npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Machine learning enabled autonomous microstructural characterization in 3D samples

Henry Chan[1*], Mathew Cherukara[1], Troy D. Loeffler[1], Badri Narayanan[1,2] and Subramanian K. R. S. Sankaranarayanan[1,3*]

We introduce an unsupervised machine learning (ML) based technique for the identification and characterization of microstructures in three-dimensional (3D) samples obtained from molecular dynamics simulations, particle tracking data, or experiments. Our technique combines topology classification, image processing, and clustering algorithms, and can handle a wide range of microstructure types including grains in polycrystalline materials, voids in porous systems, and structures from self/directed assembly in soft-matter complex solutions. Our technique does not require a priori microstructure description of the target system and is insensitive to disorder such as extended defects in polycrystals arising from line and plane defects. We demonstrate quantitively that our technique provides unbiased microstructural information such as precise quantification of grains and their size distributions in 3D polycrystalline samples, characterizes features such as voids and porosity in 3D polymeric samples and micellar size distribution in 3D complex fluids. To demonstrate the efficacy of our ML approach, we benchmark it against a diverse set of synthetic data samples representing nanocrystalline metals, polymers and complex fluids as well as experimentally published characterization data. Our technique is computationally efficient and provides a way to quickly identify, track, and quantify complex microstructural features that impact the observed material behavior.

npj Computational Materials (2020)6:1 ; https://doi.org/10.1038/s41524-019-0267-z

INTRODUCTION

Characterization of microstructural and nanoscale features in full 3D samples of materials is emerging to be a key challenge across a range of different technological applications. These microstructural features can range from grain size distribution in metals, voids and porosity in soft materials such as polymers to hierarchical structures and their distributions during self- and directed-assembly processes. It is well known that there is a strong correlation between microstructural/nanoscale features in materials and their observed properties. For the most part, however, grain size characterization is performed on 2D samples and the information from 2D slices is collated to derive the 3D microstructural information, which is inefficient and leads to potential loss of information. As such, a direct 3D classification approach for arbitrary polycrystalline microstructure is crucial and highly desirable, especially given the advancement in 3D characterization techniques such as tomography,[1] high energy diffraction microscopy (HEDM),[2] and coherent diffraction X-ray imaging.

Most industry relevant structural materials are polycrystalline in nature, and often contain thousands or millions of grains. Within each grain, the lattice arrangement of atoms is nearly identical, but the atomic orientations are different for each adjoining grain. Grain boundaries are interfaces where two grains or crystallites having different orientations meet without a disruption in the continuity of the material. Note that the thermodynamic equilibrium state of these polycrystalline materials is single crystal.[3] It is, however, well known that materials are often arrested or trapped in local minima, i.e., in the polycrystalline state. Grain formation in polycrystalline films during their growth and processing is a complex process and is highly sensitive to several parameters such as temperature, deposition rate, dopant concentration, pressure, and impurity concentration to name a few. Nuclei when formed are nanoscopic – critical sizes start from tens of atoms – and lead to nanocrystalline solids that subsequently consolidate into larger grains. These ubiquitous phenomena, from "rare events" such as nucleation to the subsequent phase transformation in crystalline solids, lie at the heart of a spectrum of physico-chemical processes that govern nanoscale material transformation. They have been a fundamental problem in materials science and are also relevant to a broad range of energy applications.

Average grain size and grain distribution are critical microstructural features that impact several physical, mechanical, optical, chemical, and thermal properties to name a few, and represent fundamental quantities to characterize polycrystalline materials.[4–9] For example, the Hall–Petch relationship[10,11] states that the final average grain size after the transformation is directly related to the strength, hardness, stress–strain properties and fatigue of a material. Several previous investigations have shown that grain size distribution has a significant effect on mechanical properties. For example, Berbenni et al.[12] showed that for a given average grain size, broadening of the grain size dispersion reduces the strength of a material. The classification and quantification of polycrystalline microstructure is therefore critically important in predicting material responses. A microstructural understanding is also important for the design and discovery of new materials with tailored properties, such as stronger materials that minimize fatigue failures of a machine component during their operation lifetime.

The ubiquitous connection between microstructure (mainly, grain-size distribution) of a material and its physical properties has motivated numerous studies on developing robust techniques to analyze microscopy/tomography images.[13–18] ASTM outlines the industry standard for grain identification in 2D data,[16] which consists of methods such as matching, planimetric, and intercept

[1]Center for Nanoscale Materials, Argonne National Laboratory, Argonne, IL, USA. [2]Department of Mechanical Engineering, University of Louisville, Louisville, KY, USA. [3]Department of Mechanical and Industrial Engineering, University of Illinois at Chicago, Chicago, IL, USA. *email: hchan@anl.gov; skrssank@uic.edu

methods. These methods, albeit can achieve high accuracy (±0.25 grain size units) and reproducibility, can be severely impaired when the intersection criterion (for distinguishing grains) is poorly chosen or the grain-size distribution is non-uniform.[16] In addition, these technique often require tedious manual measurements, and automation is challenging due variability in etching level or contrast differences although electron back scattering diffraction methods have been recently proposed to eliminate subjectivity surrounding existence/location of grain boundaries.[15,19] Automated methods for grain identification in 2D data have been developed over the years. For example, there are supervised convolutional neural network (CNN) based methods,[20] as well as unsupervised clustering or Voronoi based methods. Supervised methods once trained can achieve high accuracy, but the required prior training to target data makes them specific to the material system that they are trained for. Unsupervised methods based on a combination of histogram thresholding, watershed algorithms, and k-mean clustering can sometimes perform on par with supervised methods when a priori information (e.g., number of grains, crystal structure/orientation) and optimized hyperparameters are given, but in that case they inherent the same specificity of a material system due to the required information from specific dataset or experimental technique. Unsupervised methods that rely on just local density of atoms/electrons are applicable to a much wider range of material systems and experimental techniques, but at the expense of accuracy. Nevertheless, existing grain-analysis techniques are largely focused on 2D images and extending them to 3D images is not trivial. Extension of 2D based techniques to 3D is routinely done via stack of 2D image slices, which can be impacted by number of slices or orientation of slices and often leads to time-consuming processing. Evidently, a fast, general, reliable, and accurate way of identifying and analyzing grains in 3D images is still elusive.

