

ARTICLE OPEN



Efficient and universal characterization of atomic structures through a topological graph order parameter

James Chapman¹✉, Nir Goldman^{1,2} and Brandon C. Wood¹

A graph-based order parameter, based on the topology of the graph itself, is introduced for the characterization of atomistic structures. The order parameter is universal to any material/chemical system and is transferable to all structural geometries. Four sets of data are used to validate both the generalizability and accuracy of the algorithm: (1) liquid lithium configurations spanning up to 300 GPa, (2) condensed phases of carbon along with nanotubes and buckyballs at ambient and high temperature, (3) a diverse set of aluminum configurations including surfaces, compressed and expanded lattices, point defects, grain boundaries, liquids, nanoparticles, all at nonzero temperatures, and (4) eleven niobium oxide crystal phases generated with ab initio molecular dynamics. We compare our proposed method to existing, state-of-the-art methods for the cases of aluminum and niobium oxide. Our order parameter uniquely classifies every configuration and outperforms all studied existing methods, opening the door for its use in a multitude of complex application spaces that can require fine structure-level characterization of atomistic graphs.

npj Computational Materials (2022)8:37; <https://doi.org/10.1038/s41524-022-00717-7>

INTRODUCTION

Atomic structure–property relationships form the basis of modern materials science, underlying both the discovery of new materials and optimization of existing systems^{1–5}. At the heart of these relationships is the fundamental principle that the arrangement of atoms dictates the behavior of the material throughout a spectrum of length and time scales. In the computational domain, reliably capturing atomic structure–property relationships is vital for crystal structure prediction⁶, the dynamic evolution of complex defect networks⁷, and the construction of interatomic potentials^{8–10} among others. A crucial roadblock in this endeavor is the ability to characterize complex atomic arrangements in materials in a computationally efficient and physically meaningful way, through the use of order parameters or similar mathematical quantities. For instance, order parameters that can uniquely capture local atomic geometries are necessary to adequately characterize phase transitions from molecular dynamics simulations^{11,12} and nucleation parameters using free energy calculations^{13,14}.

However, this characterization is often nontrivial, especially for atomically disordered systems in which the underlying symmetry of the atomic geometries is difficult to determine^{15,16}. The atomic disorder is intrinsic to features such as defects, surfaces, grain boundaries, and heterogeneous interfaces, which have been critically linked to transport, mechanical, electronic, and optical properties^{17–20}. Moreover, advances in synthetic protocols for nanostructuring and mesostructuring create interface-dominated materials with local coordination, structural arrangement, and strain that deviate significantly from equilibrium bulk assumptions²¹. These applications are among many that would draw significant benefit from improved mathematical and physical capabilities to describe the disorder.

Throughout the decades, many schemes have been proposed to capture various portions of this ordered-disordered spectrum as scalar values such as the common neighbor analysis (CNA)²², adaptive CNA (A-CNA)²³, centrosymmetry parameter (CNP)

analysis²⁴, Voronoi analysis²⁵, bond order analysis such as the Steinhardt order parameter (SP)²⁶ and the bond angle analysis (BAA)²⁷. Each method has shown varying degrees of success, with each scheme playing a vital role in capturing specific classes of materials phases²³. Voronoi, SP, and other bond-order algorithms generally fail to capture the differences in crystalline systems with compressed and/or expanded lattices as well as those experiencing atomic perturbations close to the melting temperature of the material phase²⁸. While methods such as CNA and A-CNA overcome these pitfalls with a more robust underlying algorithm, they ultimately break down in situations wherein the material symmetry is lost or difficult to comprehend²⁹. In fact, all of the above algorithms struggle to capture the subtle differences in the local coordination environment when the underlying geometric symmetry is either broken or exists only at short-range such as the environments encountered in grain boundaries, surfaces, liquids, and amorphous structures³⁰.

More mathematically involved methods, such as the smooth overlap of atomic positions (SOAP)³¹, the Behler–Parrinello symmetry functions (BP)³², atomic cluster expansion (ACE)³³, Chebyshev polynomial representations (CPR)³⁴, and the adaptive generalizable neighborhood informed features (AGNI)^{9,35}, rely on sophisticated symmetry functions with a plethora of tunable parameters to map an atom's local environment to an invariant mathematical space^{36,37}. While these methods are often accurate^{38–42}, they are also computationally cumbersome when compared to the previous order parameters^{4,43} and require the nontrivial tuning of their corresponding parameter sets for a material system. Such methods also generally operate on atomic environments, not structural ones, forcing one to use statistical modifications to average the atomic feature vectors into a single configurational descriptor. Importantly, structural information obtained in this way is always tied to the atomic representation, an inherently local property, and therefore does not explicitly capture global information such as the shape and connectivity of the atomic network over long distances. The feature vectors

¹Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA, USA. ²Department of Chemical Engineering, University of California, Davis, California, CA 95616, United States. ✉email: chapman37@llnl.gov

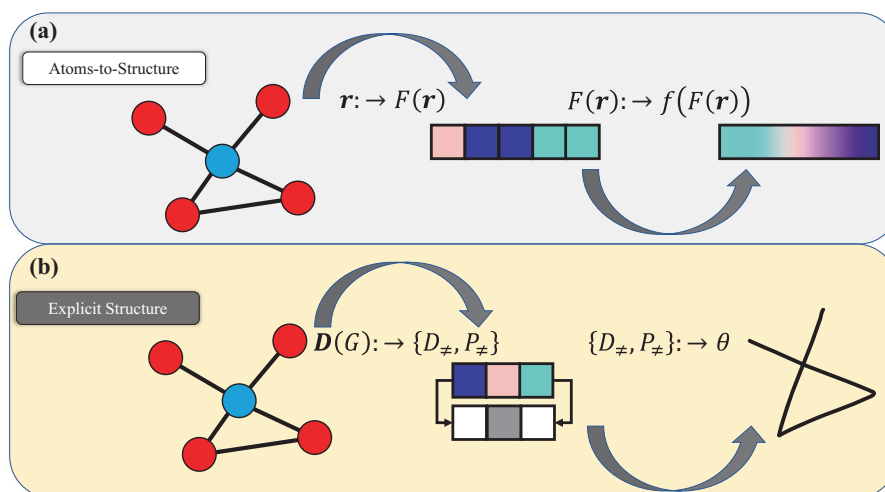


Fig. 1 Visual depiction of two structure-level characterization methods. **a** Atomic local neighborhoods are converted to feature vectors, which are modified statistically (usually taken as the N th moment) to construct a final configuration descriptor that represents the entire structure. $\mathbf{r} \rightarrow F(\mathbf{r})$ represents the mapping from atomic positions to atomic descriptors. $F(\mathbf{r}) \rightarrow f(F(\mathbf{r}))$ represents the mapping from atomic descriptors into a single structural descriptor through statistical modification of the atomic descriptors. **b** Our proposed methodology, in which global information about the atomic network is explicitly captured and passed through our order parameter to represent the structure of the network itself. The center portion of **(b)** represents the set of unique degrees captured within the graph placed above their corresponding probabilities of occurrence within the network. $D(G) \rightarrow \{D_{\neq}, P_{\neq}\}$ represents the mapping of the set of all degrees in G to the set of unique degrees in G , along with their corresponding probabilities. $\{D_{\neq}, P_{\neq}\} \rightarrow \theta$ represents the mapping of the set of unique degrees and corresponding probabilities in G to the structure-level order parameter of the graph.

generated by these methods are also not guaranteed to be unique⁴⁴, which in turn makes the atoms-to-structure mapping even more precarious. A visual depiction of this process can be seen in Fig. 1a.

Methods such as convolutional neural networks (CNN)^{45,46}, graph neural networks (GNN)^{47–50}, and variational autoencoders (VEA)^{51,52} can alleviate both the cost and manual parameter fitting of symmetry functions, but require large amounts of reference data to train the models. In particular, these methods can be difficult to train for materials with complex phase spaces where obtaining enough reliable reference data is challenging (e.g., detonations of energetic materials⁵³, nanostructures generated under explosive conditions⁵⁴, and the irradiation damage of complex glasses⁵⁵). This hinders both the generalizability and transferability of the models to both new configurations and material systems which are not characterized within the training set. Such methods can require a large number of tuned parameters in order to achieve sufficient accuracy, which in turn can present high computational costs for large system sizes. Graph theoretical methods such as those employed in MolecuLarNetworks⁵⁶ and ChemNetworks⁵⁷ have been used to analyze small molecules with good success. However, such methods rely on properties of the graphical representations that are not unique, such as the geodesic distance of the graph. These approaches can have difficulty characterizing materials classes with subtle differences, such as oxides, metals, ceramics, and/or structural environments depending on extreme pressures, grain boundaries, surfaces, and the presence of nanoparticles.

