Materials Fingerprinting Classification

Adam Spannaus[a], Kody J. H. Law[b], Piotr Luszczek[c], Farzana Nasrin[d], Cassie Putman Micucci[e], Peter K. Liaw[f], Louis J. Santodonato[g], David J. Keffer[f,*], Vasileios Maroulas[h,*]

[a]*Oak Ridge National Laboratory, Oak Ridge, TN 37830*
[b]*School of Mathematics, University of Manchester, Manchester, UK*
[c]*Innovative Computing Laboratory, University of Tennessee, Knoxville, TN 37996*
[d]*Department of Mathematics, University of Hawaii at Manoa, Honolulu, HI 96822*
[e]*Eastman Chemical Company, Kingsport, TN 37662*
[f]*Department of Materials Science and Engineering, University of Tennessee, Knoxville, TN 37996*
[g]*Advanced Research Systems, Inc., Macungie, PA 18062*
[h]*Department of Mathematics, University of Tennessee, Knoxville, TN 37996*

Abstract

Significant progress in many classes of materials could be made with the availability of experimentally-derived large datasets composed of atomic identities and three-dimensional coordinates. Methods for visualizing the local atomic structure, such as atom probe tomography (APT), which routinely generate datasets comprised of millions of atoms, are an important step in realizing this goal. However, state-of-the-art APT instruments generate noisy and sparse datasets that provide information about elemental type, but obscure atomic structures, thus limiting their subsequent value for materials discovery. The application of a materials fingerprinting process, a machine learning algorithm coupled with topological data analysis, provides an avenue by which here-to-fore unprecedented structural information can be extracted from an APT dataset. As a proof of concept, the material fingerprint is applied to high-entropy alloy APT datasets containing body-centered cubic (BCC) and face-centered cubic (FCC) crystal structures. A local atomic configuration centered on an arbitrary atom is assigned a topological descriptor, with which it can be characterized as a BCC or FCC lattice with near perfect accuracy, despite the inherent noise in the dataset. This successful identification of a fingerprint is a crucial first step in the development of algorithms which can extract more nuanced information, such as chemical ordering, from existing datasets of complex materials.

Keywords: Atom Probe Tomography, High Entropy Alloy, Machine Learning, Topological Data Analysis, Materials Discovery

1. Introduction

Recent advancements in computing and contemporary machine-learning technologies have yielded new paradigms in computational materials science that are accelerating the pace of materials research and discovery [1, 2, 3, 4, 5]. For example, researchers have used a neural network to predict materials properties, clustering them into groups consistent with those found on the periodic table [4] and data-driven materials design is an area now available to researchers due to advances in machine-learning algorithms and computational materials science databases [3, 6, 7]. These developments in computational materials science

*Corresponding author
Email addresses: dkeffer@utk.edu (David J. Keffer), vmaroula@utk.edu (Vasileios Maroulas)

have led researchers to begin exploring structure-property relationships for disordered materials, such as entropy-stabilized oxides and high-entropy alloys (HEAs) [8, 9]. Considering the number of atomic configurations in a disordered crystal structure, such as those found in HEAs [10], the number of possible atomic combinations of even a single unit cell, the smallest collection and ordering of atoms from which an entire material can be built, quickly becomes computationally intractable for existing algorithms [1]. In the present work, we propose an automated machine learning methodology for determining the lattice structure of a noisy and sparse materials dataset, e.g., the type retrieved from atom probe tomography (APT) experiments, for materials with disordered lattice structures, such as HEAs.

One of the fundamental properties of a crystalline material is the structure of its unit cell. Indeed, knowledge of the chemical ordering and geometric arrangement of the atoms of any material is essential for developing predictive structure-property relationships. As materials become more complex and the ordering of atoms amongst lattice sites becomes increasingly disordered, such as is the case with HEAs [11], these structure-property relationships have yet to be developed. Indeed, the high-configurational entropy of HEAs yields a distribution of lattice parameters and cell compositions, as opposed to a single unit cell and lattice constant found in more traditional materials.

For many classes of materials, the lattice structure is either well-known, e.g., sodium chloride (salt) is body-centered cubic, or it can be discovered via X-ray diffraction (XRD) or neutron scattering techniques [12]. XRD is a routine technique for the determination of crystal structures of metals, ceramics, and other crystalline materials. These techniques do not yield atomic level elemental distinctions or resolve local lattice distortions on a scale of less than 10Å [12], which are crucial to researchers working with highly-disordered materials, such as HEAs. Moreover, XRD cannot provide the correlation between atom identity and position in a material. This chemical ordering of atoms is essential to developing predictive relationships between the composition of an HEA and its properties.

High entropy alloys are a relatively new class of metallic alloys, first synthesized in the mid 2000's by [11]. As defined by [13], HEAs are composed of at least five atomic elements, each with an atomic concentration between 5% and 35%. These novel alloys have remarkable properties, such as: corrosion resistance [9, 14], increased strength at extreme temperatures, ductility [15, 16, 17], increased levels of elasticity [18], strong fatigue and fracture resistance [15, 19, 20], and enhanced electrical conductivity [21, 22]. HEAs are amenable to the APT analysis as the process is able to recover elemental type in addition to approximate the lattice sites in a material where the atoms sit.

