Classification of atomic environments via the Gromov–Wasserstein distance

Sakura Kawano

Department of Chemical Engineering, University of California, Davis, Davis, CA 95616, USA

Jeremy K. Mason*

Department of Materials Science and Engineering, University of California, Davis, Davis, CA 95616, USA

Abstract

Interpreting molecular dynamics simulations usually involves automated classification of local atomic environments to identify regions of interest. Existing approaches are generally limited to a small number of reference structures and only include limited information about the local chemical composition. This work proposes to use a variant of the Gromov–Wasserstein (GW) distance to quantify the difference between a local atomic environment and a set of arbitrary reference environments in a way that is sensitive to atomic displacements, missing atoms, and differences in chemical composition. This involves describing a local atomic environment as a finite metric measure space, which has the additional advantages of not requiring the local environment to be centered on an atom and of not making any assumptions about the material class. Numerical examples illustrate the efficacy and versatility of the algorithm.

Keywords: Molecular dynamics, structure identification, Gromov–Wasserstein distance

1. Introduction

Contemporary molecular dynamics simulations can involve millions of atoms, though the atoms participating in the phenomenon of interest (e.g., phase nucleation, shear band nucleation, surface adsorption) are generally many fewer. Some automated procedure to classify local atomic environments is therefore indispensable to initially identify these regions so that the researcher can perform additional analysis. Given the difficulty of precisely defining what an exceptional atomic environment would be in the absence of crystalline order, many of the procedures already proposed apply almost exclusively to crystalline solids. More specifically, the assumption is often made that most atoms are nearly on simple cubic (SC), body-centered cubic (BCC), face-centered cubic (FCC), or hexagonal close-packed (HCP) lattice sites, and the classification problem is reduced to assigning atoms to one of these classes (or to one other class containing all defected atomic environments).

Existing approaches can roughly be grouped as topological or geometric. Topological approaches construct either the network of bonds connecting neighboring atoms, or the Voronoi tessellation with the atomic positions as seeds. Atoms are assigned to a class by considering the number and arrangement of nearby bonds in the bond network or nearby faces of the Voronoi polyhedra. This intentionally disregards some information about the relative positions of the atoms to make the classification more robust to perturbations of the positions at finite temperatures (i.e., thermal noise). Topological approaches often have the advantages of computational efficiency, simplicity of exposition, and well-defined criteria for an atom to belong to a particular class. Examples in the literature include common neighbor analysis [1, 2, 3], crystal analysis [4], neighborhood graph analysis [5], Voronoi analysis [6], topological fingerprints [7], and Voronoi cell topology [8].

Geometric approaches instead map the relative positions of atoms in a local atomic environment to a continuous feature space. Each class is associated with a region of the feature space, and atoms whose feature vectors fall within one of these regions are assigned to that class. The regions are usually not defined a priori, but rather are constructed after observing the distribution of feature vectors of atoms in reference environments. Geometric approaches can provide information about the atomic environment that is not readily accessible to topological approaches, e.g., point symmetry groups or elastic strain tensors, but can suffer more from thermal noise and be more expensive to calculate. Examples in the literature include the centrosymmetry parameter [9], bond-orientational order parameters [10, 11], the Minkowski structure metric [12], bond angle analysis [13], neighbor distance analysis [3], and polyhedral template matching [14].

Of the approaches above, adaptive common neighbor analysis (ACNA) [3] and polyhedral template matching (PTM) [14] are perhaps the most frequently used to iden-

*Corresponding author

Email addresses: skawano@ucdavis.edu (Sakura Kawano), jkmason@ucdavis.edu (Jeremy K. Mason)

tify atomic environments in crystalline solids. They perform particularly well for molecular dynamics simulations of single-component systems, and the procedure proposed here is not necessarily intended for such applications. That said, there are still several respects in which they could be improved.

First, they are effectively limited to consider only one or two nearest neighbor shells around a central atom. This is a consequence of the way the local bond network is constructed for ACNA, and of the use of a convex hull as part of the matching algorithm for PTM. As the accuracy of interatomic potentials in two and three component systems continues to improve and simulations of materials with more complex crystal structures become more common, methods able to handle extended environments will likely become more relevant.

Second, the methods are sensitive to atoms entering or leaving the local environment; this is related but not entirely equivalent to being robust to thermal noise. ACNA reduces the frequency of such events by varying the radius of the local environment with the reference environment and the atomic positions, while PTM uses a topological ordering of nearby atoms to make the classification resistant to perturbations in the atomic positions. Nevertheless, a shear strain applied to a large atomic environment could still displace some of the atoms enough to leave the region being considered and frustrate the analysis.

Third, they can only include limited information about the chemical composition of the local environment, at least in the forms currently in the literature. ACNA could be adapted to include chemical information by appending the species of the atoms along bond chains [15], though this would be unwieldy for three or more chemical species. PTM has been used for binary alloys [14], but apparently requires considerable symmetry in the arrangement of the chemical species. A more flexible approach would be valuable, particularly if molecular dynamics simulations of two- and three-component systems become more common.

The procedure proposed here is based on the Gromov–Wasserstein (GW) distance recently defined by Memoli [16, 17, 18], which up to now has mostly been used for shape matching in the field of computer vision [19, 20, 21]. For example, the GW distance can be used to match an object represented as an incomplete point cloud to one of a set of reference objects, perhaps in a difference pose. This is not dissimilar to matching a local atomic environment to one of a set of reference environments, possibly with perturbed atomic positions or some of the atoms missing. Apart from resolving the three limitations above, our approach has the additional advantages of not requiring the local environment to be centered on an atom (e.g., for the identification of vacancies) and of providing a metric on the space of all local atomic environments. That said, the GW distance is more complicated to define and is substantially slower to calculate than ACNA and PTM, and for that reason is intended to be complementary to them.

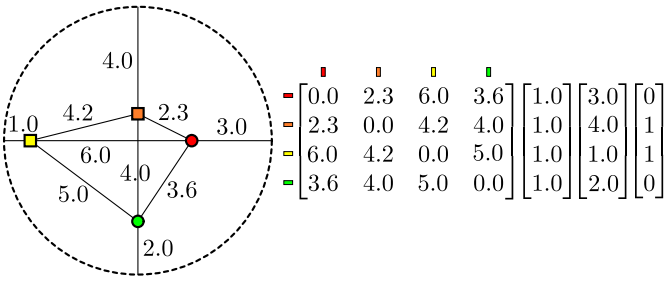

Figure 1: A description of a local atomic environment as a finite metric measure space. Color indicates the distinct points of the finite space, and circles and squares indicate the two chemical species. Numbers in the local atomic environment (left) are dimensionless Euclidean distances. The metric (second from left) indicates the pairwise distances between atoms, the measure (middle) indicates the fraction of an atom associated with each point, the distances to the boundary (second from right) are the distances from each atom to the closest point on the boundary, and the species labels (right) indicate the chemical species of the atoms.

2. Finite Metric Measure Spaces

A local atomic environment is often described by a set of vectors from the central atom to the surrounding atoms. The GW distance instead requires that a local atomic environment be described as a *finite metric measure space*. As the name implies, this involves the construction of a finite space, a metric describing distances in the space, and a measure describing the distribution of atoms in the space. Figure 1 is a concrete example of the construction for a spherical region. While the region is not required to be spherical, this simplifies some of the analysis and will be assumed throughout.

A *finite space* is a topological space that contains only a finite number of points. The natural choice for a local atomic environment is one point for each atomic center, as indicated by the red, orange, yellow and green points in Figure 1. For our purposes, a *metric* is a symmetric matrix of pairwise Euclidean distances between points, and a *measure* is a function that assigns values to points. The GW distance uses these as weights to indicate the relative importance of points in the space, but otherwise does not specify their interpretation. Here, the measure will be used to indicate the number of atoms associated with a point. Since atoms are indivisible and the position of each atom is unique, all of the entries will be 1.0.

While not part of the definition of a finite metric measure space, our description of a local atomic environment includes a vector of *distances to the boundary* for each atom and a vector of *species labels* that indicates the chemical species of the atoms associated with each point. Distances to the boundary are used to penalize the departure of atoms from the environment. This is envisioned as involving the motion of an atom to the boundary, and hence is proportional to the distance to the boundary. The chemical species is well-defined since each point is associated with a single atom. By convention, the chemical species are labeled with increasing integers starting with zero.

Describing a local atomic environment as a finite metric measure space instead of as a set of bond vectors has several advantages. First, Figure 1 shows that the local atomic environment does not need to be centered on an atom. This allows the GW distance to be used to find, e.g., the precise locations of vacancies or interstitial sites in a finite temperature system. Second, the distance matrix is invariant to translations, rotations, and reflections of the local atomic environment; these symmetries do not need to be handled in a separate calculation as with PTM.

3. Gromov–Wasserstein Distance

The GW distance is a metric [17] that allows the comparison of finite metric measure spaces. More specifically, let X be a finite space with metric $\boldsymbol{d}^X$ and measure $\boldsymbol{\mu}^X$; the triple $\mathbb{X} = \{X, \boldsymbol{d}^X, \boldsymbol{\mu}^X\}$ is a finite metric measure space. The GW distance is then a function $\mathcal{G}(\mathbb{X}, \mathbb{Y})$ with the following properties for all finite metric measure spaces $\mathbb{X}$, $\mathbb{Y}$ and $\mathbb{Z}$ with the same total measures:

1. $\mathcal{G}(\mathbb{X}, \mathbb{Y}) \geq 0$,
2. $\mathcal{G}(\mathbb{X}, \mathbb{Y}) = 0$ if and only if $\mathbb{X} = \mathbb{Y}$,
3. $\mathcal{G}(\mathbb{X}, \mathbb{Y}) = \mathcal{G}(\mathbb{Y}, \mathbb{X})$, and
4. $\mathcal{G}(\mathbb{X}, \mathbb{Z}) \geq \mathcal{G}(\mathbb{X}, \mathbb{Y}) + \mathcal{G}(\mathbb{Y}, \mathbb{Z})$.

