IOP Publishing Nanotechnology
Nanotechnology 26 (2015) 344006 (15pp) doi:10.1088/0957-4484/26/34/344006

Application of data science tools to quantify and distinguish between structures and models in molecular dynamics datasets

Surya R Kalidindi[1,2], Joshua A Gomberg[2], Zachary T Trautt[3] and Chandler A Becker[4]

[1] George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[2] School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[3] Materials Measurement Science Division, National Institute of Standards and Technology, Gaithersburg, MD, USA
[4] Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD, USA

E-mail: surya.kalidindi@me.gatech.edu

Received 11 May 2015, revised 21 June 2015
Accepted for publication 2 July 2015
Published 3 August 2015

Abstract

Structure quantification is key to successful mining and extraction of core materials knowledge from both multiscale simulations as well as multiscale experiments. The main challenge stems from the need to transform the inherently high dimensional representations demanded by the rich hierarchical material structure into useful, high value, low dimensional representations. In this paper, we develop and demonstrate the merits of a data-driven approach for addressing this challenge at the atomic scale. The approach presented here is built on prior successes demonstrated for mesoscale representations of material internal structure, and involves three main steps: (i) digital representation of the material structure, (ii) extraction of a comprehensive set of structure measures using the framework of *n*-point spatial correlations, and (iii) identification of data-driven low dimensional measures using principal component analyses. These novel protocols, applied on an ensemble of structure datasets output from molecular dynamics (MD) simulations, have successfully classified the datasets based on several model input parameters such as the interatomic potential and the temperature used in the MD simulations.

Keywords: multiscale modeling, principal component analysis, molecular dynamics

(Some figures may appear in colour only in the online journal)

Introduction

Multiscale modeling [35, 43, 52, 61] has been identified as the most promising avenue for accelerating the design, development, and deployment of new/improved materials in emerging technologies [1, 2, 64, 80, 89]. A number of recently announced national research strategic initiatives (e.g., [1, 2, 89]) are being built on the premise that an increased use of multiscale materials modeling can dramatically reduce the need for extensive (and often expensive) experimentation that dominates the current materials development efforts. However, the main factors impeding the highly desired increased utilization of multiscale modeling can be collected into three groups [45]: (i) model maturity (i.e., the accuracy and reliability of available models), (ii) model interoperability (i.e., ability of the models covering multiple scales and physics to be strung together to work seamlessly), and (iii) model inversion (i.e., ability to address high value problems of interest in materials and process design that target improvements in specific performance needs). It should be noted that

tremendous progress has indeed been made in being able to numerically simulate a broad range of materials phenomena using sophisticated physics-based modeling approaches [13, 23, 35, 43, 46, 49, 51, 52, 61, 78, 81, 97, 99, 102, 105]. However, it is essential to address the main impediments described above, if we are to realize the full benefits from these modeling approaches in advanced materials development efforts.

Modern data science tools and concepts offer a promising new avenue for addressing most of the impediments described above. Data science [10, 22, 25, 38] is mainly focused on extracting high value information (might be labeled as knowledge or wisdom) from all available data (generated by either experiments or computations). This emerging cross-disciplinary field is being built on the foundations of statistical sciences, computational sciences, systems theory, and applied mathematics, and is envisioned to have a broad range of potential applications. Indeed, data science has already enjoyed many remarkable successes in disparate application domains, including recommendation systems (e.g., Amazon [57]), personal informatics (e.g., [56]), drug discovery (e.g., [39]), decision systems (e.g., [95]), and healthcare (e.g., [101]). At its core, data science is comprised of two primary components. The first component can be broadly identified as Data Management and includes robust and reliable storage, aggregation, archival, retrieval, and sharing protocols for all kinds of data (potentially generated in the broadest variety of formats possible). The second component (more pertinent to the present discussion) centers around data analytics, and is aimed at mining hidden (embedded) high value knowledge or understanding from large collections of data.

In the context of advanced materials development efforts, the central goal of data analytics is the extraction of robust and reliable process-structure-property (PSP) linkages that capture quantitatively the roles of different unit manufacturing (or processing) steps on the salient measures of the material hierarchical structure that in turn control the properties of interest (or performance characteristics desired in service). In this regard, it is extremely important to cast the desired PSP linkages in computationally efficient forms that allow direct integration into the tools typically employed by practitioners in the product design and manufacturing fields. In other words, the PSP linkages of interest are not likely to be employed in the forms developed in the advanced numerical tools [35, 43, 52, 61] or the sophisticated homogenization theories [5, 19, 30, 68, 87, 106], but more likely in the reduced-order forms (also called surrogate models or metamodels) that allow practical solutions to inverse problems of materials and process design. In recent years, a data-centered framework has emerged for capturing highly accurate PSP linkages relevant to a broad range of materials phenomena [6, 7, 20, 28, 29, 44, 47, 48, 54, 55, 92, 108]. Almost all of the applications demonstrated so far have focused on meso-length scales in the material internal structure. For example, the relationship of mesoscale porous structures on effective transport properties has been investigated [20, 21, 58, 103, 104]. In this paper, we extend this prior framework to atomic-scale molecular dynamics (MD) datasets and demonstrate its viability as a tool for improved hierarchical modeling and as a means to characterize and distinguish between datasets used in atomistic simulations. Indeed, our goal is to use the same structure quantification techniques at the atomic scale as those used previously at the mesoscale. Consequently, the approach presented here paves the way for the development of an universal approach for the rigorous quantification of the material structure at multiple hierarchical length/structure scales.

A distinctive feature of the materials data science approach presented here is its focus on a rigorous, statistical, quantification of the material structure and its usage in arriving at PSP linkages. The underlying hypothesis in such an approach is that only a sufficiently comprehensive description of the material structure can facilitate the capture of robust and reliable PSP linkages (e.g., [4, 45, 48, 63, 79]). The central challenge, therefore, lies in the quantification of the material internal structure. A complete and rigorous description of the material internal (hierarchical) structure can be very complex, demanding very high dimensional representations. This challenge is readily appreciated when one recognizes the need to include not only the details of an idealized structure in the materials of interest, but also the inherent defects (including disorder) and their spatial distribution in the structure. For example, most materials being explored for structural applications exhibit multiphase polycrystalline microstructures at the mesoscale [32, 84, 88, 91]. A rigorous description of such material structures should include quantification of the spatial distributions of the chemical composition, thermodynamic phases, crystal lattice orientations and various hierarchical defect populations (e.g., point defects, dislocations, grain boundaries, phase boundaries, pores, microcracks). Fortunately, the field of materials science and engineering has already taught us that only certain salient features of the material internal structure dominate the macroscale performance characteristics of interest for any selected application. Therefore, the main challenge in the development of materials with improved/enhanced properties reduces to identifying and tracking only the salient microstructure features that are important to a specific engineering or technology application. In general, these salient features of the material structure are not known *a priori*, and need to be identified from an extremely large list of potential measures. This is precisely where a data-driven approach offers many advantages. In a data-driven approach, the decision on exactly what constitutes the set of important salient features is not taken in a static manner—instead it is taken objectively based on the actual available data. It is continuously refined as more data becomes available.

A major goal of this work was to test whether the methods previously developed for mesoscale structure quantification could be applied to atomistic 'samples' produced by MD simulations. In particular, our goal was to explore if these methods can objectively distinguish between atomic configurations in a way that would support multiscale modeling. In this work, the results using different interatomic potentials (models of energies and forces between atoms) were considered a surrogate for different processing methods. It is

important to distinguish objectively between results generated by different models and/or under different simulation conditions. Another important factor is that, by making use of robust global characterization methods, it is possible to establish greater confidence in the multiscale use of the results from classical MD simulations.

The structure quantification approach presented in this paper, and applied rigorously to MD datasets for the first time, comprises three essential steps. In the first step, the output from the MD simulations, presented as expected positions of the atom centers, is transformed into a digital (uniformly tessellated) structure. In the second step, the digital representation of the material structure is quantified using the framework of n-point spatial correlations (or n-point statistics) [18, 32, 33, 48, 72, 96]. Although a number of other ad hoc measures of material structure are possible, only the n-point spatial correlations provide the most complete set of measures that are naturally organized by increasing amounts of structure information. For example, the most basic of the n-point statistics are the 1-point statistics, and they reflect the probability of finding a specific discrete local state of interest at any randomly selected single point (or voxel) in the material structure. In other words, they essentially capture the information on volume fractions of the various distinct local states present in the material system. The next higher level of structure information is contained in the 2-point statistics, denoted $f_r^{hh'}$, which capture the probability of finding discrete local states h and h' at the tail and head, respectively, of a prescribed vector r randomly placed into the microstructure. This idea is closely related to the commonly used concept of pair correlation functions (PCFs) [8] that reflect, for a selected or representative atom, the probability of finding atoms (generically or of a given type) as a function of radial distance. The main difference between the PCFs and the 2-point correlation functions is that the latter capture the directional dependence, i.e., the difference between the points examined is expressed as a vector and not just a simple scalar distance.

The third and final step of structure quantification involves the objective identification of reduced-order representations of the structure using techniques such as the principal component analysis (PCA) [48, 76]. PCA provides a linear transformation of high dimensional data in a new orthogonal frame where the axes are ordered according to the observed variance among the elements of the dataset. Consequently, a truncated PCA representation provides an objective (data-driven) reduced-order representation of the original data. It is emphasized here that although PCA dimensionality reduction techniques have been explored in materials problems in prior literature [23, 94], they have only recently been employed on 2-point spatial correlations of microstructure in attempts to successfully extract high fidelity PSP linkages [20, 48, 74, 76]. The main contribution of this paper is a demonstration of the application of these computational toolsets on MD datasets, and to compare and contrast the results with those obtained using the simpler structure measures used currently. Although further development of the ideas presented here is needed before they can be broadly adopted, this work demonstrates the viability and advantages of employing spatial statistics and PCA protocols on the MD datasets.

