MRS Communications (2019), *9*, 874–881
© Materials Research Society, 2019
doi:10.1557/mrc.2019.94

MRS Communications

Artificial Intelligence Research Letter

Robocrystallographer: automated crystal structure text descriptions and analysis

Alex M. Ganose and **Anubhav Jain,** Lawrence Berkeley National Laboratory, Energy Technologies Area, 1 Cyclotron Road, Berkeley, CA 94720, USA
Address all correspondence to Anubhav Jain at **ajain@lbl.gov**

(Received 14 January 2019; accepted 1 July 2019)

Abstract

Our ability to describe crystal structure features is of crucial importance when attempting to understand structure–property relationships in the solid state. In this paper, the authors introduce robocrystallographer, an open-source toolkit for analyzing crystal structures. This package combines new and existing open-source analysis tools to provide structural information, including the local coordination and polyhedral type, polyhedral connectivity, octahedral tilt angles, component-dimensionality, and molecule-within-crystal and fuzzy prototype identification. Using this information, robocrystallographer can generate text-based descriptions of crystal structures that resemble descriptions written by human crystallographers. The authors use robocrystallographer to investigate the dimensionalities of all compounds in the Materials Project database and highlight its potential in machine learning studies.

Introduction

The crystal structure of a material plays a fundamental role in determining its properties.[1,2] This is best exemplified by carbon allotropes[3]—diamond is extremely hard and electrically insulating, whereas graphite is soft and semi-metallic. Even minor structural modifications can have a profound effect on a broad array of properties, ranging from ferroelectricity and piezoelectricity (e.g., phase-dependent ferroelectricity in $BaTiO_3$)[4] to conductivity (e.g., metal to insulator Peierls distortions)[5] and photocatalytic activity (e.g., observed only in Rutile rather than Anatase TiO_2).[6] Accordingly, our ability to describe and understand such structural features is of crucial importance when characterizing new and existing materials.[7]

There now exists an increasing number of packages for programmatically analyzing crystal structures. These mainly fall into two categories. The first provides information on the global structure, such as the dimensionality,[8–11] the symmetry information,[12] and whether the structure matches a known mineral prototype.[13] The second analyzes local coordination environments and site geometries (e.g., the recent ChemEnv[14] and LocalEnv[15] packages). Currently, however, there are not yet tools for describing semi-local structure, i.e., how the local geometry connects throughout space to form the overall structure (Fig. 1). Furthermore, to our knowledge, no package provides human-readable descriptions of crystal structure resembling that found in a journal article.

In this paper, we introduce robocrystallographer, an open-source toolkit for analyzing crystal structures. We illustrate how robocrystallographer can extract local, semi-local, and global structure features, and use these to generate human-readable text descriptions and machine learning features. Next, we showcase the code on several example structures and use it to investigate the dimensionally of all materials in the Materials Project database. Lastly, we demonstrate how robocrystallographer can be used in statistical machine learning models to improve the accuracy of predictions of elastic properties.

Overview and code design

Robocrystallographer follows a modular design, with individual sub-packages for different types of analysis. In this way, individual components may be used as standalone tools. Robocrystallographer builds upon the many open-source tools that already exist for analyzing crystal structures, including pymatgen[16] for manipulating structure objects and local environment analysis, spglib[12] for symmetry analysis, matminer[17] for structure fingerprinting, the AFLOW[13] prototype library for framework analysis, OpenBabel[18] for processing molecular structures, and PubChemPy[19] for interfacing with the PubChem website.[20] The three primary functions of the code (Fig. 1) are: (i) to condense a crystal structure into a descriptive, machine-readable Javascript Object Notation (JSON) representation containing the structure features; (ii) to generate a human-readable text description of the structure from the JSON format; and (iii) to extract a set of machine learning features from the descriptive JSON. We briefly outline how these functions are achieved in robocrystallographer.

https://doi.org/10.1557/mrc.2019.94 Published online by Cambridge University Press