These conditions are designed to ensure that every metric conform to our usual intuitions about distance in Euclidean space. In particular, the fourth condition is known as the triangle inequality, and is required for the clustering of points to be defined in a meaningful way; without this, even if $\mathbb{X}$ is close to $\mathbb{Y}$ and $\mathbb{Y}$ is close to $\mathbb{Z}$, $\mathbb{X}$ and $\mathbb{Z}$ could still be arbitrarily far apart. The definition of a metric is provided here because our use of the word is somewhat more restricted than elsewhere in the materials science literature [22].

The notion of a *measure coupling* will be useful when describing the calculation of the GW distance. Given finite metric measure spaces $\mathbb{X}$ and $\mathbb{Y}$ with n and m points, an admissible measure coupling between them is an $n \times m$ matrix $\boldsymbol{\mu}$ with non-negative entries. Intuitively, this provides a correspondence of points in X with points in Y that allows for partial matching. Denote the row and column sums as $\nu_i^X = \sum_j \mu_{ij}$ and $\nu_j^Y = \sum_i \mu_{ij}$ for all $i \in [1, n]$ and $j \in [1, m]$. A measure coupling can be balanced or unbalanced, where a balanced measure coupling is one for which the row sums equal $\boldsymbol{\mu}^X$ and the column sums equal $\boldsymbol{\mu}^Y$, i.e., $\nu_i^X = \mu_i^X$ and $\nu_j^Y = \mu_j^Y$. Let the set of all admissible unbalanced measure couplings for the finite metric measure spaces $\mathbb{X}$ and $\mathbb{Y}$ be indicated by $\mathcal{M}(\boldsymbol{\mu}^X, \boldsymbol{\mu}^Y)$.

Given an admissible measure coupling $\boldsymbol{\mu}$, define the quantity

$$J(\boldsymbol{\mu}|\boldsymbol{d}^X, \boldsymbol{d}^Y) = \sum_{i',i=1}^{n} \sum_{j',j=1}^{m} |d^X_{i'i} - d^Y_{j'j}| \mu_{i'j'} \mu_{ij}$$

and let λ_i^X be the distance to the boundary of the ith point of X. Then the unbalanced GW distance[1] between $\mathbb{X}$ and $\mathbb{Y}$ is defined here as

$$\mathcal{G}(\mathbb{X}, \mathbb{Y}) = \min_{\boldsymbol{\mu} \in \mathcal{M}} \left[\frac{1}{2} J(\boldsymbol{\mu}|\boldsymbol{d}^X, \boldsymbol{d}^Y) + \sum_{i=1}^{n} \lambda_i^X \left| \nu_i^X - \mu_i^X \right| + \sum_{j=1}^{m} \lambda_j^Y \left| \nu_j^Y - \mu_j^Y \right| \right], \quad (1)$$

following the same approach as for the unbalanced Wasserstein distance of Chizat et al. [23]. The motivation for the unbalanced GW distance is that it is not always possible to find a balanced measure coupling, e.g., when there are unequal number of atoms between the reference and local environments.

Figure 2 provides several examples that are intended to help the reader develop an intuition for this definition. Let the reference environment $\mathbb{X}$ be the leftmost in the figure. The local atomic environment $\mathbb{Y}_1$ second from the left is identical to $\mathbb{X}$ except for a permutation of the atomic labels, and the $\boldsymbol{\mu}^*$ that achieves the minimum in Eq. 1 is a permutation matrix that maps one set of atoms to the other (e.g., the orange atom in $\mathbb{X}$ is mapped to the yellow atom in $\mathbb{Y}_1$). Since $\mathbb{X}$ and $\mathbb{Y}_1$ differ only by a symmetry of the physical system, the GW distance $\mathcal{G}(\mathbb{X}, \mathbb{Y}_1)$ vanishes. The local atomic environment $\mathbb{Y}_2$ second from the right additionally has a small perturbation applied to the green atom's position, visible as changes of distance to the yellow and orange atoms. In this case $\boldsymbol{\mu}^*$ remains the same, but $\mathcal{G}(\mathbb{X}, \mathbb{Y}_2)$ is the sum of the magnitudes of the distance changes between every pair of atoms. Since the distances in $\mathbb{X}$ and $\mathbb{Y}_2$ are the same except for those around the perturbed atom, $\mathcal{G}(\mathbb{X}, \mathbb{Y}_2)$ is the sum of the magnitudes of the distance changes from the green to the yellow and orange atoms in $\mathbb{Y}_2$. The rightmost local atomic environment $\mathbb{Y}_3$ instead has a missing atom, requiring that the corresponding column of $\boldsymbol{\mu}^*$ be removed. The resulting discrepancy between $\boldsymbol{\mu}^X$ and $\boldsymbol{\nu}^X$ makes $\mathcal{G}(\mathbb{X}, \mathbb{Y}_3)$ the distance the missing atom would have traveled to reach the boundary and leave the local atomic environment. The minimum in Eq. 1 allows the GW distance to remain continuous as the magnitude of the perturbation increases and removing the atom becomes the less expensive option.

There are several other conditions that should be satisfied by the finite metric measure spaces before the GW distance is applied. First, the measures should be strictly positive. Any points for which the measures are zero (i.e., that are not occupied by atoms) should be removed from the spaces, the corresponding rows and columns removed from the distance matrices, and the corresponding entries removed from the measures, distances to the boundary, and species labels. Second, the algorithm is more stable

[1] This is actually the unbalanced 1-Gromov–Wasserstein distance. p-Gromov–Wasserstein distances can be defined for any $p \in [1, \infty)$.

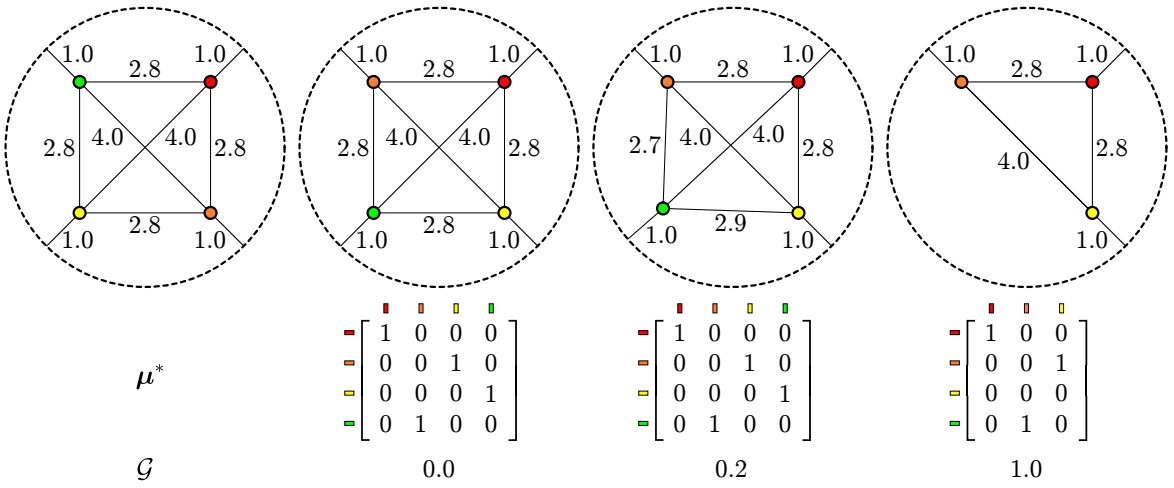

Figure 2: Examples of local atomic environments intended to clarify the meaning of Eq. 1. The leftmost environment $\mathbb{X}$ is the reference environment, with the numbers indicating the distances between pairs of atoms. The second from the left environment $\mathbb{Y}_1$ differs only by a permutation of the atomic labels, the second from the right environment $\mathbb{Y}_2$ additionally has a perturbation applied to the green atom's position, and the rightmost environment $\mathbb{Y}_3$ instead has a missing atom. The second row gives the measure couplings $\boldsymbol{\mu}^*$ that realize the minimum of Eq. 1, with the rows corresponding to atoms of $\mathbb{X}$ and the columns to atoms of $\mathbb{Y}_i$. The third row gives the GW distance $\mathcal{G}(\mathbb{X}, \mathbb{Y}_i)$ to the same precision as the distances.

when the median off-diagonal entry of the distance matrices is of order one. If κ is the median of these entries, then the distance matrices and distances to the boundary should be divided by κ before the calculation, and the GW distance multiplied by κ after the calculation.

This leaves the problem of finding a measure coupling $\boldsymbol{\mu}^*$ that realizes the minimum in Eq. 1. Formally, this is at least as difficult as a nonconvex quadratic optimization problem with linear constraints, and for which there is no known polynomial-time algorithm to find the global minimum [24]. In practice, the approach followed in the literature [16, 17, 20, 21] is to approximate $\boldsymbol{\mu}^*$ by successive linear optimization problems, and the same approach is followed here:

1. Initialize $\boldsymbol{\mu}$ with some admissible measure coupling $\boldsymbol{\mu}^0$ and set $k = 0$.
2. Solve the linear optimization problem

$$c_{ij}^k = \frac{1}{2}\sum_{i'=1}^{n}\sum_{j'=1}^{m} |d_{ii'}^X - d_{jj'}^Y| \mu_{i'j'}^k,$$

$$\boldsymbol{\mu}^{k+1} = \underset{\boldsymbol{\mu}\in\mathcal{M}}{\operatorname{argmin}} \left[\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}^k \mu_{ij} + \sum_{i=1}^{n} \lambda_i^X \left|\nu_i^X - \mu_i^X\right| + \sum_{j=1}^{m} \lambda_j^Y \left|\nu_j^Y - \mu_j^Y\right| \right]. \quad (2)$$

3. If the stopping criterion is satisfied, set $\boldsymbol{\mu}^* = \boldsymbol{\mu}^{k+1}$ and exit. If not, set $k = k + 1$ and return to Step 2.