Background: MD datasets

MD has been used for a wide range of applications where atomistic mechanisms and large system sizes (relative to quantum mechanical methods) are both important. In general, modern MD simulations performed on compute clusters are capable of producing very large amounts of data for a variety of simulation conditions and configurations. As a specific example, in one of the more comprehensive studies of grain boundary (GB) motion, 15 unique trends were observed among 388 GBs simulated in nickel [77] using the synthetic driving force method [41]. These observations resulted from only exploring the effect of the five geometric degrees of freedom and temperature. The volume of data would skyrocket if one were to undertake a systematic investigation of the effect of: (i) different pure (elemental) materials, (ii) different interatomic potential for a given element [14], (iii) different driving forces, and (iv) different methods of quantifying GB properties. More generally, it is noted that the availability of high performance computing has allowed the generation of massive amounts of simulation data, but the data analysis methods have not kept pace.

Various methods have been employed in prior literature to distinguish and describe the salient features observed in the data sets produced by the MD simulations. Particularly common are the PCFs [8], as well as local order parameters. These include centrosymmetry descriptors [50] and common neighbor analysis [27, 40, 100]. Besides these order parameters, one can also use quantities such as the energy per atom or atomic volume to identify or quantify salient local features of interest (e.g., grain boundaries or distinct phases). It should be noted that all of the order parameters described above capture very local descriptions of the material structure, and can be suitably re-interpreted as selected mappings of the short-range n-point spatial correlations mentioned earlier. While these various metrics can yield very valuable information (particularly in identification of defects and defect regions), they are generally inadequate for providing a systematic hierarchical description of the material structure observed in the MD datasets.

In this paper, we demonstrate the tremendous potential of the framework of n-point statistics for a rigorous, statistically meaningful, comprehensive quantification of the material structure. This step is then followed by a clear demonstration of the potential of PCA in arriving at objective low-dimensional representations of the material structure. Towards these goals, we have selected a simple case study that involves a quantitative comparison of material structure simulated by a selected set of interatomic potentials to model the inter-atom energies and forces. Although this case study represents a fairly simple set of MD datasets (i.e., involving only homogeneous phases), it provides a clear test case for the methods

and it is hoped that it demonstrates the tremendous potential the new data-driven protocols described here hold for a systematic application to a broader set of MD datasets.

Background: spatial correlations

As noted earlier, structure quantification is central to the extraction of transferrable materials knowledge needed in multiscale materials modeling efforts. A digital signal representation of the material structure serves as a natural starting point for the ensuing discussion. In particular, it has been proposed to represent the discretized material internal structure as m_s^h [3], which denotes the probability that a specified spatial bin (or voxel) indexed by s is physically occupied by a potential local state indexed by h. Since the values of m are bounded between zero and one (in many cases it can be just binary [3]), it produces a generalized representation for a broad range of materials systems at different length/structure scales. The information on the different length scales is encoded into the properties associated with the spatial bins, while the information on the local state of the material (e.g., chemical composition, phase identifiers, order parameters, tensorial representations of different defect configurations of interest) is encoded into the properties associated with the bins in the local state space. The digital signal representation of structure offers many computational advantages in a broad range of materials data transformations and knowledge extractions [17, 28, 29, 31–33, 44, 48, 55, 72–75, 84, 85, 103].

The material structure representation described above is particularly well suited for the computations of spatial correlations (i.e., information on the relative placement of local states in the material structure) [18, 32, 33, 48, 72, 96]. Based on the earlier definitions, the 2-point spatial correlations (or 2-point statistics) can be mathematically expressed as [33, 72]

$$f_r^{hh'} = \frac{1}{S_r}\sum_{s=1}^{S_r} m_s^h m_{s+r}^{h'}, \tag{1}$$

where r indexes the bins in the space of vectors (generally the same binning scheme as that used for the spatial domain). In equation (1), S_r denotes the number of spatial bins for which the bins indexed s and $s+r$ both lie within the spatial domain of the material structure instantiation being studied. If assumptions of periodicity of the material structure are invoked (as routinely done in MD simulations), then $S_r = S$, where S is the total number of spatial bins in the microstructure instantiation. It is also pointed out that computationally efficient schemes for computing the spatial correlations using discrete Fourier transforms (DFTs) have been developed and utilized successfully [33, 72].

For most structural material systems of interest in advanced technologies, the set of n-point statistics is an extremely large unwieldy set even for $n = 2$. Rigorous analysis of these datasets is only possible with the application of data science tools. For example, it was recently demonstrated that techniques such as PCA [37, 42, 86], can be used to obtain objective low dimensional representations of the 2-point statistics [48, 76]. PCA provides a linear transformation of high dimensional data in a new orthogonal frame where the axes are ordered according to the observed variance among the elements of the dataset. Consequently, a truncated PCA representation provides an objective (data-driven) reduced-order representation of the original data. It is emphasized here that although PCA dimensionality reduction techniques have been explored in materials [23, 94] and biology [9, 16, 24, 34, 62] problems in prior literature, they have only recently been employed on 2-point spatial correlations of microstructure in attempts to successfully extract high fidelity structure-property linkages [20, 48, 74, 76].

As an example, let $\left\{ f_r \,\middle|\, r = 1,\ 2,\ \ldots,\ R \right\}$ denote the truncated set of independent 2-point statistics [72] of interest in a specific application. Let $i = 1,\ 2,\ \ldots,\ I$ enumerate the elements of an ensemble of material structures being studied. It is generally expected that $I \leqslant R$. In such situations, PCA identifies a maximum of $(I-1)$ orthogonal directions in the R-dimensional space that are arranged by decreasing levels of variance in the given ensemble of structures. Mathematically, the PCA representation of any member of the selected ensemble (of structures), labeled by superscript (k), can be expressed as

$$f_r^{(k)} = \sum_{i=1}^{\min((I-1),\,R)} \alpha_i^{(k)} \varphi_{ir} + \bar{f}_r, \tag{2}$$

where $\bar{f}_r$ is simply the averaged 2-point statistics for the entire ensemble, and $\alpha_i^{(k)}$ (referred as PC weights) provide an objective representation of the (k)th structure in the new orthogonal reference frame identified by φ_{ir} (from PCA). Another important output from the PCA is the significance of each principal component, b_i, obtained in the eigenvalue decomposition performed as a part of the PCA [37, 42, 86]. The values of b_i provide important measures of the inherent variance among the members of the ensemble of structures [76]. More importantly, by retaining only the components associated with the most significant eigenvalues, it is often possible to obtain an objective reduced-order representation of the structure with only a handful of parameters. Mathematically, this reduced-order representation can be expressed as

$$f_r^{(k)} \approx \sum_{i=1}^{R^*} \alpha_i^{(k)} \varphi_{ir} + \bar{f}_r, \tag{3}$$

where $R^* \ll \min((I-1), R)$. Selection of R^* will depend on the specific properties that need to be correlated to the structure metrics. Note also that the concepts described above can be easily extended to include higher-order statistics of the structure (e.g., 3-point spatial correlations). The PCA representations of the n-point statistics have been successfully used in automated and efficient classification of various ensembles of structures [48, 74].

In most prior examples presented to date in literature, the local state was defined at the continuum scale and identified as a specific thermodynamic phase found in the micrograph.

However, the same methodology can be applied to material structures at other length scales. In a recent paper, this approach was successfully applied to quantify the semi-crystalline polymer structure datasets produced by MD simulations [26].

Extension of spatial correlations to MD datasets

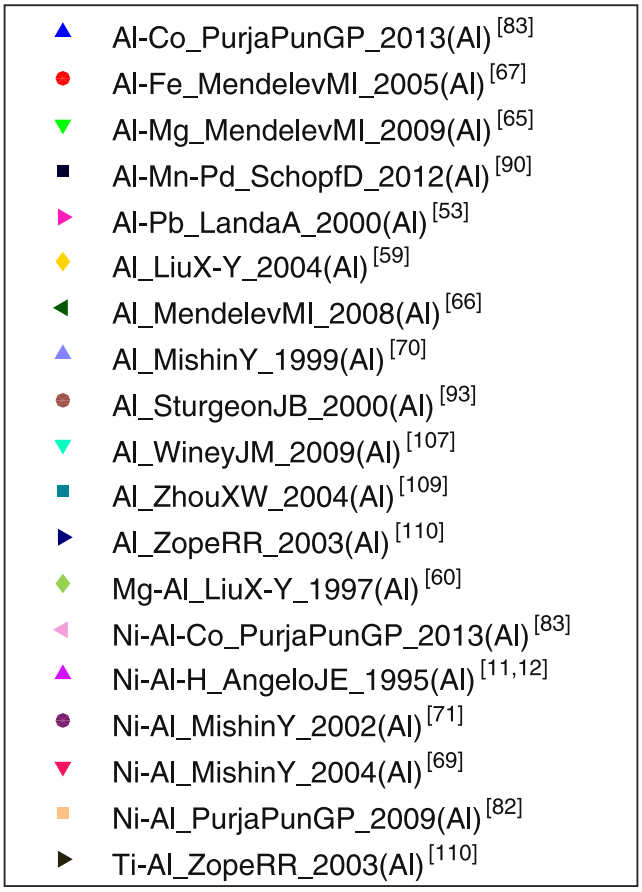
One challenge of applying 2-point statistics to atomistic configuration datasets is the subjective choice of how to transform the discrete set of atomic points into a regular three-dimensional (3D) grid of voxels. This choice is likely to be driven by the nature of the application. For example, in simulations encompassing a relatively large number of atoms, it may be preferable for a single voxel to encompass multiple atoms and the local state in each voxel is defined by measures such as the density or the mean orientation of the enclosed atoms (e.g., [26]). Alternatively, it may be preferable to quantify structural variations at the atomic scale, in which case the voxel size should be selected to be smaller than the atomic radius; we will focus our discussion here to these cases.

