, P. C.; McMillan, R. K.; Pearson, W. B. New refinements of the γ brass type structures Cu_5Zn_8, Cu_5Cd_8 and Fe_3Zn_{10}. *Acta Crystallogr. B* **1974**, *30*, 1412–1417, DOI: 10.1107/S0567740874004997.

(48) Ondruš, P.; Veselovský, F.; Gabašová, A.; Hloušek, J.; Šrein, V.; Vavřín, I.; Skála, R.; Sejkora, J.; Drábek, M. Primary minerals of the Jáchymov ore district. *J. Geosci.* **2003**, *48*, 19–147.

(49) Fleet, M. E. The structure of magnetite. *Acta Crystallogr. B* **1981**, *37*, 917–920, DOI: 10.1107/S0567740881004597.

(50) Holtzberg, F.; Methfessel, S. Rare-Earth Compounds with the Th_3P_4-Type Structure. *J. Appl. Phys.* **1966**, *37*, 1433–1435, DOI: 10.1063/1.1708500.

(51) Fayek, M. K.; Leciejewicz, J. Neutron-diffraction Study of Pb_3O_4. *Z. Anorg. Allg. Chem.* *336*, 104–109, DOI: 10.1002/zaac.19653360115.

(52) Goiffon, A.; Jumas, J.-C.; Maurin, M.; Philippot, E. Etude comparée à diverses températures (173, 293 et 373° K) des structures de type quartz α des phases MIII-IXVO4 (MIII= Al, Ga et XV= P, As). *J. Solid State Chem.* **1986**, *61*, 384 – 396, DOI: 10.1016/0022-4596(86)90047-2.

(53) Achary, S.; Jayakumar, O.; Tyagi, A.; Kulshresththa, S. Preparation, phase transition and thermal expansion studies on low-cristobalite type $Al_{(1-x)}Ga_xPO_4$ (x=0.0, 0.20, 0.50, 0.80 and 1.00). *J. Solid State Chem.* **2003**, *176*, 37–46, DOI: https://doi.org/10.1016/S0022-4596(03)00341-4.

(54) Haines, J.; Cambon, O.; Astier, R.; Fertey, P.; Chateau, C. Crystal structures of α-quartz homeotypes boron phosphate and boron arsenate: structure-property relationships. *Z. Kristallogr. Cryst. Mater.* **2004**, *219*, 32–37, DOI: 10.1524/zkri.219.1.32.25397.

(55) Muller, O.; Roy, R. *The major ternary structural families*; Springer Verlag, 1974.

(56) Abrahams, S. C.; Reddy, J. M. Crystal Structure of the Transition-Metal Molybdates. I. Paramagnetic Alpha-$MnMoO_4$. *J. Chem. Phys.* **1965**, *43*, 2533–2543, DOI: 10.1063/1.1697153.

(57) Bulou, A.; Nouet, J. Structural phase transitions in ferroelastic $TlAlF_4$: DSC investigations and structures determinations by neutron powder profile refinement. *J. Phys. C: Solid State Phys.* **1987**, *20*, 2885–2900, DOI: 10.1088/0022-3719/20/19/014.

(58) Fábry, J.; Pérez-Mato, J. M. Some stereochemical criteria concerning the structural stability of A_2BX_4 compounds of type β-K_2SO_4. *Phase Transit.* **1994**, *49*, 193–229, DOI: 10.1080/01411599408201174.

(59) Rannev, N. V.; Shchedrin, B. M.; Venevtsev, Y. N. Crystal structure of ferrielectric stibiumniobite $SbNbO_4$. *Ferroelectrics* **1976**, *13*, 523–525, DOI: 10.1080/00150197608236657.

(60) Zachariasen, W. H. Crystal chemical studies of the 5f-series of elements. XXI. The crystal structure of magnesium orthouranate. *Acta Crystallogr. A* **1954**, *7*, 788–791, DOI: 10.1107/S0365110X54002459.

(61) Baran, E. J. Materials belonging to the $CrVO_4$ structure type: preparation, crystal chemistry and physicochemical properties. *J. Mater. Sci.* **1998**, *33*, 2479–2497, DOI: 10.1023/A:1004380530309.

(62) Wildner, M.; Giester, G. Crystal structure refinements of synthetic chalcocyanite ($CuSO_4$) and zincosite ($ZnSO_4$). *Mineral. Petrol.* **1988**, *39*, 201–209, DOI: 10.1007/BF01163035.

(63) Errandonea, D.; Pellicer-Porres, J.; Manjón, F. J.; Segura, A.; Ferrer-Roca, C.; Kumar, R. S.; Tschauner, O.; Rodríguez-Hernández, P.; López-Solano, J.; Radescu, S.; Mujica, A.; Muñoz, A.; Aquilanti, G. High-pressure structural study of the scheelite tungstates $CaWO_4$ and $SrWO_4$. *Phys. Rev. B* **2005**, *72*, 174106, DOI: 10.1103/PhysRevB.72.174106.

(64) Larsson, A.-K.; Withers, R. L.; Perez-Mato, J.; Fitz Gerald, J.; Saines, P.; Kennedy, B.; Liu, Y. On the microstructure and symmetry of apparently hexagonal $BaAl_2O_4$. *J. Solid State Chem.* **2008**, *181*, 1816–1823, DOI: 10.1016/j.jssc.2008.03.043.

(65) Hazen, R. M.; Au, A. Y. High-pressure crystal chemistry of phenakite (Be_2SiO_4)

and bertrandite ($Be_4Si_2O_7(OH)_2$). *Phys. Chem. Miner.* **1986**, *13*, 69–78, DOI: 10.1007/BF00311896.

(66) Morgan, B. Mechanistic Origin of Superionic Lithium Diffusion in Anion-Disordered Li_6PS_5X Argyrodites. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.12349703.v1, 2020.

(67) Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat.* **1901**, *37*, 547–579, DOI: 10.5169/seals-266450.

(68) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864, DOI: 10.1103/PhysRev.136.B864.

(69) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865, DOI: 10.1103/PhysRevLett.77.3865.

(70) Langreth, D. C.; Mehl, M. J. Beyond the local-density approximation in calculations of ground-state electronic properties. *Phys. Rev. B* **1983**, *28*, 1809, DOI: 10.1103/PhysRevB.28.1809.

(71) Dudarev, S. L.; Botton, G. A.; Savrasov, S. Y.; Humphreys, C. J.; Sutton, A. P. Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA+U study. *Phys. Rev. B* **1998**, *57*, 1505, DOI: 10.1103/PhysRevB.57.1505.

(72) Csonka, G. I.; Perdew, J. P.; Ruzsinszky, A.; Philipsen, P. H. T.; Lebègue, S.; Paier, J.; Vydrov, O. A.; Ángyán, J. G. Assessing the performance of recent density functionals for bulk solids. *Phys. Rev. B* **2009**, *79*, 155107, DOI: 10.1103/PhysRevB.79.155107.

(73) Ganose, A. M.; Jain, A. Robocrystallographer: Automated crystal structure text descriptions and analysis. *MRS Commun.* **2019**, *9*, 874–881, DOI: 10.1557/mrc.2019.94.

For Table of Contents Only

Automated bonding algorithms that detect coordination environments are useful in both traditional and emerging aspects of materials science. In this work, we introduce MaterialsCoord, an open-source software package for comparing bonding algorithms by determining how well they match literature descriptions of bonding in elemental, binary, and ternary crystal structures. We also detail a novel algorithm called CrystalNN, which we compare against existing algorithms on a diverse set of prototypical crystal structures.