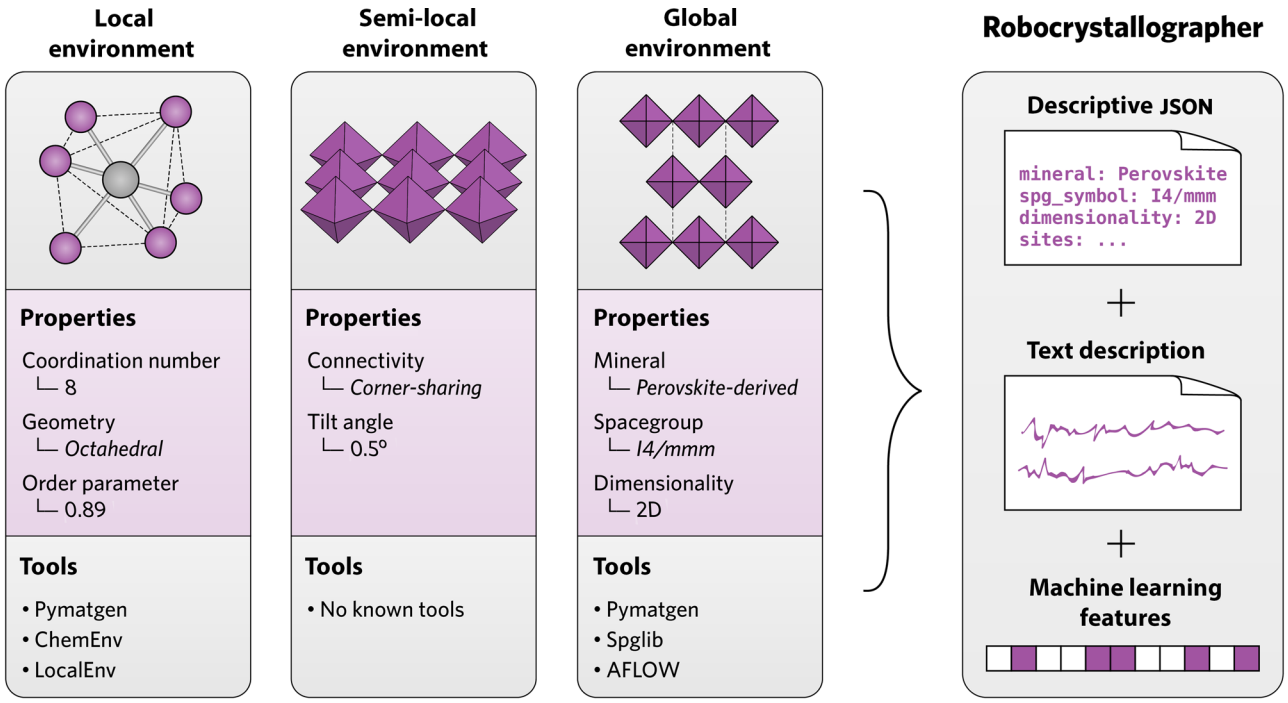


Figure 1. A crystal structure can be broken down into the local environment (e.g., site coordination number and geometry), the semi-local environment detailing how the sites connect throughout space (e.g., polyhedral connectivity and tilt angles), and the global environment (e.g., mineral type and symmetry information). Robocrystallographer analyzes each of these components and compiles the information into machine-readable (JSON), human-readable (text), and machine learning formats.

Generating the descriptive JSON

The conversion from crystal structure to descriptive JSON is handled within the Condense submodule of robocrystallographer (Fig. 2). The first step is to assign oxidation states to all sites in the structure. These are used for descriptive purposes and to increase accuracy when determining the structural bonding. The oxidation states are assigned based on (i) achieving charge balance and (ii) using oxidation states most consistent with statistics of oxidation states found in the International Crystal Structure Database. While this method performs well for most structures, it can fail when unusual oxidation states are present (e.g., peroxides and persulfides) and in cases of disproportionation (e.g., Pb^{2+} and Pb^{4+} in Pb_3O_4—in which an averaged oxidation state of 2.67+ is assigned to all Pb sites).

Next, global structure properties are evaluated including the symmetry information (e.g., space group using the spglib[12] library) and whether the structure matches any known mineral prototypes. Mineral matching is first attempted using the AFLOW structure library prototype matcher available in pymatgen.[13,16] This algorithm relies on StructureMatcher, an affine mapping technique with tunable tolerances. Matches are determined by first aligning the crystal lattices (considering different settings and unit cells) and then computing atomic distances. Robocrystallographer has the option to simplify zero-dimensional (0D) clusters of sites to a single site, allowing for the correct identification of prototypes containing organic molecules such as the mixed organic–inorganic hybrid perovskites (see sections on dimensionality analysis and molecule analysis in Supplementary Material). If no match is found, "fuzzy" matching will be performed using the geometric structure fingerprint calculated using matminer.[17] The structure fingerprint (SiteStatsFingerprint in matminer) encodes information about the different types of local environments (e.g., "tetrahedral") present in a crystal structure. We describe structures that do not directly match a prototype but possess a small Euclidean fingerprint distance to that prototype, as being similar to or "like" that structure (e.g., "perovskite-like"). Furthermore, if a structure matches a known prototype but contains a different number of atomic elements, the structure is described as "derived" from that structure (e.g., "perovskite-derived"). Additional details on the fuzzy matching algorithm are provided in Supplementary Material.

Next, the structural bonding is determined using one of the nearest-neighbor routines provided by pymatgen.[21] It is essential that bonding is assigned correctly due to its importance throughout all subsequent analysis steps. Robocrystallographer defaults to using the CrystalNN bonding algorithm, which, in our testing, produced the most reasonable bonding interpretation for a wide array of systems. Details of this approach can be found in the pymatgen documentation[21] and will be the subject of a future work. In essence, CrystalNN uses Voronoi decomposition[22] and solid angle weights to determine the

https://doi.org/10.1557/mrc.2019.94 Published online by Cambridge University Press