An experimental process that unambiguously determines the position, identity of each atom, and structure of a material is currently nonexistent [1, 23]. Indeed, quantification of different lattice parameters and unit-cell compositions have not previously been reported due to data quality issues inherent to APT [24, 25]. While these experiments are able to discern elemental types at a high resolution, the process has two drawbacks, *(i) sparsity*: empirically, approximately 65% of the atoms from a sample are not registered by the detector [12]; and *(ii) noise*: the atoms that are observed have their spatial coordinates corrupted by experimental noise [25]. As noted by [25], the spatial resolution of the APT process is up to 3Å (0.3 nm) in the xy-horizontal plane, which is approximately the length of an unit cell. This experimental noise has a two-fold impact on the data retrieved by a typical experiment. First, the noise prevents materials science researchers from extracting elemental atomic distributions, which are essential for developing the necessary interaction potentials for molecular dynamics simulations. Secondly, the experimental noise is significant enough to make atoms that are first neighbors in an atomic neighborhood, i.e., those atoms which occupy adjacent lattice sites, appear as second or third neighbors and vice versa [25]. Furthermore, the experimental noise is only one source of distortion to the lattice structure. HEAs exhibit local lattice deformations due to the random distribution of atoms throughout the material and atoms of differing size sitting at adjacent lattice points [10].

This deformation of the local crystal structure makes any determination of the lattice a challenging

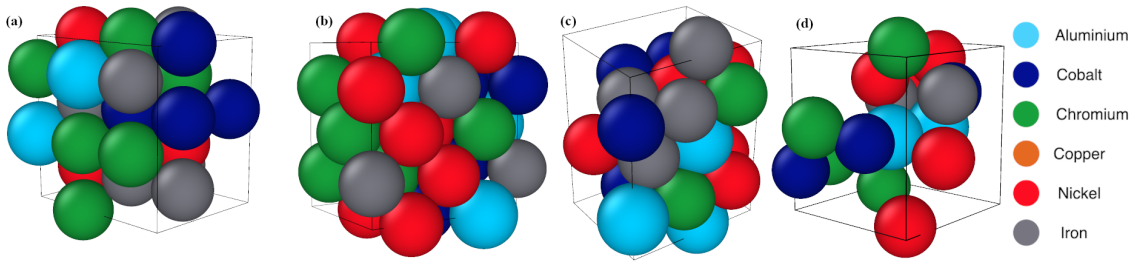

Figure 1: Examples of the lattice structures that we consider viewed with the visualization software Ovito [30] which uses empirical atomic radii in its visualizations. We consider three different crystals: (a) body-centered cubic (BCC), (b) face-centered cubic (FCC), and (c) hexagonal close packed (HCP) lattices showing their similarities and differences with complete, noiseless data. The FCC and HCP structures have only a subtle difference in their geometry. The HCP structure forms an identifying parallelogram (c), whereas the FCC forms a square (b) when all atoms within a radius of the center atom are collected. (d) Example of an FCC structure retrieved from an APT analysis of the HEA $Al_{0.3}CoCrFeNi$ [31] demonstrating the sparsity and atomic displacements due to the resolution of APT process. The noise and sparsity from the APT process obscures this difference between the FCC and HCP structures.

problem for any symmetry-based algorithm, such as [26, 27, 28]. The field of atom probe crystallography has emerged in recent years [23, 29] and existing methodologies in this area seek to discover local structures when the global structure is known *a priori*. In the case of HEAs, the global lattice structure is unknown and must be discovered. Indeed, drawing correct conclusions about the material's crystal structure is virtually impossible from the APT analysis using current techniques [25].

A recent method relying on a convolutional neural network [5] classified synthetic crystal structures that are either noisy or sparse by creating a diffraction image from a lattice structure and using this image as input data for the neural network. The authors of [5] claim that their methodology could be applied to data with both experimental noise and significant levels of sparsity, as is typically retrieved by APT experiments, but without showcasing any such instances. Briefly, diffraction images are diffraction patterns generated by simulating the results of an X-ray diffraction experiment. In particular, they create the interference pattern that is generated when a series of waves encounter a crystal lattice and either pass through unobstructed or encounter an atom and bend around the atom.

Here we propose a machine-learning approach, a materials fingerprint, to classify the crystal structure of a material by looking at local atomic neighborhoods through the lens of topological data analysis (TDA). TDA is a field that uses topological features within data for machine learning tasks [32, 33, 34]. It has found other applications in materials science, such as the characterization of amorphous solids [35], equilibrium phase transitions [36], and similarity of pore-geometry in nanomaterials [37]. Our motivation is to encode the geometric peculiarities of HEAs by considering atomic positions within a neighborhood and looking at the neighborhood's topology. Key differences between atomic neighborhoods are encoded in the empty space, e.g., holes and voids, between atoms, as well as clusters of atoms in the neighborhood. These identifying topological features of an atomic neighborhood can be calculated through the concept of homology, which is the mathematical study of 'holes' in different dimensions and differentiate the shape and structure of the neighborhoods. Extracting this homological information from each atomic neighborhood, we can distinguish between the different lattice structures that we consider; figure 1 shows idealized versions of these crystal structures. A typical lattice retrieved from an APT experiment is in figure 1(d).