This is known as the alternate convex search algorithm [17, 25], and converges to a local minimum of the original problem. In principle, the quality of the result could be improved by repeatedly running the algorithm with randomized initial conditions. In practice, the measure coupling is initialized to a constant matrix and heuristic perturbations are regularly applied to break any symmetries. The efficacy of this approach is visible in Section 6.

The linear optimization problem in Eq. 2 is identical to the one used to calculate an unbalanced Wasserstein distance [23]. Let $\epsilon > 0$ be a regularization parameter and replace the linear optimization problem in Eq. 2 with

$$\tilde{\boldsymbol{\mu}}^{k+1} = \underset{\boldsymbol{\mu}\in\mathcal{M}}{\operatorname{argmin}} \left[\sum_{i=1}^{n}\sum_{j=1}^{m} (c_{ij}^k + \epsilon \log \mu_{ij})\mu_{ij} + \sum_{i=1}^{n} \lambda_i^X \left|\nu_i^X - \mu_i^X\right| + \sum_{j=1}^{m} \lambda_j^Y \left|\nu_j^Y - \mu_j^Y\right| \right].$$

This can be solved efficiently with a modified Sinkhorn-Knopp algorithm [23] as follows:

1. Initialize $a_i^0 = 1$ for all $i \in [1, n]$, $b_j^0 = 1$ for all $j \in [1, m]$, $\gamma_{ij} = \exp(-c_{ij}^k/\epsilon)$, and set $\ell = 0$.
2. Set $a_i^{\ell+1} = \min[e^{\lambda_i^X/\epsilon}, \max(e^{-\lambda_i^X/\epsilon}, \mu_i^X / \sum_j \gamma_{ij} b_j^\ell)]$.
3. Set $b_j^{\ell+1} = \min[e^{\lambda_j^Y/\epsilon}, \max(e^{-\lambda_j^Y/\epsilon}, \mu_j^Y / \sum_i a_i^{\ell+1} \gamma_{ij})]$
4. If the stopping criterion is satisfied, set $\mu_{ij}^{k+1} = a_i^{\ell+1}\gamma_{ij}b_j^{\ell+1}$ and exit. If not, set $\ell = \ell + 1$ and return to Step 2.

Decreasing ϵ reduces the regularization and drives $\tilde{\boldsymbol{\mu}}^{k+1}$ toward the solution of Eq. 2, but can introduce numerical instabilities.

Our implementation uses the log-domain stabilization and ϵ-scaling of Schmitzer [26, 23]. These modifications to the basic Sinkhorn-Knopp algorithm require the introduction of several additional parameters; $\tau = 1000$ regulates the frequency of absorption iterations for the log-domain stabilization, and ϵ is scaled by factors of 4 from an initial

value of the median of the c_{ij} to a final value of 0.001 times the median distance between distinct points. The various heuristics used to escape local minima are described in Appendix A. The algorithm is written as a library in portable C11 with Python and MATLAB interfaces, is open source, and is available on request.

4. Classification of Atomic Environments

While the calculation of the unbalanced GW distance introduced in Section 3 is relevant to general finite metric measure spaces, this section instead describes the application of the unbalanced GW distance to the classification of local atomic environments. In particular, Section 3 does not introduce the chemical species of the atoms. The extension of the unbalanced GW distance to the case of multiple species is called the composition-restricted Gromov–Wasserstein (CRGW) distance.

First, the user should specify a region of Euclidean space to be used for the definition of all local atomic environments. Since crystal structure and orientation can vary throughout a simulation cell, a spherical region with a radius of 1.5 to 2.5 times the average atomic spacing is a reasonable choice.

Second, the user should provide a reference atomic environment for each class being considered. The finite metric measure spaces of the reference atomic environments are then constructed and stored for subsequent use. These take the form of sets $\mathbb{X} = \{X, \boldsymbol{d}^X, \boldsymbol{\mu}^X, \boldsymbol{\lambda}^X, \boldsymbol{\delta}^X\}$ where $\boldsymbol{\lambda}^X$ and $\boldsymbol{\delta}^X$ contain the distances to the boundary and the species labels.

Third, the local atomic environment to be classified is identified, and the corresponding finite metric measure space $\mathbb{Y} = \{Y, \boldsymbol{d}^Y, \boldsymbol{\mu}^Y, \boldsymbol{\lambda}^Y, \boldsymbol{\delta}^Y\}$ is constructed using the same region as before. The CRGW distance from $\mathbb{Y}$ to each of the reference environments is calculated, and a user-specified criterion is used to classify the local atomic environment on the basis of these distances. Part of the advantage of this approach is that the classification criterion can be as simple or as complex as the user desires; the local atomic environment could be assigned to the class with the smallest distance, or assigned to the most likely class using the probability distributions of distances developed in Section 5.

This leaves the calculation of the CRGW distance itself. Let $\mathbb{X}$ and $\mathbb{Y}$ be the finite metric measure spaces of the reference and local atomic environments, and have n and m atoms respectively. Let $\boldsymbol{R}$ be an $n \times m$ matrix with R_{ij} equal to one if the ith atom of $\mathbb{X}$ and the jth atom of $\mathbb{Y}$ have the same species label, and zero otherwise. Then the CRGW distance $\mathcal{D}(\mathbb{X}, \mathbb{Y})$ is still defined by means of Eq. 1, but with the minimization performed over the restricted set of measure couplings with the same zero entries as $\boldsymbol{R}$ (atoms of different chemical species cannot be coupled). Within the context of Section 3, this restriction can be realized by replacing the initial measure coupling

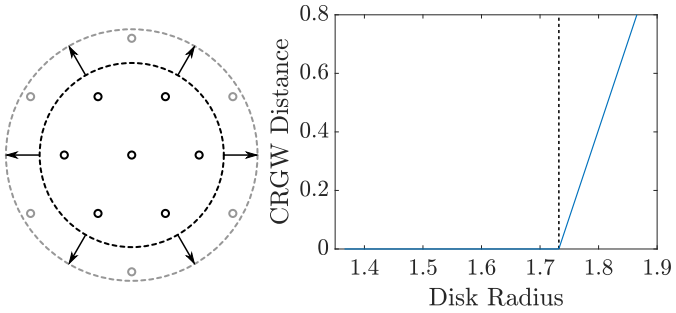

Figure 3: The CRGW distance is continuous with respect to atoms entering and leaving a local atomic environment. The radius of the environment on the left is increased from $(1+\sqrt{3})/2$ to $(2+\sqrt{3})/2$ in units of the atomic spacing. The distance to the initial condition on the right is continuous, with a discontinuous first derivative at $\sqrt{3}$.

μ_{ij}^0 with $R_{ij}\mu_{ij}^0$ and replacing γ_{ij} in Step 1 of the modified Sinkhorn-Knopp algorithm with $\gamma_{ij} = R_{ij}\exp(-c_{ij}/\epsilon)$.

Note that the calculation of the CRGW distance actually increases in efficiency with the number of chemical species for a fixed number of atoms. The reason for this is that the sparsity of $\boldsymbol{R}$ increases with the number of chemical species, dramatically reducing the set of possible measure couplings in Eq. 1. That said, any efficiency gains would likely be offset by an increase in the number of reference atomic environments defined by the user.

With the CRGW distance defined, the rest of this section consists of illustrative examples where the procedure is applied to local atomic environments in two dimensions. This simplification is used only for clarity of the figures; since the CRGW distance does not explicitly depend on the dimension of the ambient space, the calculation is precisely the same in two and three dimensions.

Figure 3 shows that the CRGW distance is continuous with respect to atoms entering and leaving the local atomic environment. The atoms are arranged on a triangular lattice with unit spacing, and the radius of the local atomic environment on the left is increased from $(1+\sqrt{3})/2$ to $(2+\sqrt{3})/2$. The CRGW distance to the environment is a continuous function of the radius, even though six atoms enter the region at a radius of $\sqrt{3}$ and cause a discontinuous derivative at the dashed vertical line.

Figure 4 shows that the CRGW distance is continuous with respect to displacements of the local atomic environment. The atoms are arranged on a triangular lattice with unit spacing as before, and the center of a local atomic environment of radius $(1+\sqrt{3})/2$ is moved along a straight line between neighboring atoms. The CRGW distance to the initial environment is a continuous function of the displacement, passing through a maximum halfway between the atoms before returning to zero. The first two dashed vertical lines indicate discontinuous derivatives caused by the bottom atom leaving the environment and two of the uppermost atoms entering the environment, respectively. The environment briefly contains eight atoms before two corresponding events occur in reverse order as the distance

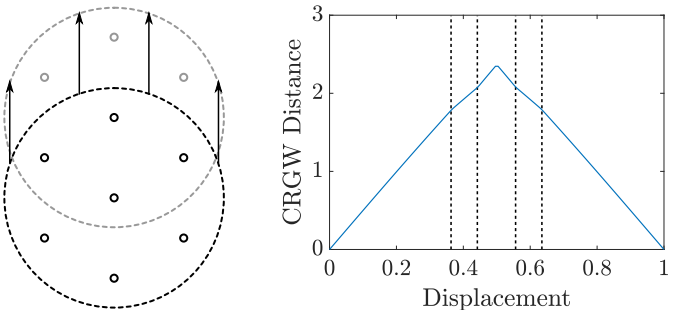

Figure 4: The CRGW distance is continuous with respect to displacements of the local atomic environment. The center of the environment on the left is moved in the vertical direction by one atomic spacing. The distance to the initial condition on the right is continuous, with discontinuous first derivatives at $(\sqrt{3}-1)/2$ and $(3-\sqrt{1+2\sqrt{3}})/2$, and the reflection of these quantities about 0.5.