With the advent of fourth generation synchrotron X-ray sources which possess extreme brightness and increased coherence, it has become possible to image materials over time in 3D (i.e., 4D imaging). Such advanced imaging is particularly invaluable especially when seeking information about material response under in-situ or operando conditions. For polycrystalline materials, a few imaging modalities including diffraction contrast tomography (DCT), Laue diffraction and HEDM have been used to create 3D maps of the polycrystalline state of the sample.[21] Segmentation or the appropriation of the resulting image into discrete domains is often a challenge especially in tomographic images. When the contrast between regions or segments is faint, simple thresholding is often insufficient and more advanced techniques such as clustering, deformable models or gradient based techniques are required, which have been employed with varying degrees of success.[21] The ability to rapidly and accurately segment images not just for polycrystalline materials but also to identify inclusions and precipitates within a matrix would be invaluable for real-time characterization of materials.

Here, we present a method that combines topology classification, image processing, and unsupervised machine learning including clustering algorithms to enable rapid microstructural characterization in 3D samples. Our method provides grain size distribution of samples derived from either simulations or experiments. We demonstrate the method on synthetic data of several representative polycrystal types – metals (fcc, bcc, hcp) and ice (hexagonal/cubic), as well as experimentally collected data of Ni-based superalloy. The method is insensitive to the presence of extended defect structures such as stacking faults and semi-amorphous domains which stymie standard classification methods. We have also extended the method to the characterization other microstructural features such as voids in porous materials[22] (i.e., polymer matrices) and micellar distribution in complex solutions. The technique is computationally efficient and enables fast identification, tracking, and quantification of microstructural features that affect material properties. We envision this approach to be vital for future real time analysis of data obtained from large characterization facilities such as synchrotrons and broadly applicable to any 3D crystallographic data. The approach also enables characterization across a broad class of materials from polycrystalline inorganics such as metals and ceramic to soft materials such as polymers and self/directed assembled structures in complex fluids.

RESULTS

Microstructural characterization

Figure 1 illustrates the major steps in our ML method for autonomous microstructural characterization. These steps can be loosely organized into three main processes, analogous to that in

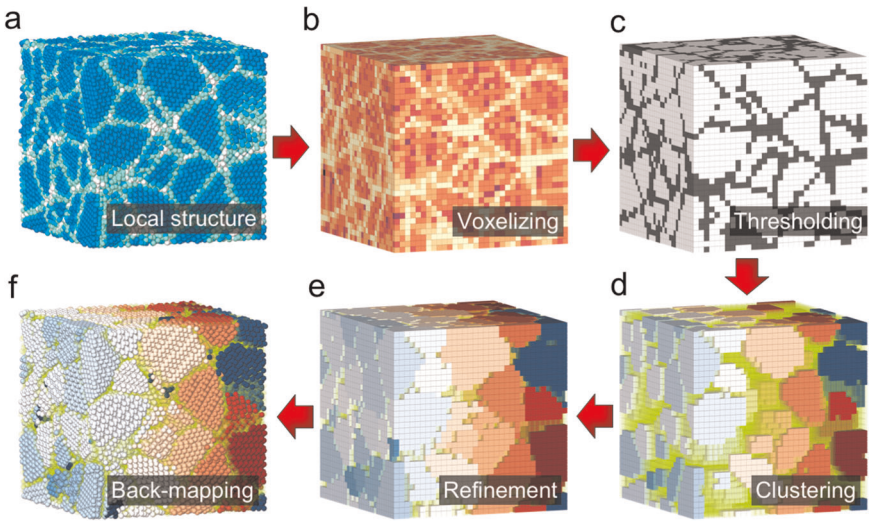

Fig. 1 A schematic showing the major steps of our ML method for autonomous microstructural characterization of 3D polycrystalline samples. a Identification of local structures using topological classifiers. **b** Voxelization improves the processing efficiency and enables image-based processing techniques. **c** Thresholding enhances the distinction between microstructures and boundaries. **d** Clustering algorithm identifies individual microstructures. **e** Refinement process improves the size estimation and distribution of identified microstructures. **c** An optional back-mapping step transforms voxel data back to atomistic representation.

a data science process (i.e., data collection and cleaning, data analysis, and data finishing).

Process 1: Preconditioning and topological classifiers

The first step in our microstructural analysis is to distinguish between the microstructures (e.g., grains) and their boundaries (Fig. 1a). For atomistic polycrystalline systems, this can be done via local structure identification using topological classifiers, such as common neighbor analysis (CNA) for fcc, bcc, hcp structure types that require topological information up to 1st nearest neighbors, and extended CNA for diamond (hexagonal/cubic) structure types that require up to 2nd nearest neighbors. These classifiers assign local structure labels to atoms based on their topological relationships with nearby neighbors. Unknown atom types (amorphous) or unlabeled atoms are typically excluded from the microstructural analysis. For soft materials, the labeling can be done via atom type assignment based on chemical elements, bond topology, and local charges, etc. Next, voxelization (Fig. 1b) is performed on the labeled (e.g., crystalline) atoms/beads, which makes possible efficient data preconditioning using standard image processing techniques. Lastly, preconditioning procedures such as image filters (e.g., uniform blur, local variance, etc.) and thresholding are applied on the voxelized data or experimental images to identify boundaries of microstructures (Fig. 1c).

Process 2: Unsupervised machine learning

Microstructural analysis is performed via clustering of the preconditioned voxels (Fig. 1d). Voxels of similar local structure labels are clustered. The number of clusters and their volumes, e.g., number of grains and their sizes, provide an estimate of the size distribution. Furthermore, individual microstructure is assigned a unique cluster label that can be utilized for visualization purpose. The choice of clustering algorithms (e.g., K-Means, DBSCAN, Mean-Shift, Gaussian mixture models) depends on the amount of pre-existing knowledge about the system, which can include the number of microstructures, characteristics of the boundaries, etc. Although in the results section, we demonstrate that density-based clustering algorithms (e.g., DBSCAN) can effectively handle all the tested polycrystal types and soft material systems even with limited pre-existing knowledge.

Process 3: Refinements and back-mapping

The number and size estimate of microstructures obtained from the unsupervised machine learning process can be improved via a refinement step. Techniques such as label propagation and label spreading can effectively be used to assign cluster label to unlabeled voxels/atoms nearby the boundaries (Fig. 1e). This step recovers information that might have been lost during the preconditioning process (e.g., thresholding and blur filters). Finally, for atomistic systems, a quick back-mapping based on the spatial relationship between voxels and atom coordinates can be used to transform voxels back to their corresponding atomistic representation (Fig. 1f).