Using these topologically-derived features, we are able to classify the crystal structure of HEAs from the APT data with accuracy approaching 100%. To test the robustness of our proposed method, we combine levels of sparsity and noise on synthetic data and find our method accurately classifies the crystal structure. Our novel methodology couples the power of topological data analysis to extract the intrinsic topology of

these crystal lattices with a machine learning classification scheme to differentiate between lattice structures and classify them with a high degree of precision.

The outline of this paper is as follows. In Section 2 we describe the APT experimental process and the details related to the analysis of the HEAs that we consider. Section 3 provides details of the classification model for recognizing crystal structures. Numerical results are presented in section 4 and we conclude with discussion in section 5.

2. Atom Probe Tomography

In this section we discuss the APT experimental process and the postprocessing employed to create the data. Furthermore, we discuss the resulting data and its characteristics.

2.1. APT Process

APT was conducted using a Local Electrode Atom Probe (LEAP) 4000 XHR instrument at the Center for Nanophase Materials Sciences of the Oak Ridge National Laboratory [31, 38]. The process systematically evaporates ions from a specimen's hemispherical surface using voltage or laser pulses. A position sensitive detector collects the ions, and the timing between the pulse and detection events gives the time-of-flight, which identifies each species based on unique mass-to-charge ratios. A reconstruction algorithm is used to create a tomographic dataset from the x, y detector data and the sequence of detection gives the z-dimension in the reconstruction. Sample specimens for APT experiments are typically sharp, conical tips with a maximum diameter of less than 100 nm and a length of several hundred nanometers typically. Thus all APT experiments investigate nanoscale structures and samples that contain nanoparticles embedded in a matrix can be examined as well as layered heterostructures.

2.2. APT Data

For our problem, the data consists of spatial coordinates of approximately 10^8 atoms with elemental type [25], constituting a highly-disordered metallic alloy that is composed of BCC or FCC lattice structures. The sample [12] was chosen because it has been previously well-characterized. This alloy consists of three phases, a Cu-rich FCC phase, an Fe-Cr rich BCC phase, and a remaining phase that incorporates all six elements, though the proportions of Cu, Fe, and Cr are depleted due to accumulation in the other phases. Importantly all three phases are present in the APT sample. When viewing the entire data set with atoms identified by color, some nanoscale information is immediately evident. The eye perceives elemental segregation of the Cu-rich and Fe-Cr rich phases into nanoscale domains. The orange copper-rich area is especially evident, as seen in figure 2(a). However, one cannot infer any meaningful structure at a finer scale when viewing the entire dataset from a typical APT experiment and further analysis requires that individual atomic neighborhoods be extracted from the larger sample. Viewing each neighborhood individually, figure 2(b), we can see that they contain a wealth of information about the shape of the material under investigation, despite the noise and sparsity present in a typical APT experiment.

3. Methods

In this section we give the mathematical background necessary for our method, detailed introductions can be found in [39, 40].