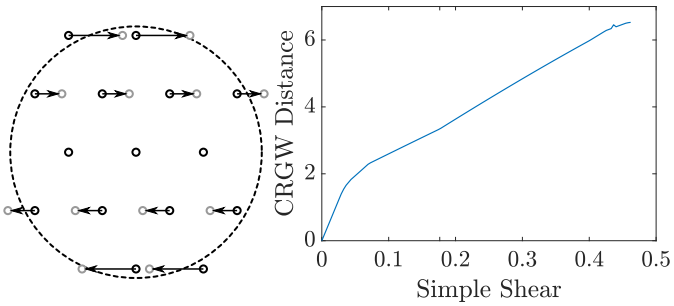

Figure 5: The CRGW distance is continuous with respect to elastic deformations of the local atomic environment. The environment on the left is subjected to a simple shear of $4/(5\sqrt{3})$. The distance to the initial condition on the right is continuous.

returns to zero.

Figure 5 shows that the CRGW distance is continuous with respect to elastic deformations of the local atomic environment. The atoms are again arranged on a triangular lattice with unit spacing, and a local atomic environment of radius $(2+\sqrt{3})/2$ is subjected to a simple shear that increases to a maximum of $4/(5\sqrt{3})$. The CRGW distance to the initial environment is a continuous function of the shear. The shoulder around 0.5 is caused by atoms in the local environment being displaced to the boundary as the distance to the boundary decreases and the cost of matching to a reference atom increases.

The remaining figure in this section considers the performance of the CRGW distance for a defected material with multiple chemical species and phases. The leftmost image in Figure 6 shows the atomic positions in a simulation cell with periodic boundary conditions. The atomic shape (circle or square) and color (red or blue) indicate the chemical species and phase, where the phases can be distinguished by chemical composition and lattice type. The blue phase additionally contains two vacancies on distinct hexagonal unit cell sites. Distances are expressed in units of the interatomic spacing, which is assumed to be the same for all chemical species and phases. The radius of all local atomic environments is set to 1.75.

The six reference atomic environments appear at the bottom of the figure, and are divided into three groups. From left to right, these correspond to atomic sites in the red phase, atomic sites in the blue phase, and vacancies in the blue phase. Local atomic environments of the same radius are constructed on a grid throughout the simulation cell, and the CRGW distances to the six reference atomic environments are calculated for each one. The right three images of Figure 6 show the smallest distance to any of the reference environments in the respective group, with smaller distances indicating more similarity. The atoms belonging to the red phase, the blue phase, and the interface can be identified by visual inspection of the middle images, and the location of the vacancies is clearly indicated in the rightmost image.

5. Thermal Noise

All of the examples in Section 4 positioned the atoms on lattice sites, whereas molecular dynamics simulations are generally performed at finite temperatures with perturbed atomic positions. While molecular dynamics simulations can be quenched to return the atoms to their lattice sites, this requires additional computation and can complicate the observation of temperature-dependent phenomena. Hence, any approach to classify local atomic environments would ideally be robust to such perturbations. As described in Section 1, existing geometric approaches handle this by identifying each class with some region of a feature space, with the regions defined by observation and convention rather than more fundamental considerations. This is not entirely necessary though; one could model atomic displacements as independent random variables, and derive a probability distribution of feature vectors for a given reference environment. Classification of an environment would then be reduced to, e.g., comparison with a set of prediction intervals.

This is the approach developed in the current section. Let $\mathbb{X}$ be a given reference environment and $\mathbb{Y}$ be the same environment subject to random thermal displacements of a given magnitude. The predicted distribution of CRGW distances $\mathcal{G}(\mathbb{X},\mathbb{Y})$ is constructed below, and allows one to test the hypothesis that a test environment $\mathbb{Z}$ is also derived from $\mathbb{X}$ by the application of random thermal displacements. This procedure is used to classify local atomic environments in molecular dynamics simulations in Section 6.

Let there be a reference environment where all of the atoms are on the interior of the region and are not too close to the boundary. Suppose that the potential energy ϕ of the ith atom can be approximated in the vicinity of the minimum by a parabolic function

$$\phi(\boldsymbol{r}_i) = \frac{1}{2}a|\boldsymbol{r}_i|^2 + b$$

where $\boldsymbol{r}_i$ is the atomic displacement of the ith atom from the position of minimum potential energy. Assuming that

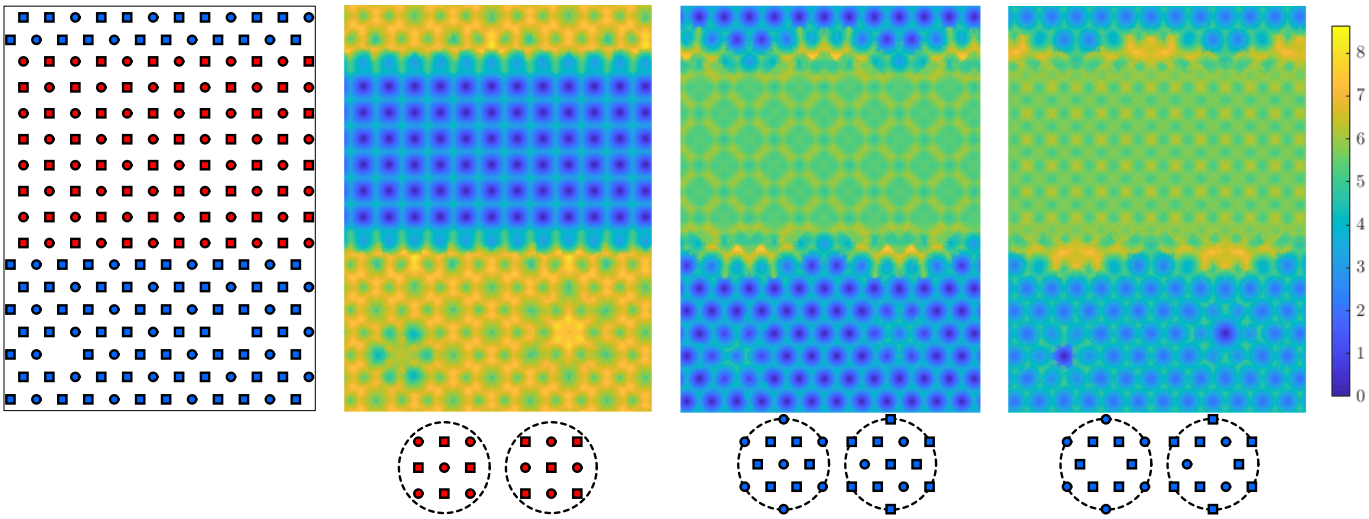

Figure 6: Performance of the CRGW distance in a defected material with multiple chemical species and phases. The leftmost image shows atomic positions, with chemical species indicated by circles or squares and phases indicated by red or blue. The remaining three images show the smaller of the CRGW distances to the the two local atomic environments below the respective figure, and indicate, from left to right, atomic sites in the red phase, atomic sites in the blue phase, and vacancies in the blue phase.

atomic displacements are independent, the probability distribution $p(\boldsymbol{r}_i)$ of a displacement of the ith atom in the canonical ensemble is a product of normal distributions

$$p(\boldsymbol{r}_i) = \left(\frac{a}{2\pi k_B T}\right)^{3/2} \exp\left(-\frac{a|\boldsymbol{r}_i|^2}{2k_B T}\right)$$

where k_B is Boltzmann's constant and T is the absolute temperature. Let $\sigma_r^2 = k_B T/a$ indicate the variance of the atomic displacements.

Let $\mathbb{X}$ be a reference environment, and $\mathbb{Y}$ a perturbation of that environment. Suppose that the atomic perturbations are small enough that all of the n atoms remain in the environment, and that each atom in $\mathbb{X}$ can be unambiguously identified with an atom in $\mathbb{Y}$. For any natural ordering of atoms in $\mathbb{X}$ and $\mathbb{Y}$, μ_{ij} is a diagonal matrix with ones and zeros on the diagonal. Let ξ_i^h be the ith entry of the diagonal, with $h \in \mathcal{H}$ indicating which of the possible 2^n binary vectors is chosen. Each $\boldsymbol{\xi}^h$ corresponds to a particular subset of atomic pairs in $\mathbb{X}$ and $\mathbb{Y}$ being mapped to the boundary. Equation 1 reduces for this case to

$$\begin{aligned}\mathcal{G}(\mathbb{X},\mathbb{Y}) &= \min_{h\in\mathcal{H}}\left[\frac{1}{2}\sum_{i,j}^{n}|d_{ij}^X - d_{ij}^Y|\xi_i^h\xi_j^h + \sum_i^n \lambda_i^X(1-\xi_i^h)\right.\\ &\qquad\left. + \sum_i^n \lambda_i^Y(1-\xi_i^h)\right]\\ &= \min_{h\in\mathcal{H}} D_h. \end{aligned} \tag{3}$$

Observe that the D_h are correlated random variables, constructed as sums of the random variables $|d_{ij}^X - d_{ij}^Y|$ and λ_i^Y. The joint probability distribution of the D_h will be modeled as a multivariate normal distribution using the multivariate central limit theorem. The probability distribution of $\mathcal{G}(\mathbb{X},\mathbb{Y})$ can then be constructed by explicitly sampling from the joint distribution of the D_h and finding the minimum D_h for each sample. The problem is thereby reduced to the calculation of the means and covariance matrix of the D_h that define the multivariate normal distribution. These are found in Appendix B to be

$$\langle D_h\rangle = \sum_{i,j}^{n}\frac{\sigma_r}{\sqrt{\pi}}\xi_i^h\xi_j^h + 2\sum_i^n \lambda_i^X(1-\xi_i^h) \tag{4}$$

$$\begin{aligned}\mathrm{cov}(D_h, D_g) = \Bigg[&\sum_i^n (1-\xi_i^h)(1-\xi_i^g)\\ &+\left(1-\frac{2}{\pi}\right)\sum_i^n\sum_{j\neq i}^n \xi_i^h\xi_j^h\xi_i^g\xi_j^g\\ &+\sum_i^n\sum_{j\neq i}^n\sum_{k\neq i,j}^n \xi_i^h\xi_j^h\xi_i^g\xi_k^g f(\theta_{ijk})\Bigg]\sigma_r^2\end{aligned} \tag{5}$$

where $\langle\cdot\rangle$ indicates the mean of a quantity and $f(\theta_{ijk})$ is defined by Eq. B.6. The covariance matrix depends on the geometry of the reference environment via the angles θ_{ijk} between triplets of atoms in the reference environment.