As a proxy for more complex atomic structures, we here consider MD simulations of atomic volumes with a single chemical species as a function of temperature. These simulations represent relatively simple MD calculations that are being used as part of the NIST Interatomic Potentials Repository project to help establish a set of reference calculations to help researchers select interatomic potentials (models of how the atoms interact, also called force fields) that are most appropriate for a given application [15]. Except for choice of interatomic potential, the methodology is kept fixed for every simulation, which is: (i) determine the 0 K equilibrium fcc lattice constant via a molecular statics simulation, (ii) create a $10 \times 10 \times 10$ face centered cubic (fcc) unit cell (4000 atoms) using the equilibrium lattice constant, (iii) create a uniform distribution of atomic velocities at the desired simulation temperature, and (iv) perform an isothermal–isobaric (NPT) simulation at the desired temperature for 2 000 000 time steps using a 1 fs time step. Data analysis described here takes place within the final 1 000 000 time steps. Instantaneous coordinates were recorded every 50 000 fs, and these were used in the analysis presented here. Average reported pressures, volumes, temperatures, energies, etc, reach steady state well within that equilibration time for all simulations. The long simulations were done (instead of shorter ones that may have been adequate), primarily for two reasons. The first was to minimize the chance of a particular trajectory not being in equilibrium while running the same duration for all simulations (to make comparisons more robust). The second was to allow more time for first-order phase transitions to occur to thermodynamically favorable states. While this is not an issue for low homologous temperatures (T/T_M), it is more significant near the melting temperature of the interatomic potential where phase transitions (melting) were observed for several of the interatomic potentials. Melting is identified by local structural disorder and a significant increase in atomic volume. The python scripts used to generate the simulations and the data itself are available on the NIST Interatomic Potentials Repository site (http://ctcms.nist.gov/potentials). While calculations have been performed for a number of different interatomic potentials defining elemental interactions for Al, Ni, Cu, Ag, and Au, here we are focusing on just the Al results. The interatomic potentials included in this study are summarized in table 1, along with the appropriate references [8, 11, 12, 53, 59, 60, 65–67, 69–71, 82, 83, 90, 93, 107, 109, 110].

It is important to note that these calculations include some simulations well outside the intended usage of the interatomic potentials (e.g., using the pure elements of a potential only fit for use with compounds and thus they may not give the most accurate values for single-element atomic volumes). However, users often use interatomic potentials well outside the range of where they were fit, and it is important to understand how that choice affects the answers obtained. This is discussed in much more detail in [14, 15, 98]. In this work, several interatomic potentials have melting temperatures for pure aluminum that are significantly lower than the experimental value of 933 K, which will be discussed in more detail later.

Figure 1(a) shows a MD simulation dataset typical of those included in this study. In this dataset, the center positions of the atoms were taken directly from the results of the MD simulations (as instantaneous coordinates) and a sphere of radius $a = 1.18$ Å was constructed around the center to denote the atom. The entire volumetric domain used in the simulation was then discretized into a uniform grid and the material structure was converted to a simple digital signal, denoted as m_s^h (as introduced earlier). In this notation, the local state descriptor, h, was allowed only two values: $h = 1$ was used to refer to the atomic species and $h = 0$ was used to refer to the empty space between the atomic species. As mentioned earlier, s serves as an index for the spatial bin. For 3D space, it is convenient to think of s as an integer array, i.e. $s = \{s_1, s_2, s_3\}$, with each s_i taking only integer values. The level of discretization employed is typically a variable parameter in the data-driven explorations. In the present study, based on a few trials we established a spatial bin size of approximately $0.252\ \text{Å} = 0.214a$ since further refinement did not influence the computed spatial correlations in any significant manner. The value assigned to m_s^h denotes the volume fraction of local state h found in the spatial bin s. Although, in principle, the value of m_s^h can range between zero and one, we have only allowed this variable to take either the value zero or one in this study; such structures have been referred as eigen structures in prior literature [3]. More specifically, if the distance between the center of a given voxel and the center of the voxel containing the coordinates of the atom center is less than or equal to the radius, that voxel is assigned a value of one (i.e., the voxel is included in the atom). For eigen microstructures, f_0^{11} would actually be the volume fraction occupied by the atomic species in the total volume being studied. Furthermore, since there are only two local states in the datasets considered here, only one

Table 1. List of Al force fields used and their corresponding notation and references.



autocorrelation is enough to capture all of the non-redundant 2-point spatial correlations [4, 31, 36, 72, 96]. In this paper, we will therefore only focus on f_r^{11}, and simply refer to these as f_r.

Next, we discuss the computation of f_r from m_s^1. A specific challenge encountered arises from the fact that the overall simulation volume in the MD results is not kept constant. In other words, results from different potentials or even different snapshots from a single potential are expected to result in different simulation volumes. Since we have fixed the spatial bin size (described above), this would lead to fractional voxels at the edges of the volume. Furthermore, since the MD simulations conducted for this study have employed periodic boundary conditions, we wish to rigorously account for these boundary conditions in computing the spatial correlations. The strategy devised and employed in this study, to address the considerations described above, consisted of the following steps: (i) the microstructure signal, m_s^1, is expanded by employing the same periodic boundary assumptions that were utilized in the MD simulations. As an example, this expansion is shown in figure 1(b) for a representative 2D section through the simulation volume in figure 1(a). For this example, the domain volume size is increased from $L_o = 40.5$ Å to $L_e = 73.08$ Å (in each of the three-dimensions). Note that this expansion serves two purposes: (a) while the initial volume size (output from the MD simulation) is unlikely to be an exact integer multiplier of the selected spatial bin size, the size of expanded region is selected to ensure that it is indeed an exact multiplier of the spatial bin size (this feature is essential to allow the use of DFT algorithms). (b) The increase in size is needed to allow the placement of all vectors of interest in computing the spatial correlations (the tails of the vectors of interest will lie within the original volume, but the heads of these vectors may lie in the expanded volume). For all the MD volumes included in the study, the expansion size was selected to include all vectors up to a size of 59 spatial bins (this number was selected after a few trials and noting that vectors larger than this do not carry any additional salient information in the computed 2-point statistics for the volumes studies here); the corresponding number of statistics will be 119^3 (59 positive, 59 negative, and the zero vector components in each of the three-dimensions). Discretization using finer grids was seen to have a negligible effect on the clustering (i.e., classification) of interest for the present study (visualized later as dendrograms; cf figure 4). It is important to note that the discretization level is an important parameter of the protocols described here, and has to be adjusted suitably for different studies. (ii) A second microstructure signal $\tilde{m}_s^1$ of the same extent as m_s^1 is created by copying the values of m_s^1 for all of the spatial bins corresponding to atoms whose centers fit inside the original volume (of size L_o) and assigning zeros for

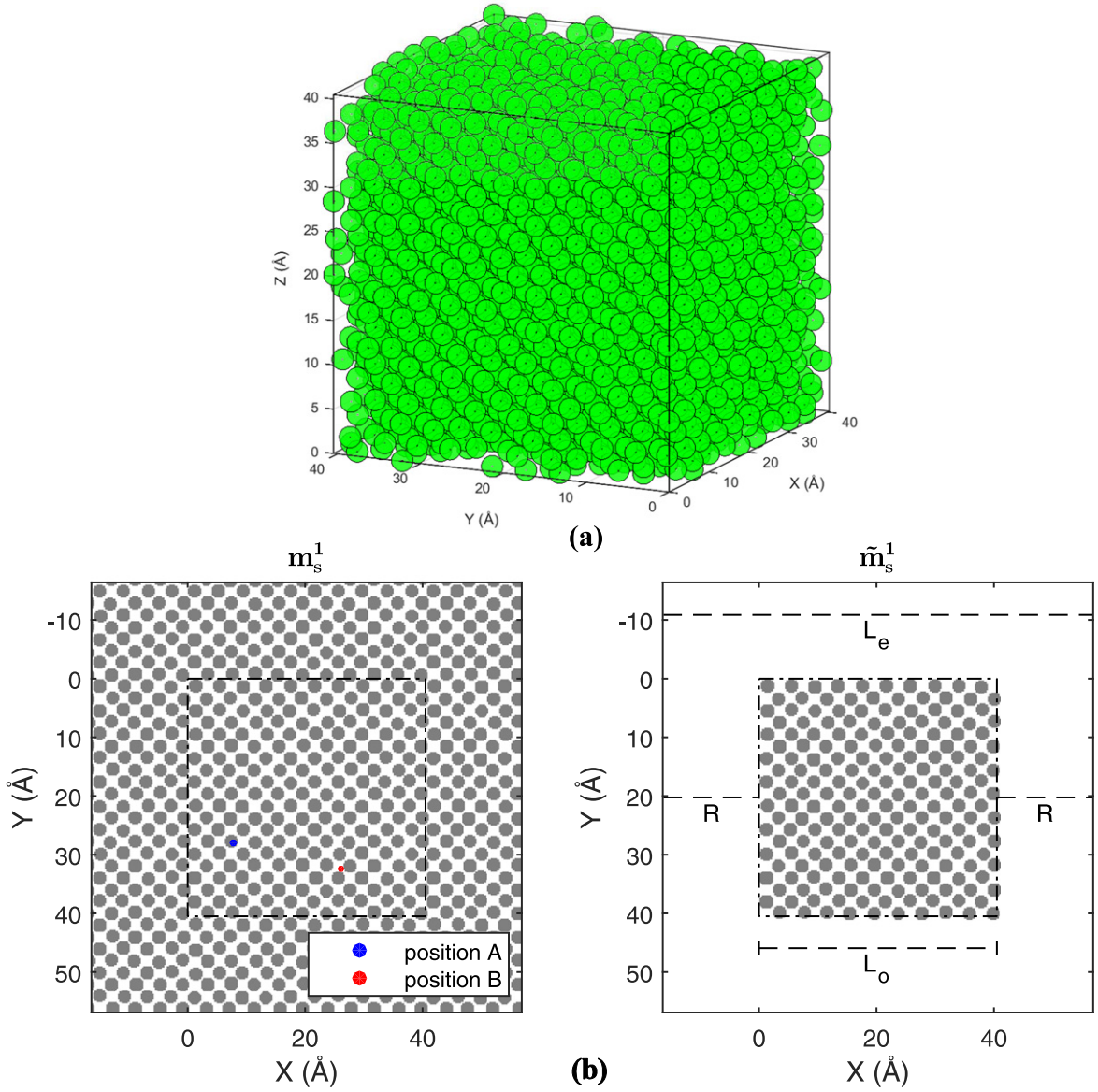


Figure 1. (a) Coordinates of a 4000 atom Al equilibrium simulation at 300 K at 10 ps using the force field 'Al–Pb_LandaA_2000'. Dots represent atomic centers as generated by the simulation. For the purpose of 2-point statistics each atom was assigned a radius of 1.18 A, as depicted by the green circles. Though not clear in this figure, the structure is crystalline (face centered cubic) as expected. (b) Cross section corresponding to $Z = 20.24$ Å of the corresponding discretized microstructure signals constructed in the novel protocols described in this paper. The full 3D discretized images are used to calculate the 2-point statistics.

the rest of the spatial bins (also shown in figure 1(b)). The number of spatial bins copied from the original volume is denoted as S_r. (iii) The 2-point spatial correlations of interest are computed as the convolution of the two microstructure signals, m_s^1 and $\tilde{m}_s^1$ (i.e., using these instead of m_s^h and $m_s^{h'}$ in equation (1)), truncated to include only vectors whose 3D components are smaller than R.