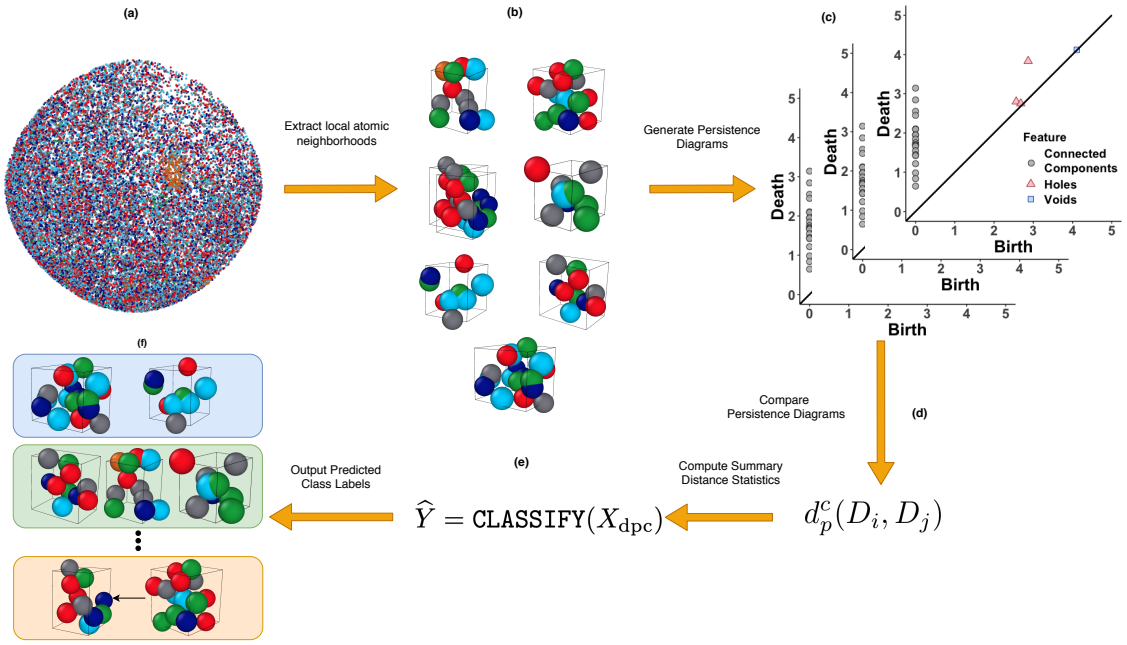

Figure 2: Flowchart of the materials fingerprinting methodology. (a) The APT data is processed as outlined in section 2.1. (b) Individual atomic neighborhoods are extracted from an APT dataset as described in section 2.2. (c) We create a collection of persistence diagrams, each diagram associated with an atomic neighborhood, as explained in section 3.1. (d) Similarity metrics between these persistence diagmras are computed via the d_p^c-distance as defined in equation (3.1). (e) We create a feature matrix composed of the summary statistics of these distances, which is used as input in algorithm 1 to classify the persistence diagrams. (f) Output from algorithm 1 classifying the structures under investigation, section 3.

3.1. Topological Data Analysis

To extract the salient topological information from the atomic neighborhoods, we turn to topological data analysis, particularly persistent homology. Persistent homology describes connectedness and void space present within an object and allows one to infer global properties of space from local information [41]. Instead of considering only clusters of atoms, homology also incorporates information about the regions enclosed by the atoms. This approach yields topological features of the data in different homological dimensions. In the case of these atomic neighborhoods created by APT experiments, 0−dim homological features are connected components, 1−dim homological features are holes, and 2−dim homological features are voids, 2-dim holes, i.e., the space enclosed by a sphere.

To study the persistent homology of atomic structures extracted by HEAs, such as the atomic neighborhoods in figure 2(b), we create spheres of increasing radii around each atom in a neighborhood, detect when homological features emerge and disappear, and record these radii in a persistence diagram, see figures 2(c) and 3(e).Taking the atoms' spatial positions in the xyz-coordinate system recovered by the APT experimental process, we begin by considering a sphere of radius ϵ centered at each atom, see figure 3(a). The algorithm starts at radius $\epsilon = 0$ and this is the reason why all points start at 0 in the persistence diagram associated with clusters and connected components (grey circles in figures 2(c) and 3(e)). Indeed, all atoms within a structure are initially treated as different clusters. Increasing the radii, the algorithm starts clustering atoms together by examining if their spheres intersect at a certain radius. If they do, the these atoms form a cluster and that signifies the 'death' of the members of clusters as being considered separately. Meanwhile, as spheres