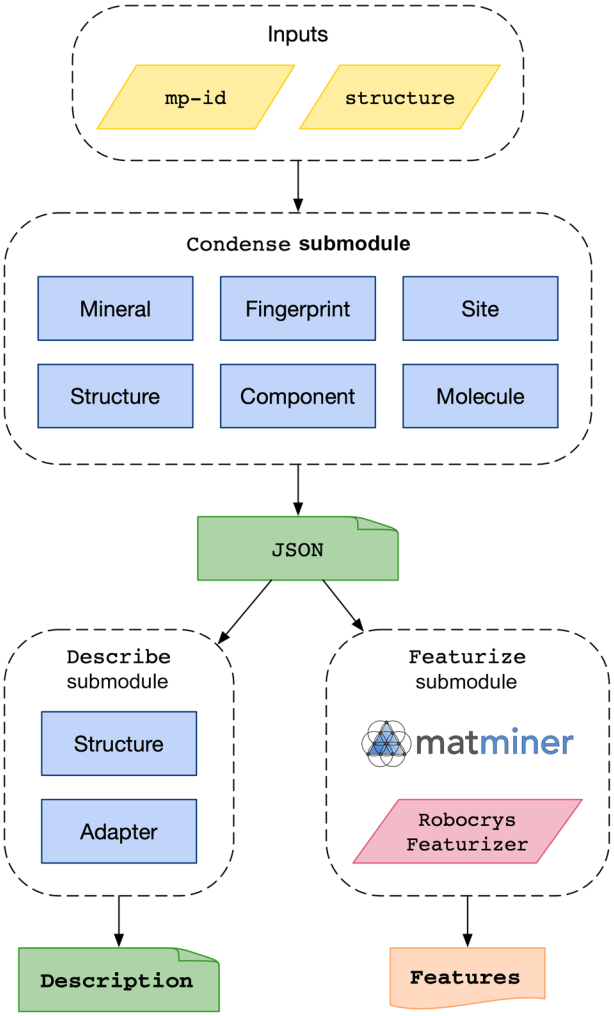


Figure 2. Schematic of the robocrystallographer program indicating code components within the core submodules (Condense, Describe, and Featurize) and program flow. The robocrystallographer machine learning featurizer extends the matminer code.[17]

probability of various coordination environments and selects the one with highest probability.

We next identify the bonded components (all sets of sites that are connected by bonds) that comprise the structure using the StructureGraph module in pymatgen.[16] For each structure component, we identify its dimensionality using the modified breadth-first search approach detailed by Larsen et al.[9] and as implemented in pymatgen. While this method is, in principle, exact, the performance depends on the accuracy of the assigned bonding. A comparison between different bonding algorithms and their effect on dimensionality for 40 reference structures is provided in Supplementary Material. If the component is one- or two-dimensional (1D or 2D), its orientation is determined using singular value decomposition. By describing the dimensionalities of individual components, we obtain a much deeper understanding of the material geometry than available from pre-existing tools. When a single structure contains multiple components with different dimensionalities (e.g., 1D chains within a three-dimensional (3D) framework), we take the largest dimensionality of all components as the dimensionality of the overall structure. Lastly, if the cell contains two or more 2D components with the same Miller index and different compositions, the structure is labeled as a van der Waals heterostructure and the minimum repeating sequence of the components is identified.

Any 0D components are compared against the PubChem database to determine if they match known molecules. As mentioned previously, such components can be reduced to a single site for prototype identification. Thus, hybrid systems in which molecules are embedded in a crystal (such as the hybrid metal-organic halide systems) are correctly identified as being perovskite, and in many cases, the molecular component can be explicitly matched with a known molecule (see Supplementary Material for such an example).

Having covered the more global aspects of structure, we next analyze the local and semi-local structure. Using spglib[12] and pymatgen,[16] we identify the inequivalent components and, within each component, the symmetrically inequivalent atoms. The LocalEnv module of pymatgen is used to determine the site geometry and obtain information on the nearest and next-nearest neighbors. This module uses specially designed order parameters[15] to match the geometry of a near-neighbor configuration to known patterns such as "octahedral," "tetrahedral," and "square planar" (as of this writing, 24 distinct motifs are available). These order parameters can detect and report distortions from perfect motifs in a continuous manner. We note that the ChemEnv[14] package is also available for this purpose; we use the order parameter technique for speed as well as compatibility with custom near-neighbor algorithms. To assess the connectivity between a site and its next-nearest neighbor(s) (i.e., polyhedral connectivity), we determine the number of nearest neighbors shared between the two sites. If the two sites share one, two, or more nearest neighbors, the connectivity is determined as corner-, edge-, and face-sharing, respectively. The corner-sharing octahedral tilt angles (θ_{tilt}) are calculated as follows:

$$\theta_{\text{tilt}} = 180 - \angle(X_{\text{oct}} - Y_{\text{NN}} - Z_{\text{oct}}),$$

where X_{oct} and Z_{oct} are two octahedral sites, and Y_{NN} is a shared nearest-neighbor site.

Most materials chemistry packages represent chemical formulae in a reduced form based on the overall composition of the system, with the elements ordered by their electronegativities. This method specifically ignores the connectivity of sites within the structure. This is useful when the structural properties of a compound are not known but result in confusing or unrecognizable formulae for some materials, for example "$Li(N_3)(H_2O)$" is reduced to "LiH_2N_3O." In robocrystallographer, the formulae of the individual bonded components are used to reconstruct the overall chemical formula, allowing for more

https://doi.org/10.1557/mrc.2019.94 Published online by Cambridge University Press

understandable formulae that more faithfully describe the structural bonding.