In this work, we build upon these efforts through the development of a physically intuitive and computationally efficient framework, henceforth referred to as the scalar graph order parameter (SGOP), which serves as a semi-empirical graph topology metric. One distinct advantage of SGOP lies in the ability to explicitly capture global information about the graph by looking at the set of unique node degrees contained within the graph and the probability of these degrees occurring throughout it. We also discuss a vector graph order parameter (VGOP), which allows for a set of different SGOP values in order to add a high degree of fidelity to our analysis. In general, SGOP characterizes

the graph representation of an atomic network by determining two physical characteristics of the graph: the entropy of the graph, and its connectivity. This characterization can be broken down into three parts: (1) identification of subgraphs contained within the system, (2) determination of the shape of the subgraphs, which is motivated to resemble the entropy of the subgraph, and (3) calculations of the connectivity of the subgraphs, which is determined via the subgraph's degree matrix. A visual demonstration of this workflow can be found in Fig. 1b.

Importantly, SGOP explicitly captures global information about the graph itself such as its shape and connectivity rather than operating at the node-level of the graph. It is this difference that allows for not only a reduction in the complexity of our algorithm's functional form when compared to existing methodologies. This also provides a more physically intuitive understanding of the relationship between configurations, based on the resulting order parameter value. Importantly, SGOP explicitly captures global information about the graph itself such as its shape and connectivity, placing it in a distinct category from existing methodologies such as symmetry functions which operate at the node-level of the graph.

The remainder of the paper is as follows. We provide the reader with a simplistic and intuitive theoretical validation of SGOP by observing how SGOP characterizes several simple graphs. We then perform atomic structure characterization over a multitude of distinct systems such as liquid lithium, elemental carbon⁵⁸, and aluminum⁵¹, and finally several crystal phases of niobium oxide under dynamic conditions. A discussion about our proposed methodology and its future implications are then discussed. We conclude this work by providing a detailed theoretical discussion regarding the configurational graph construction, SGOP, and VGOP formalisms.

RESULTS

Theoretical validation

We validate the SGOP formalism by characterizing simple and intuitive graphs to showcase what is meant by a physically-intuitive predictive capability. Figure 2 shows three graphs, each

containing four vertices. One can think of these as atomic systems containing four atoms each. Due to their edge connections, however, each graph represents a unique composition. Our argument is that these graphs can be judged by their entropy and connectivity observed within their degree sets. For the first graph on the left with a degree set of $D_G = 1, 2$, each degree has a probability $P(d_m)$ of 0.5. For the remaining two graphs each degree has a probability $P(d_m)$ of 1, with degree sets of $D_G = 2$ and $D_G = 3$, respectively.

The entropic term contained within SGOP will be zero for the second and third graphs, and their degree sets contain only a single unique value. This makes physical sense, as the disorder contained within the set of unique degrees is zero. For the first graph the entropic term will be nonzero, again making intuitive sense. One can immediately see why the connectivity

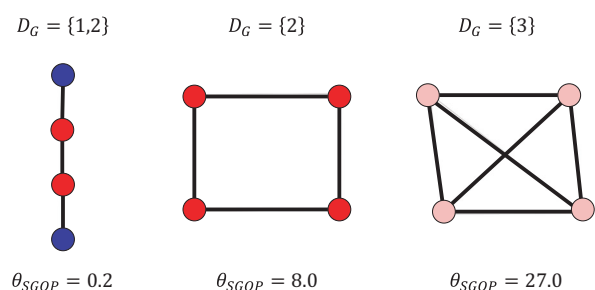


Fig. 2 Theoretical validation of the SGOP formalism. The coloring of the nodes corresponds to their respective degree value. The degree set used to calculate the SGOP is shown above each graph, while the calculated SGOP value for the graph is shown below it. Here, the degree value is simply taken as the number of edges into each node.

term is included, as the entropic terms of the second and third graphs are identical. This is expected, as the property of the graph that we are measuring is indeed the same between the two graphs. The connectivity term will then distinguish the graphs, as the unique degree values themselves are used to calculate this term. From Fig. 2 this can be seen as the SGOP value for the second graph is 8 while the SGOP for the third graph is 27, even though they both have the same entropic term of zero. The SGOP values also make physical sense, as a less connected graph will have a smaller SGOP. This makes the SGOP ideal for providing a physically informative prediction, allowing one to not only cluster unique systems from one another in an unsupervised manner but also provide physical intuition for what the SGOP value represents.

Liquid lithium

Previous works have indicated that the configuration space of the liquid phases of lithium spans a vast domain, with each liquid phase showing structural differences when compared to results from a different pressure⁵⁹. These structural dissimilarities result in strong differences in properties such as the vibrational density of states, which ultimately govern the self-transport behavior of the material⁶⁰. Previous density functional theory calculations have shown that, within a given temperature range, there is a strong linear correlation between the self-diffusion constant and the density of the liquid phase⁶⁰. The coupling of these two properties allows one to make predictions on unknown phases at high pressures without the need for performing nontrivial and expensive simulations and/or experiments.

Figure 3 highlights the ability of a single SGOP value to characterize the complexity of the lithium liquid phase space. The SGOP values shown here were calculated from a graph coordination network (GCN), discussed within the methods section, which employed an R_c of 2.5 Å. This corresponds to the maximum value of the first peak of the radial distribution function (RDF) position

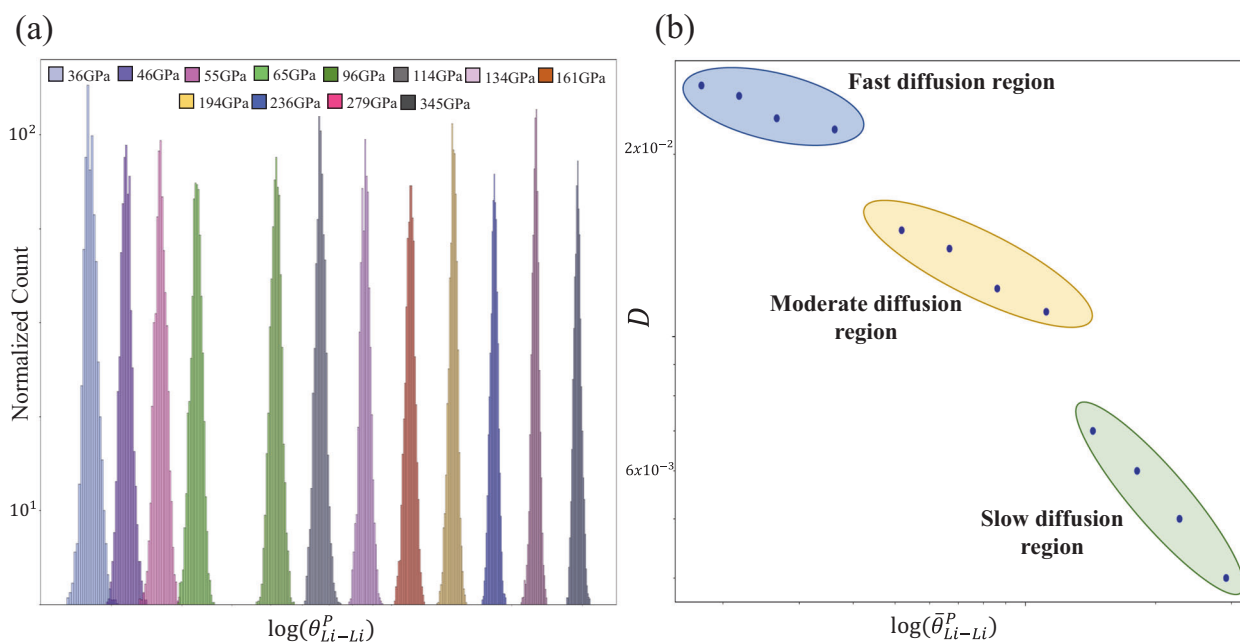


Fig. 3 Characterization of liquid Li using SGOP. **a** Histograms representing the distribution in SGOP values for each liquid phase of lithium. Colors represent the different phases and are defined in the plot by the external pressure on the simulation box obtained from DFT. **b** Self-diffusion constant values, obtained from the mean squared displacement of the ab initio molecular dynamics trajectories plotted as a function of the mean of each histogram, shown in **(a)**. Three diffusion regions are highlighted and are determined by the underlying structure of the vibrational density of states of each trajectory, calculated from the velocity auto-correlation function. θ^P is defined as the SGOP for a specific phase P .

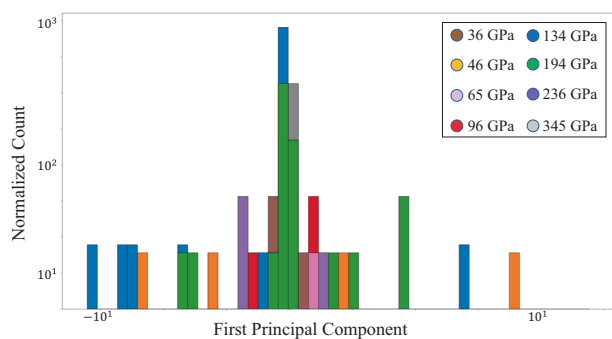


Fig. 4 PageRank clustering of liquid Li structures. The distribution of the first PCA component of various liquid Li phases, was calculated originally from the PageRank algorithm. Each color represents a unique liquid phase. Large gaps exist between bins due to the relatively discrete predictions made by PageRank.

over all liquid phases used. Figure 3a shows histograms of the SGOP values for each liquid phase. One can clearly see the separation of each phase, indicating that the SGOP values are capable of characterizing the unique differences in local geometry encountered within each phase, but also the spread within a specific phase. The structures encountered within a liquid phase, over some period of time, will oscillate about an equilibrium point, assuming that all external conditions are held constant. Figure 3a provides a visual representation of these perturbations, with the width of each normal distribution representing the extent of the spread for a given phase.