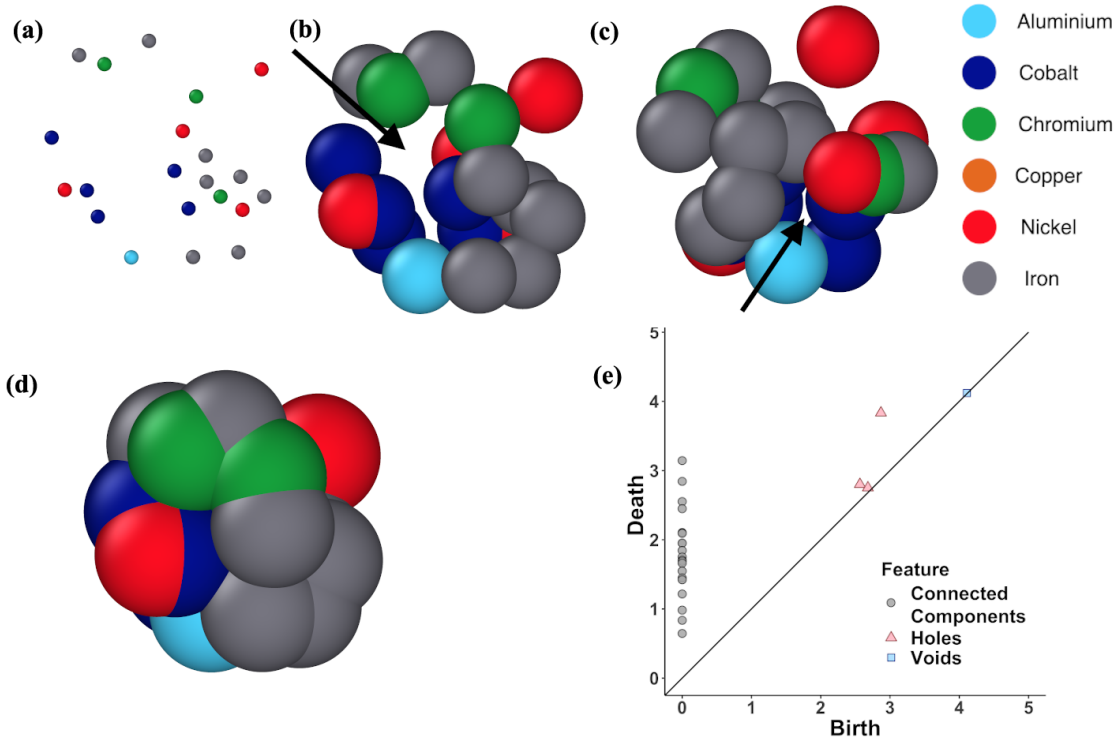

Figure 3: Atomic neighborhood from an APT experiment [12] with the alloy $Al_{1.3}CoCrCuFeNi$. The atomic type is illustrated by the color, and is visualized with [30]. (a) Shows each atom in a neighborhood as a point cloud in $\mathbb{R}^3$. We begin by drawing a radius centered at each atom. As the radius of these spheres increases in (b), a 1−dim hole forms in the atomic structure. Increasing the radii further, in (c) the formation of a 2−dim hole, a void, is evident. Continuing to increase the radii, in (d) the radii have increased such that all atoms form one cluster. The persistence diagram for this structure is shown in (e). In the persistence diagram, the birth and death axes denote the emergence or disappearance of topological features as the radii of the spheres centered on each atom increase and start to intersect.

grow holes and voids (2-dim holes) are created, see figures 3(b) and 3(c). By the same token, these holes and voids get filled in due to increasing the radii, and are represented in a persistence diagram by their death time (radius-wise). Indeed, such topological features are recorded in a persistence diagram using a different label (color). Eventually, at some radius, all spheres will intersect, which means that all atoms belong to the same cluster and any hole or void has been covered. This yields the end of the algorithm for creating a persistence diagram. These homological features summarized in a persistence diagram capture information about the shape of the neighborhood itself. This type of multiscale analysis is key to bypassing the noise and sparsity present in the data and to extract meaningful details about the configuration of each neighborhood. For example, the corresponding diagram for the atomic neighborhood in figure 3(a) is shown in figure 3(e). The persistence diagram encodes information about the structure of each neighborhood by providing insight about the number of atoms, the size and distance among atoms, possible configuration of the faces, and 3−dimensional structure. The persistence diagram then functions as a proxy for the APT data by reducing an atomic neighborhood to its most pertinent qualities.

As the extracted persistence diagrams generated by APT experiments summarize the shape peculiarities of each atomic neighborhood, different types of lattice structures yield persistence diagrams with various identifying features [42]. Indeed, examining the homological features, we see the essential structural

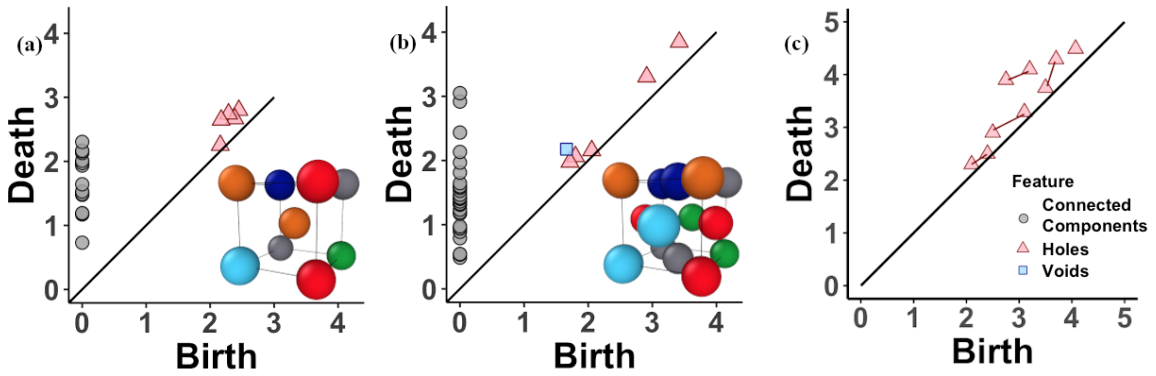

Figure 4: Sample persistence diagrams of a material from the APT analysis of the alloys $Al_{1.3}CoCrCuFeNi$ and $Al_{0.3}CoCrFeNi$ for two of the lattice types considered here: BCC (a) and FCC (b), respectively [12, 31]. Notice the distinguishing 2−dim feature, the blue square, in the diagram derived from an FCC lattice. Additionally, the diagram generated from the BCC structure has fewer 0−dim features. (c) The d_p^c metric computes the distance between two persistence diagrams generated by atomic neighborhoods, both containing 1-dim features, denoted by the red triangles. The d_p^c metric measures the distance between the diagrams by first finding the best matching between points, given by the lines between the triangles. Any unmatched points, e.g., the remaining triangle, are then penalized by the constant term c. The birth and death axes denote the emergence or disappearance of topological features, as a function of distance between atoms in a neighborhood.

differences between crystal lattices in different dimensions. Consider figure 4, which displays the difference between persistence diagrams for BCC and FCC structures. From the viewpoint of topology, the inside of an FCC cell contains a void, whereas the BCC cell does not, thus yielding an important contrast. In the case of noiseless and complete data, the presence of a void separates the BCC and FCC cells when juxtaposing their crystal structures, as we see in the insets of figure 4 (a,b). The persistence diagrams capture differences in (i) the number of neighbors (8 for BCC and 12 for FCC), (ii) the spacing between neighbors, i.e., density, and (iii) the arrangement of neighbors.