Microstructural characterization applied to example systems

To demonstrate the generality of the described approach, we apply our ML method for the characterization of microstructural features in both polycrystalline materials and in soft materials such as polymers and micelles. In the former, the goal is to characterize the grain size distribution in 3D polycrystalline samples whereas in the latter, the ML algorithm is used to identify porosity and voids in soft materials such as polymer matrix and micellar distribution during a typical aggregation process in complex fluids. To adapt the ML method for these systems, mainly the preconditioning process (local structure classification, voxelization bin size, etc.) needs to be customized, and the details are discussed case-by-case. Below, we first describe our approach for the clustering and refinement processes.

In all the above described systems, the number of microstructures is not known, and the microstructures can be irregularly shaped, so we choose to use a local density-based clustering algorithm, DBSCAN, for the microstructural analysis. The algorithm has two hyperparameters in its clustering criterion: neighborhood cutoff (ε) and the minimum required number of neighbors (N_{min}). For simplicity, we define ε to include only 1st nearest voxels of each voxel and start with the strictest criterion ($N_{min} = 27$ for 3D or 9 for 2D) and loosen it until the total number of clusters is maximized. Refinement of the clusters is done by assigning unlabeled voxels to neighboring cluster labels of maximum occurrence, with priority given to unlabeled voxels close to smaller microstructures. Finally, to recover an atomistic representation from voxels, atoms are assigned cluster labels of the voxels that they are located in.

Case 1: Grain size distribution in metal polycrystals

Four metal samples (aluminum, iron, silicon, and titanium) are chosen as representatives of common polycrystal types (fcc, bcc, diamond, and hcp). For benchmarking, we prepared synthetic polycrystalline samples with a known size distribution (see Methods). The preconditioning process begins with local structure identification of atoms using standard CNA for fcc, bcc, and hcp structures, and extended CNA for diamond structures. The atoms are classified as either "crystalline" or "boundary" types. Voxelization of atoms is done based on number densities of crystalline atoms using a uniform bin size (4.5 Å for fcc Al, 4.1 Å for bcc Fe, 4.0 Å for diamond Si, and 4.4 Å for hcp Ti). A 40-percentile thresholding of non-zero voxels is applied in all samples to exclude grain boundary voxels from the clustering process.

Results of the ML grain analysis for the four metal samples are shown in Fig. 2. For each polycrystal type, a plot shows the target (in red) and predicted (in blue) grain size distributions sampled using gaussian kernel density estimation. The snapshots next to each plot visualize the polycrystallinity of these samples, where individual grains are colored by their sizes (smallest in red, largest in blue). Comparison between the target and predicted distributions indicates that our unsupervised method has achieved >94% accuracy in predicting the number of grains, and correctly identifying grains that are larger than ~200 atoms in size.

Case 2: In situ visualization and 3D analysis of simulation trajectories

The high computational efficiency of our ML method makes it suitable for in-situ post-processing of molecular dynamics (MD) trajectories. To demonstrate this, we apply the grain analysis on the entire > 1 μs MD simulation trajectory (with a frame every 0.1 ns) of a polycrystalline ice sample, which was previously performed using a coarse-grained (CG) model of water.[23] The preconditioning process is similar to that of diamond Si, where extended CNA was used to identify hexagonal, cubic, and stacking disordered phases of ice. Due to the larger sizes of CG beads compared to atoms, a larger bin size of 5 Å was used in the voxelization process. The voxelization is done based on number densities of cubic and hexagonal beads.

Figure 3 shows two representative snapshots at $t = 330$ ns (smaller grains) and $t = 669$ ns (larger grains). The bottom left of each snapshot shows the result of the grain analysis, where individual grains are colored by their sizes. Despite performing the analysis on an uncorrelated frame-by-frame basis, the coloring is relatively consistent due to the sorting by grain sizes. However, changes in number of grains across frames can lead to inconsistent assignment of cluster labels, which makes it difficult to isolate one grain and track its time evolution. We envision this

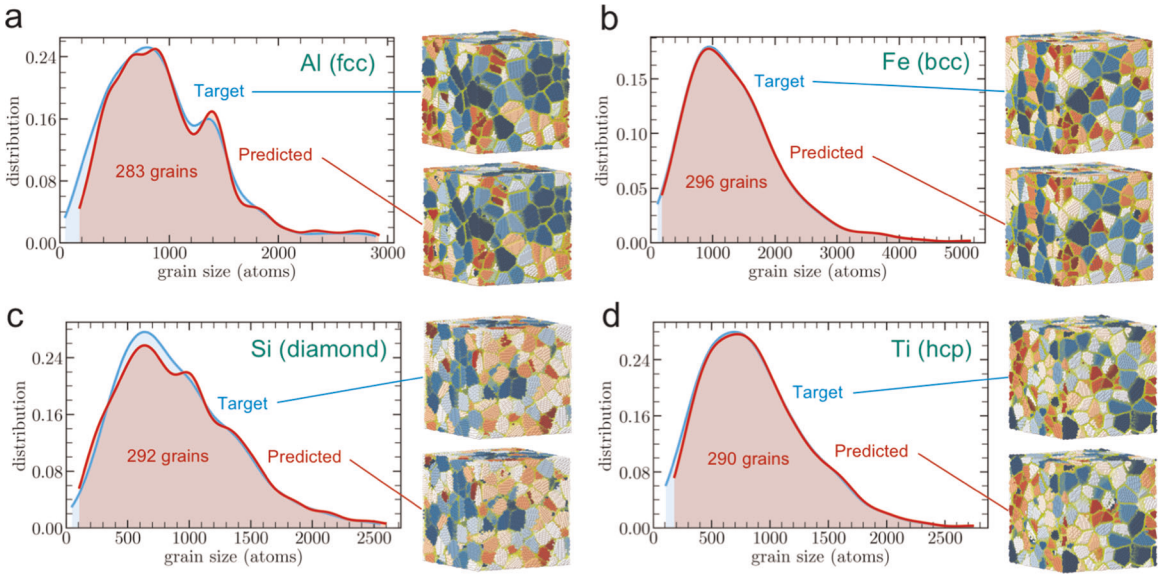

Fig. 2 Application of our ML method on several representative polycrystalline metal samples. Each of the samples (aluminum, iron, silicon, and titanium) is ~20 nm × 20 nm × 20 nm in size (~500,000 atoms). All samples have 300 grains. The plots show the target (in red) and predicted (in blue) grain size distributions. The distributions are normalized such that the shared area equates to the total number of grains. Polycrystallinity of these samples are visualized by snapshots shown next to the plots, where individual grains are colored by their sizes (smallest in red, largest in blue). The sample set consists of common polycrystal types: **a** fcc, **b** bcc, **c** diamond, **d** hcp.