Figure 3b showcases the SGOP's ability to reproduce the underlying trend of density versus self-diffusion constant. Diffusion constant values were calculated from the mean square displacement for each liquid phase using a simple Fickian diffusion model⁶¹. Figure 3b tracks the changes in the self-diffusion constant as a function of the mean SGOP value from each phase's histogram, shown in Fig. 3a. From this relationship we can correctly identify three diffusion regions: (1) fast diffusion occurs in low-density phases, (2) moderate diffusion occurs in phases that are more dense than the low-density regime but do not exhibit "crystal-like" properties, and (3) slow diffusion occurs in highly-compressed phases which behave more closely to a crystal phase than a liquid one. The SGOP histogram averages are able to classify not only the structures within each liquid phase, but also correctly identify unique self-diffusion regions across a vast configuration space. This clearly indicates that one can use the SGOP values as inputs to predictive models.

In order to further test the relative accuracy of our method against another graph-based methodology, we also compare our SGOP predictions with those from the PageRank (PR) method⁶². PR has been used previously to study molecules⁵⁶ and chemical solutions⁵⁷ and represents a state-of-the-art graph-based methodology to study complex chemical transformations and their energy landscapes. For our PR analysis, we employ a weighted connectivity matrix based on the real-space distance between atoms^{56,63}. PR yields an output of an eigenvector corresponding to the maximum connectivity between nodes in a graph. We then condense this information into a scalar through the use of a principal component analysis (PCA) decomposition, where the first PCA component yields a least-squares approximate representation of the data and can be used for structural analysis⁶⁴. Further details regarding our use of PR can be found in the Supplementary Methods section, along with Supplementary Fig. 5.

Figure 4 shows the distribution of the first PCA component values of various liquid Li phases. As one can see, PR's ability to uniquely classify the various liquid Li phases is limited, with significant overlap existing between all considered phases. We

also note that the resulting histograms tend to be very discrete, indicating that the resulting PR information is not unique enough to distinguish the small differences between certain liquid phases. We can therefore conclude that SGOP can classify the subtle structural differences found within this disordered phase space with a much higher level of fidelity than that of PR.

Carbon

While the structures encountered within the liquid lithium phase space are highly complex, they required only a single SGOP to classify the phases. This was in part due to the density acting as a sole property needed to characterize the local coordination environment. However, as one aims to characterize the multitude of unique structural motifs within a material's phase space, a single SGOP may not be unique enough to differentiate between local atomic geometries. One example of this is elemental carbon, which exhibits a rich configuration space that includes both two-dimensional and three-dimensional structures, nanotubes containing varying amounts of free volume, and nanoparticles that exist in many shapes and sizes. This diversity of structures and coordination numbers readily indicates the need for multiple SGOPs to adequately represent various portions of the coordination environment.

Here we use an extension of the SGOP formalism, called the vector graph order parameter (VGOP), which is discussed within the methods section, to characterize a previously created and highly diverse carbon dataset⁵⁸. Due to the presence of different phases with subtle structural differences, such as graphite vs. diamond, an R_c set of (3, 4, 5, and 6 Å) was chosen, which was determined via the peaks of the radial distribution functions from each material in the dataset. This set of R_c ensures that all peak positions are uniquely represented via the corresponding GCN and is intended to be general in nature and not tailored to the specific peak positions, thereby making the selection of values transferable to any material system. The resulting VGOP can be thought of as a feature set, similar to those discussed earlier, but with the significant advantage of both small size and easy physical interpretability. As discussed in the methods section, each VGOP was normalized and decomposed using PCA. Information regarding the PCA metrics can be found in the Supplementary Methods section, along with Supplementary Figs. 6–8.

Figure 5a indicates the VGOP's ability to characterize the various bulk phases of elemental carbon. All phases (graphene, graphite, diamond, and lonsdaleite) are clearly differentiated, and perhaps more importantly, are clustered in a physically-intuitive manner. Graphene is clustered near graphite, but far from both diamond and lonsdaleite. Graphite is clustered between graphene and diamond, while lonsdaleite (hexagonal diamond) finds itself clustered near diamond but far from both graphene and graphite. Amorphous diamond and lonsdaleite also cluster near one another but are located in a unique portion of the PCA space, when compared to the ordered crystal phases. As was the case with lithium (using the SGOP), the VGOP not only characterizes each phase correctly but also classifies them in a physically-informed manner, providing the user with an intuitive interpretation pathway. One important aspect of these results can be seen in the classification of the amorphous configurations. The VGOP correctly classifies the structures encountered during each trajectory as similar despite the large disparity in the temperature used to generate each amorphous phase (2000 vs. 4000 K). The high fidelity of the VGOP framework can allow for precise analysis of phenomena such as phase transitions and/or for free energy calculations, where an easy and clear distinction between material phases is vital.

Similar clustering trends exist in Fig. 5b, c for nanotubes and buckyball-like nanoparticles respectively. In Fig. 5b, nanotubes with small radii are isolated from those with large radii in the PCA

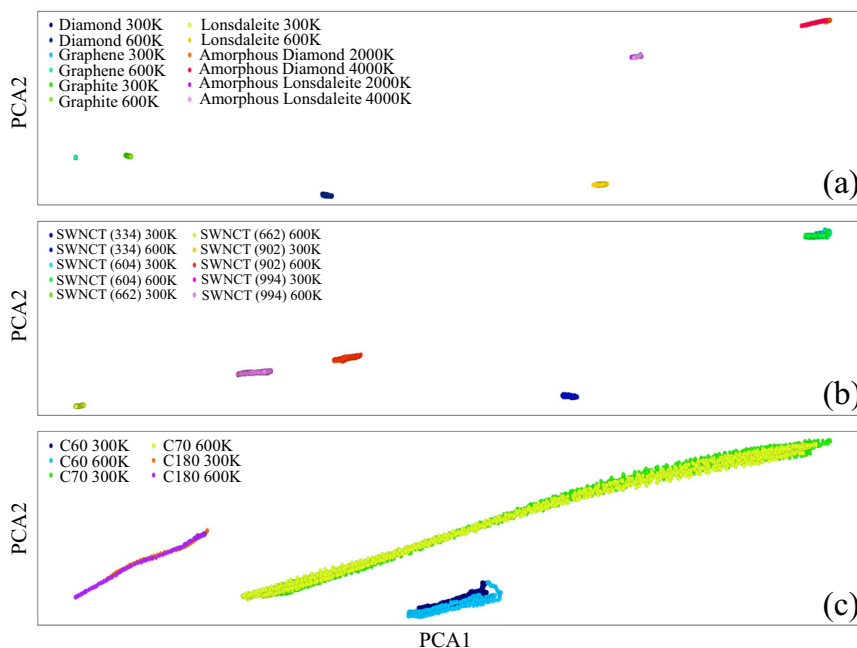


Fig. 5 Visual clustering of carbon structures using VGOP. Principal component analysis plots, calculated from the VGOPs of the carbon dataset's GCNs, indicate the unsupervised clustering of a multitude of structural motifs: **a** the bulk phases of carbon (graphene, graphite, diamond, and lonsdaleite), **b** single-walled carbon nanotubes of varying length and radius, and **c** buckyballs of varying diameter.

space. This again makes intuitive sense, as the local atomic coordination environment will change as a function of the nanotube's radius. This can also be seen in Fig. 5c for the case of small buckyball-like nanoparticles where particles with a fewer number of atoms are more densely packed than those with a larger number of atoms. This relationship is captured accurately in our VGOP calculations, with small particles clustered to the right and large particles clustered to the left of Fig. 5c. It is important to note here that the number of timesteps in each trajectory is not identical, so while the C_{70} buckyball appears to extend further right than the C_{60} particle, the C_{70} trajectory explores a significantly larger portion of its phase space than the C_{60} trajectory.

Aluminum

The previous cases of lithium and carbon provided insight into the ability of the SGOP and VGOP frameworks to characterize both structural disorder and geometric diversity. For the case of aluminum, this coupling of complexity and heterogeneity is obtained by observing a multitude of nonzero temperature structural environments including surfaces, compressed and expanded lattices, point defects, grain boundaries, liquids, nanoparticles, all calculated previously via *ab initio* molecular dynamics⁶⁵. For all environments except for the bulk FCC, BCC, and HCP trajectories, the structures were obtained from NVE simulations with an initial temperature of 1000 K.

Here we compare our results to those of SP and the AGNI crystal fingerprint. SP represents a mathematically robust, though fixed with respect to any parameterization, characterization scheme that has been used to determine structural similarities for several decades. AGNI, on the other hand, represents a relatively new class of characterization schemes, in which structures are represented as a vector of highly parameterized functions, with each vector element capturing distinct parts of an atom's local geometric environment. By taking the PCA of SP, AGNI, and VGOP, we can create a level playing field, in which a direct comparison can be made between all three methodologies and their ability to characterize the same set of structural environments. Taking the

PCA of feature sets has been used previously to visualize AGNI's ability to characterize atomic structures⁹.