3.2. Persistence Diagram Similarity Metric

Different crystal structures produce different size point clouds [42]. To properly account for differences in the number of points when comparing two persistence diagrams, we employ the d_p^c distance, introduced in [32]. For a given configuration, the persistence diagram can be compared to a reference persistence diagram via a similarity metric, for BCC and FCC structures as an example. Suppose $D_1 = \{d_1^1, \ldots, d_n^1\}$ and $D_2 = \{d_1^2, \ldots, d_m^2\}$ are two persistence diagrams associated with two local atomic neighborhoods such that $n \leq m$. Let $c > 0$ and $1 \leq p < \infty$ be fixed parameters. Then the d_p^c distance between D_1 and D_2 is

$$d_p^c(D_1, D_2) = \left(\frac{1}{m} \left(\min_{\pi \in \Pi_m} \sum_{i=1}^{n} \min(c, \|d_i^1 - d_{\pi(i)}^2\|_\infty)^p + c^p |m - n| \right) \right)^{\frac{1}{p}} \tag{3.1}$$

where Π_m is the set of permutations of $(1, \ldots, m)$. If $n > m$, define $d_p^c(D_1, D_2) := d_p^c(D_2, D_1)$.

This distance matches points between the persistence diagrams being compared, and those that are unmatched are penalized by a regularization term c. Figure 4(c) shows an example of how the distance between two persistence diagrams is computed. We first find the optimal matching, denoted by the red lines between triangles. This matching between points corresponds to the summation term in the distance. If the matched distance is greater than c, then we add c to the matching distance, otherwise, we add the distance

between matched points. The unmatched 1-dim feature, denoted by the red triangle, is penalized by the regularization term c in the second part of the definition. In developing the materials fingerprint, we compare persistence diagrams with respect to 0, 1, and 2−dim homological features, i.e., connected components, holes, and voids, employing this distance. We then compute summary statistics (mean, variance) from these distances to create features for the classification algorithm.

3.3. Classification Model

We write D_i as the persistence diagram generated by atom positions in an atomic neighborhood retrieved by the APT experiment as seen in figure 2. Note that the number of atoms in a neighborhood is not constant, but varies between atomic neighborhoods in a sample. For the multiclass classification problem, we are interested in modeling the conditional probability $\pi(Y = \ell \mid X)$ for a given input X, which encapsulates features of persistence diagrams and a class label $Y = \ell$. We write the classification model as a generalized additive regression model [43, 44]. Choosing this type of model gives us the flexibility to let our data determine the correct functional form, as opposed to imposing a linear model as in traditional logistic regression. Accordingly, an L-class model is written

$$\begin{aligned}
\log\left(\frac{\pi(Y=1\mid X)}{\pi(Y=L\mid X)}\right) &= \alpha_1 + F_1(X),\\
\log\left(\frac{\pi(Y=2\mid X)}{\pi(Y=L\mid X)}\right) &= \alpha_2 + F_2(X),\\
&\vdots\\
\log\left(\frac{\pi(Y=L-1\mid X)}{\pi(Y=L\mid X)}\right) &= \alpha_{L-1} + F_{L-1}(X),
\end{aligned}$$

where $F_i(X) = \sum_{j=1}^{P} \alpha_j f_j(X)$ is a linear combination of smooth functions f_j. Here $\mathbf{X} \in \mathbb{R}^{N\times P}$ and $N = \sum_{i=1}^{L} N_i$ is such that for $1 \le i \le N$ an arbitrary row of $\mathbf{X}$ is

$$\mathbf{X}_i = (\mathbb{E}^0_{i,\lambda_1}, \mathbb{E}^1_{i,\lambda_1}, \mathbb{E}^2_{i,\lambda_1}, \mathrm{Var}^0_{i,\lambda_1}, \mathrm{Var}^1_{i,\lambda_1}, \mathrm{Var}^2_{i,\lambda_1}, \ldots, \mathbb{E}^0_{i,\lambda_L}, \mathbb{E}^1_{i,\lambda_L}, \mathbb{E}^2_{i,\lambda_L}, \mathrm{Var}^0_{i,\lambda_L}, \mathrm{Var}^1_{i,\lambda_L}, \mathrm{Var}^2_{i,\lambda_L}), \quad (3.2)$$

where $\mathbb{E}^k_{i,\lambda_\ell} = \frac{1}{N_\ell}\sum_{j=1}^{N_\ell} d_p^c(D_i^k, D_j^k)$ and $\mathrm{Var}^k_{i,\lambda_\ell} = \frac{1}{N_\ell - 1}\sum_{j=1}^{N_\ell} (d_p^c(D_i^k, D_j^k) - \mathbb{E}^k_{i,\lambda_\ell})^2$ are the mean and variance respectively of the d_p^c distance, equation (3.1), between any diagram D_i^k and the collection of all persistence diagrams in the class $\lambda_\ell \in \Lambda$, $1 \le \ell \le L$ and homological dimension $k = 0, 1, 2$. The pseudocode for our algorithm is presented in algorithm 1 and is visually represented in figure 2.