To sample from this distribution, define the matrix elements $\Sigma_{hg} = \mathrm{cov}(D_h, D_g)$ and find any real matrix $\boldsymbol{A}$ such that $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$. Let z_g be a random variable distributed according to the standard normal distribution. Then

$$Y = \min_{h\in\mathcal{H}}\left[\langle D_h\rangle + \sum_g A_{hg} z_g\right]$$

samples from the distribution of $\mathcal{G}(\mathbb{X},\mathbb{Y})$ implicitly defined by Eq. 3.

Figure 7 provides some numerical evidence that samples of $\mathcal{G}(\mathbb{X},\mathbb{Y})$ can be used to construct the empirical distribution. The reference environment on the left resembles

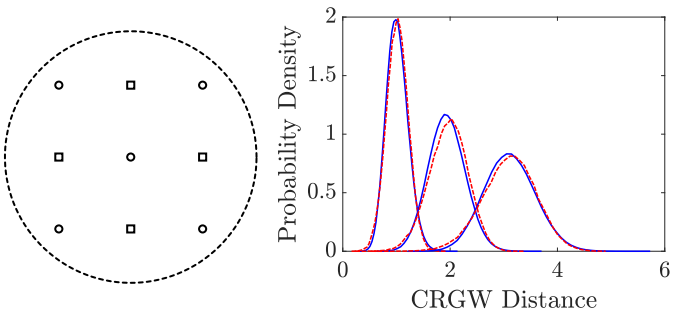

Figure 7: The local atomic environment on the left contains five circle atoms and four square atoms on a square lattice with unit spacing in a region of radius $(\sqrt{2}+2)/2$. The measured (solid blue) and predicted (dashed red) distributions of CRGW distances for $\sigma_r = 0.025$, 0.05, and 0.1 are on the right.

one in Figure 6. The plot on the right shows that the predicted probability distribution (dashed red) is a good approximation for the measured one (solid blue), even when the standard deviations of the atomic displacements are as large as one-tenth the average atomic spacing. The small offset of the mean is likely the result of three sources of error; the atomic displacements are assumed to be small relative to the atomic spacing, D_h is a sum of random variables that are not identically distributed, and the number of random variables in some of the D_h is relatively small.

6. Applications to Molecular Dynamics

This section describes the use of the CRGW distance to classify atomic environments in several molecular dynamics (MD) simulations performed in LAMMPS [27]. The initial application compares the ability of the CRGW distance to distinguish simple crystal structures (i.e., BCC, FCC, and HCP) with that of ACNA and PTM. To that end, BCC tungsten [28], FCC copper [29], and HCP magnesium [30] single crystals were simulated at temperatures up to melting in the isothermal-isobaric ensemble (NPT). The simulated systems respectively contained 4394, 8788, and 8788 atoms. The BCC and FCC unit cells were cubic, while the non-standard HCP unit cell was length a in the x-direction, $\sqrt{3}a$ in the y-direction, and $\sqrt{8/3}a$ in the z-direction. A single crystal of each material was quenched to 0 K, then heated in increments of 20 K up to melting with an equilibration of 3 ps at each temperature. The exceptions to this are that tungsten was heated in increments of 50 K and equilibrated for 5 ps ps above 4000 K, and aluminum and magnesium were heated in increments of 10 K below 300 K. The pressure was set to 0 bar throughout.

The σ_r values used to construct the predicted CRGW distance distributions were found by directly measuring atomic displacements in the MD simulations after accounting for translation, rotation, and expansion of the local environments. This is effectively a measure of the magnitude of thermal displacements, and is predicted in Section 5 to increase as $\sqrt{T}$. Figure 8 shows that σ_r follows this expectation reasonably accurately for temperatures below one-third of the melting point. Lindemann's criterion [31, 32] further suggests that melting occurs if σ_r exceeds a critical value. The melting points of the potentials were identified by discontinuities in the potential energy per atom at 5000 K for tungsten, 1200 K for aluminum, and 1200 K for magnesium, and generally occurred when $\sigma_r \approx 0.1$ in units of the average atomic spacing.

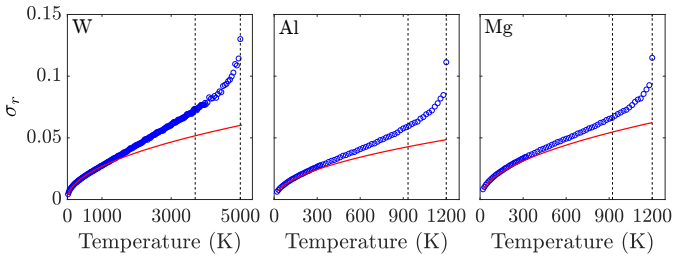

Figure 8: Blue circles show σ_r measured from simulations, and red curves show a $\sigma_r \propto \sqrt{T}$ trend line fit. For each figure, the dashed vertical line to the left indicates the true melting point of the crystal, and the dashed vertical line to the right indicates the apparent melting point for an average heating rate of 6.66×10^{12} K/s.

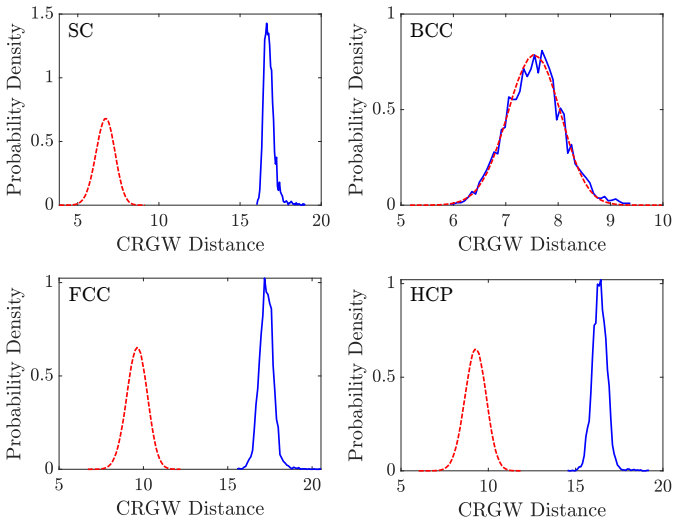

Figure 9: The CRGW distances calculated for tungsten local environments at 1000 K (solid blue lines) compared to the distributions predicted for SC, BCC, FCC, and HCP reference environments (dashed red lines). 97.2% of tungsten atoms were correctly classified as BCC at this temperature.

The classification of atomic environments in this section is based on p-values, or the probability of obtaining a CRGW distance at least as extreme as the one observed given that the local atomic environment actually derives from the specified reference environment. If the CRGW distance falls below the median of a predicted distribution, the mass of the predicted distribution to the left of that distance is the p-value. If the CRGW distance falls above the median reference environment, the mass of the predicted distribution to the right of that distance is the p-value. The maximum possible p-value is 0.5 when the CRGW distance is exactly the median value.

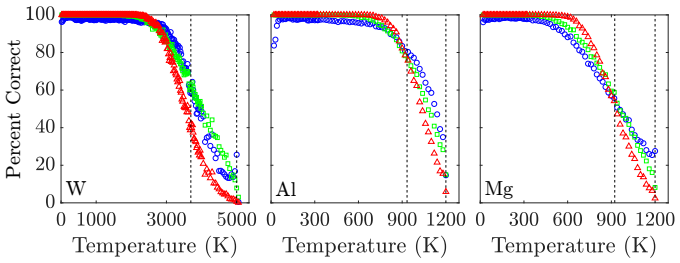

Figure 10: Percent of tungsten classified as BCC (left), percent of aluminum classified as FCC (middle), and percent of magnesium classified as HCP (right) as a function of temperature. Blue circles are for the CRGW distance with a p-value criterion of 0.01, green squares are for ACNA, and red triangles are for PTM. For each figure, the dashed vertical line to the left indicates the true melting point of the crystal, and the dashed vertical line to the right indicates the melting point of the potential.

Consider the classification of atomic environments in tungsten at 1000 K in Figure 9. The p-values for each local environment were calculated for SC, BCC, FCC, and HCP reference environments. If a local environment's p-value was greater than 0.01 for a particular reference environment and was lower for all other reference environments, then the local environment was classified accordingly. This two-part condition ensures that, e.g., the environment sufficiently resembles a perturbed BCC environment and is more likely to be a perturbed BCC environment than any other reference environment. This classification scheme is much more rigorous than those used in the past, and effectively provides the user with an uncertainty in addition to the classification.

Figure 9 more specifically plots the measured CRGW distance distributions between a local atomic environment and a given reference environment (solid blue), and the probability distributions that would be predicted if the local atomic environment really were a perturbation of that reference environment (dashed red). That the probability distributions coincide for the BCC structure indicates that the vast majority of atoms should be classified as BCC. For this particular simulation, 97.2% of atoms were correctly classified as BCC.

Figure 10 shows the percent of tungsten classified as BCC, the percent of aluminum classified as FCC, and the percent of magnesium classified as HCP as functions of temperature. The classification scheme described in this section (blue circles) correctly classifies more than 95% of the atoms up to two-thirds of the melting point for BCC and FCC. The slight dip at the lower temperatures are perhaps due to low-frequency phonons being mistaken as rotations in the measurement of σ_r, and the earlier decline for the HCP structure could be caused by the approximation of spherically-symmetric atomic displacements being less valid for noncentrosymmetric materials. Nevertheless, the method correctly classifies more than 90% of the atoms at half the melting point for HCP. This performance is comparable to that of ACNA (green squares) and PTM (red triangles) as implemented in OVITO [33], though the CRGW distance is considerably more expensive to calculate. Specifically, informal measurements suggest that ACNA, PTM, and the CRGW distance require 1 μs/atom, 5 μs/atom, and 0.25 s/atom to classify atomic environments. That is, the CRGW distance is roughly 10^5 times slower, making real-time analysis impractical.