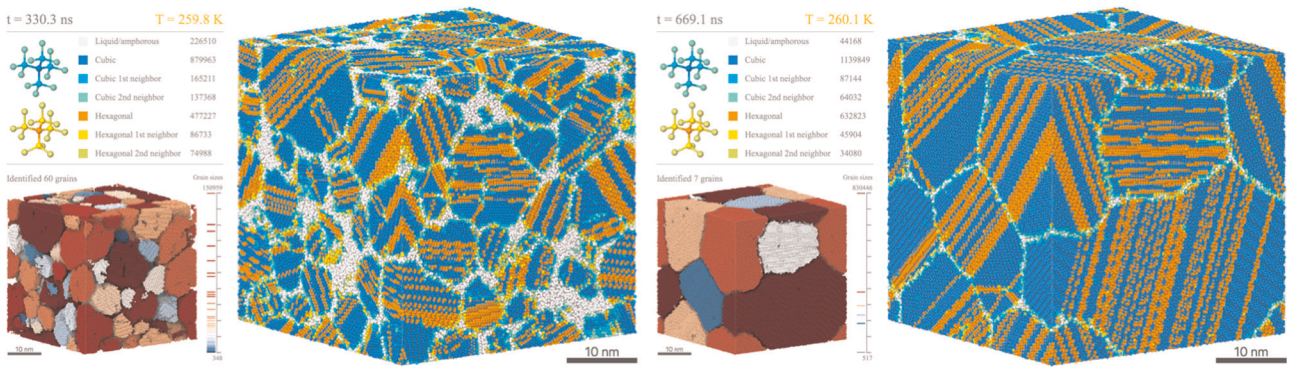

Fig. 3 Demonstration of our unsupervised ML grain analysis method on large-scale MD simulations. Snapshots from a 2-million molecules simulation of polycrystalline ice performed using a CG model of water.[23] The right side of each snapshot shows the hexagonal/cubic stacking disordered ice grains and their grain boundaries. Bottom left of each snapshot shows the result of the grain analysis, where individual grains are colored by their sizes.

to be resolved in future works by introducing correlation across frames based on spatial proximity and lattice orientation of the individual grains.

Case 3: Characterization of porosity and voids in polymer matrix

The described ML based approach can be easily extended for void analysis in porous material samples. To demonstrate this, high density polysiloxane and polyethylene samples were prepared (see Methods). These samples were equilibrated and densified using MD simulations. The preconditioning process is simply voxelization of the system based on the number densities of the polymer atoms. A bin size of 3 Å was used in the voxelization step to sample larger void spaces, although smaller bin sizes (higher resolution) can be used to sample much smaller spaces. Results of the void analysis are shown in Fig. 4. The method can handle large voids (Fig. 4a) as well as small voids (Fig. 4c). The method provides size distributions of voids, such as those shown in Fig. 4b, d, which can be used to characterize the porosity nature of the matrix samples.

Case 4: Characterization of micellar size distribution in complex fluids

The described ML based approach is also suitable for the structural analysis of hierarchical soft materials in complex fluids. The dynamics of ions and mesoscale structure in complex organic fluids is a fascinating fundamental science problem with deep implications for many important energy, chemical, and biological systems. Many recent studies[24–27] have indicated that ion dynamics and transport can be strongly influenced by the hierarchical mesoscale ordering and internal interfaces that often occur in these systems.[28,29] The formation of such hierarchical structures provides a broad opportunity to design new materials with outstanding performance for diverse applications such as battery electrolytes, MRI contrast reagents, sensors, catalysts, and solvent extraction systems.[30] Although the equilibrium structure and phase behavior of complex fluids has been the subject of much study, there is a need to characterize the dynamics to understand and control ion transport, complexation, and aggregation processes. Here, we use our ML algorithm to characterize the micellar size distribution during the

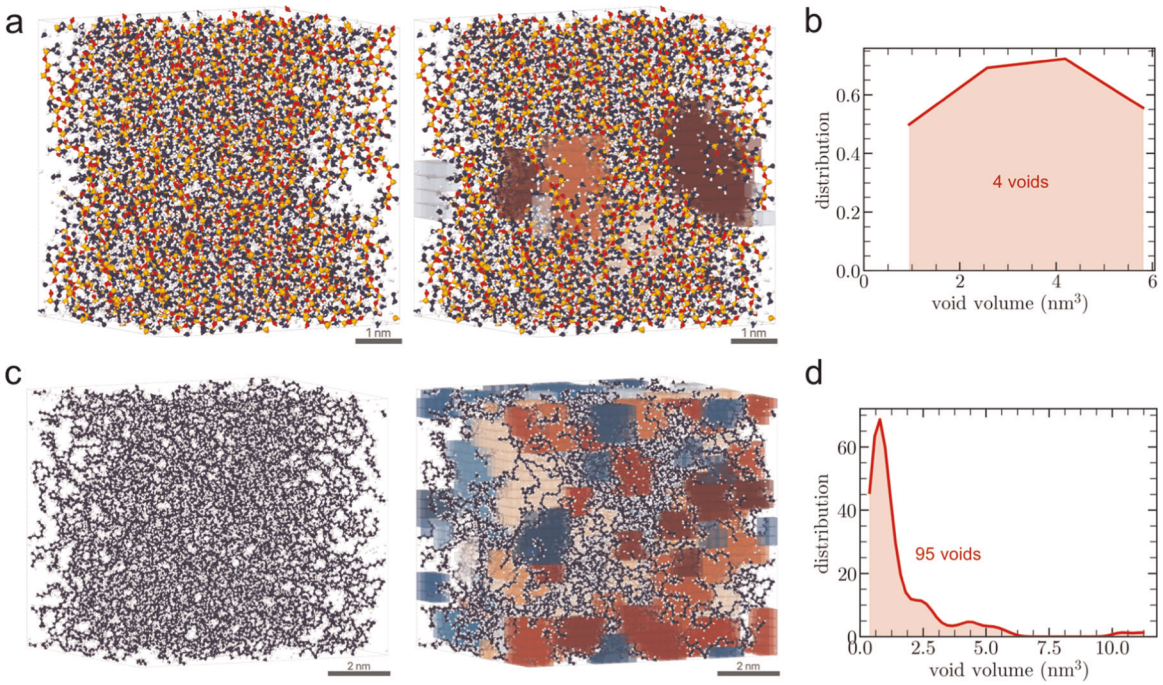

Fig. 4 Demonstration of our unsupervised ML method on the analysis of voids in polymeric systems. The figure shows polysiloxane sample (top) and polyethylene sample (bottom). **a**, **c** Snapshots from atomistic MD simulations showing the identified void spaces. Individual voids are colored by their sizes. **b**, **d** Plots showing the size distribution of the voids. The distributions are normalized such that the shaded area equates to the total number of voids.