3.4. Computational and Storage Considerations

Computing entries of the feature matrix $\mathbf{X}$, equation (3.2), requires computing the mean and variance of d_p^c distances with k−dim persistence homology, $k = 0, 1, 2$. For example, in the case of binary classification between BCC and FCC lattice types, with N_1 and N_2 neighborhoods respectively, for each BCC persistence diagram, each $\mathbb{E}^k_{i,\lambda_1}$ computation requires N_1 steps and for FCC, it is N_2 steps. Similarly, computing the variance accurately in a numerically stable fashion, e.g., when the size of the dataset is large and the variance is small, each BCC diagram takes $2 \times N_1$ steps for the two pass algorithm [45]. In total, each row of $\mathbf{X}$ has complexity $O_i(N_1, N_2) = 9 \times (N_1 + N_2)$ and the entire feature matrix ends up with quadratic complexity: $O(N_1, N_2) = 9 \times (N_1 + N_2)^2$. With the atomic counts on the order of hundreds of thousands: $N_1, N_2 \approx O(10^5)$, the quadratic component clearly dominates with 10^{10} computational steps.

Algorithm 1 Materials Fingerprinting

Training Step

Read in labeled data (training set) with L classes and compute persistence diagrams in the training set $\mathcal{D}_{train}$, which has N_ℓ diagrams from the ℓth class, and set $N = \sum_{\ell=1}^{L} N_\ell$.

Read in response vector $Y = (1 \cdot \mathbf{1}, \ldots, \ell \cdot \mathbf{1}, \ldots, L \cdot \mathbf{1})^T$ where $\mathbf{1}$ is a vector of 1's in $\mathbb{R}^{N_\ell}$.

for $i = 1, \ldots, N$ **do**

Compute feature matrix $\mathbf{X}$ according to equation (3.2)

$$\mathbf{X}_i = (\mathbb{E}^0_{i,\lambda_1}, \mathbb{E}^1_{i,\lambda_1}, \mathbb{E}^2_{i,\lambda_1}, \mathrm{Var}^0_{i,\lambda_1}, \mathrm{Var}^1_{i,\lambda_1}, \mathrm{Var}^2_{i,\lambda_1}, \ldots, \mathbb{E}^0_{i,\lambda_L}, \mathbb{E}^1_{i,\lambda_L}, \mathbb{E}^2_{i,\lambda_L}, \mathrm{Var}^0_{i,\lambda_L}, \mathrm{Var}^1_{i,\lambda_L}, \mathrm{Var}^2_{i,\lambda_L})$$

where

$$\mathbb{E}^k_{i,\lambda_\ell} = \frac{1}{N_\ell} \sum_{j=1}^{N_\ell} d_p^c(D_i^k, D_j^k), \quad \mathrm{Var}^k_{i,\lambda_\ell} = \frac{1}{N_\ell - 1} \sum_{j=1}^{N_\ell} (d_p^c(D_i^k, D_j^k) - \mathbb{E}^k_{i,\lambda_\ell})^2,$$

for $\lambda_\ell \in \Lambda, k \in \{0, 1, 2\}$.

end for

$C(\mathbf{X}) = \texttt{ADABOOST}(\mathbf{X}, Y)$ ▷ Obtain a classification rule C from the AdaBoost ensemble classification algorithm

Testing Step

Read in unlabeled APT point cloud data and compute persistence diagrams $\mathcal{D}_{test} = \{\widehat{D}_j\}_{j=1}^J$.

for $j = 1, \ldots, J$ **do**

Compute

$$\widehat{\mathbf{X}}_j = (\widehat{\mathbb{E}}^0_{j,\lambda_1}, \widehat{\mathbb{E}}^1_{j,\lambda_1}, \widehat{\mathbb{E}}^2_{j,\lambda_1}, \widehat{\mathrm{Var}}^0_{j,\lambda_1}, \widehat{\mathrm{Var}}^1_{j,\lambda_1}, \widehat{\mathrm{Var}}^2_{j,\lambda_1}, \ldots, \widehat{\mathbb{E}}^0_{j,\lambda_L}, \widehat{\mathbb{E}}^1_{j,\lambda_L}, \widehat{\mathbb{E}}^2_{j,\lambda_L}, \widehat{\mathrm{Var}}^0_{j,\lambda_L}, \widehat{\mathrm{Var}}^1_{j,\lambda_L}, \widehat{\mathrm{Var}}^2_{j,\lambda_L})$$

where

$$\widehat{\mathbb{E}}^k_{j,\lambda_\ell} = \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} d_p^c(\widehat{D}_j^k, D_n^k), \ \widehat{\mathrm{Var}}^k_{j,\lambda_\ell} = \frac{1}{N_\ell - 1} \sum_{n=1}^{N_\ell} (d_p^c(\widehat{D}_j^k, D_n^k) - \widehat{\mathbb{E}}^k_{j,\lambda_\ell})^2,$$

for $\lambda_\ell \in \Lambda, k \in \{0, 1, 2\}$.