We use the VGOP framework, with an R_c set of (3, 4, 5, 7, and 8 Å) determined via the aluminum RDF peaks (with each value in the set capturing a unique peak in the RDF). A visual representation of the GCN for $R_c = 3$ Å for several of the Aluminum structures is shown in Fig. 6. From Fig. 6 one can see how the GCNs capture unique information about the structure. In the case of bulk Al, the GCN indicates high but uniform connectivity amongst the nodes, while for the case of the grain boundary there exists two distinct regions of the graph, one corresponding to the bulk-like region and the other representing the interface region. A similar graph structure exists in the surface, though the surface region is far more chaotic and randomized than the fairly ordered grain boundary interface region. These structural differences within the graph provide a unique mapping from structure to VGOP, implying that structures with similar VGOP must have similar structural environments (provided one captures all relevant information via the cutoff radii).

For each method, a PCA decomposition was performed on the initial feature vector (i.e., computed SGOPs for each R_c value), with the first two principal components chosen for visualization purposes. Such a procedure has been shown previously to be an accurate way of visualizing the high-dimensional spaces used in the structure characterization techniques used here⁹. Figure 7 showcases each method's ability to accurately characterize each class of aluminum environments. One should note here that each subplot's axes have been normalized between zero and one for visualization purposes, and that the absolute axis values between subplots are not shared. The absolute PCA representation can be found in Supplementary Fig. 9.

The first column in Fig. 7 represents the Steinhardt order parameter PCA classification. From Fig. 7a one can see that the SP PCA eigenvectors can clearly distinguish between the low and high-temperature BCC, FCC, and HCP phases. It also performs well when classifying the FCC liquid phase as distinct from the ordered crystal phases. However, the SP struggles to identify high-temperature BCC as having the same underlying coordination environment as low-temperature BCC. We know from the length

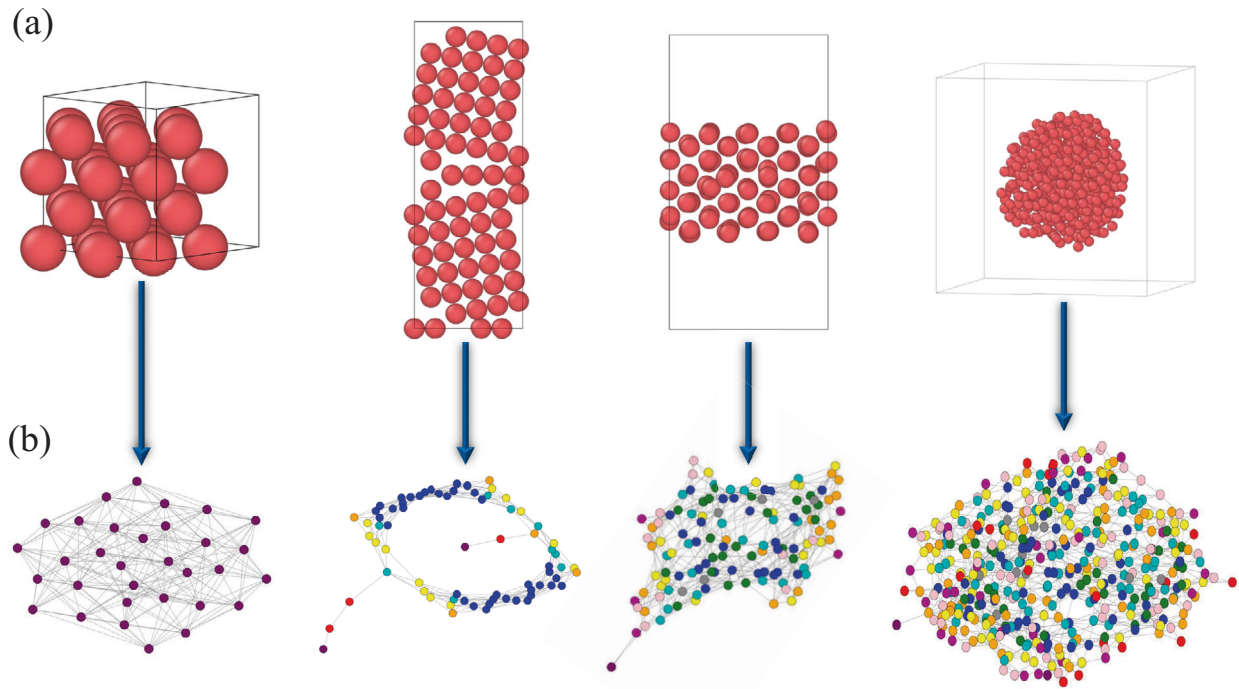


Fig. 6 Visual depiction of graph coordination networks for various Al structures. **a** Aluminum atomic configurations from left to right: bulk FCC, $\Sigma(510)$ grain boundary, (110) surface, and a 12 Å diameter nanoparticle. **b** The corresponding graph coordination network at an R_c of 3 Å. Vertices represent the atoms within their respective structures shown in **(a)**, while the vertex colors represent the degree of the vertex. It should be noted that the colors are not universal, but are relative to the smallest vertex degree within the graph, with purple representing the smallest degree and with blue being indicative of the largest degree.

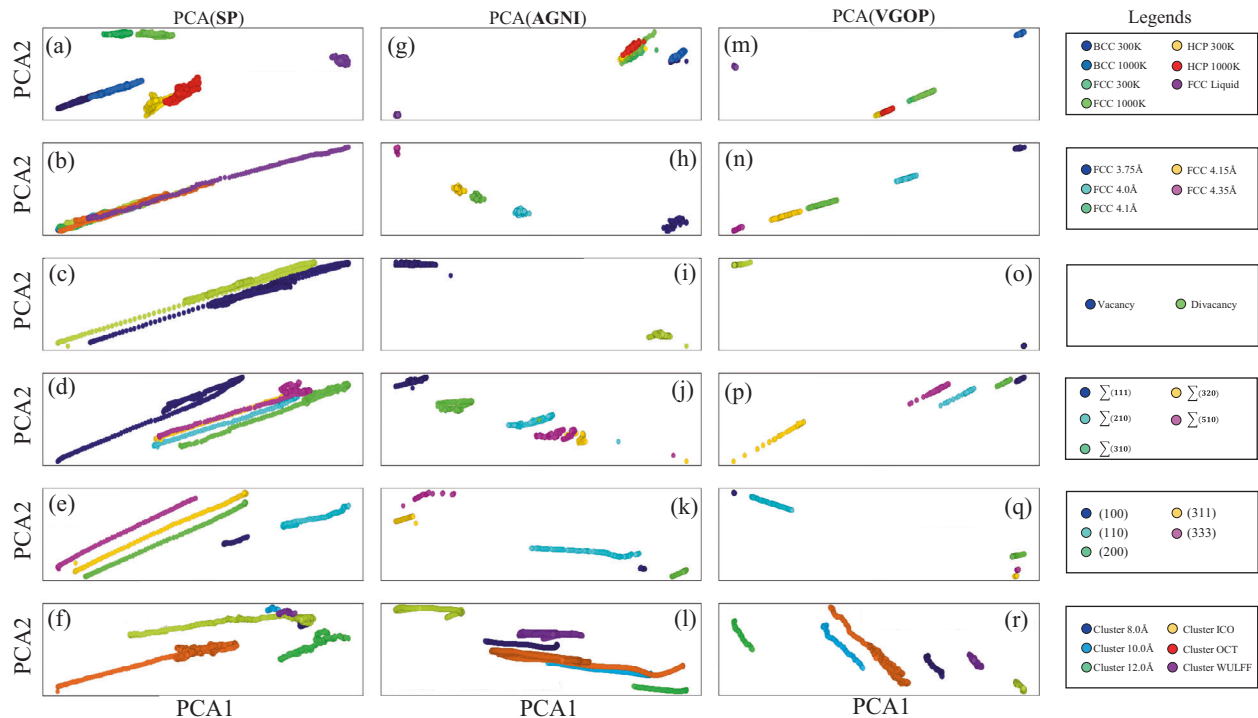


Fig. 7 Spatial clustering of Al between the Steinhardt order parameter, AGNI, and VGOP. Comparison of the PCA reduction of three different methods, using a robust aluminum database: (column 1) the Steinhardt order parameter, (column 2) the AGNI crystal fingerprint, and (column 3) the VGOP. Each row represents a unique subset of the aluminum configuration space. Colors are uniform across the columns. Each PCA subplot's axis have been normalized to the data present within the plot to allow for better visualization of the data. Row labeling is as follows: **a, g, m** Bulk, **b, h, n** Bulk with strain, **c, i, o** Bulk with vacancies, **d, j, p** Grain boundaries, **e, k, q** Surfaces, and **f, l, r** Clusters.

of the trajectories that the high-temperature structure has large thermal fluctuations of the ions that can mask its symmetry. In addition, Fig. 7b indicates the SP's inability to correctly identify the structural differences between compressed and expanded FCC lattices, effectively characterizing all cases as a single entity. Figure 7c also highlights the SP's difficulty when attempting to differentiate between a single vacancy within a pristine bulk environment and that of a divacancy in an otherwise identical geometry.