Figures 2(a)–(c) present selected 2D sections of the 3D contour plots of 2-point spatial correlations (these are visualized as the contours of the values of f_r in the 3D vector space of r, with $r = (0, 0, 0)$ at the center of the plot). The sections shown in this figure depict, as expected, a roughly periodic pattern consistent with the crystalline structure reflected in the spatial positioning of the atoms in the actual volumetric domain analyzed by the MD simulations (shown in figure 1(a)). It is important to recognize that the f_r values plotted in figures 2(a)–(c) are actually statistics denoting the probability of finding two voxels separated by the vector r and occupied by the atomic species. As a reference, the reader might take note that in a perfectly disordered (i.e., random) spatial distribution of local states (not shown), the 2-point spatial correlations show a single spike at the center (for $r = (0, 0, 0)$) and then immediately asymptote to a uniform value as one moves away from the center. The reader should also note that the value of $f_{(0,0,0)}$ at the center of these plots corresponds to the atomic volume fraction.

Figure 2(d) presents the more familiar PCF used extensively in literature for quantifying the material structure in the

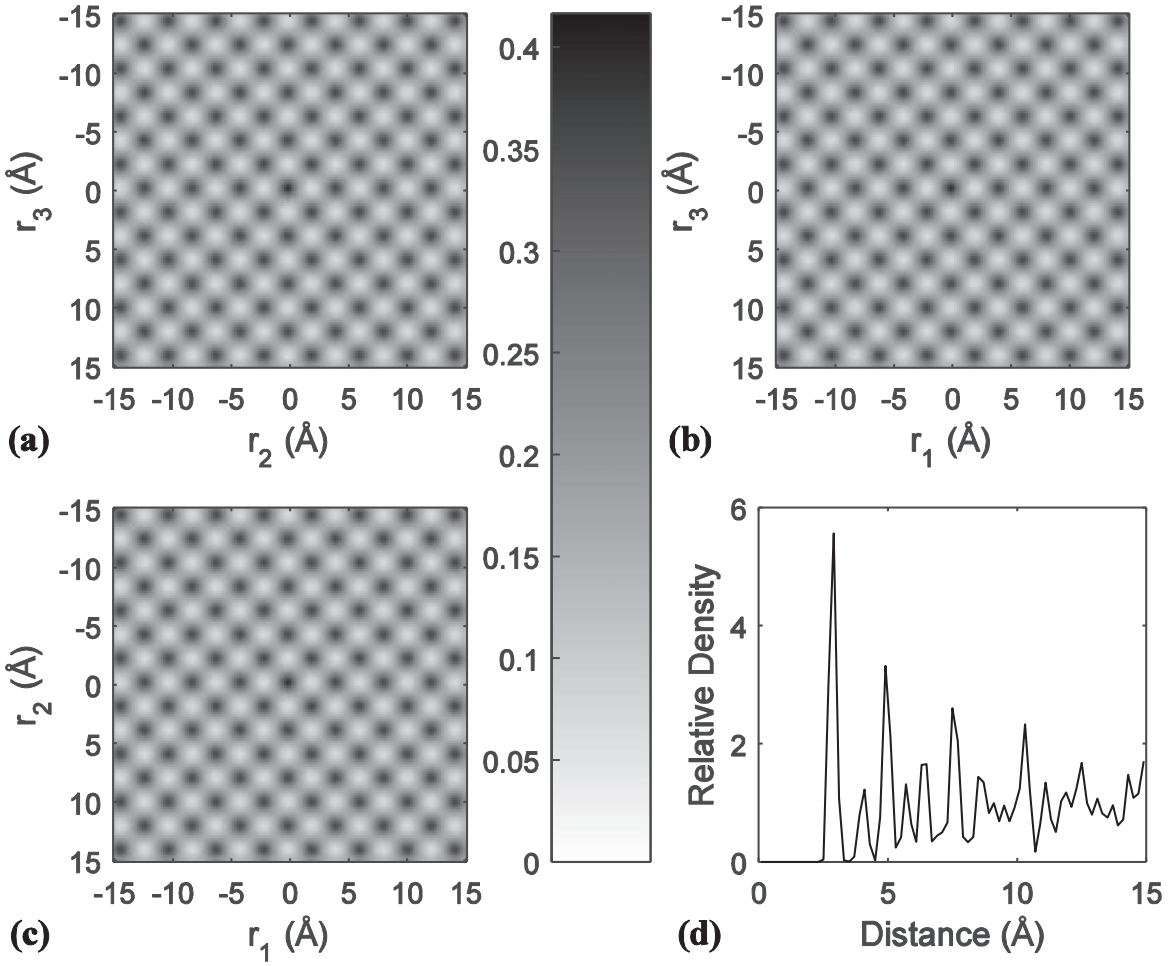


Figure 2. (a)–(c) The cross sections corresponding to $r_1 = 0$, $r_2 = 0$, and $r_3 = 0$, respectively, of the 2-point statistics of the dataset presented in figure 1. (d) The pair correlation function of the same structure dataset, only considering atom centers.

MD simulation results. As one might infer, the peaks in the PCF plot correspond to suitably integrated (and normalized) values of the 3D 2-point spatial correlations over the orientation variables defining the vector r. In other words, PCF is expressed only as a function of the magnitude of r, while the 2-point spatial correlations retain explicitly the dependence on both magnitude and direction of r.

Application of spatial correlations to MD datasets

Figure 3 presents a classification of the MD simulation datasets in the PCA space (following the protocols described earlier) for the MD simulated atomic structures at 300 and 900 K, respectively, using the 19 potentials selected for this study. For each potential, the study included twenty atomic structures (taken at different times in the simulation after reaching an equilibrium state). Therefore, a total of 380 atomic structures were included in this analysis at each simulation temperature. Each data point in figures 3(a) and (b) represents the first three PC scores (or weights) for each MD simulated atomic structure included in the analyses. The computation of course provides many more dimensions (or PC scores), but it also indicated that these three PC scores account for 99.8% of the important differences in the entire ensemble of atomic structures included in the study. This massive degree of dimensionality reduction is fully consistent with the prior experience involving mesoscale systems. In this regard, it is also satisfying to note that the range of the PC scores is systematically decreasing for the higher-ranked PC scores (for example, the range for PC1 was about −90 to about 20, whereas the range for PC2 was about −15 to about 15), further confirming that the higher-ranked PC scores are indeed less important in capturing the salient features of the structures included in the ensemble.

Keeping in mind that the PCA representation in figure 3 denotes a dimensionality reduction from $119^3 = 1\,685\,159$ to just three, it is indeed remarkable that this representation effectively captures both the intra-class and the inter-class variations within the entire ensemble. This result is even more remarkable when one notes that this classification was performed in a completely unsupervised manner. In other words, the PCA computation was not informed in any way about the different potentials used in the MD simulations in producing the atomic structures included in the study. This is a clear testament to the power of the 2-point spatial correlations and

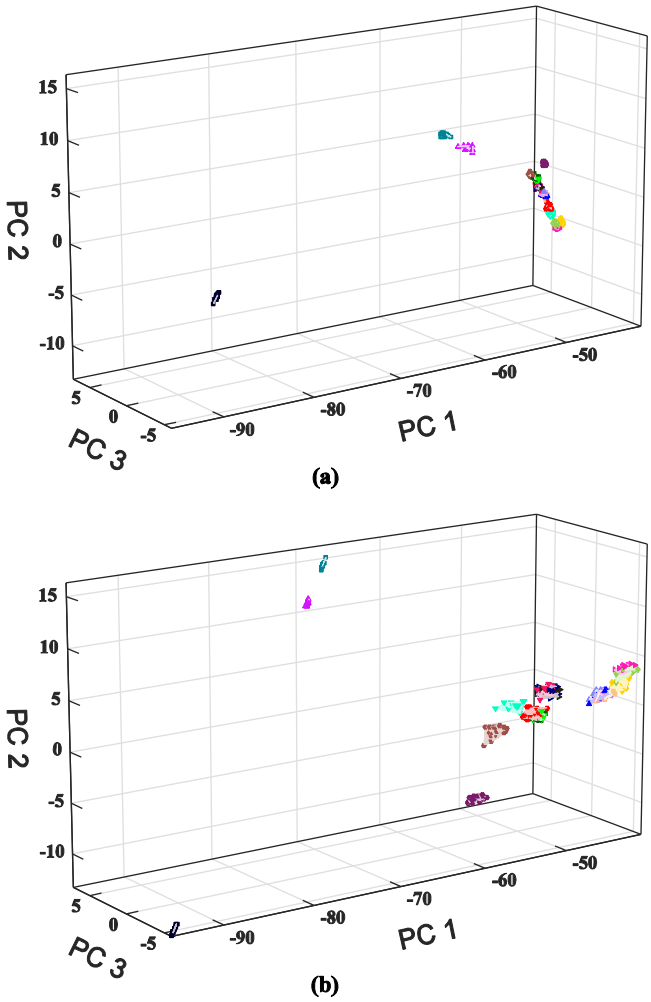


Figure 3. The 2-point statistics every 50 ps from 1.05 to 2.0 ns of Al simulations using the force fields in table 1 projected onto the first three principal components at 300 K (a) and and 900 K (b).

principal component analyses in capturing the salient features of the material structure in a rigorous stochastic framework. It is also very satisfying to note that the intra-class variance (reflected in the size of the cluster associated with each potential) in the simulated structures is significantly smaller than the inter-class variance. Moreover, the intra-class variance seems to be of roughly the same order of magnitude for all the different potentials included in this study, and is slightly higher for the datasets produced at the higher simulation temperature. All of these observations are consistent with expectations, and provide strong support to our claim that the protocols used in this study produce high value, low dimensional, measures of the material structure.