Together, the full analysis is compiled into a descriptive JSON representation, containing keys for each property including the mineral, formula, space group information, and dimensionality. A summary of this process along with some example JSON data is provided in Supplementary Material. The conversion of the descriptive JSON to human-readable text description and machine learnable features is handled within the Describe and Featurize submodules of robocrystallographer, respectively (Fig. 2). More details of these procedures are provided in Supplementary Material.

General information and usage modes

Robocrystallographer is compatible with Python 3.6+. The application programming interface (API) is modular and fully documented, with 60 unit-tests covering the entire codebase. A schematic indicating how the components interact is provided in Fig. 2. The code is released under a modified BSD license and is available open source at https://github.com/hackingmaterials/robocrystallographer. Robocrystallographer can be used from either the command line or via the Python API with the full details provided in the package documentation. The time needed to generate structural descriptions is discussed in Supplementary Material; a rough estimate is 0.22 s per site in the structure (e.g., 4–5 s for a structure with 20 sites).

Results

Crystal structure descriptions

To highlight the functionality of robocrystallographer, we have generated automated text descriptions (using the Describe module of robocrystallographer) for several crystal structures of varying complexity. In this section, we reproduce verbatim these descriptions, albeit with slight typesetting adjustments. Note that in each case, oxidation states are inferred automatically using the routines available in pymatgen. In the simplest case, the output of robocrystallographer on GaAs[23] (Fig. 3(a)) is: "GaAs is zincblende structured and crystallizes in the cubic $F\bar{4}3m$ space group. Ga^{3+} is bonded to four equivalent As^{3-} atoms to form corner-sharing $GaAs_4$ tetrahedra. All Ga–As bond lengths are 2.49 Å. As^{3-} is bonded in a tetrahedral geometry to four equivalent Ga^{3+} atoms." Despite the simple structure, this description demonstrates many of the core features of robocrystallographer. The mineral prototype is matched correctly, as is the space group information. The geometry of each site is determined and the presence of corner-sharing tetrahedra is identified. Lastly, the code keeps track of which bond lengths have been described such that each length is described only once. Overall, the description is clearly understandable and resembles a human description of the structure. We note that this description, as well as the ones to follow, involved no human intervention; the only information provided to robocrystallographer was the structural information (i.e., lattice parameters and site positions) of GaAs.

For more complex structures with multiple inequivalent sites and a mixture of site connectivities (corner, edge, or face-sharing), the output is longer but remains readable. To increase clarity, we can display the symmetry labels for each site. For example, for $CrVO_4$[24] (Fig. 3(b)), the output is as follows: "$CrVO_4$ crystallizes in the orthorhombic *Cmcm* space group. $V(1)^{5+}$ is bonded to two equivalent $O(1)^{2-}$ and two equivalent $O(2)^{2-}$ atoms to form VO_4 tetrahedra that share corners with six equivalent $Cr(1)O_6$ octahedra. The corner-sharing octahedra tilt angles range from 47–54°. Both V(1)–O(1) bond lengths are 1.82 Å. Both V(1)–O(2) bond lengths are 1.69 Å. $Cr(1)3^+$ is bonded to two equivalent $O(2)^{2-}$ and four equivalent $O(1)^{2-}$ atoms to form CrO_6 octahedra that share corners with six equivalent $V(1)O_4$ tetrahedra and edges with two equivalent $Cr(1)O_6$ octahedra. Both Cr(1)–O(2) bond lengths are 1.99 Å. All Cr(1)–O(1) bond lengths are 2.04 Å. There are two inequivalent

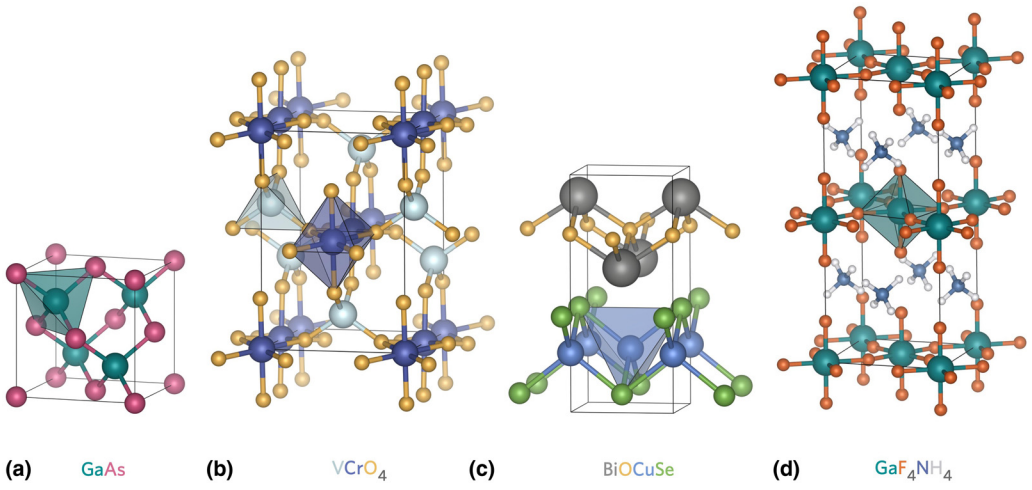


Figure 3. Crystal structures of (a) GaAs, (b) $VCrO_4$, (c) BiOCuSe, and (d) GaF_4NH_4. Only cation polyhedral are drawn.

https://doi.org/10.1557/mrc.2019.94 Published online by Cambridge University Press

O^{2-} sites. $O(1)^{2-}$ is bonded in a distorted trigonal planar geometry to one $V(1)^{5+}$ and two equivalent $Cr(1)^{3+}$ atoms. $O(2)^{2-}$ is bonded in a distorted bent 120 degrees geometry to one $V(1)^{5+}$ and one $Cr(1)^{3+}$ atom."