The utility of the CRGW distance instead lies in the ability to classify more complicated atomic environments. Yu et al. proposed an interatomic potential for zirconia [34] modeled on the well-known BKS potential for silica [35]. They found the potential to be suitable for simulations of cubic and monoclinic zirconia, with the monoclinic phase being slightly preferred by 0.11 eV/ZrO_2 (the tetragonal phase spontaneously transforms to cubic). A simulation cell containing a single crystal of cubic zirconia with seven unit cells along each coordinate direction was prepared and relaxed at 0 K and 0 bar. The simulation then proceeded in the isothermal-isobaric (NPT) ensemble, with the temperature raised in intervals of 28 K every 3 ps and the pressure maintained at 0 bar. Given the lower enthalpy of the monoclinic phase, a phase transformation from cubic to monoclinic was expected. CRGW distances were calculated for zirconium- and oxygen-centered reference environments of radius 4.04 Å in the cubic and monoclinic phases relaxed at 0 K and 0 bar. While the approximation that the zirconium and oxygen atoms experience the same magnitude thermal vibrations is poor, the same classification criterion was used as above with p-values in the interval of 10^{-2} to 10^{-4} depending on the phase of interest.

The expected transformation occurred in three stages, shown in Figure 11. The cubic phase in (a-b) remained stable up to 786 K, with distributed disorder developing over a period of 2.9 ps to give (c-d). This involved [010] columns of zirconium atoms displacing along [100] directions, as revealed by (c) where only zirconium atoms are shown. The disordered structure subsequently developed three monoclinic nuclei over a period of 1.1 ps, visibly extending along the [010] direction in (e-f). The positioning of the nuclei suggests that they are not energetically independent, but interact mechanically as a consequence of the transformation strain and the periodic boundary conditions. These remained stable for several tens of picoseconds, but eventually merged and grew to give the monoclinic system in (g-h) after 62.5 ps. The transformation did not result in a single crystal though, with two distinct regions differing by a non-lattice translation in the [010] direction. These regions can be identified in (g) either by the pattern of the columns of zirconium atoms or by unclassified oxygen atoms that occur at the interfaces. The existence of these interfaces is intimately related to the use of periodic boundary conditions, and should not be construed as a general feature of the transition.

Indeed, a careful study of the cubic to monoclinic zirconia phase transition would require investigating size effects, homogeneous and heterogeneous nucleation barriers, the elastic strains in the transformed structure, and the slight differences between the relaxed monoclinic structure

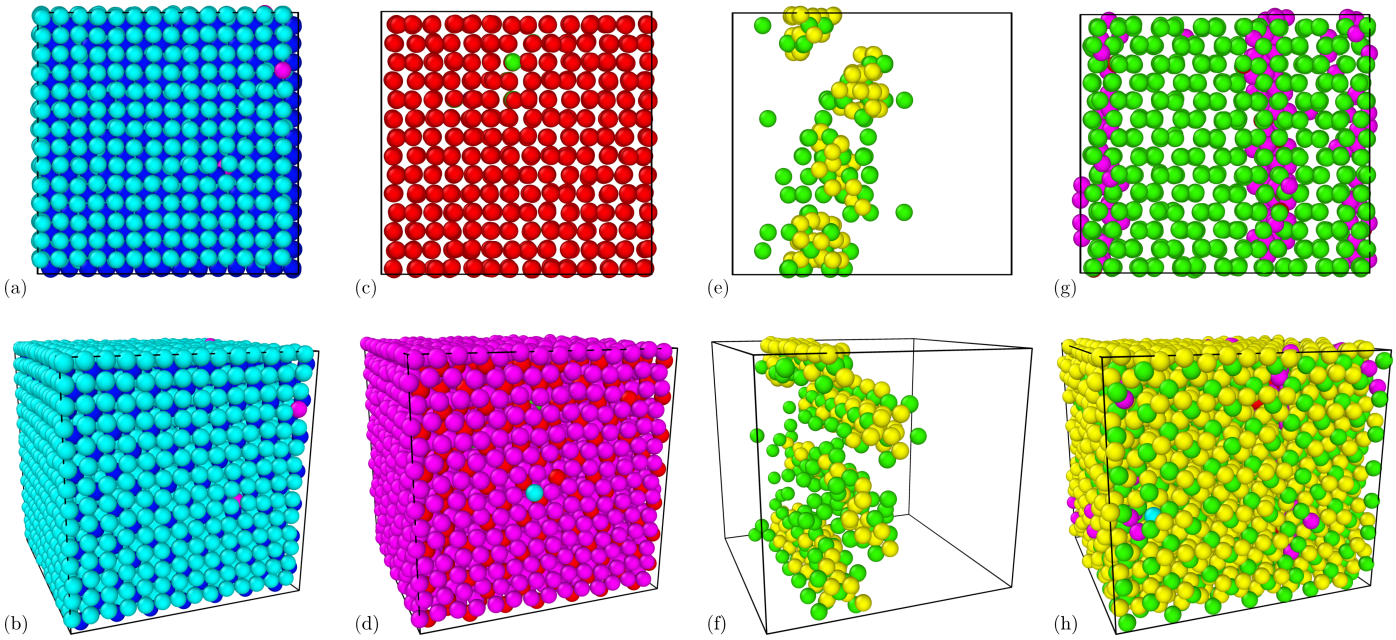

Figure 11: Molecular dynamics simulation of a phase transformation from cubic to monoclinic zirconia. Zirconium and oxygen atoms are dark blue and light blue in the cubic phase, green and yellow in the monoclinic phase, and red and purple otherwise. The [100] and [010] directions of the cubic and monoclinic phases are respectively to the left and out of the page in the top row. (a-b) Cubic zirconia at 786 K. (c-d) Intermediate structure at 812 K. Only zirconium is shown in (c) to reveal the incipient symmetry breaking. (e-f) The monoclinic phase is nucleated at 812 K, 1.1 ps after the structure in (c-d). (g-h) The transformation is completed at 1400 K, 62.5 ps after the structure in (e-f). Only zirconium in the monoclinic phase and unclassified oxygen is shown in (g) to reveal the interfacial defects.

and that published in the literature [36]. This is not undertaken here since the purpose of this study is instead to show that the CRGW distance can be used to classify atomic environments in systems at elevated temperatures with more species and more complicated crystal structures than can be handled by standard ACNA and PTM.

7. Conclusion

An automated method to classify local atomic environments via the composition-restricted Gromov–Wasserstein (CRGW) distance is proposed. Advantageous properties of this method include that it is invariant to translations, rotations, and reflections of the local atomic environment, and that it does not require the local atomic environment to be centered on an atom. The method does not make any assumption about the material class, making it applicable with minimal modification to materials with multiple chemical species and general crystal structures. Molecular dynamics results for single crystals verify that the method is a reliable approach to classifying local atomic environments in pure metals at temperatures up to half the melting point, albeit less efficiently than for techniques already available in the literature. The strength of the method is instead its applicability to general atomic systems, as is demonstrated by preliminary analysis of a cubic to monoclinic phase transition in zirconia.

Acknowledgements

J.K.M. was supported by the National Science Foundation under Grant No. DMR 2003849.

Appendix A. Heuristics

The minimization problem in Eq. 1 is difficult because of the presence of many local minima, some of them introduced by symmetries in the reference environment. Specifically, the algorithm described in Section 3 can split an atom's mass between several reference atoms related by a symmetry operation. The implementation handles this by forcefully breaking the symmetry and assigning the first such atom to precisely one other atom after each step of alternate convex search. This gradually forces the coupling matrix to be a (0, 1)-matrix, where the atoms of the reference and local structures are either matched or sent to the boundary and partial matching is disallowed. Second, the algorithm for the unbalanced GW distance often finds a local minimum by sending all atoms to the boundary. This is discouraged by beginning with artificially high values of $\boldsymbol{\lambda}^X$ and $\boldsymbol{\lambda}^Y$ in Eq. 1, and gradually relaxing them to their final values. Third, a central atom is sometimes inserted with a species label that differs from all other atoms in the environment. Forcing the center atom in the reference environment to be assigned to that in the local environment empirically helps the other atoms to be assigned consistently. While the resulting algorithm cannot guarantee a unique minimum distance coupling, the results in Sections

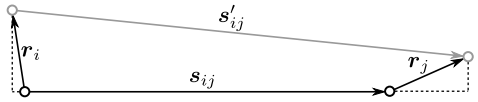

Figure B.12: If the lengths of $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ are small relative to $\boldsymbol{s}_{ij}$, then $\boldsymbol{s}_{ij}$ and $\boldsymbol{s}'_{ij}$ are nearly parallel, and the difference in the length of $\boldsymbol{s}'_{ij}$ and $\boldsymbol{s}_{ij}$ is approximately the difference of the projections of $\boldsymbol{r}_j$ and $\boldsymbol{r}_i$ onto $\boldsymbol{s}_{ij}$.

5 and 6 strongly suggest that the minimum is achieved in almost every case.