An effective tool for visualizing distances in high-dimensional spaces is a dendrogram, which depicts the hierarchy of the distances between the data points. Figures 4(a) and (b) depict the inter-class distances (between the cluster-means) as dendrograms for the same dataset that was depicted in figure 3. Broadly, the PCA has identified the following clustering of potentials based on the differences in the structures produced by the MD simulations: the first group corresponds to the force fields referenced in [53, 59, 60, 67, 107], the second group corresponds to the force fields referenced in [8, 65, 66, 69, 70, 82, 83, 93, 110], including both force fields referenced in each of [83] and [110]. The four force fields referenced in [71, 90, 109] and [11, 12] are distinctly far away from the two groups identified above. The groupings of these results will be discussed in more detail in a later section. Here we reiterate that interatomic potentials are fit with different types of reference data and optimized for particular applications. Potentials fit for particular compounds, e.g., the B2 phase in Ni–Al, may not be the best option for treating the full Ni–Al phase diagram, though they may be the best available for the intended application.

Additional insights from the analysis presented here can be obtained from the plots of the PCs obtained in the analysis described above. Plots of $\bar{f}_r$ and φ_{ir} (for different values of i;

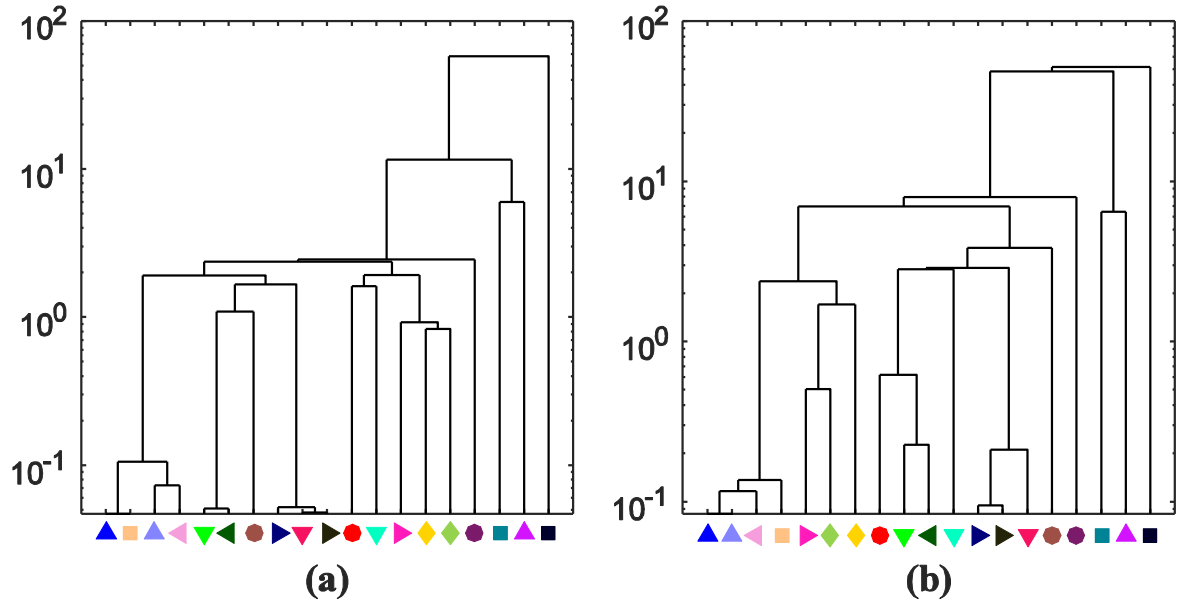


Figure 4. The dendrograms of centroid distances of the data depicted in figure 3 at 300 K (a) and 900 K (b).

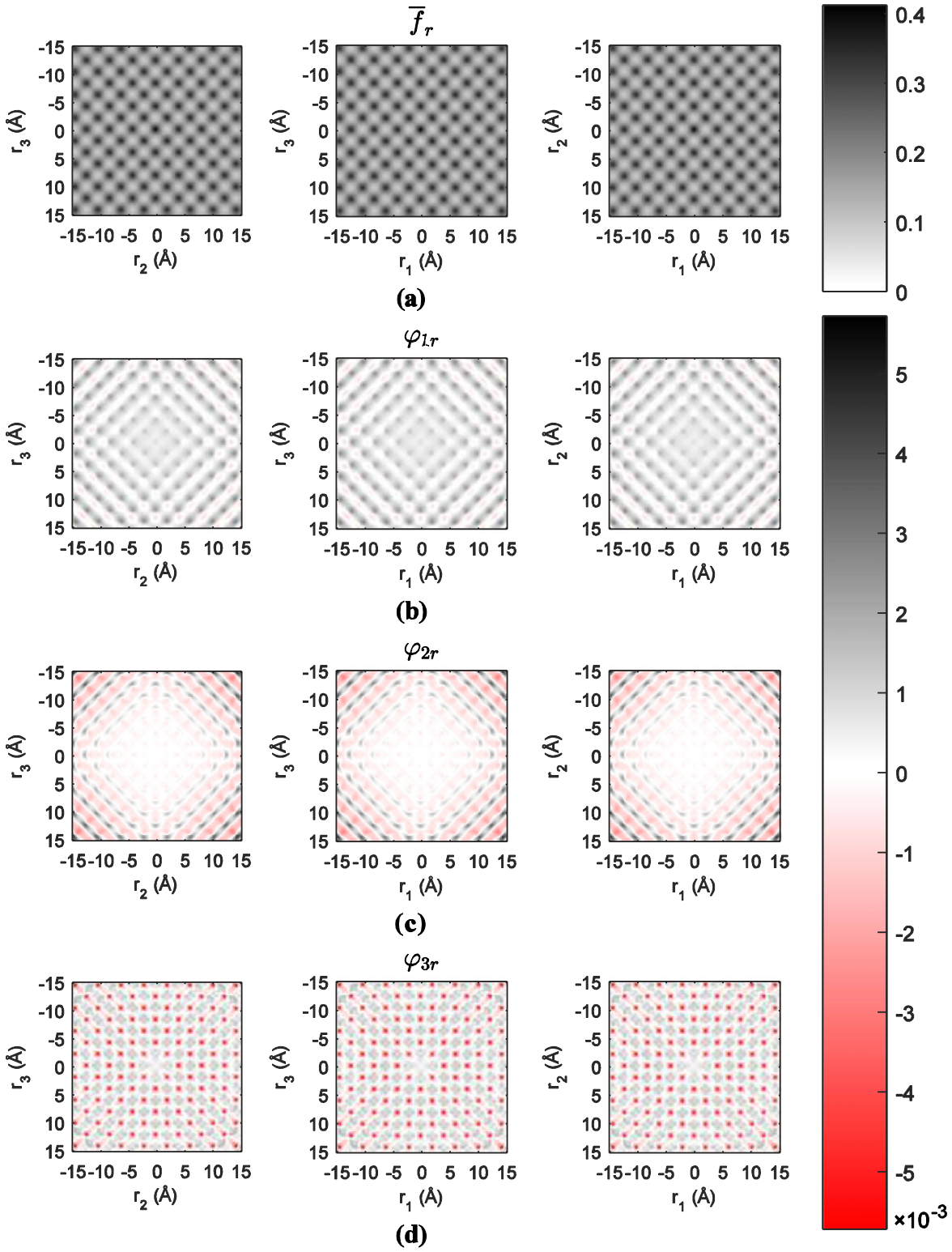


Figure 5. Contour plots of the ensemble averaged spatial correlations and the PCA basis (eigen vectors) for the datasets shown in figure 3(a), each shown as three orthogonal cross-sections.

see equation (3) are presented in figures 5(a)–(d). As with the plots shown in figures 2(a)–(c), r indexes the discretization of the vector space used in defining the 2-point spatial correlations. The plots of $\bar{f}_r$ (figure 5(a)) simply reflect the averaged auto-correlations for the entire ensemble of atomic structures included in the study. As expected, the averaged auto-correlation reflects an arrangement of the atoms on a highly periodic lattice. One can judge the degree of periodicity by comparing intensities of the different peaks in these plots with the intensity of the center peak. For a perfectly periodic

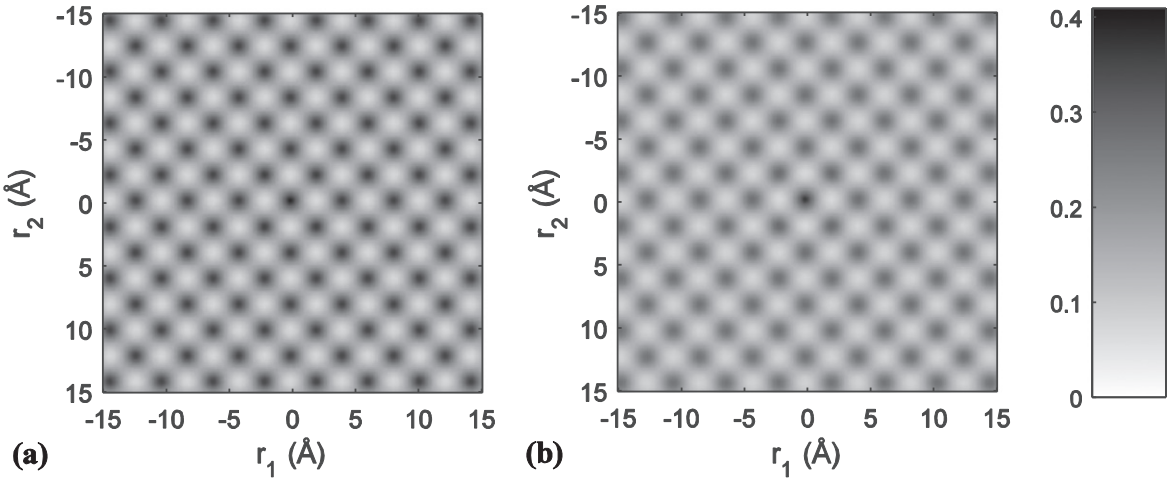


Figure 6. $r_3 = 0$ cross sections of the 2-point statistics of the force field 'Al_SturgeonJB_2000(Al)' at (a) 300 K and (b) 900 K.

arrangement, the peak intensity will be the same for all peaks in the entire plot. As the arrangement becomes less periodic, the peak intensities drop systematically as one moves away from the center peak. As mentioned earlier, for a random arrangement, this drop in the peak intensity will be rather abrupt. In the present study, we will see a more significant drop in the peak intensities for the atomic structures simulated at higher temperatures (described later) compared to the ones simulated at lower temperature.