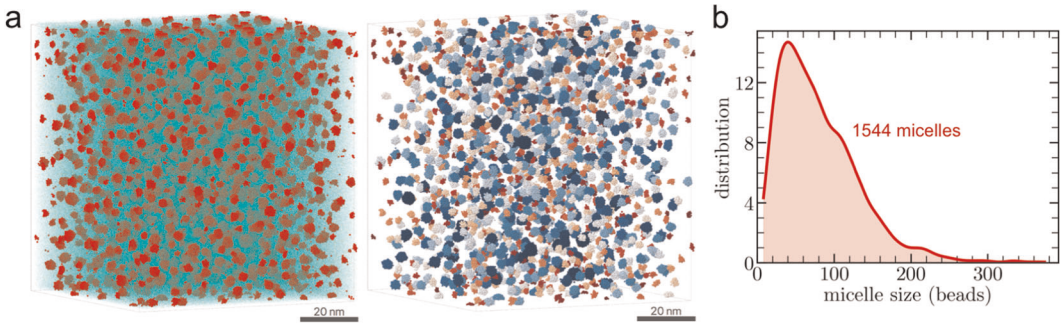

Fig. 5 Demonstration of our unsupervised ML method on the size distribution analysis of reverse micelles in a complex solution. **a** Snapshots from CG MD simulations showing cluster of water beads within individual micelles colored by their sizes. **b** Plot showing the size distribution of the micelles as a function of the water cluster sizes. The distribution is normalized such that the shaded area equates to the total number of micelles.

aggregation process in a 3D colloidal sample obtained from molecular simulation trajectory.

To demonstrate this, we obtained a configuration of reverse micelles from CG MD simulations (see Methods). The preconditioning process includes voxelization of the system based on the number densities of water beads. Due to the high 4:1 CG ratio of this model, a large bin size of 8 Å is used in the voxelization step. Figure 5a shows the water clusters within individual equilibrated micelles colored by their sizes, and Fig. 5b shows the micellar size distribution as a function of these water cluster sizes.

Case 5: Grain size distribution in superalloy sample from experiment

The described ML based approach can be applied to images collected from experiments. Unlike voxelized atomistic data, 3D images obtained from experimental characterization techniques, such as tomography and coherent diffraction X-ray imaging, contain more noise and artifacts. Furthermore, these images can be of bright-field or dark-field types and the grains can span a range of pixel/voxel intensity values, which require grain boundaries detection techniques beyond thresholding with just one cutoff. Figure 6 shows examples of such images and demonstrates the use of a local variance filter to effectively identify boundaries of microstructures. Our method utilizes the same boundary detection method as outlined earlier. Figure 7 demonstrates the use of our method on an IN100 Ni-based superalloy sample collected from serial-sectioning experiments. The processing pipeline on such data is illustrated in Fig. 7a and the resulting grain size distribution is shown in Fig. 7b–d demonstrate the same processing pipeline applied to input images of lower resolutions, which result in significant speedup in processing albeit at the expense of lower feature detection resolution.

DISCUSSION

The robustness of our microstructural analysis method can be assessed by the deviation in results upon introducing variations to

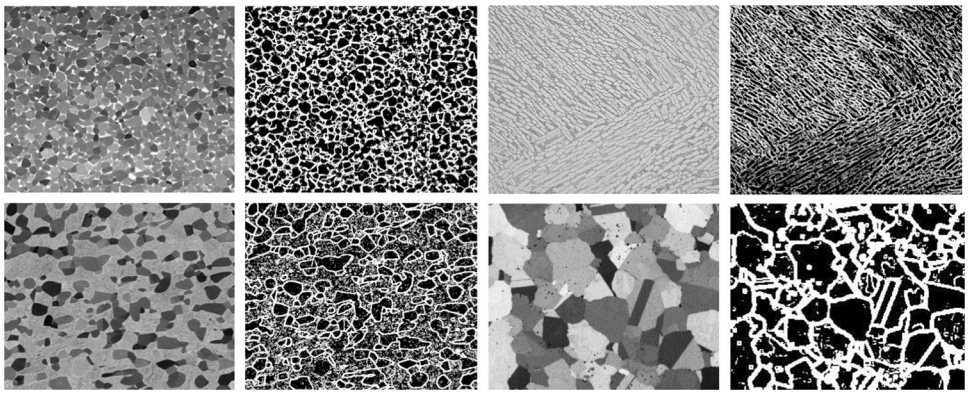

Fig. 6 Examples illustrating the use of local variance filter for grain boundary identification. 2D images of polycrystalline grain samples are reproduced with permission from Campbell et al.[36] and Groeber et al.[15] The method can handle both bright field and dark field images and is only sensitive to local variance of pixel intensity which eliminates problems associated with direct thresholding based on absolute pixel intensity.