Appendix B. Mean and Covariance of the D_h

Initially consider $\langle D_h \rangle$, the mean of D_h for normally-distributed atomic displacements. From the definition in Eq. 3, the relevant equation is

$$\langle D_h \rangle = \frac{1}{2}\sum_{i,j}^{n} \langle |d_{ij}^X - d_{ij}^Y| \rangle \xi_i^h \xi_j^h + \sum_{i}^{n} \lambda_i^X (1-\xi_i^h) + \sum_{i}^{n} \langle \lambda_i^Y \rangle (1-\xi_i^h)$$

where λ_i^X is the constant distance to the boundary in the reference environment. As described in Section 5, the probability distribution $p(\boldsymbol{r}_i)$ of a displacement of the ith atom in the canonical ensemble is assumed to be

$$p(\boldsymbol{r}_i) = \left(\frac{1}{2\pi\sigma_r^2}\right)^{3/2} \exp\left(-\frac{|\boldsymbol{r}_i|^2}{2\sigma_r^2}\right)$$

with $\sigma_r^2 = k_B T/a$ indicating the variance of the atomic displacements.

Since $p(\boldsymbol{r}_i)$ is spherically symmetric and $|\lambda_i^Y - \lambda_i^X| \ll \lambda_i^X$ is assumed, the distance to the external boundary is distributed as

$$p(\lambda_i^Y) = \left(\frac{1}{2\pi\sigma_r^2}\right)^{1/2} \exp\left(-\frac{(\lambda_i^Y - \lambda_i^X)^2}{2\sigma_r^2}\right),$$

from which the mean and variance of λ_i^Y are found to be

$$\langle \lambda_i^Y \rangle = \lambda_i^X \tag{B.1}$$

$$\mathrm{var}(\lambda_i^Y) = \sigma_r^2. \tag{B.2}$$

This specifies the terms in the third sum in the equation for $\langle D_h \rangle$ above.

Now consider $|\delta_{ij}| = |d_{ij}^X - d_{ij}^Y|$ for $i \neq j$. As described in Section 3, the GW distance between the reference environment and a perturbed environment is effectively the sum of the magnitudes of the changes in the distances between all pairs of atoms. Let $\boldsymbol{s}_{ij}$ be the vector from the ith atom to the jth atom; the central quantity of interest is the change in the length of this vector with the application of perturbations. Figure B.12 suggests that if the perturbations are small relative to $\boldsymbol{s}_{ij}$, then

$$\delta_{ij} = (\boldsymbol{r}_j - \boldsymbol{r}_i) \cdot \hat{\boldsymbol{s}}_{ij}$$

is the approximate change in the length of $\boldsymbol{s}_{ij}$, where $\hat{\boldsymbol{s}}_{ij}$ is the unit vector $\boldsymbol{s}_{ij}/|\boldsymbol{s}_{ij}|$. Since $p(\boldsymbol{r}_i)$ is spherically symmetric, the probability distribution of the projected displacement $\boldsymbol{r}_i \cdot \hat{\boldsymbol{s}}_{ij}$ is the normal distribution

$$p(\boldsymbol{r}_i \cdot \hat{\boldsymbol{s}}_{ij}) = \frac{1}{\sqrt{2\pi\sigma_r^2}} \exp\left(-\frac{|\boldsymbol{r}_i \cdot \hat{\boldsymbol{s}}_{ij}|^2}{2\sigma_r^2}\right). \tag{B.3}$$

The probability distribution $p(\delta_{ij})$ can be found from Eq. B.3 by a change of variables; if $\epsilon_{ij} = (\boldsymbol{r}_j + \boldsymbol{r}_i) \cdot \hat{\boldsymbol{s}}_{ij}$ is the counterpart to δ_{ij}, and $p(\boldsymbol{r}_j \cdot \hat{\boldsymbol{s}}_{ij}) p(\boldsymbol{r}_i \cdot \hat{\boldsymbol{s}}_{ij})$ is the joint distribution of $\boldsymbol{r}_j \cdot \hat{\boldsymbol{s}}_{ij}$ and $\boldsymbol{r}_i \cdot \hat{\boldsymbol{s}}_{ij}$, then

$$\begin{aligned} p(\delta_{ij}, \epsilon_{ij}) &= \frac{1}{2} p\left(\frac{\epsilon_{ij} + \delta_{ij}}{2}\right) p\left(\frac{\epsilon_{ij} - \delta_{ij}}{2}\right) \\ &= \frac{1}{4\pi\sigma_r^2} \exp\left(-\frac{\epsilon_{ij}^2 + \delta_{ij}^2}{4\sigma_r^2}\right) \end{aligned}$$

is the joint distribution of ϵ_{ij} and δ_{ij}, where the factor of $1/2$ is the Jacobian determinant of the transformation. Integrating over ϵ_{ij} and observing that $p(\delta_{ij})$ is a symmetric function gives

$$p(|\delta_{ij}|) = \frac{1}{\sqrt{\pi\sigma_r^2}} \exp\left(-\frac{|\delta_{ij}|^2}{4\sigma_r^2}\right)$$

for the probability distribution of the magnitude of the change in the distance between the ith and jth atoms. The resulting mean and variance are

$$\langle |\delta_{ij}| \rangle = \frac{2\sigma_r}{\sqrt{\pi}} \tag{B.4}$$

$$\mathrm{var}(|\delta_{ij}|) = \left(2 - \frac{4}{\pi}\right)\sigma_r^2. \tag{B.5}$$

Using Eqs. B.1 and B.4 allows the equation for $\langle D_h \rangle$ introduced at the beginning of this section to be reduced to Eq. 4.

This only leaves the calculation of the covariance matrix. From the definition of the covariance:

$$\mathrm{cov}(D_h, D_g) = \langle D_h D_g \rangle - \langle D_h \rangle \langle D_g \rangle.$$

Expanding all the products and cancelling terms gives

$$\begin{aligned} \mathrm{cov}(D_h, D_g) = &\frac{1}{4}\sum_{i,j}^{n}\sum_{i',j'}^{n} \xi_i^h \xi_j^h \xi_{i'}^g \xi_{j'}^g \mathrm{cov}(|\delta_{ij}|, |\delta_{i'j'}|) \\ &+ \frac{1}{2}\sum_{i,j}^{n}\sum_{i'}^{n} \xi_i^h \xi_j^h (1-\xi_{i'}^g) \mathrm{cov}(|\delta_{ij}|, \lambda_{i'}^Y) \\ &+ \frac{1}{2}\sum_{i',j'}^{n}\sum_{i}^{n} \xi_i^g \xi_j^g (1-\xi_{i'}^h) \mathrm{cov}(\lambda_i^Y, |\delta_{i'j'}|) \\ &+ \frac{1}{2}\sum_{i,i'}^{n} (1-\xi_i^h)(1-\xi_{i'}^g) \mathrm{cov}(\lambda_i^Y, \lambda_{i'}^Y). \end{aligned}$$

We start with the last term. λ_i^Y and $\lambda_{i'}^Y$ for $i \neq i'$ are independent by inspection, so this reduces to $\sum_i (1-\xi_i^h)(1-$

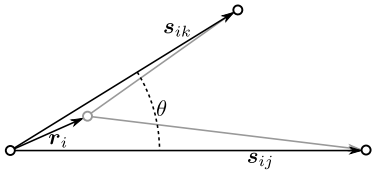

Figure B.13: The changes in the lengths of the vectors $\boldsymbol{s}_{ij}$ and $\boldsymbol{s}_{ik}$ with a displacement $\boldsymbol{r}_i$ are correlated, with the strength of the correlation depending on the angle θ.

$\xi_i^g)\sigma_r^2$ by Eq. B.2. Now consider $\mathrm{cov}(|\delta_{ij}|, |\delta_{i'j'}|)$. If $i = i'$ and $j = j'$, then this reduces to $\mathrm{var}(|\delta_{ij}|)$ as given in Eq. B.5. If all the indices are distinct, then $\mathrm{cov}(|\delta_{ij}|, |\delta_{i'j'}|)$ vanishes by inspection. The only remaining case is for $\mathrm{cov}(|\delta_{ij}|, |\delta_{ik}|)$ for $j \neq k$.

With reference to Figure B.13, a coordinate system is constructed in the plane of the page with the x-axis along $\hat{\boldsymbol{s}}_{ij}$ and the y-axis in the vertical direction. The joint distribution of the $\zeta = \boldsymbol{r}_i \cdot \hat{\boldsymbol{s}}_{ij}$ and $\eta = \boldsymbol{r}_i \cdot \hat{\boldsymbol{s}}_{ik}$ is found from that of x and y by the change of variables

$$x = \zeta \qquad y = -\cot(\theta)\zeta + \csc(\theta)\eta$$

with the Jacobian determinant $\csc(\theta)$. The resulting distribution is

$$p(\zeta, \eta) = \frac{\csc\theta}{2\pi\sigma_r^2} \exp\left[-\frac{\csc\theta^2(\zeta^2 - 2\cos\theta\zeta\eta + \eta^2)}{2\sigma_r^2}\right].$$

This is multiplied by a normal distribution of $\boldsymbol{r}_j \cdot \hat{\boldsymbol{s}}_{ij}$ like the one in Eq. B.3, a change of variables

$$\boldsymbol{r}_i \cdot \hat{\boldsymbol{s}}_{ij} = (\epsilon_{ij} - \delta_{ij})/2 \qquad \boldsymbol{r}_j \cdot \hat{\boldsymbol{s}}_{ij} = (\epsilon_{ij} + \delta_{ij})/2$$

with the Jacobian determinant $1/2$ is performed, and the dependence on ϵ_{ij} is integrated out to find the joint distribution of δ_{ij} and $\boldsymbol{r}_i \cdot \hat{\boldsymbol{s}}_{ik}$. This procedure is repeated with $\boldsymbol{r}_k \cdot \hat{\boldsymbol{s}}_{ik}$ to find the joint distribution of δ_{ij} and δ_{ik}:

$$p(\delta_{ij}, \delta_{ik}) = \frac{1}{\pi\sqrt{14 - 2\cos(2\theta)}\sigma_r^2} \exp\left\{-\frac{2[\delta_{ij}^2 + \delta_{ik}^2 - \delta_{ij}\delta_{ik}\cos(\theta)]}{[7 - \cos(2\theta)]\sigma_r^2}\right\}.$$