The plots of φ_{ir} in figures 5(b)–(d) reflect a prioritized set of orthogonal deviations from the averaged autocorrelation. In other words, φ_{1r} reflects the most dominant deviation, φ_{2r} is the next most dominant deviation, and so on. Note the difference in signs between the red and black peaks in these plots. Consequently, a combination of closely placed pair of red and black spots on these plots reflects shift of the peak from its position in the ensemble average. The overall plot of φ_{1r} therefore captures systematic shifts in the interatomic distances between any selected atom and its neighbors, with the shifts being higher for far away neighbors compared to those that are nearby. Therefore, φ_{1r} appears to capture well the overall volume differences among the snapshots of the atomic structure. In the most general case, each of the φ_{ir} captures a certain scaled deviation in the intensities of all of the statistics included in the PCA analyses. Because of the large number of the statistics included in the PCA (each structure is characterized by 1 685 159 2-point statistics), it is often very difficult to assign a simple interpretation for what detail of the structure is captured by each individual φ_{ir}. It should also be noted from figure 5 that the structure detail captured by the different φ_{ir} exhibit different levels and types of directional dependence.

As implied in equation (3), one can construct the autocorrelation of any specific atomic structure included in the study by starting with the averaged autocorrelation and adding weighted contributions from each of the principal components. These weights are precisely the weights depicted in the low dimensional PCA representations of figures 3(a) and (b). It should be noted that such a reconstruction typically involves a truncation error when the higher-order principal components are ignored. However, since PCA provides a prioritized list of principal components, one can make the decision on an appropriate truncation level for a specific study in a very objective manner.

Figures 6(a) and (b) compare the 2-point statistics for the atomic structures predicted by one force field at two temperatures, respectively: 300 and 900 K. As mentioned earlier, one of the salient differences in these plots is in the rate of decay of the peak intensities as one moves from the center peak, indicating the existence of a higher level of disorder (thermal noise) in the atomic structure at the higher temperature. It should be noted that this is a statistically rigorous evaluation of the difference in the atomic structures at the two temperatures. There is also a difference in the lattice parameter at the two temperatures, which can be easily inferred by looking closely at the positions of the peaks (with respect to the center) in the plots presented in figure 6.

It is also instructive to examine the variation of the PC scores as a function of temperature for the different force fields. This is shown in figures 7(a) and (b) after performing a PCA on all of the averaged 2-point statistics for each force field at each simulation temperature. Of particular interest are the four force fields corresponding to [71, 90, 109] and [11, 12], which show significantly different behavior compared to the rest of the data sets. Indeed, as shown in figure 8, this difference in the predicted results from these four force fields is also evident in the plots of the averaged atomic volume. The force field used in [71] was strongly weighted to reproduce the properties of B2-NiAl, which may explain its poor behavior for pure aluminum. The other three interatomic potentials ([90, 109] and [11, 12]) were found to melt in the course of the simulations. Further investigation is needed to determine the cause of the low temperature melting phenomenon predicted by these force fields. If one looks at the volumes in figure 8 at 300 K, there are several bands of volumes. Close examination of figures 3(a) and 8 reveals that the groupings of average atomic volumes, determined from overall simulation size fluctuations, map directly to the

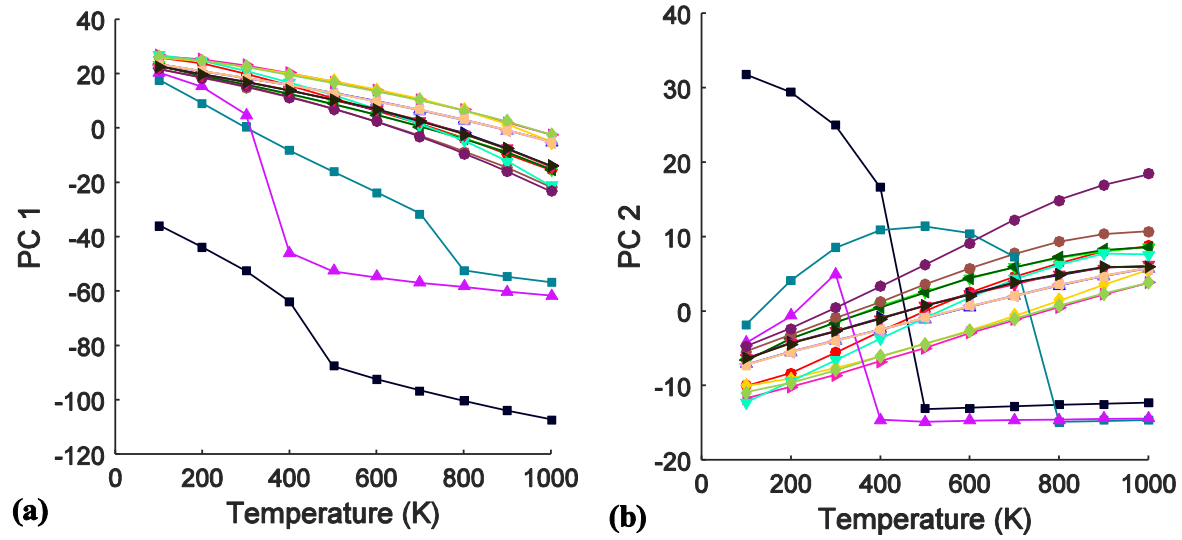


Figure 7. The variation of (a) first principal component and (b) second principal component for the averaged 2-point statistics at each temperature. Only the mean 2-point statistics at each temperature for each force field were included in this PCA.

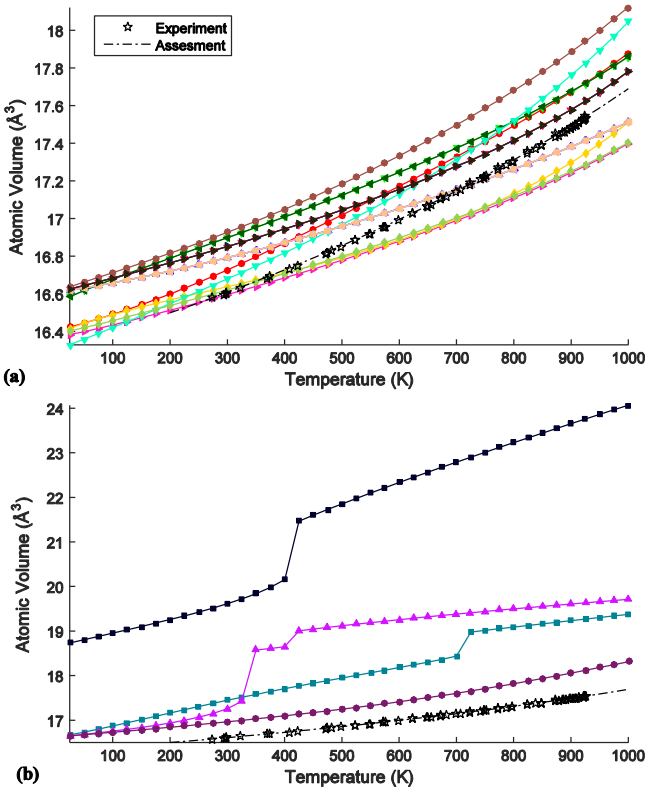


Figure 8. Average atomic volumes from MD simulations of the (a) interatomic potentials closest to the experimental reference data, and (b) the four interatomic potentials exhibiting the largest deviation from the reference values. The discontinuities reflect phase changes associated with melting.

groupings determined from the *n*-point statistics and PCA analysis. Similar clustering is evident at 900 K, where there is a greater spread in average volumes for the simulations conducted with the different interatomic potentials. The fact that the PC scores automatically capture this effect, without *a priori* information about the phases, bodes well for their utility in capturing high values structure-property linkages. While a simple measure such as the atomic volume would also capture a similar effect, there is no guarantee that it captures all of the significant differences seen in the predicted MD structures. The protocols presented here ensure that all of the salient differences in the ensemble of predicted structures are indeed captured to a high degree of completeness (note that the two PCs referenced in figure 7 capture 96.3% of the differences in the elements of the ensemble).

Conclusions

This initial study demonstrates the utility and the viability of utilizing rigorous structure quantification protocols to results predicted by MD. Of particular significance is the fact that similar protocols were previously applied successfully to material structure datasets at the mesoscale. This study reinforces the possibility that a consistent set of structure quantification tools can be designed and applied to a broad range of materials systems at vastly different length/structure scales, and paves the way forward for the formulation and validation of such a universal framework. Furthermore, since the framework employs data-driven approaches, it leads to rigorous, practically useful, low dimensional, representations and visualizations. These are central to our goals for creating high value materials knowledge systems.

Acknowledgments

SRK and JAG acknowledge support from NIST 70NANB14H191. ZTT and CAB acknowledge support from the NIST Materials Genome Initiative program. No approval or endorsement of any commercial product by NIST is

intended or implied. Certain commercial software systems are identified in this paper to facilitate understanding. Such identification does not imply that these software systems are necessarily the best available for the purpose.