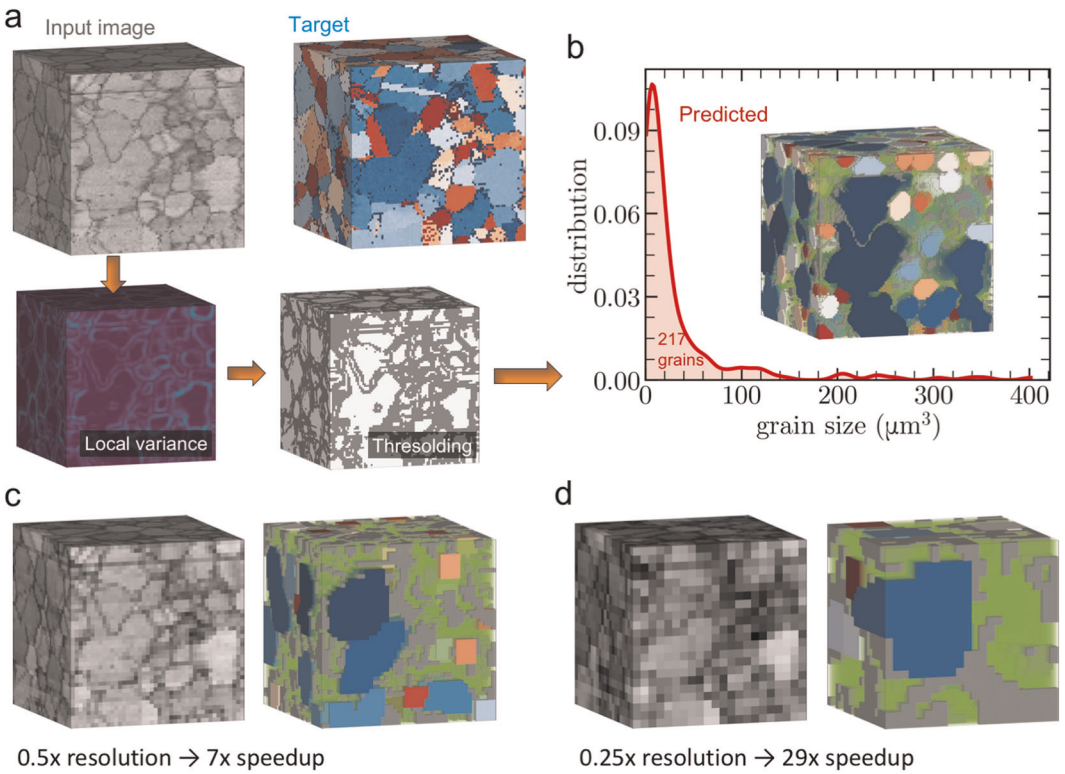

Fig. 7 Demonstration of our unsupervised ML method on grain identification of an IN100 Ni-based superalloy sample collected from serial-sectioning experiments. a 3D input image reconstructed from electron backscatter diffraction (EBSD) data[15] and the corresponding target grain segmentation labeled using inverse pole figure (IPF) coloring. In our method, the input image is pre-processed using a local variance filter and thresholding prior to the clustering and refinement step. **b** The predicted grain size distribution and grain segmentation obtained using our method. Boundary and unidentified voxels are colored by green and gray. **c**, **d** Lower resolution input images obtained by down sampling and the corresponding grain segmentation predicted by our method. The effect of down sampling is analogous to using a large bin size in the voxelization step for atomistic data. Down sampling significantly speeds up the processing but at the expense of accuracy (i.e., ability to detect small and fine features).

the data and hyperparameters at each of the major steps in the processing pipeline (Fig. 1). The cascade of operations leads to the possibility of errors earlier in the processing pipeline propagating downstream. Out of the 6 major steps, only the first 4 steps (i.e., local structure classification, voxelization, thresholding, and clustering) are likely to be affected by the variations in data or the choice of hyperparameters. Furthermore, hyperparameters in the clustering step can be optimized on-the-fly based on the number of identified clusters.

Here, we provide a quantitative assessment of error sensitivities associated with the remaining 3 steps. We use the fcc Al system as an example and manually introducing variations to the data and hyperparameters. Figure 8a–c shows snapshots of a sample with ~0%, 15%, and 25% randomly perturbed local structure labels. These incorrect labels in atomistic data affect the identification of grain boundaries (i.e., crystalline versus amorphous atoms), which is analogous to introducing noise to images from experimental measurements. The grain size distribution plots in Fig. 8d–e

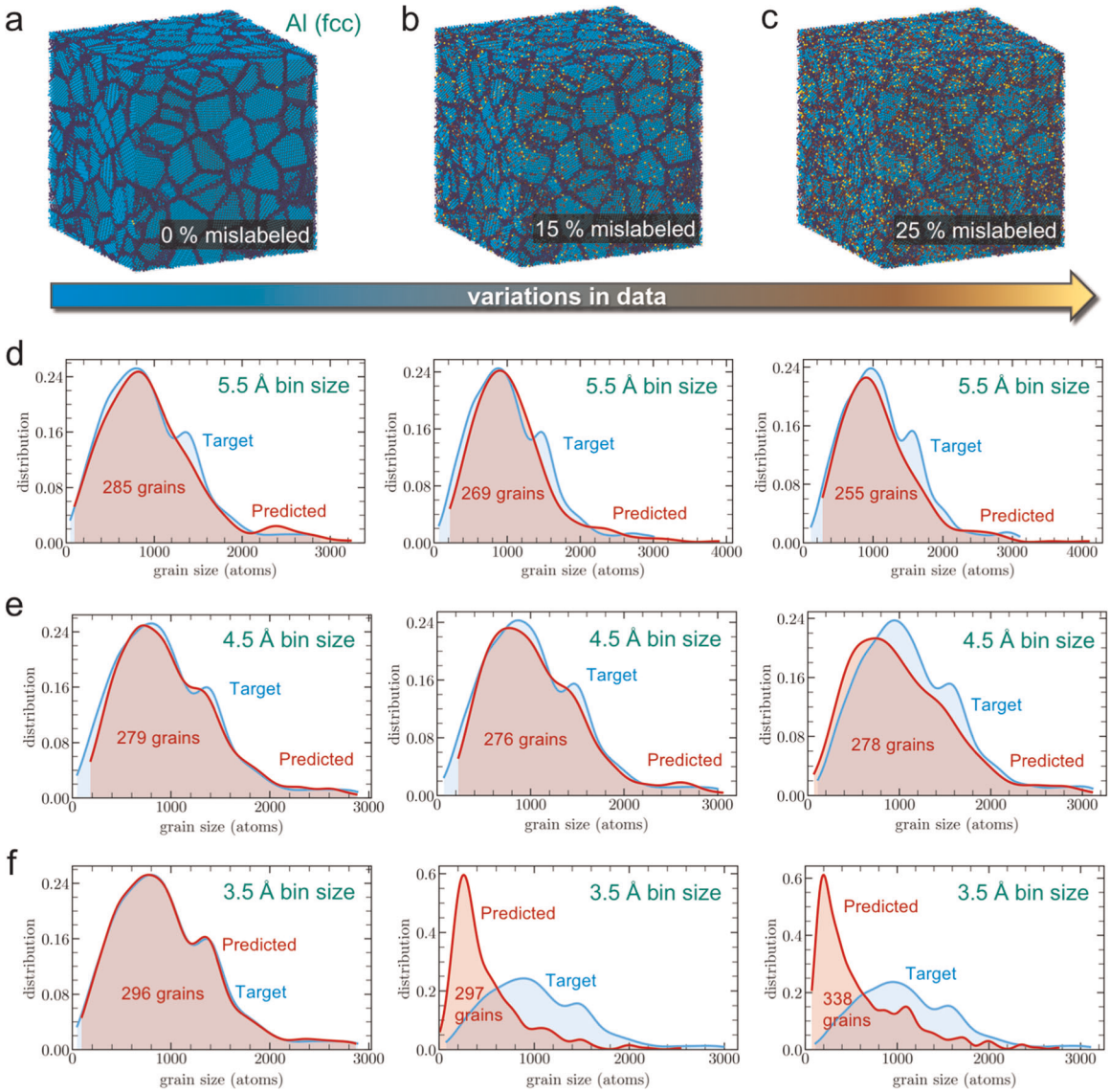

Fig. 8 Error sensitivity of our unsupervised ML method on grain identification. a–c Snapshots of the atomistic fcc Al sample with varying number of randomly perturbed local structure labels. **d–f** Grain size distribution plots of the system. Local variance filter with a 90-percentile thresholding is used for grain boundary identification, which alleviates the error sensitivity of the method to different voxelization bin sizes. Plots from left to right correspond to the amount of data variation in **a–c**, whereas from top to bottom, the voxelization bin size changes from 5.5 Å to 4.5 Å to 3.5 Å. As the bin size increases, the method become more resilient to variations in data due to more data averaging from down-sampling. This however comes at the expense of losing fine structures in the grain size distribution.