The joint distribution of $|\delta_{ij}|$ and $|\delta_{ik}|$ is constructed from $p(\delta_{ij}, \delta_{ik})$ by adding together the four variants with each combination of signs for δ_{ij} and δ_{ik}. Given $p(|\delta_{ij}|, |\delta_{ik}|)$, the covariance of $|\delta_{ij}|$ and $|\delta_{ik}|$ is found to be

$$\begin{aligned}\mathrm{cov}(|\delta_{ij}|, |\delta_{ik}|) &= \frac{1}{\pi}\left\{2\arctan\left[\frac{\sqrt{2}\cos(\theta)}{\sqrt{7 - \cos(2\theta)}}\right]\cos(\theta) \right. \\ &\qquad \left. + \sqrt{14 - 2\cos(2\theta)} - 4\right\}\sigma_r^2 \\ &= f(\theta)\sigma_r^2. \end{aligned} \tag{B.6}$$

The remaining terms in the equation for $\mathrm{cov}(D_h, D_g)$ are those involving $\mathrm{cov}(|\delta_{ij}|, \lambda_{i'}^Y)$. Since this vanishes by inspection for $i' \neq \{i, j\}$, only $\mathrm{cov}(|\delta_{ij}|, \lambda_i^Y)$ need be considered further. Suppose that the probability of the ith atom leaving the environment is vanishing small. Then a procedure analogous to that followed for $p(\delta_{ij}, \delta_{ik})$ gives

$$p(\delta_{ij}, \lambda_i^Y) = \frac{1}{\pi\sqrt{6 - 2\cos(2\theta)}\sigma_r^2} \exp\left\{-\frac{\delta_{ij}^2 + 2\omega_i^2 - 2\delta_{ij}\omega_i\cos(\theta)}{[\cos(2\theta) - 3]\sigma_r^2}\right\}$$

for the joint distribution of δ_{ij} and λ_i^Y, where $\omega_i = \lambda_i^Y - \lambda_i^X$. The joint distribution of $|\delta_{ij}|$ and λ_i^Y is constructed from $p(\delta_{ij}, \delta_i)$ by adding the two variants with each sign of δ_{ij}. Remarkably, the covariance of $|\delta_{ij}|$ and λ_i^Y is found to vanish.

At this point, the covariance of D_h and D_g can be given explicitly as

$$\begin{aligned}\mathrm{cov}(D_h, D_g) = &\sum_i^n (1 - \xi_i^h)(1 - \xi_i^g)\sigma_r^2 \\ &+ \frac{1}{2}\sum_{i',j'}^n \xi_i^h\xi_j^h\xi_i^g\xi_j^g \mathrm{var}(|\delta_{ij}|) \\ &+ \sum_{i,j}^n \sum_{k\neq j}^n \xi_i^h\xi_j^h\xi_i^g\xi_k^g \mathrm{cov}(|\delta_{ij}|, |\delta_{ik}|).\end{aligned}$$

where the multipliers for the second and third terms arise from the number of ways to assign the shared indices. Substituting Eqs. B.5 and B.6 for $\mathrm{var}(|\delta_{ij}|)$ and $\mathrm{cov}(|\delta_{ij}|, |\delta_{ik}|)$ then gives Eq. 5.

Data availability

The raw data required to reproduce these findings cannot be shared at this time due to technical or time limitations. The processed data required to reproduce these findings cannot be shared at this time due to technical or time limitations.

References

[1] J. D. Honeycutt, H. C. Andersen, Molecular dynamics study of melting and freezing of small Lennard-Jones clusters, Journal of Physical Chemistry 91 (1987) 4950–4963.

[2] D. Faken, H. Jónsson, Systematic analysis of local atomic structure combined with 3D computer graphics, Computational Materials Science 2 (1994) 279–286.

[3] A. Stukowski, Structure identification methods for atomistic simulations of crystalline materials, Modelling and Simulation in Materials Science and Engineering 20 (2012) 045021.

[4] A. Stukowski, Computational analysis methods in atomistic modeling of crystals, JOM 66 (2014) 399–407.

[5] W. F. Reinhart, A. W. Long, M. P. Howard, A. L. Ferguson, A. Z. Panagiotopoulos, Machine learning for autonomous crystal structure identification, Soft Matter 13 (2017) 4733–4745.

[6] C. Hsu, A. Rahman, Interaction potentials and their effect on crystal nucleation and symmetry, The Journal of Chemical Physics 71 (1979) 4974–4986.

[7] T. Schablitzki, J. Rogal, R. Drautz, Topological fingerprints for intermetallic compounds for the automated classification of atomistic simulation data, Modelling and Simulation in Materials Science and Engineering 21 (2013) 075008.
[8] E. A. Lazar, J. Han, D. J. Srolovitz, Topological framework for local structure analysis in condensed matter, Proceedings of the National Academy of Sciences 112 (2015) E5769–E5776.
[9] C. L. Kelchner, S. Plimpton, J. Hamilton, Dislocation nucleation and defect structure during surface indentation, Physical Review B 58 (1998) 11085.
[10] P. J. Steinhardt, D. R. Nelson, M. Ronchetti, Bond-orientational order in liquids and glasses, Physical Review B 28 (1983) 784.
[11] S. Winczewski, J. Dziedzic, J. Rybicki, A highly-efficient technique for evaluating bond-orientational order parameters, Computer Physics Communications 198 (2016) 128–138.
[12] W. Mickel, S. C. Kapfer, G. E. Schröder-Turk, K. Mecke, Shortcomings of the bond orientational order parameters for the analysis of disordered particulate matter, The Journal of Chemical Physics 138 (2013) 044501.
[13] G. Ackland, A. Jones, Applications of local crystal structure measures in experiment and simulation, Physical Review B 73 (2006) 054104.
[14] P. M. Larsen, S. Schmidt, J. Schiøtz, Robust structural identification via polyhedral template matching, Modelling and Simulation in Materials Science and Engineering 24 (2016) 055007.
[15] N. Lümmen, T. Kraska, Common neighbour analysis for binary atomic systems, Modelling and Simulation in Materials Science and Engineering 15 (2007) 319.
[16] F. Mémoli, On the use of Gromov-Hausdorff distances for shape comparison, in: Eurographics Symposium on Point-Based Graphics, The Eurographics Association, 2007.
[17] F. Mémoli, Gromov–Wasserstein distances and the metric approach to object matching, Foundations of Computational Mathematics 11 (2011) 417–487.
[18] F. Mémoli, Distances Between Datasets, in: Modern Approaches to Discrete Curvature, Springer, 2017, pp. 115–132.
[19] B. Schmitzer, C. Schnörr, Modelling convex shape priors and matching based on the Gromov-Wasserstein distance, Journal of Mathematical Imaging and Vision 46 (2013) 143–159.
[20] J. Solomon, G. Peyré, V. G. Kim, S. Sra, Entropic metric alignment for correspondence problems, ACM Transactions on Graphics (TOG) 35 (2016) 72.
[21] G. Peyré, M. Cuturi, J. Solomon, Gromov-Wasserstein averaging of kernel and distance matrices, in: International Conference on Machine Learning, 2016, pp. 2664–2672.
[22] A. S. Keys, C. R. Iacovella, S. C. Glotzer, Characterizing complex particle morphologies through shape matching: Descriptors, applications, and algorithms, Journal of Computational Physics 230 (2011) 6438–6463.
[23] L. Chizat, G. Peyré, B. Schmitzer, F.-X. Vialard, Scaling algorithms for unbalanced optimal transport problems, Mathematics of Computation 87 (2018) 2563–2609.
[24] P. M. Pardalos, S. A. Vavasis, Quadratic programming with one negative eigenvalue is NP-hard, Journal of Global Optimization 1 (1991) 15–22.
[25] J. Gorski, F. Pfeuffer, K. Klamroth, Biconvex sets and optimization with biconvex functions: a survey and extensions, Mathematical Methods of Operations Research 66 (2007) 373–407.
[26] B. Schmitzer, Stabilized sparse scaling algorithms for entropy regularized transport problems, arXiv:1610.06519 (2016).
[27] S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, Technical Report, Sandia National Labs., Albuquerque, NM (United States), 1993.
[28] M.-C. Marinica, L. Ventelon, M. Gilbert, L. Proville, S. Dudarev, J. Marian, G. Bencteux, F. Willaime, Interatomic potentials for modelling radiation defects and dislocations in tungsten, Journal of Physics: Condensed Matter 25 (2013) 395502.
[29] M. Pascuet, J. Fernández, Atomic interaction of the MEAM type for the study of intermetallics in the Al–U alloy, Journal of Nuclear Materials 467 (2015) 229–239.
[30] Z. Wu, M. Francis, W. Curtin, Magnesium interatomic potential for simulating plasticity and fracture phenomena, Modelling and Simulation in Materials Science and Engineering 23 (2015) 015004.
[31] C. Chakravarty, P. G. Debenedetti, F. H. Stillinger, Lindemann measures for the solid-liquid phase transition, The Journal of chemical physics 126 (2007) 204508.
[32] S. Sarkar, C. Jana, B. Bagchi, Breakdown of universal Lindemann criterion in the melting of Lennard-Jones polydisperse solids, Journal of Chemical Sciences 129 (2017) 833–840.
[33] A. Stukowski, Visualization and analysis of atomistic simulation data with OVITO-the Open Visualization Tool, Modeling and Simulation in Materials Science and Engineering 18 (2010).
[34] J. Yu, R. Devanathan, W. J. Weber, Unified interatomic potential for zircon, zirconia and silica systems, Journal of Materials Chemistry 19 (2009) 3923–3930.
[35] B. Van Beest, G. J. Kramer, R. Van Santen, Force fields for silicas and aluminophosphates based on ab initio calculations, Physical Review Letters 64 (1990) 1955.
[36] K. Whittle, G. Lumpkin, S. Ashbrook, Neutron diffraction and MAS NMR of Cesium Tungstate defect pyrochlores, Journal of Solid State Chemistry 179 (2006) 512–521.