A similar trend emerges when characterizing the subtle differences in grain boundary structures, shown in Fig. 7d. The $\Sigma(210)$, $\Sigma(310)$, and $\Sigma(510)$ grain boundaries should yield some underlying similarities, but are technically unique environments. However, the SP has difficulty in distinguishing between the configurations and also classifies the $\Sigma(510)$ and $\Sigma(320)$ grain boundaries as identical coordination environments, which is incorrect. Interestingly the SP performs well when characterizing the differences between surface environments in Fig. 7e, perhaps due to the well-defined uniqueness in the surface layers. For the case of the nanoparticles, shown in Fig. 7f, the SP is able to clearly differentiate between the ordered clusters (icosahedral, octahedral, and Wulff particles), but fails to correctly capture the differences inherent in the disordered particles (8.0, 10.0, and 12.0 Å particles). All told, the SP cannot be reliably used to characterize the complexity of the aluminum configuration space.

The second column represents the AGNI crystal fingerprint classification. From Fig. 7g one can see that the AGNI PCA eigenvectors can clearly distinguish between the low and high-temperature BCC and HCP/FCC but fails to correctly capture the differences between HCP and FCC. However, it does perform well when classifying the FCC liquid phase as distinct from the ordered bulk phases. Figure 4h indicates AGNI's ability to correctly identify the structural differences between compressed and expanded FCC lattices. Figure 7i highlights AGNI's capabilities in differentiating between the vacancy and divacancy environments. Unlike the SP, AGNI performs much better when characterizing the subtle differences between grain boundaries, though does encounter some overlap between the $\Sigma(510)$ and $\Sigma(320)$ structures. AGNI also performs well when characterizing the differences between surface environments in Fig. 7k. However, for the case of the nanoparticles, shown in Fig. 7l, AGNI fails to properly distinguish between the ordered and disordered clusters, similar to the problematic characterization of the SP. Overall, while AGNI can correctly capture a much larger portion of the aluminum configuration space, it breaks down in several areas, some of which could be correctly captured by the SP.

The third column represents the VGOP classification. From observing Fig. 7m–r one can see that the VGOP framework predicts a unique characterization for every structural environment encountered in the dataset. Perhaps equally as important is the VGOP's ability to cluster similar coordination environments together, providing an intuitive and natural unsupervised clustering. In principle, if one did not know what the structures being characterized were, they could identify geometric similarities, or differences, between them. Having this ability could make the VGOP a powerful tool for enhancing sampling methods during model development. One could use the VGOP to indicate structures that a model does not need to be parameterized on, due to the underlying similarities with other environments.

Niobium oxide

The previous three examples showed the power of the SGOP formalism to characterize a vast configuration space under dynamic conditions. However, each study contained a single chemical element within the system, resulting in a single atomic network differentiated by varying the GCN cutoff values. Here, we extend our efforts to multicomponent systems by characterizing

the complexity encountered in the niobium oxide (Nb_xO_y) phase space from ab initio molecular dynamics (AIMD) simulations. Here, AIMD simulations were performed on 11 unique crystal phases Nb_xO_y phases (see the methods section for more details).

For the case of VGOP, three interaction types were used to construct unique GCN: (1) Nb–O, (2) O–O, and (3) Nb–Nb. The Nb–O VGOP was calculated using an R_c set of (2, 3, 4 Å), while the O–O and Nb–Nb VGOPs both used an R_c set of (3, 4, 5, 6 Å). Again, we emphasize here the general nature of these cutoff values, chosen intuitively to capture all possible atomic interaction regions within the system. As there are peaks in the RDF between each of these values (ex: a peak exists in the Nb–O RDF between 2 and 3 Å), these sets of R_c effectively bound all possible coordination environments within the maximum cutoff value for each interaction type. The O–O and Nb–Nb networks do not use a cutoff of 2 Å as there is little-to-no interaction information contained between oxygen or niobium atoms at those distances across all Nb_xO_y phases studied in this work. Further details regarding the VGOP and AGNI parameters used can be found in Supplementary Table 1.

Figure 8 shows the results of the VGOP (left) and AGNI (right) characterization. The AGNI characterization, while providing the ability to distinguish between different stoichiometries of niobium, fails to quantitatively separate distinct phases with identical stoichiometries. For example, Nb_4O_4 is clustered uniquely from the various Nb_2O_5 phases, but all Nb_2O_5 phases overlap with one another. This can be attributed, in part, to the fact that each local atomic environment is similar between the phases, due to the Nb–O bonding environments, which is what symmetry functions such as AGNI are designed to capture. However, the global structure of the atomic networks is not captured by this approach, with the averaging of local information not being sufficient to fully capture the uniqueness of the global atomic network.

The VGOP characterization represents a far more robust classification than AGNI, with little-to-no overlap existing between any of the Nb_xO_y phases. From Fig. 8 (left) one can see that each phase is isolated in its own region of the 2D PCA space. As not all AIMD simulations have the same amount of simulated time we cannot expect that each phase will cover the same portion of the phase space. However, it is still interesting that some Nb_xO_y phases explore a much larger region of their configuration space, indicating that some phases encounter larger perturbations from their initial configuration than other phases. Regardless, these types of deviations are all clearly characterized by our VGOP analysis.

Finally, VGOP can be leveraged to determine the correlation of the distance between cluster centroids within the PCA space to similarities within each constituent atomic network. For example, in the bottom left of Fig. 8 (left) we see two phases that exist in close proximity ($\alpha\text{-Nb}_2\text{O}_5$ and $\alpha\text{-NbO}_2$). At first glance, these phases may seem quite different, though upon further inspection one will see that the Nb–O and O–O atomic networks are very similar, even at a cutoff distance of 6 Å. Further details regarding the similarities and differences of the atomic networks can be found in Supplementary Fig. 4. These results can be further refined by calculating the VGOP with a different R_c set. Even with the relatively coarse set of R_c used in this work, we still observe that the VGOP clustered the two phases uniquely, with little overlap, and tells the user that the two phases share similarities in their underlying atomic network.

DISCUSSION

Structure–property relationships, which have always served a fundamental role in materials science, have become critically important due to the ever-increasing need for new materials with targeted properties and chemistries. Frequently, a high

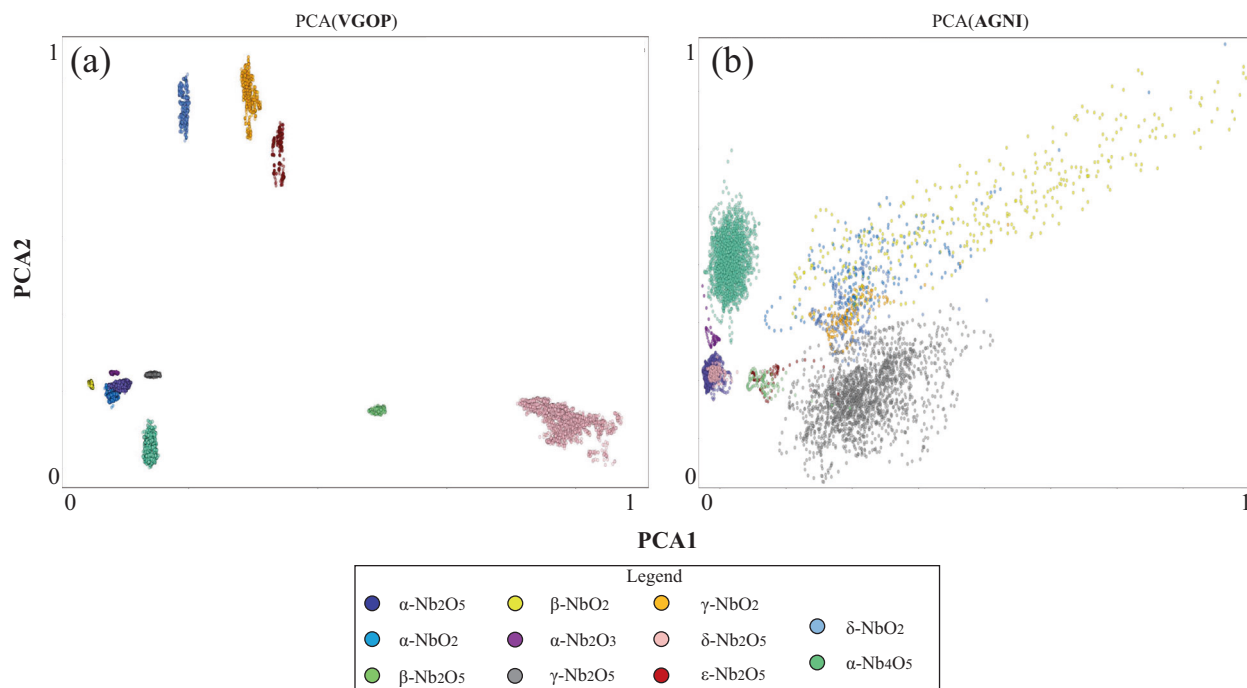


Fig. 8 Spatial clustering of Nb_xO_y between AGNI and VGOP. Comparison of the PCA reduction of the AGNI (b) and VGOP (a) generated feature vectors for various crystalline phases of niobium oxide. The inserted legend indicates the colors differentiating the phases. The PCA axes have been normalized and shifted to be all positive.