Support for low dimensionality structures can be observed in the description for BiOCuSe[25] (Fig. 3(c)): "BiOCuSe is parent of FeAs superconductors structured and crystallizes in the tetragonal *P*4/*nmm* space group. The structure is two-dimensional and consists of one BiO sheet oriented in the (0, 0, 1) direction and one CuSe sheet oriented in the (0, 0, 1) direction. In the BiO sheet, Bi^{3+} is bonded in a 4-coordinate geometry to four equivalent O^{2-} atoms. All Bi–O bond lengths are 2.35 Å. O^{2-} is bonded in a tetrahedral geometry to four equivalent Bi^{3+} atoms. In the CuSe sheet, Cu^{1+} is bonded to four equivalent Se^{2-} atoms to form a mixture of edge and corner-sharing $CuSe_4$ tetrahedra. All Cu–Se bond lengths are 2.52 Å. Se^{2-} is bonded in a 4-coordinate geometry to four equivalent Cu^{1+} atoms." Here, robocrystallographer identifies the two separate structure components and reports their dimensionality and orientation. The use of components is further reflected in the chemical formula. Namely, if the elements were solely ordered by their electronegativity, the formula would instead be reported as "CuBiSeO." By describing each component separately, the description becomes more readable and easier to understand in terms of structural components. We note, however, that such ordering preferences may depend on the type of analysis being performed (this particular structure is usually called "BiCuSeO" in the literature).

Robocrystallographer can process structures with mixed dimensionalities, with 0D structures matched against the PubChem database.[20] This is exemplified by the description for GaF_4NH_4[26] (Fig. 3(d)): "GaF_4NH_4 crystallizes in the tetragonal *I*4/*mcm* space group. The structure is two-dimensional and consists of two ammonium molecules and one GaF_4 sheet oriented in the (0, 0, 1) direction. In the GaF_4 sheet, Ga^{3+} is bonded to six F^{1-} atoms to form corner-sharing GaF_6 octahedra. The corner-sharing octahedral tilt angles are 27°. There are two shorter (1.87 Å) and four longer (1.95 Å) Ga–F bond lengths. There are two inequivalent F^{1-} sites. In the first F^{1-} site, F^{1-} is bonded in a bent 150° geometry to two equivalent Ga^{3+} atoms. In the second F^{1-} site, F^{1-} is bonded in a single-bond geometry to one Ga^{3+} atom." We have thus demonstrated that the notable aspects of a diverse set of structures can be captured automatically and reported to a user in a highly readable manner by robocrystallographer.

Statistical analysis of structure dimensionalities

We have employed the component-based dimensionality finding routines implemented in robocrystallographer to investigate the statistics of structure dimensionalities in the Materials Project database[27] (Fig. 4). At the time of writing, this database comprises 133,688 crystal structures which we filter to remove very unstable entries—those with an energy above the hull of >500 meV/atom—providing 114,300 final materials. We ran robocrystallographer on these structures and used the JSON representation to analyze and search through the