demonstrate that our method is resilient to such noise. For instance, the method can handle up to ~25% variation in the fcc Al data when the voxelization bin size is 5.5 Å, and up to ~15% data variation when the bin size is 4.5 Å. This robustness is attributed to the various down-sampling operations (e.g., voxelization and local variance/uniform filters) and the use of a density-based clustering algorithm that can handle noise (i.e., DBSCAN). As the voxelization bin size increases, the amount of data averaging increases, which makes it more resilience to variations in the data. There is, however, a trade-off. Larger bin size also leads to more efficient processing (Fig. 7c, d), but this comes at the expense of losing information of small features or fine details in the grain size distribution. Also note that the use of local variance filter for boundary identification alleviates the sensitivity of the method to different bin sizes, but the method remains sensitive to the thresholding cutoff value in the thresholding step. We found that a 90-percentile thresholding of non-zero voxels works well for atomistic data from simulations, and a 40-percentile thresholding generally works for images from experiment. This thresholding cutoff, similar to the hyperparameters (ε and N_{min}) associated with the DBSCAN clustering algorithm, can potentially be optimized on-the-fly based on the number of identified grains. Future studies might investigate techniques analogous to Otsu thresholding for choosing such cutoff in a deterministic manner.

The efficiency of our method can be analyzed based on the time complexity of the steps in the workflow. Excluding the time that it takes to load the input data, the major time-consuming steps are voxelization, clustering, and refinement. The voxelization step has a time complexity of $O(n)$ since each atom/bead is processed once during the conversion into voxels. However, this operation provides a significant time saving in return for the remaining steps in the workflow since the voxelized system is typically ~25% of the original system size which is further reduced via subsequent preconditioning and thresholding. The clustering step, in particular DBSCAN clustering in 3D space, has a typical time complexity of $O(n \log(n))$, where n is the remaining number

of voxels after thresholding in the preconditioning process. The clustering step is supported by a *k*-d tree (incorporating periodic boundary conditions) for fast nearest neighbor search, which has a worst-case time complexity of $O(n \log(n))$ to build and average time complexity of $O(\log(n))$ for each neighbor query. The same *k*-d tree is used in the refinement step, where repeated query is performed on voxels with no cluster labels. The time spent in the refinement step varies depending on nature of the (grain) boundaries.

In conclusion, we summarized the importance and challenges in microstructural analysis of polycrystalline samples in full 3D and outlined an unsupervised ML approach to solve this major problem. Our ML method starts with data preconditioning using local structure topological classifier, voxelization, and image processing. An unsupervised ML clustering algorithm was then used to obtain statistics and distribution of the microstructures. Finally, techniques such as label spreading was used to refine the results and back-mapping is performed to recover the atomistic representation. We demonstrated the efficacy of our method on several different classes of materials ranging from polycrystalline solids to soft materials such as polymers and complex fluids. The technique is applicable for the characterization of grain size distribution, voids, porosity and similar microstructural features across a broad class of inorganic and soft material systems. The technique can be applied to synthetic data samples, as well as experimentally measured data. We also highlighted the computational efficiency and error sensitivity of the method and emphasized its suitability for future real time analysis of data from large characterization facilities.

METHODS

Polycrystal sample preparation

Synthetic polycrystalline metal samples of a fixed number of grains (300) were prepared using Voronoi tessellation. Each sample is ~20 nm × 20 nm × 20 nm in size (~500,000 atoms) with periodic boundaries applied in the *x*-, *y*-, and *z*- directions. Grain size distribution curves of these samples were obtained for the purpose of benchmarking, where atoms at the grain (Voronoi) boundaries were identified and excluded from the grain size distribution curves to provide a more accurate grain size count. The identification of boundary atoms was done using standard CNA for fcc, bcc, and hcp lattice types and extended CNA for diamond lattice types.

Polycrystalline ice simulation

Polycrystalline ice samples were obtained from previously performed CG MD simulations of homogeneous nucleation runs[23] using LAMMPS.[31] The sample size is ~40 nm × 40 nm × 40 nm (~2-million water molecules) and the microstructure analysis was performed on the entire trajectory for up to $t = 1.2$ μs for a frame every 0.1 ns.

Polymer sample preparation

Two types of polymer matrix samples, polysiloxane and polyethylene, were prepared using atomistic fixed bond models. The sample sizes were ~5 nm × 6 nm × 6 nm box (~17k atoms) and ~8 nm × 9 nm × 8 nm box (~33k atoms), respectively. These samples were minimized and equilibrated for up to 200 ns in LAMMPS[31] using an empirical class2 potential with parameters from the COMPASS and PCFF force fields.[32,33] An isothermal-isobaric (NPT) ensemble at T = 300 K and varying pressures was used to densify the samples.

Micelle sample preparation

A sample of complex solution was prepared using a CG model (4:1 mapping). The sample size was ~82 nm × 82 nm × 90 nm, containing 125,000 water molecules, 1,500,000 dodecane molecules, and 120,400 surfactant-like molecules. The sample was minimized and equilibrated for up to 200 ns in NAMD[34] using the MARTINI force field[35] to obtain a configuration of reverse micelles. An isothermal-isobaric (NPT) ensemble at T = 300 K and $P = 1$ bar was used for the equilibration.

DATA AVAILABILITY

The data that support the findings of this study are available from the authors upon reasonable request.

CODE AVAILABILITY

Code and workflow developed in this study are available from the authors upon reasonable request.