degree of precision and accuracy can be needed to uniquely characterize new structural environments and ultimately map those unique geometries to properties of interest. Our efforts here discuss a structural characterization scheme that can uniquely classify the local atomic coordination environments present in atomistic configurations through a semi-empirical graph order parameter. This order parameter aims to explicitly capture global information about the underlying atomic graph, such as the shape and connectivity of the graph, presenting an improvement over existing structural characterization methods which represent the global structure of the system as some combination of atomic environments, which are inherently local. Our formalism is computationally efficient and mathematically robust, providing the ability to characterize subtle differences in atomic structure over a wide range of dynamic conditions. While the SGOP formalism requires minimal user-adjusted parameters (such as the graph R_c and SGOP exponent), they are physically intuitive and require only a limited understanding of the underlying system to be appropriately chosen. The computational efficiency combined with the uniqueness and physically-informed nature of the SGOP formalism allows it to be applied to a plethora of challenging application spaces including enhanced sampling, and unsupervised clustering, which generally requires the ability to determine subtle distinctions between underlying phases or structures.

We propose the use of the SGOP as a valuable complement to methods that are chiefly focused on local atomic structure. Because our proposed methodology operates on global information about the structure of the atomic network, it is especially ideal for the study of problems such as the classification of unique material phases. SGOP also operates on the graph itself, and not on atomic coordinates, thereby making it possible to employ SGOP to characterize more abstract concepts that are best represented as graphs, such as the pathway an ion takes during diffusion, the morphology of grains and voids, and the shape of surface features.

METHODS

Ab initio molecular dynamics

All AIMD data was obtained using the Vienna ab initio simulation package (VASP)^{66,67}. AIMD simulations were performed only for the case of the niobium oxide crystal phases. The Perdew–Burke–Ernzerhof (PBE) functional⁶⁸ was used to calculate the electronic exchange–correlation interaction. Projector augmented wave (PAW) potentials⁶⁹ and plane-wave basis functions up to a kinetic energy cutoff of 600 eV were used. Γ -centered k-point meshes were carefully calibrated for each atomic configuration to ensure numerical convergence in both energy and atomic forces. Each AIMD step’s electronic loop was considered converged at 10^{-7} eV. Simulations were performed in the NVE ensemble, with all phases having an initial temperature of 1000 K, using a 0.5 fs timestep. Each MD run’s initial configuration was obtained from Materials Project^{70,71} and was subsequently relaxed using the same parameters used during AIMD, though with the appropriate relaxation parameters, with an energy tolerance of 10^{-7} eV and force tolerance of 10^{-3} eVÅ⁻¹. During the relaxation, the volume was allowed to change along with the ion positions. Visual depictions of all atomic structures used in this work can be found in Supplementary Figs. 1–3.

Graph coordination networks

The diversity and complexity of atomic structures necessitate the efficient and intuitive characterization of these environments. In this work, we employ a graph-based characterization scheme, which we call the graph coordination network, to identify pairwise atomic networks contained within a configuration of atoms. GCNs begin by sorting the chemical identities of the atoms in the configuration into separate categories. Depending on the pairwise interaction one aims to capture, the various species lists are then scanned to find atomic interactions that occur within some cutoff radius. The GCN is similar to a radial distribution function, as it aims to capture the unique coordination environments encountered by each atom, with respect to a particular chemical interaction environment. The GCN can be represented by an adjacency matrix, with matrix elements defined by:

$$G_{k_i, k_j}^{i-j} = \frac{1}{d_{k_i, k_j}} \ni d_{k_i, k_j} \leq R_c \quad (1)$$

Here, i and j represent the chemical identities of the atoms contained in the GCN. k_i and k_j are the atomic index of a given atom from chemical

specie i and j respectively. d_{k_i, k_j} is defined as the L^2 -norm between two atoms. R_c is the cutoff radius specified when constructing the GCN. A visual depiction of how a GCN is constructed from various aluminum atomic structures can be found in Fig. 6. Each matrix element, $\frac{1}{d_{k_i, k_j}}$, represents the weight of a given edge for a given pair of connected nodes in the graph. The degree of each node is then given by the sum of the elements in a node's edge set. It should be noted that the matrix representation of the GCN is equivalent to a Coulomb matrix^{72,73}, which has been used previously to characterize molecular environments.

For the case of multicomponent systems, such as those encountered in alloys, polymers, and oxides, the GCN formalism can be extended to incorporate unique interactions between chemical species. For instance, in the case of a system containing two elements, three distinct GCN can exist (1) A-A, (2) A-B, and (3) B-B. Depending on the problem being studied, one may wish to employ specific graphs and there is no requirement that all three graphs be used to describe a particular system. For example, in the case of a two-component oxide, one may wish to observe how the network of metal ions changes over time, which can give an indication of how diffuse the metal network is. In this case, one would construct a metal-metal GCN where the corresponding nodes in the graph do not include any oxygen atoms.

Scalar graph order parameter

Here, we introduce the scalar graph order parameter to characterize the atomic coordination networks contained within the GCNs described in the previous section. Generally speaking, one can think of SGOP as a semi-empirical physically-informed graph similarity metric. We define this SGOP as:

$$\theta_{i-j, R_c} = \sum_s \left(\sum_m^{D_s} P(d_m) \log_b P(d_m) + d_m P(d_m) \right)^3 \quad (2)$$

Here, i and j represent the chemical identities of the atoms contained in the GCN. R_c is the cutoff radius specified when constructing the GCN. We make the assumption that a particular GCN is disconnected, and that the underlying network exists as a set of subgraphs, S , with s indexing a particular subgraph. Note that in the event a GCN is fully connected the outer sum disappears and no further changes are required to the formalism. D_s is the set of unique node degrees in a subgraph, with P_{d_m} being the probability of a given degree, d_m , occurring in the subgraph.

The underlying formalism of SGOP provides physical intuition about a graph: (a) $P(d_m) \log_b P(d_m)$ uses entropy to approximate a graph's shape and (b) $d_m P(d_m)$ characterizes a graph's connectivity. The entropic term can be easily identified as capturing the amount of chaos present in a graph, providing a unique mapping to the underlying shape. The connectivity term represents an empirical approximation of the density of a graph. It can be insufficient to compare more standard graph properties such as the maximum degree, minimum degree, and average degree, as these metrics can be not unique enough to capture the diversity present in a material's phase space. Therefore, the connectivity term was crafted to identify not only the degrees present within a graph but also the likelihood of occurrence of those degrees.

The cubed exponent of the inner summation provides a heuristic weighting mechanism to compare the sum of entropy and connectivity that was determined through trial and error. It is important to remember that the SGOP value is simply the sum of subgraph SGOPs if multiple subgraphs are present within a configuration of atoms. If the exponent is too large, highly connected and chaotic subgraphs will always be weighted too heavily when compared to smaller, poorly connected subgraphs. If the exponent is too small the opposite becomes true, in which subgraphs that are explicitly distinct run the risk of becoming indistinguishable during unsupervised clustering. Our experimentation has indicated that a cubed exponent provides a strong balance between these two extreme scenarios. In this way, SGOP can capture both similarities and subtle differences between graphs in a computationally efficient manner.

We note that the SGOP formalism is generalizable and transferable to any graph characterization, and is not restricted to the study of atomic configurations. It should also be noted that Eq. 2 is invariant under permutation, translation, replication (system size), and rotation operations. We also note that there exists a multitude of graph-based formalisms in the literature that aim to characterize atomic structures⁷⁴⁻⁷⁷, and the primary distinction between such methods and those prescribed in this work is the computational cost, mathematical completeness, and universal

transferability of our method. While further details regarding the software formalism and cost of the SGOP calculations can be found in Supplementary Table 2, we will indicate here that an SGOP for a 32,000 atom Aluminum system was computed in less than 0.5 s using only a serial execution. The low cost of the algorithm allows for the efficient characterization of not only complex structural systems but also the study of systems on the order of tens of nanometers in size.

Vector graph order parameter

While the SGOP formalism prescribed in the previous section accurately characterizes the graph network encoded within an atomic environment, the resulting value encodes local geometric information within a coordination sphere of radius R_c . Many atomic environments share underlying similarities in their local structure, which leads to overlapping values within the order parameter space. As a result, a single scalar is often not sufficient to distinguish between the complexity of a material's configuration space due to seemingly small but important differences encountered between atomic systems.