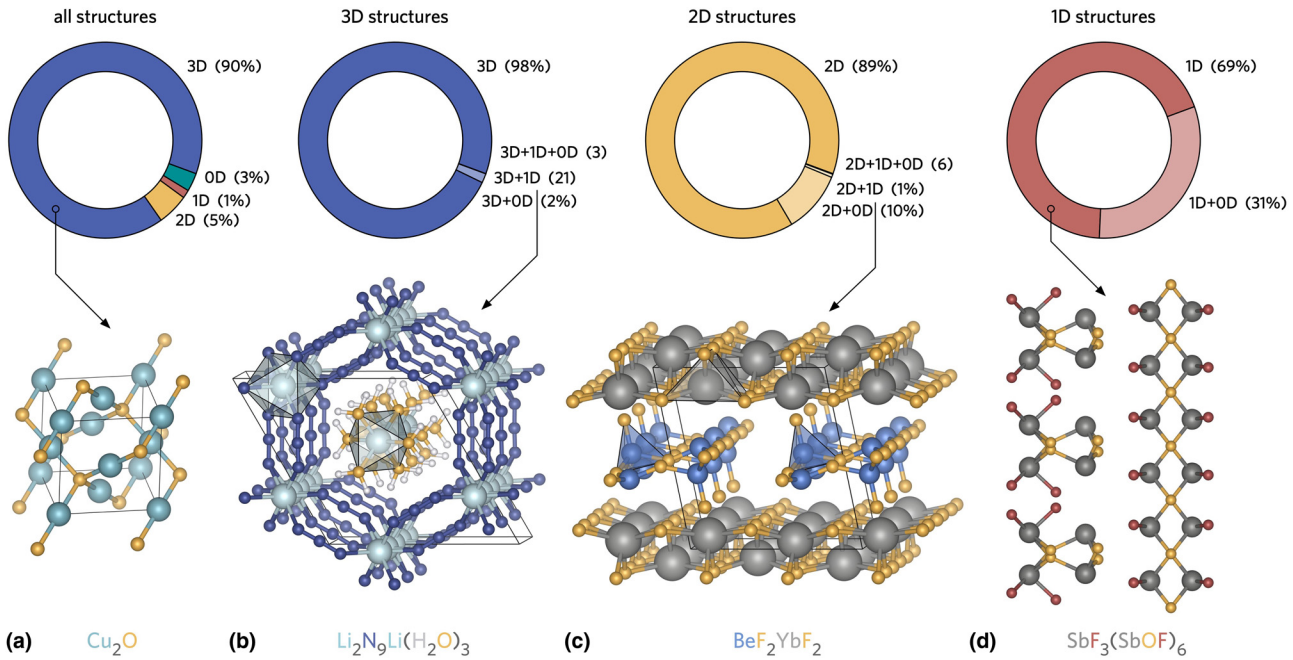


Figure 4. Statistics of structure dimensionalities in the Materials Project database and example crystal structures. For clarity, only a portion of the $SbF_3(SbOF)_6$ crystal structure is displayed.

https://doi.org/10.1557/mrc.2019.94 Published online by Cambridge University Press

database. We note that, while the dimensionality finding algorithm is exact for a given bonded structure, limitations in determining the crystal bonding mean that the results presented here will only provide a qualitative guide to the true dimensionality statistics. An assessment of different bonding schemes and their impact on dimensionality is provided in Supplementary Material. In general, the bonding scheme used in this work (CrystalNN) tends to slightly overestimate the bonding, leading to an artificial increase in the number of 3D structures (see Supplementary Material).

We find that the vast majority of compounds in the Materials Project (90%) are 3D. 2D, 1D, and 0D structures comprise 5%, 1%, and 3% of the database, respectively. The component-based analysis allows for an enhanced understanding of structure dimensionalities. For example, by filtering the structures for those containing more than one 3D component, we can easily identify interpenetrated structures such as Cu_2O (Fig. 4(a)), with additional perspectives provided in Supplementary Material.

The component analysis can also be used to investigate the statistics of mixed dimensionality structures. As the overall dimensionality of a structure is determined by the largest component dimensionality, an *n*-dimensional structure can also include components of lower dimensionality. The breakdown for each set of structure dimensionalities is displayed in Fig. 4. For 3D structures, the majority contain only 3D components, with a small percentage (2%) containing intercalant atoms (3D + 0D). We find that twenty one 3D compounds also contain 1D ribbons passing through open channels in the structure, for example $Li_2N_9Li(H_2O)_3$ (Fig. 4(b)), and three compounds possess a combination of 3D, 1D, and 0D components.

Similarly, for 2D structures, the majority are comprised of solely 2D components (89%), with smaller subsets containing mixed 2D/0D components (10%), mixed 2D/1D components (1%), and mixed 2D/1D/0D components (six materials). An example of a structure with mixed 2D/1D dimensionality, BeF_2YbF_2, is illustrated in Fig. 4(c). For 2D structures containing more than one 2D component, the orientations of the 2D sheets are always aligned in the same direction. 1D structures exhibit the largest percentage of mixed component-dimensionality compounds (31%); however, again structures with only a single component-dimensionality compound dominate (69%). For 1D structures, however, several compounds possess ribbons bonded in orthogonal directions (as identified through the orientation analysis in robocrystallographer), such as $SbF_3(SbOF)_6$ (Fig. 4(d)).

Using robocrystallographer for machine learning

To demonstrate the utility of robocrystallographer in statistical machine learning, we have tested the effect of including robocrystallographer features when modeling elastic properties. Specifically, we have modeled the density functional theory (DFT) calculations of bulk modulus (K) produced by de Jong,[28] provided as the "elastic_tensor_2015" dataset in the matminer package.[17] This dataset consists of 1181 inorganic compounds. Three models were trialed, the first included only composition-based features,[29] the second included composition and sine matrix structural features (implemented in the SineColoumbMatrix featurizer in matminer[30]), and the final model included composition and robocrystallographer features. As the purpose of this study is to demonstrate the potential of robocrystallographer features, we used the same simple regression model (random forest with 100 estimators) for all feature sets. The total number of features in each model was reduced to 10 using the MultiSURF algorithm, as implemented in the Scikit-Rebate package.[31,32] Full details are provided in Supplementary Material. The model including the robocrystallographer features produced slightly more accurate predictions of bulk modulus (mean squared error [MSE] of 537 GPa) relative to the model containing just composition-based features (MSE of 544 GPa) and that containing both structure and composition features (MSE of 544 GPa). Perhaps, more importantly, a plot of predicted versus DFT calculated bulk modulus demonstrates that many outliers are reduced with the addition of the robocrystallographer features (see Supplementary Material). In addition to improved performance, the features produced by robocrystallographer allow for interpretable models through the analysis of feature importance plots (as provided in the Supplementary Material). In contrast, the sine matrix structural features do not have a clear physical interpretation which makes it difficult to rationalize models containing these features.