Received: 13 May 2019; Accepted: 2 December 2019;
Published online: 06 January 2020

REFERENCES

1. Ludwig, W. et al. Three-dimensional grain mapping by x-ray diffraction contrast tomography and the use of Friedel pairs in diffraction data analysis. *Rev. Sci. Instrum.* **80**, 033905 (2009).
2. Lienert, U. et al. High-energy diffraction microscopy at the advanced photon source. *JOM* **63**, 70–77 (2011).
3. Holm, E. A. & Foiles, S. M. How grain growth stops: a mechanism for grain-growth stagnation in pure materials. *Science* **328**, 1138–1141 (2010).
4. McFadden, S. X., Mishra, R. S., Valiev, R. Z., Zhilyaev, A. P. & Mukherjee, A. K. Low-temperature superplasticity in nanostructured nickel and metal alloys. *Nature* **398**, 684–686 (1999).
5. Uchic, M. D., Dimiduk, D. M., Florando, J. N. & Nix, W. D. Sample dimensions influence strength and crystal plasticity. *Science* **305**, 986 (2004).
6. Van Swygenhoven, H. Grain boundaries and dislocations. *Science* **296**, 66 (2002).
7. Offerman, S. E. et al. Grain nucleation and growth during phase transformations. *Science* **298**, 1003 (2002).
8. Arzt, E. Size effects in materials due to microstructural and dimensional constraints: a comparative review. *Acta Mater.* **46**, 5611–5626 (1998).
9. Chu, Z. et al. Impact of grain boundaries on efficiency and stability of organic-inorganic trihalide perovskites. *Nat. Commun.* **8**, 2230 (2017).
10. Hall, E. O. The deformation and ageing of mild steel: III discussion of results. *Proc. Phys. Soc. Sect. B* **64**, 747–753 (1951).
11. Petch, N. J. The cleavage strength of polycrystals. *J. Iron Steel Inst.* **174**, 25–28 (1953).
12. Berbenni, S., Favier, V. & Berveiller, M. Impact of the grain size distribution on the yield stress of heterogeneous materials. *Int. J. Plast.* **23**, 114–142 (2007).
13. Bennett E. G., Roebuck B. *The metallographic measurement of hardmetal grain size*. (National Physical Laboratory, 2000).
14. Engqvist, H. & Uhrenius, B. Determination of the average grain size of cemented carbides. *Int. J. Refractory Met. Hard Mater.* **21**, 31–35 (2003).
15. Groeber, M. A., Haley, B. K., Uchic, M. D., Dimiduk, D. M. & Ghosh, S. 3D reconstruction and characterization of polycrystalline microstructures using a FIB–SEM system. *Mater. Charact.* **57**, 259–273 (2006).
16. Peregrina-Barreto, H. et al. Automatic grain size determination in microstructures using image processing. *Measurement* **46**, 249–258 (2013).
17. Roebuck, B. Measuring WC grain size distribution. *Met. Powder Rep.* **54**, 20–24 (1999).
18. ISO 449-2. *Hardmetals-metallographic determination of microstructure. Part 2: Measurement of WC grain size* (International Standards Organization, Geneva, 2008).
19. Mingard, K. P. et al. Comparison of EBSD and conventional methods of grain size measurement of hardmetals. *Int. J. Refractory Met. Hard Mater.* **27**, 213–223 (2009).
20. Ma, B. et al. Deep learning-based image segmentation for al-la alloy microscopic images. *Symmetry* **10**, 107 (2018).
21. Maire, E. & Withers, P. J. Quantitative X-ray tomography. *Int. Mater. Rev.* **59**, 1–43 (2014).
22. Narayanan, B. et al. Machine learnt bond order potential to model metal–organic (Co–C) heterostructures. *Nanoscale* **9**, 18229–18239 (2017).
23. Chan, H. et al. Machine learning coarse grained models for water. *Nat. Commun.* **10**, 379 (2019).
24. Kofu, M. et al. Heterogeneous slow dynamics of imidazolium-based ionic liquids studied by neutron spin echo. *J. Phys. Chem. B* **117**, 2773–2781 (2013).
25. Burankova, T., Hempelmann, R., Wildes, A. & Embs, J. P. Collective ion diffusion and localized single particle dynamics in pyridinium-based ionic liquids. *J. Phys. Chem. B* **118**, 14452–14460 (2014).
26. Zheng, Z.-P. et al. Ionic liquids: not only structurally but also dynamically heterogeneous. *Angew. Chem. Int. Ed.* **54**, 687–690 (2015).

27. Lee, K.-K. et al. Ultrafast fluxional exchange dynamics in electrolyte solvation sheath of lithium ion battery. *Nat. Commun.* **8**, 14658 (2017).
28. Hayes, R., Warr, G. G. & Atkin, R. Structure and nanostructure in ionic liquids. *Chem. Rev.* **115**, 6357–6426 (2015).
29. Prévost, S., Gradzielski, M. & Zemb, T. Self-assembly, phase behaviour and structural behaviour as observed by scattering for classical and non-classical microemulsions. *Adv. Colloid Interface Sci.* **247**, 374–396 (2017).
30. Ellis, R. J. et al. Complexation-induced supramolecular assembly drives metal-ion extraction. *Chem. – A Eur. J.* **20**, 12796–12807 (2014).
31. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).
32. Sun, H. COMPASS: an ab initio force-field optimized for condensed-phase applicationsoverview with details on alkane and benzene compounds. *J. Phys. Chem. B* **102**, 7338–7364 (1998).
33. Sun, H., Mumby, S. J., Maple, J. R. & Hagler, A. T. An ab Initio CFF93 all-atom force field for polycarbonates. *J. Am. Chem. Soc.* **116**, 2978–2987 (1994).
34. Phillips, J. C. et al. Scalable molecular dynamics with NAMD. *J. Computational Chem.* **26**, 1781–1802 (2005).
35. Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P. & de Vries, A. H. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **111**, 7812–7824 (2007).
36. Campbell, A., Murray, P., Yakushina, E., Marshall, S. & Ion, W. New methods for automatic quantification of microstructural features using digital image processing. *Mater. Des.* **141**, 395–406 (2018).

ACKNOWLEDGEMENTS

Use of the Center for Nanoscale Materials was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357. This research used resources of the National Energy Research Scientific Computing Center; a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We also acknowledge the use of the Carbon cluster and LCRC facilities at Argonne.

AUTHOR CONTRIBUTIONS

H.C., M.J.C., B.N. and S.K.R.S. conceived and designed the project. H.C., B.N. and M.J.C. developed the machine learning based method for 3D microstructural characterization. S.K.R.S. prepared the polycrystalline metal samples. M.J.C. and H.C. performed the large-scale simulations of ice grain formation and growth. H.C. prepared the polymer matrix samples. T.D.L. performed the large-scale simulations of micelle. S.K.R.S. supervised the overall project. All the authors performed the data analysis and contributed to the preparation of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to H.C. or S. K. R. S.S.

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

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