References

[1] National Science and Technology Council 2011 Materials Genome Initiative for Global Competitiveness www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf

[2] National Science and Technology Council 2012 A National Strategic Plan for Advanced Manufacturing www.whitehouse.gov/sites/default/files/microsites/ostp/iam_advancedmanufacturing_strategicplan_2012.pdf

[3] Adams B L, Gao X and Kalidindi S R 2005 Finite approximations to the second-order properties closure in single phase polycrystals *Acta Mater.* **53** 3563–77

[4] Adams B L, Kalidindi S R and Fullwood D 2012 *Microstructure Sensitive Design for Performance Optimization* (Waltham, MA: Butterworth-Heinemann)

[5] Adams B L and Olson T 1998 Mesostructure—properties linkage in polycrystals *Prog. Mater. Sci.* **43** 1–88

[6] Al-Harbi H F and Kalidindi S R 2015 Crystal plasticity finite element simulations using a database of discrete Fourier transforms *Int. J. Plast.* **66** 71–84

[7] Al-Harbi H F, Landi G and Kalidindi S R 2012 Multi-scale modeling of the elastic response of a structural component made from a composite material using the materials knowledge system *Modelling Simul. Mater. Sci. Eng.* **20** 055001

[8] Allen M P and Tildesley D J 1987 *Computer Simulation of Liquids* (Oxford, UK: Clarendon)

[9] Amadei A, Linssen A B M and Berendsen H J C 1993 Essential dynamics of proteins *Proteins: Struct. Funct. Bioinformatics* **17** 412–25

[10] Anderson C 2008 The end of theory: the data deluge makes the scientific method obsolete Wired Magazine updated 6/23/2008), available at: (www. wired. com/science/discoveries/magazine/16-07/pb_theory)

[11] Angelo J E, Moody N R and Baskes M I 1995 Trapping of hydrogen to lattice defects in nickel *Modelling Simul. Mater. Sci. Eng.* **3** 289

[12] Baskes M I, Sha X, Angelo J E and Moody N R 1997 Trapping of hydrogen to lattice defects in nickel *Modelling Simul. Mater. Sci. Eng.* **5** 651

[13] Becker C A, Ågren J, Baricco M, Chen Q, Decterov S A, Kattner U R, Perepezko J H, Pottlacher G R and Selleby M 2014 Thermodynamic modelling of liquids: CALPHAD approaches and contributions from statistical physics *Phys. Status Solidi* (b) **251** 33–52

[14] Becker C A, Tavazza F and Levine L E 2011 Implications of the choice of interatomic potential on calculated planar faults and surface properties in nickel *Phil. Mag.* **91** 3578–97

[15] Becker C A, Tavazza F, Trautt Z T and Buarque de Macedo R A 2013 Considerations for choosing and using force fields and interatomic potentials in materials science and engineering *Curr. Opin. Solid State Mater. Sci.* **17** 277–83

[16] Berendsen H J C and Hayward S 2000 Collective protein dynamics in relation to function *Curr. Opin. Struct. Biol.* **10** 165–9

[17] Bochenek B and Pyrz R 2004 Reconstruction of random microstructures: a stochastic optimization problem *Comput. Mater. Sci.* **31** 93–11

[18] Brown W F 1955 Solid mixture permittivities *J. Chem. Phys.* **23** 1514–7

[19] Castañeda P P, Telega J J and Gambin B 2004 *Nonlinear Homogenization and its Applications to Composites, Polycrystals and Smart Materials* (Berlin: Springer)

[20] CeCen A, Fast T, Kumbur E C and Kalidindi S R 2014 A data-driven approach to establishing microstructure-property relationships in porous transport layers of polymer electrolyte fuel cells *J. Power Sources* **245** 144–53

[21] Çeçen A, Wargo E A, Hanna A C, Turner D M, Kalidindi S R and Kumbur E C 2012 3D microstructure analysis of fuel cell materials: spatial distributions of tortuosity, void size and diffusivity *J. Electrochem. Soc.* **159** B299–307

[22] Cleveland W S 2001 Data science: an action plan for expanding the technical areas of the field of statistics *Int. Stat. Rev.* **69** 21–6

[23] Curtarolo S, Morgan D, Persson K, Rodgers J and Ceder G 2003 Predicting crystal structures with data mining of quantum calculations *Phys. Rev. Lett.* **91** 135503

[24] David C C and Jacobs D J 2014 *Protein Dynamics* ed D R Livesay (New York: Humana Press) pp 193–226

[25] Dhar V 2013 Data science and prediction *Commun. ACM* **56** 64–73

[26] Dong X, McDowell D L, Kalidindi S R and Jacob K I 2014 Dependence of mechanical properties on crystal orientation of semi-crystalline polyethylene structures *Polymer* **55** 4248–57

[27] Faken D and Jonnson H 1994 Systematic analysis of local atomic structure combined with 3D computer graphics *Comput. Mater. Sci.* **2** 279–86

[28] Fast T and Kalidindi S R 2011 Formulation and calibration of higher-order elastic localization relationships using the MKS approach *Acta Mater.* **59** 4595–605

[29] Fast T, Niezgoda S R and Kalidindi S R 2011 A new framework for computationally efficient structure-structure evolution linkages to facilitate high-fidelity scale bridging in multi-scale materials models *Acta Mater.* **59** 699–707

[30] Fullwood D T, Adams B L and Kalidindi S R 2008 A strong contrast homogenization formulation for multi-phase anisotropic materials *J. Mech. Phys. Solids* **56** 2287–97

[31] Fullwood D T, Kalidindi S R, Niezgoda S R, Fast A and Hampson N 2008 Gradient-based microstructure reconstructions from distributions using fast Fourier transforms *Mater. Sci. Eng.* A **494** 68–72

[32] Fullwood D T, Niezgoda S R, Adams B L and Kalidindi S R 2010 Microstructure sensitive design for performance optimization *Prog. Mater. Sci.* **55** 477–562

[33] Fullwood D T, Niezgoda S R and Kalidindi S R 2008 Microstructure reconstructions from 2-point statistics using phase-recovery algorithms *Acta Mater.* **56** 942–8

[34] García A E 1992 Large-amplitude nonlinear motions in proteins *Phys. Rev. Lett.* **68** 2696–9

[35] Ghosh S, Lee K and Moorthy S 1995 Multiple scale analysis of heterogeneous elastic structures using homogenization theory and voronoi cell finite element method *Int. J. Solids Struct.* **32** 27–62

[36] Gokhale A M, Tewari A and Garmestani H 2005 Constraints on microstructural two-point correlation functions *Scr. Mater.* **53** 989–93

[37] Halko N, Martinsson P-G, Shkolnisky Y and Tygert M 2011 An algorithm for the principal component analysis of large data sets *SIAM J. Sci. Comput.* **33** 2580–94

[38] Hey T, Tansley S and Tolle K 2009 *The Fourth Paradigm Data-Intensive Scientific Discovery* (Redmond, WA: Microsoft Research)

[39] Hohman M, Gregory K, Chibale K, Smith P J, Ekins S and Bunin B 2009 Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery *Drug Discovery Today* **14** 261–70