Discussion

Robocrystallographer brings together many previously unconnected analysis packages into a centralized toolkit for better understanding crystal structures. Once a structure is condensed into a descriptive JSON representation, the user can perform many different types of analyses and searches on the data. For example, the data can be used to classify materials by properties such as polyhedral connectivity that were previously difficult to calculate automatically. Furthermore, the JSON data are easily searchable and therefore could be used to create a database of structures that can be filtered based on materials properties, including site geometry and connectivity, or by dimensionality as we demonstrated in this work. An additional opportunity is the ability to make a library of crystal structure components that could be used to compose novel materials or to understand trends in crystal structures formed by various compositions. In a similar vein, the condensed structure can be used to assess the overall structural and chemical diversity of crystal structure databases. Lastly, the analysis may be of use in analyzing defective crystal structures; indeed, the mineral matching routines implemented are somewhat robust to the presence of vacancies in the structure, as discussed in Supplementary Material. Nevertheless, small changes in the algorithms (e.g., excluding "standard deviation" from the

https://doi.org/10.1557/mrc.2019.94 Published online by Cambridge University Press

fuzzy structure fingerprints) may further increase robustness to finding defect structures.

Robocrystallographer is, to our knowledge, the first tool that can produce text-based descriptions of crystal structures. The descriptions can be automatically generated for libraries of compounds and used to provide additional easy to understand the analysis of structural properties. If integrated into online databases such as the Materials Project,[27] the text descriptions can help provide a greater context and improved understanding of the structures, as well as to improve accessibility for visually impaired users. Disseminating these text descriptions may also encourage a feedback cycle in which users can help improve the performance of existing structure analysis tools. For example, if robocrystallographer produces an unexpected result, it will be immediately apparent to a user reading the text description, who may report such discrepancies to the software package authors to improve the underlying algorithms. Robocrystallographer can thus facilitate the improvement of structure analysis routines by making it easy to validate their results.

Another area in which robocrystallographer may be used is in text mining the literature to extract materials information. Current text mining approaches can parse materials compositions accurately[33,34] but are unable to transform text descriptions of structures into any usable information. Using text similarity algorithms[35] to compare the reported structure descriptions in a text report with robocrystallographer descriptions for a set of candidate structures, it may be possible to better understand what structure is being described in a particular piece of text, making progress towards the longstanding issue of extracting structure from materials text.

Conclusion

We have introduced robocrystallographer, an open-source toolkit for analyzing crystal structures. The package can condense a structure into a descriptive JSON representation amenable to analysis and provides insights into the local coordination and polyhedral type, polyhedral connectivity, octahedral tilt angles, and component dimensionality. Additional analysis enables heterostructure and molecule-within-crystal identification, and fuzzy prototype matching. Using this information, robocrystallographer can generate text-based descriptions of crystal structures that resemble descriptions written by human crystallographers. We used the output of robocrystallographer to investigate the dimensionalities of all compounds in the Materials Project database, finding a rich degree of structural diversity and highlighting interesting cases such as mixed dimensionality. Lastly, we highlight the potential of robocrystallographer in machine learning studies.

Supplementary Material

The supplementary material for this article can be found at https://doi.org/10.1557/mrc.2019.94.

Acknowledgments

The authors acknowledge many useful discussions with Matt Horton regarding structure dimensionality. The authors additionally acknowledge Matt Horton for his work on the StructureGraph and BondedStructure components of pymatgen and for parsing the AFLOW prototype library. The authors acknowledge Evan Spotte-Smith for his work on the MoleculeGraph functionality in pymatgen. The authors acknowledge useful discussions with Leigh Weston regarding materials science text mining. The authors acknowledge Donny Winston for facilitating the calculation of robocrystallographer on all structures in the Materials Project database. The authors acknowledge useful conversations with Alex Dunn regarding machine learning model optimization. This work was intellectually led and funded by the U.S. Department of Energy (DOE) Basic Energy Sciences (BES) program—the Materials Project—under Grant No. KC23MP. Lawrence Berkeley National Laboratory is funded by the DOE under award DE-AC02-05CH11231.