Here, we introduce the vector graph order parameter, which is simply a set of SGOP values, calculated using a unique, user-chosen set of R_c . By taking a set of coordination sphere radii, one can ensure that various portions of an atom's local geometry are properly encoded. We note here that, unlike previous methodologies which rely on explicitly capturing all unique atomic environments present in the system, VGOP merely requires that the global graph information at each cutoff is captured. This allows one to choose R_c in a less restrictive manner, and even a qualitative guess as to the important connectivity information is often all that is required. Figure 9 shows a visual workflow for how the VGOP is determined for the case of a carbon nanoparticle. Principle component analysis⁷⁸ is used to reduce the number of features and allow for the visual inspection of the underlying data. Z-score normalization⁷⁹ was used to normalize the VGOPs as a preprocessing measure to aid in the PCA decomposition, though in principle is not necessary. For the material systems studied in this work, the first two principle components comprised at least 95% of the underlying variance, and therefore the remaining components were discarded. Further information regarding the PCA decomposition for all systems studied in this work can be found in the Supplementary Methods section, along with Supplementary Figs. 6-8.

For the case of multicomponent systems, the final VGOP vector has a length of $\sum_n^N k_n$, where N is the number of GCN used to describe the system, with an index of n , and k_n is the number of cutoff values used in that particular GCN's VGOP vector. There is no restriction requiring that each GCN used to describe a unique A-B chemical interaction must have the same set of cutoff values. In fact, it makes intuitive sense that each interaction type would have its own set of cutoff values corresponding to the interaction distances encountered within that specific atomic network. The final VGOP for a system containing N interaction types would be given as:

$$\theta_{\text{final}} = \{\theta_0, \theta_1, \dots, \theta_k\} \quad (3)$$

where θ_k is the VGOP vector of interaction type k . As stated previously, all θ_k are mutually exclusive, with no requirement that an overlap exists between them. Each θ_k can have its own length and set of R_c . If one wishes to employ a normalization technique, one can either perform it over the entire set or on each θ_k individually, with each scenario representing a different statistical modification of the original θ_{final} vector. It is important to emphasize here that not all possible interaction types in the system have to be explicitly represented in θ_k , and it is up to the user to determine which interaction types are important for the problem being studied.

DATA AVAILABILITY

All data required to reproduce this work can be requested by contacting the corresponding author. All Al and C structures can be found in the Khazana data repository, while all initial Nb_xO_y structures can be found at Materials Project. We forward the reader to Sabri Elatresh and Stanimir Bonev for access to all Li data.

CODE AVAILABILITY

The SGOP formalism will be made publicly available in LAMMPS.

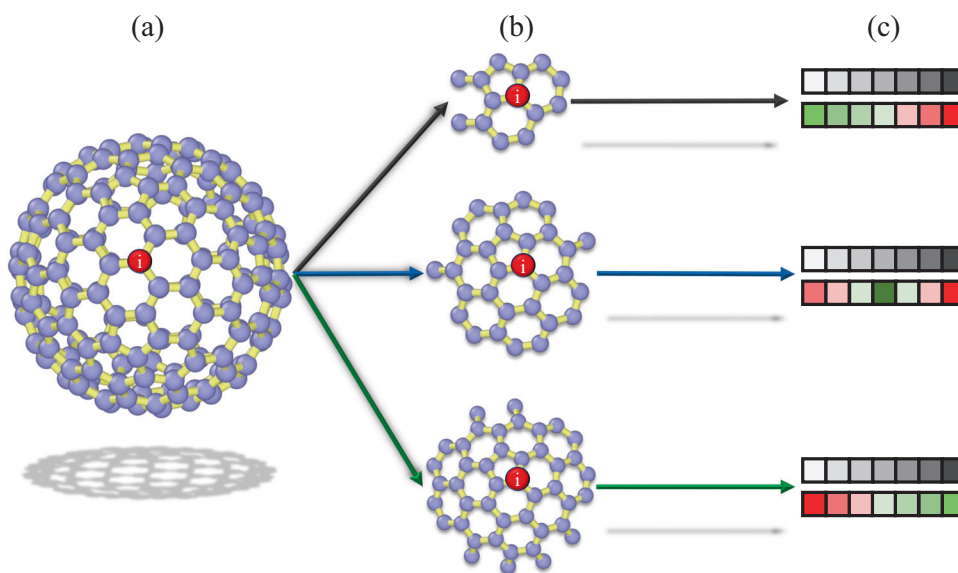


Fig. 9 Visualization of the VGOP framework. **a** A 180-atom buckyball is shown, with a specific atom i highlighted in red. **b** Three substructures of the buckyball, determined by selecting atoms varying cutoff radii away from atom i , that form the basis of unique GCNs. **c** Three sets of two bars, which each subset representing the degree set (top) of each GCN and the degree probability set (bottom). The colors of the degree set are defined as white representing low connectivity and black representing high connectivity. The colors of the probability set are specified as green representing high likelihood and red representing low likelihood of occurrence in the GCN. The varying colors of the probability sets indicate that the smallest substructure in **(b)** yields a large number of poorly connected nodes and little-to-no high connectivity, while the opposite is true for the largest substructure in **(b)**.

Received: 8 July 2021; Accepted: 23 January 2022;
Published online: 11 March 2022

REFERENCES

- Santiso, E. E. & Trout, B. L. A general set of order parameters for molecular crystals. *J. Chem. Phys.* **134**, 064109 (2011).
- Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 83 (2019).
- Archer, A. et al. Order parameter and connectivity topology analysis of crystalline ceramics for nuclear waste immobilization. *J. Phys. Cond. Mat.* **26**, 485011 (2014).
- Zuo, Y. et al. Performance and cost assessment of machine learning interatomic potentials. *J. Phys. Chem. A* **124**, 731–745 (2020).
- Xu, J., Cao, X. & Hu, P. Perspective on computational reaction prediction using machine learning methods in heterogeneous catalysis. *Phys. Chem. Chem. Phys.* **23**, 11155–11179 (2021).
- Fischer, C. C., Tibbetts, K. J., Morgan, D. & Ceder, G. Predicting crystal structure by merging data mining with quantum mechanics. *Nat. Mater.* **5**, 641–646 (2006).
- 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.* **4**, 34 (2017).
- Jinnouchi, R., Karsai, F., Verdi, C., Asahi, R. & Kresse, G. Descriptors representing two- and threebody atomic distributions and their effects on the accuracy of machine-learned inter-atomic potentials. *J. Chem. Phys.* **152**, 234102 (2020).
- Batra, R. et al. General atomic neighborhood fingerprint for machine learning-based methods. *J. Phys. Chem. C* **123**, 15859–15866 (2019).
- Caro, M. A. Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials. *Phys. Rev. B* **100**, 024112 (2019).
- Kawasaki, T. & Onuki, A. Construction of a disorder variable from Steinhardt order parameters in binary mixtures at high densities in three dimensions. *J. Chem. Phys.* **135**, 174109 (2011).
- Steinhardt, P. J., Nelson, D. R. & Ronchetti, M. Icosahedral bond orientational order in supercooled liquids. *Phys. Rev. Lett.* **47**, 1297–1300 (1981).
- Radhakrishnan, R. & Gubbins, K. E. Free energy studies of freezing in slit pores: an order-parameter approach using Monte Carlo simulation. *Mol. Phys.* **96**, 1249–1267 (1999).
- Eslami, H., Khanjari, N. & Muller-Plathe, F. A local order parameter-based method for simulation of free energy barriers in crystal nucleation. *J. Chem. Theory Comput.* **13**, 1307–1316 (2017).
- Gereben, O. & Pusztai, L. Determination of the atomic structure of disordered systems on the basis of limited Q-space information. *Phys. Rev. B* **51**, 5768–5772 (1995).
- Tian, Z. A., Liu, R. S., Dong, K. J. & Yu, A. B. A new method for analyzing the local structures of disordered systems. *EPL* **96**, 36001 (2011).
- Stachurski, Z. H. On structure and properties of amorphous materials. *Materials* **4**, 1564–1598 (2011).
- Li, Q. et al. Recent progress in some amorphous materials for supercapacitors. *Small* **14**, 1800426 (2018).
- Zhou, W.-X. et al. Thermal conductivity of amorphous materials. *Adv. Funct. Mater.* **30**, 1903829 (2020).
- Yan, S. et al. Research advances of amorphous metal oxides in electrochemical energy storage and conversion. *Small* **15**, 1804371 (2019).
- Leung, C. L. A. et al. Laser-matter interactions in additive manufacturing of stainless steel SS316L and 13-93 bioactive glass revealed by in situ X-ray imaging. *Addit. Manuf.* **24**, 647–657 (2018).
- Honeycutt, J. D. & Andersen, H. C. Molecular dynamics study of melting and freezing of small Lennard-Jones clusters. *J. Phys. Chem.* **91**, 4950–4963 (1987).
- Stukowski, A. Structure identification methods for atomistic simulations of crystalline materials. *Model Simul. Mat. Sci. Eng.* **20**, 045021 (2012).
- Kelchner, C. L., Plimpton, S. J. & Hamilton, J. C. Dislocation nucleation and defect structure during surface indentation. *Phys. Rev. B* **58**, 11085–11088 (1998).
- Druckfehlerverzeichnis der Arbeiten von, O. Perron (Bd. 132) und G. Voronoi (Bd. 133). en. *J. für die Reine und Angew. Math.* **1908**, 242a–242a (1908).
- Steinhardt, P. J. & Chaudhari, P. Point and line defects in glasses. *Philos. Mag. A* **44**, 1375–1381 (1981).
- Ackland, G. J. & Jones, A. P. Applications of local crystal structure measures in experiment and simulation. *Phys. Rev. B* **73**, 054104 (2006).
- Keys, A. S., Iacovella, C. R. & Glotzer, S. C. Characterizing complex particle morphologies through shape matching: Descriptors, applications, and algorithms. *J. Comp. Phys.* **230**, 6438–6463 (2011).
- Deng, L. et al. Local identification of chemical ordering: extension, implementation, and application of the common neighbor analysis for binary systems. *Comp. Mat. Sci.* **143**, 195–205 (2018).
- Snow, B. D., Doty, D. D. & Johnson, O. K. A simple approach to atomic structure characterization for machine learning of grain boundary structure-property models. *Front. Mater.* **6**, 120 (2019).
- De, S., Bartok, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).

32. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
33. Zeni, C., Rossi, K., Glielmo, A. & de Gironcoli, S. Compact atomic descriptors enable accurate predictions via linear models. *J. Chem. Phys.* **154**, 224112 (2021).
34. Artrith, N., Urban, A. & Ceder, G. Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species. *Phys. Rev. B* **96**, 014112 (2017).
35. Chandrasekaran, A. et al. Solving the electronic structure problem with machine learning. *npj Comput. Mater.* **5**, 22 (2019).
36. Zuo, Y. et al. Performance and cost assessment of machine learning interatomic potentials. *J. Phys. Chem. A* **124**, 731–745 (2020).
37. Onat, B., Ortner, C. & Kermodé, J. R. Sensitivity and dimensionality of atomic environment representations used for machine learning interatomic potentials. *J. Chem. Phys.* **153**, 144106 (2020).
38. Chapman, J. & Ramprasad, R. Multiscale modeling of defect phenomena in platinum using machine learning of force fields. *JOM* **72**, 4346–4358 (2020).
39. Deringer, V. L. et al. Realistic atomistic structure of amorphous silicon from machine-learning driven molecular dynamics. *J. Phys. Chem. Lett.* **9**, 2879–2885 (2018).
40. Deringer, V. L. & Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B* **95**, 094203 (2017).
41. Rosenbrock, C. W., Homer, E. R., Csányi, G. & Hart, G. L. W. Discovering the building blocks of atomic systems using machine learning: application to grain boundaries. *npj Comput. Mater.* **3**, 29 (2017).
42. Jose, K. V. J., Artrith, N. & Behler, J. Construction of high-dimensional neural network potentials using environment-dependent atom pairs. *J. Chem. Phys.* **136**, 194111 (2012).
43. Himanen, L. et al. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
44. Pozdnyakov, S. N., Zhang, L., Ortner, C., Csányi, G. & Ceriotti, M. Local invertibility and sensitivity of atomic structure-feature mappings. *Open Res Europe* **1**, 26 (2021).
45. Duvenaud, D. K. et al. Convolutional networks on graphs for learning molecular fingerprints. *Adv. Neural Inf. Process. Syst.* **28**, 1–9 (2015).
46. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **30**, 595–608 (2016).
47. 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.* **31**, 3564–3572 (2019).
48. Zeng, M. et al. Graph convolutional neural networks for polymers property prediction. Preprint at <https://arxiv.org/abs/1811.06231> (2018).
49. Shui, Z. & Karypis, G. Heterogeneous molecular graph neural networks for predicting molecule properties. In *Proc. 20th IEEE Conference on Data Mining* 492–500 (IEEE, 2020).
50. Pathak, Y., Mehta, S. & Priyakumar, U. D. Learning atomic interactions through solvation free energy prediction using graph neural networks. *J. Chem. Inf. Model* **61**, 689–698 (2021). PMID: 33546556.
51. Batra, R. et al. Polymers for extreme conditions designed using syntax-directed variational autoencoders. *Chem. Mater.* **32**, 10489–10500 (2020).
52. Kalinin, S. V., Dyck, O., Jesse, S. & Ziatdinov, M. Exploring order parameters and dynamic processes in disordered systems via variational autoencoders. *Sci. Adv.* **7**, eabd5084 (2021).
53. Lindsey, R. K., Bastea, S., Goldman, N. & Fried, L. E. Investigating 3,4-bis(3-nitrofurazan-4-yl)furoxan detonation with a rapidly tuned density functional tight binding model. *J. Chem. Phys.* **154**, 164115 (2021).
54. Kim, H.-J. et al. Nanostructures generated by explosively driven friction: experiments and molecular dynamics simulations. *Acta Mater.* **57**, 5270–5282 (2009).
55. Delaye, J.-M., Peugot, S., Bureau, G. & Calas, G. Molecular dynamics simulation of radiation damage in glasses. *J. Non Cryst. Solids* **357**, 2763–2768 (2011).
56. Mooney, B. L., Corrales, L. & Clark, A. E. MoleculaRnetworks: an integrated graph theoretic and data mining tool to explore solvent organization in molecular simulation. *J. Comput. Chem.* **33**, 853–860 (2012).
57. Ozkanlar, A. & Clark, A. E. ChemNetworks: a complex network analysis tool for chemical systems. *J. Comput. Chem.* **35**, 495–505 (2014).
58. Del Rio, B. G., Kuenneth, C., Tran, H. D. & Ramprasad, R. An efficient deep learning scheme to predict the electronic structure of materials and molecules: the example of graphene-derived allotropes. *J. Phys. Chem. A* **124**, 9496–9502 (2020).
59. Guillaume, C. L. et al. Cold melting and solid structures of dense lithium. *Nat. Phys.* **7**, 211–214 (2011).
60. Gorelli, F. A. et al. Lattice dynamics of dense lithium. *Phys. Rev. Lett.* **108**, 055501 (2012).
61. Berthier, L., Chandler, D. & Garrahan, J. P. Length scale for the onset of Fickian diffusion in supercooled liquids. *EPL* **69**, 320–326 (2005).
62. Page, L., Brin, S., Motwani, R. & Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web*. Stanford InfoLab, **422**, 1–17 (1999).
63. Xing, W. & Ghorbani, A. Weighted PageRank algorithm. In *Proc. Second Annual Conference on Communication Networks and Services Research* 305–314 (IEEE, 2004).
64. Drineas, P., Mahoney, M. W., Muthukrishnan, S. & Sarlós, T. Faster least squares approximation. *Numer. Math.* **117**, 219–249 (2011).
65. Pun, G. P. P., Batra, R., Ramprasad, R. & Mishin, Y. Physically informed artificial neural networks for atomistic modeling of materials. *IEEE Trans. Inf. Forensics* **10**, 2339 (2019).
66. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
67. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented wave method. *Phys. Rev. B* **59**, 1758 (1999).
68. Perdew, J. P., Burke, K. & Wang, Y. Generalized gradient approximation for the exchange-correlation hole of a many electron system. *Phys. Rev. B* **54**, 16533–16539 (1996).
69. Blochl, P. E. Projector augmented wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
70. Gatehouse, B. & Wadsley, A. The crystal structure of the high temperature form of niobium pentoxide. *Acta Crystallogr.* **17**, 1545–1554 (1964).
71. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
72. Schrier, J. Can one hear the shape of a molecule (from its Coulomb matrix eigenvalues)? *J. Chem. Inf. Model* **60**, 3804–3811 (2020).
73. Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **115**, 1094–1101 (2015).
74. Wang, X. et al. Molecule property prediction based on spatial graph embedding. *J. Chem. Inf. Model* **59**, 3817–3828 (2019).
75. Wodo, O., Tirthapura, S., Chaudhary, S. & Ganapathysubramanian, B. A graph-based formulation for computational characterization of bulk heterojunction morphology. *Org. Electron.* **13**, 1105–1113 (2012).
76. Estrada, E. Characterization of 3D molecular structure. *Chem. Phys. Lett.* **319**, 713–718 (2000).
77. Hall, L. H., Mohney, B. & Kier, L. B. The electrotopological state: structure information at the atomic level for molecular graphs. *J. Chem. Inf. Model* **31**, 76–82 (1991).
78. Karamzadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M. & Hooman, A. An overview of principal component analysis. *J. Sig. Inf. Proc.* **04**, 173–175 (2013).
79. Friedman, L. & Komogortsev, O. V. Assessment of the effectiveness of seven biometric feature normalization techniques. *IEEE Trans. Inf. Forensics Secur.* **14**, 2528–2536 (2019).

ACKNOWLEDGEMENTS

We would like to thank Sabri Elatresh and Stanimir Bonev for allowing us to use their DFT liquid lithium database. J. Chapman, N. Goldman, and B. Wood are partially supported by the Laboratory Directed Research and Development (LDRD) program (20-SI-004) at Lawrence Livermore National Laboratory. This work was performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under contract No. DE-AC52-07NA27344.

AUTHOR CONTRIBUTIONS

B.C.W. supervised the research. J.C. created the software. J.C. and N.G. designed the methodology and aided in the analysis of the data. J.C. wrote the manuscript with inputs from all authors.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-022-00717-7>.

Correspondence and requests for materials should be addressed to James Chapman.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022