[40] Honeycutt J D and Andersen H C 1987 Molecular-dynamics study of melting and freezing of small Lennard-Jones clusters *J. Phys. Chem.-Us* **91** 4950–63
[41] Janssens K G F, Olmsted D, Holm E A, Foiles S M, Plimpton S J and Derlet P M 2006 Computing the mobility of grain boundaries *Nat. Mater.* **5** 124–7
[42] Jolliffe I 2005 *Principal Component Analysis* (New York: Wiley)
[43] Kadowaki H and Liu W K 2004 Bridging multi-scale method for localization problems *Comput. Methods Appl. Mech. Eng.* **193** 3267–302
[44] Kalidindi S R 2012 Computationally-efficient fully-coupled multi-scale modeling of materials phenomena using calibrated localization linkages *ISRN Mater. Sci.* 305692
[45] Kalidindi S R 2015 Data science and cyberinfrastructure: critical enablers for accelerated development of hierarchical materials *Int. Mater. Rev.* **60** 150–68
[46] Kalidindi S R, Bhattacharya A and Doherty R 2004 Detailed analysis of plastic deformation in columnar polycrystalline aluminum using orientation image mapping and crystal plasticity models *Proc. R. Soc.* A **460** 1935–56
[47] Kalidindi S R, Niezgoda S R, Landi G, Vachhani S and Fast T 2010 A novel framework for building materials knowledge systems *Comput. Mater. Continua* **17** 103–25
[48] Kalidindi S R, Niezgoda S R and Salem A A 2011 Microstructure informatics using higher-order statistics and efficient data-mining protocols *JOM* **63** 34–41
[49] Karma A, Trautt Z T and Mishin Y 2012 Relationship between equilibrium fluctuations and shear-coupled motion of grain boundaries *Phys. Rev. Lett.* **109** 095501
[50] Kelchner C L, Plimpton S J and Hamilton J C 1998 Dislocation nucleation and defect structure during surface indentation *Phys. Rev.* B **58** 11085–8
[51] Kirane K, Ghosh S, Groeber M and Bhattacharjee A 2009 Grain level Dwell Fatigue Crack nucleation model for Ti alloys using crystal plasticity finite element analysis *Trans. ASME, J. Eng. Mater. Technol.* **131** 021003
[52] Kouznetsova V G, Geers M G D and Brekelmans W A M 2004 Multi-scale second-order computational homogenization of multi-phase materials: a nested finite element solution strategy *Comput. Methods Appl. Mech. Eng.* **193** 5525–50
[53] Landa A, Wynblatt P, Siegel D J, Adams J B, Mryasov O N and Liu X Y 2000 Development of glue-type potentials for the Al–Pb system: phase diagram calculation *Acta Mater.* **48** 1753–61
[54] Landi G and Kalidindi S R 2010 Thermo-elastic localization relationships for multi-phase composites *CMC-Comput. Mater. Continua* **16** 273–93
[55] Landi G, Niezgoda S R and Kalidindi S R 2010 Multi-scale modeling of elastic response of three-dimensional voxel-based microstructure datasets using novel DFT-based knowledge systems *Acta Mater.* **58** 2716–25
[56] Li I, Dey A and Forlizzi J 2010 *A Stage-Based Model of Personal Informatics Systems* (New York: ACM) pp 557–66
[57] Linden G, Smith B and York J 2003 Amazon. com recommendations: item-to-item collaborative filtering *IEEE Internet Comput.* **7** 76–80
[58] Litster S, Epting W K, Wargo E A, Kalidindi S R and Kumbur E C 2013 Morphological analyses of polymer electrolyte fuel cell electrodes with nano-scale computed tomography imaging *Fuel Cells* **13** 935–45
[59] Liu X-Y, Ercolessi F and Adams J B 2004 Aluminium interatomic potential from density functional theory calculations with improved stacking fault energy *Modelling Simul. Mater. Sci. Eng.* **12** 665
[60] Liu X-Y, Ohotnicky P P, Adams J B, Rohrer C L and Hyland R W Jr 1997 Anisotropic surface segregation in Al–Mg alloys *Surf. Sci.* **373** 357–70
[61] Luscher D J, McDowell D L and Bronkhorst C A 2010 A second gradient theoretical framework for hierarchical multiscale modeling of materials *Int. J. Plast.* **26** 1248–75
[62] Maisuradze G G, Liwo A and Scheraga H A 2009 Principal component analysis for protein folding dynamics *J. Mol. Biol.* **385** 312–29
[63] McDowell D L, Panchal J H, Choi H-J, Seepersad C C, Allen J K and Mistree F 2009 *Integrated Design of Multiscale, Multifunctional Materials and Products* (Amsterdam: Elsevier)
[64] McDowell D L and Story T L 1998 New directions in materials design science and engineering (MDS&E), *Report of a NSF DMR-Sponsored Workshop* (19–21 October)
[65] Mendelev M I, Asta M, Rahman M J and Hoyt J J 2009 Development of interatomic potentials appropriate for simulation of solid–liquid interface properties in Al–Mg alloys *Phil. Mag.* **89** 3269–85
[66] Mendelev M I, Kramer M J, Becker C A and Asta M 2008 Analysis of semi-empirical interatomic potentials appropriate for simulation of crystalline and liquid Al and Cu *Phil. Mag.* **88** 1723–50
[67] Mendelev M I, Srolovitz D J, Ackland G J and Han S 2005 Effect of Fe Segregation on the migration of a non-symmetric Σ5 Tilt grain boundary in Al *J. Mater. Res.* **20** 208–18
[68] Milton G W 2002 *The Theory of Composites* (Cambridge: Cambridge University Press)
[69] Mishin Y 2004 Atomistic modeling of the γ and γ'-phases of the Ni–Al system *Acta Mater.* **52** 1451–67
[70] Mishin Y, Farkas D, Mehl M J and Papaconstantopoulos D A 1999 Interatomic potentials for monoatomic metals from experimental data and *ab initio* calculations *Phys. Rev.* B **59** 3393–407
[71] Mishin Y, Mehl M J and Papaconstantopoulos D A 2002 Embedded-atom potential for B2-NiAl *Phys. Rev.* B **65** 224114
[72] Niezgoda S R, Fullwood D T and Kalidindi S R 2008 Delineation of the space of 2-point correlations in a composite material system *Acta Mater.* **56** 5285–92
[73] Niezgoda S R and Kalidindi S R 2009 Applications of the phase-coded generalized hough transform to feature detection, analysis, and segmentation of digital microstructures *CMC-Comput. Mater. Continua* **14** 79–97
[74] Niezgoda S R, Kanjarla A K and Kalidindi S R 2013 Novel microstructure quantification framework for databasing, visualization, and analysis of microstructure data *Integrating Mater. Manuf. Innov.* **2**
[75] Niezgoda S R, Turner D M, Fullwood D T and Kalidindi S R 2010 Optimized structure based representative volume element sets reflecting the ensemble-averaged 2-point statistics *Acta Mater.* **58** 4432–45
[76] Niezgoda S R, Yabansu Y C and Kalidindi S R 2011 Understanding and visualizing microstructure and microstructure variance as a stochastic process *Acta Mater.* **59** 6387–400
[77] Olmsted D L, Holm E A and Foiles S M 2009 Survey of computed grain boundary properties in face-centered cubic metals-II: grain boundary mobility *Acta Mater.* **57** 3704–13
[78] Palumbo M *et al* 2014 Thermodynamic modelling of crystalline unary phases *Phys. Status Solidi* B **251** 14–32
[79] Panchal J H, Kalidindi S R and McDowell D L 2013 Key computational modeling issues in integrated computational materials engineering *Comput. Aided Des.* **45** 4–25
[80] Pollock T M *et al* 2008 *Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security* (Washington DC: National Academies)
[81] Przybyla C P and McDowell D L 2011 Simulated microstructure-sensitive extreme value probabilities for high cycle fatigue of duplex Ti-6Al-4V *Int. J. Plast.* **27** 1871–95

[82] Purja P G P and Mishin Y 2009 Development of an interatomic potential for the Ni–Al system *Phil. Mag.* **89** 3245–67
[83] Purja P G P, Yamakov V and Mishin Y 2015 Interatomic potential for the ternary Ni–Al–Co system and application to atomistic modeling of the B2–L1 0 martensitic transformation *Modelling Simul. Mater. Sci. Eng.* **23** 065006
[84] Qidwai S M, Turner D M, Niezgoda S R, Lewis A C, Geltmacher A B, Rowenhorst D J and Kalidindi S R 2012 Estimating response of polycrystalline materials using sets of weighted statistical volume elements (WSVEs) *Acta Mater.* **60** 5284–99
[85] Roberts A P 1997 Statistical reconstruction of three-dimensional porous media from two-dimensional images *Phys. Rev.* E **56** 3203
[86] Rokhlin V, Szlam A and Tygert M 2009 A randomized algorithm for principal component analysis *SIAM J. Matrix Anal. Appl.* **31** 1100–24
[87] Roters F, Eisenlohr P, Hantcherli L, Tjahjanto D D, Bieler T R and Raabe D 2010 Overview of constitutive laws, kinematics, homogenization and multiscale methods in crystal plasticity finite-element modeling: theory, experiments, applications *Acta Mater.* **58** 1152–211
[88] Rowenhorst D J, Lewis A C and Spanos G 2010 Three-dimensional analysis of grain topology and interface curvature in a b-titanium alloy *Acta Mater.* **58** 5511–9
[89] Schmitz G J and Prahl U 2014 ICMEg–the integrated computational materials engineering expert group–a new European coordination action *Integr. Mater. Manuf. Innov.* **3** 2
[90] Schopf D, Brommer P, Frigan B and Trebin H-R 2012 Embedded atom method potentials for Al–Pd–Mn phases *Phys. Rev.* B **85** 054201
[91] Schwartz A J, Kumar M and Adams B L 2000 *Electron Backscatter Diffraction in Materials Science* (New York: Kluwer)
[92] Shaffer J B, Knezevic M and Kalidindi S R 2010 Building texture evolution networks for deformation processing of polycrystalline fcc metals using spectral approaches: applications to process design for targeted performance *Int. J. Plast.* **26** 1183–94
[93] Sturgeon J B and Laird B B 2000 Adjusting the melting point of a model system via Gibbs-Duhem integration: application to a model of aluminum *Phys. Rev.* B **62** 14720–7
[94] Suh C, Rajagopalan A, Li X and Rajan K 2002 The application of principal component analysis to materials science data *Data Sci. J.* **1** 19–26
[95] Tien J M 2003 Toward a decision informatics paradigm: a real-time, information-based approach to decision making *IEEE Trans. Syst. Man Cybern.* C **33** 102–13
[96] Torquato S 2002 *Random Heterogeneous Materials* (New York: Springer)
[97] Trautt Z T and Mishin Y 2014 Capillary-driven grain boundary motion and grain rotation in a tricrystal: a molecular dynamics study *Acta Mater.* **65** 19–31
[98] Trautt Z T, Tavazza F and Becker C A Facilitating the selection and creation of accurate interatomic potentials with robust tools and characterization (http://hdl.handle.net/11256/121)
[99] Trautt Z T, Upmanyu M and Karma A 2006 Interface mobility from interface random walk *Science* **314** 632–5
[100] Tsuzuki H, Branicio P S and Rino J P 2007 Structural characterization of deformed crystals by analysis of common atomic neighborhood *Comput. Phys. Commun.* **177** 518–23
[101] Wan T T 2006 Healthcare informatics research: from data to evidence-based management *J. Med. Syst.* **30** 3–7
[102] Wang B L, Wen Y H, Simmons J and Wang Y Z 2008 Systematic approach to microstructure design of Ni-base alloys using classical nucleation and growth relations coupled with phase field modeling *Metall. Mater. Trans.* A **39A** 984–93
[103] Wargo E A, Hanna A C, Cecen A, Kalidindi S R and Kumbur E C 2012 Selection of representative volume elements for pore-scale analysis of transport in fuel cell materials *J. Power Sources* **197** 168–79
[104] Wargo E A, Schulz V P, Çeçen A, Kalidindi S R and Kumbur E C 2013 Resolving macro- and micro-porous layer interaction in polymer electrolyte fuel cells using focused ion beam and x-ray computed tomography *Electrochim. Acta* **87** 201–12
[105] Wen Y H, Simmons J P, Shen C, Woodward C and Wang Y 2003 Phase-field modeling of bimodal particle size distributions during continuous cooling *Acta Mater.* **51** 1123–32
[106] Willis J R 1981 Variational and related methods for the overall porperties of composite materials *Adv. Appl. Mech.* **21** 2–78
[107] Winey J M, Alison K and Gupta Y M 2009 A thermodynamic approach to determine accurate potentials for molecular dynamics simulations: thermoelastic response of aluminum *Modelling Simul. Mater. Sci. Eng.* **17** 055004
[108] Yabansu Y C, Patel D K and Kalidindi S R 2014 Calibrated localization relationships for elastic response of polycrystalline aggregates *Acta Mater.* **81** 151–60
[109] Zhou X W, Johnson R A and Wadley H N G 2004 Misfit-energy-increasing dislocations in vapor-deposited CoFe/NiFe multilayers *Phys. Rev.* B **69** 144113
[110] Zope R R and Mishin Y 2003 Interatomic potentials for atomistic simulations of the Ti–Al system *Phys. Rev.* B **68** 024102