References

1. W.H. Bragg: The significance of crystal structure. *J. Chem. Soc. Trans.* **121**, 2766 (1922).
2. A. Van De Walle: A complete representation of structure-property relationships in crystals. *Nat. Mater.* **7**, 455–458 (2008).
3. H.O. Pierson: *Handbook of Carbon, Graphite, Diamonds and Fullerenes: Processing, Properties and Applications* (William Andrew, New York, 2012).
4. A. von Hippel: Ferroelectricity, domain structure, and phase transitions of barium titanate. *Rev. Mod. Phys.* **22**, 221–237 (1950).
5. J.K. Burdett and S. Lee: Peierls distortions in two and three dimensions and the structures of AB solids. *J. Am. Chem. Soc.* **105**, 1079–1083 (1983).
6. D.O. Scanlon, C.W. Dunnill, J. Buckeridge, S.A. Shevlin, A.J. Logsdail, S. M. Woodley, R.A. Catlow, M.J. Powell, R.G. Palgrave, G.W. Watson, T.W. Keal, P. Sherwood, A. Walsh, and A.A. Sokol: Band alignment of rutile and anatase TiO_2. *Nat. Mater* **12**, 798–801 (2013).
7. A. Zunger: Inverse design in search of materials with target functionalities. *Nat. Rev. Chem.* **2**, 0121 (2018).
8. P. Gorai, E.S. Toberer, and V. Stevanović: Computational identification of promising thermoelectric materials among known quasi-2D binary compounds. *J. Mater. Chem. A* **4**, 11110–11116 (2016).
9. P.M. Larsen, M. Pandey, M. Strange, and K.W. Jacobsen: Definition of a scoring parameter to identify low-dimensional materials components. (2018). *arXiv:*1808.02114 1–11.
10. L. Himanen, P. Rinke, and A.S. Foster: Materials structure genealogy and high-throughput topological classification of surfaces and 2D materials. *npj Comput. Mater.* **4**, 1–10 (2018).
11. M. Ashton, J. Paul, S.B. Sinnott, and R.G. Hennig: Topology-scaling identification of layered solids and stable exfoliated 2D materials. *Phys. Rev. Lett.* **118**, 1–6 (2017).
12. A. Togo and I. Tanaka: Spglib: a software library for crystal symmetry search. (2018). *arXiv:*1808.01590 1–11.
13. M.J. Mehl, D. Hicks, C. Toher, O. Levy, R.M. Hanson, Gus Hart, and S. Curtarolo: The AFLOW library of crystallographic prototypes: part 1. *Comput. Mater. Sci* **136**, S1–S828 (2017).
14. D. Waroquiers, Xavier Gonze, G.-M. Rignanese, C. Welker-Nieuwoudt, F. Rosowski, M. Göbel, S. Schenk, P. Degelmann, R. André, R. Glaum, and G. Hautier: Statistical analysis of coordination environments in oxides. *Chem. Mater* **29**, 8346–8360 (2017).
15. N.E.R. Zimmermann, M.K. Horton, A. Jain, and M. Haranczyk: Assessing local structure motifs using order parameters for motif recognition, interstitial identification, and diffusion path characterization. *Front. Mater.* **4**, 1–13 (2017).

https://doi.org/10.1557/mrc.2019.94 Published online by Cambridge University Press

16. S.P. Ong, W.D. Richards, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V.L. Chevrier, K.A. Persson, and G. Ceder: Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci* **68**, 314–319 (2013).
17. L. Ward, A. Dunn, A. Fahaninia, 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: Matminer: an open source toolkit for materials data mining. *Comput. Mater. Sci* **152**, 60–69 (2018).
18. N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, and G. Hutchison: Open babel: an open chemical toolbox. *J. Cheminform* **3**, 33 (2011).
19. M. Swain: PubChemPy. https://github.com/mcs07/PubChemPy (accessed January 11, 2019).
20. S. Kim, P.A. Thiessen, E.E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B.A. Shoemaker, and J. Wang: Pubchem substance and compound databases. *Nucleic Acids Res* **44**, D1202–D1213 (2016).
21. Pymatgen. http://pymatgen.org (accessed January 14, 2019): 2019.
22. G. Voronoi: 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.* **133**, 97–178 (1908).
23. G. Giesecke and H. Pfister: Präzisionsbestimmung der Gitterkonstanten von $A^{III}B^{V}$ -verbindungen. *Acta Crystallogr.* **11**, 369–371 (1958).
24. B.C. Frazer and P.J. Brown: Antiferromagnetic structure of $CrVO_4$ and the anhydrous sulfates of divalent Fe, Ni, and Co. *Phys. Rev.* **125**, 1283–1291 (1962).
25. L.N. Kholodkovskaya, L.G. Akselrud, A.M. Kusainova, V.A. Dolgikh, and B.A. Popovkin: Bicuseo: synthesis and crystal structure. *Mater. Sci. Forum* **133–136**, 693–696 (1993).
26. M. Roos and G. Meyer: Kristallstrukturen von NH_4GaF_4 und $NH_4GaF_4{\cdot}NH_3$. *Zeitschr. Anorg. Allg. Chem.* **625**, 1843–1847 (1999).
27. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, and G. Ceder: Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater* **1**, 011002 (2013).
28. M. de Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C.K. Ande, S. van der Zwagg, J.J. Plata, and C. Toher: Charting the complete elastic properties of inorganic crystalline compounds. *Sci. Data* **2**, 150009 (2015).
29. L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton: A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 16028 (2016).
30. F. Faber, A. Lindmaa, O.A. Von Lilienfeld, and R. Armiento: Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **115**, 1094–1101 (2015).
31. G. Tzanis, C. Berberidis, and I. Vlahavas: *Machine Learning and Data Mining in Bioinformatics. Machine Learning* (IGI Global, Pennsylvania, 2011).
32. R.J. Urbanowicz, R.S. Olson, P. Schmitt, M. Meeker, and J.H. Moore: Benchmarking relief-based feature selection methods for bioinformatics data mining. *J. Biomed. Inform.* **85**, 168–188 (2017).
33. M.C. Swain and J.M. Cole: Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).
34. E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti: Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater* **29**, 9436–9444 (2017).
35. W.H. Gomaa and A.A. Fahmy: A survey of text similarity approaches. *Int. J. Comput. Appl.* **68**, 13–18 (2013).

https://doi.org/10.1557/mrc.2019.94 Published online by Cambridge University Press