end for

Classify unlabeled APT data

$\widehat{Y} = C(\widehat{\mathbf{X}})$ ▷ Yields class labels for $\mathcal{D}_{test}$ as $\hat{Y} \in \{1, \ldots, \ell, \ldots, L\}^J$.

Each of these steps requires the d_p^c distance computation given by equation (3.1), which is computationally non-trivial for the majority of the diagrams due to the identification of the optimal permutation between the diagrams being compared. In order to reduce the total elapsed time of the computation, we used over 1000 x86 cores that ranged from Intel Westmere to Intel Skylake, ranging in cores per socket from 8 to 36 with up to 72 cores per node. Additional speedup of about 20% came from porting the code for computing the feature matrix from Python to C. The python code is publicly available at `https://github.com/maroulaslab/Materials-Fingerprinting`.

4. Numerical Experiments

We present here the outcome of algorithm 1 in both synthetic and real experimental data as well as provide a sensitivity analysis. We first present results of our fingerprinting process in different scenarios with synthetic data to test the robustness of our method. We consider synthetic APT data with various levels of sparsity and additive Gaussian noise, $\mathcal{N}(0, \sigma^2)$, as in real APT experimental data. In each of the experiments presented, we perform 10-fold cross validation on the entire dataset to control for overfitting of the model, randomly splitting the dataset into 10 partitions. For each partition, we create a classification rule from the other 9 partitions, and use the remaining one as a test set. Our accuracy, defined here as (1 - Misclassification rate), is recorded for each partition as it is used as the test set. The reported accuracy rate is the mean accuracy over all 10 partitions. The hyperparameters c, p were set to the same values across all experiments, $c = (1, 0.05, 1.05)$ and $p = 2$, to provide a fair basis for comparison, and were selected by a grid search to provide the highest accuracy score in the binary classification problem with 67% missing data and $\mathcal{N}(0, 1)$ additive noise. A previous work [42], discusses the role of c and choosing this parameter.

4.1. Synthetic APT Data

We first present results of our fingerprinting process in different scenarios with synthetic data to test the robustness of our method. First, we test with combinations of noise and sparsity that we expect to see in real APT data. Next, we examine the effect of class imbalances on the accuracy of our methodology in the binary classification case of BCC and FCC materials. As a final experiment with synthetic data, we repeat the scenario of varying the concentration between BCC and FCC structures, but augment the data set with a constant number of HCP lattice types. We observe the methodology is robust against different levels of noise and sparsity in the case of the binary classification problem. When the HCP structures are introduced into the dataset, the accuracy decreases, due to the similarity of the FCC and HCP structures, especially in the presence of additive noise and sparsity that we consider. These results are presented in tables 1 to 3.

4.2. Sensitivity Analysis

To understand the effect of different levels of noise and sparsity in the data, the materials fingerprint was applied to synthetic data having different levels of sparsity and noise, similar to those values found in real APT data. For each combination presented, we the dataset was composed of 400 structures, split evenly between BCC and FCC types. We observe perfect accuracy in the case of complete, noiseless data, as these lattice types differ in both their geometry and atomic density. As the data becomes increasingly degraded, the accuracy correspondingly decreases, but does not fall below 90% in this analysis. Table 1 summarizes these results. We do observe a relative decrease in accuracy with 50% sparsity and $\mathcal{N}(0, 0.75^2)$ added noise. We attribute this decrease to the choice of c and p for the distance computations. Indeed, for all the experiments presented herein, we used the same values of c and p. We may further optimize these parameters to produce higher accuracy for each combination of noise and missing data considered, at the risk of over-fitting for a specific dataset.

Table 1: Mean 10-fold cross validation accuracy, for synthetic APT data with different percentages of atoms removed and $\mathcal{N}(0, \sigma^2)$ added noise.

Sparsity \ Std. Dev.	$\sigma = 0$	$\sigma = 0.25$	$\sigma = 0.75$	$\sigma = 1$
0%	100%	99.67%	98%	97.67%
33%	100%	99.32%	96.67%	94.67%
50%	97.33%	100%	92.67%	94%
67%	98.67%	100%	99.33%	92%

Table 2: Mean 10-fold cross validation accuracy, for synthetic APT data with $\mathcal{N}(0, \sigma^2)$ added noise and 50% missing, the dataset is comprised of 5,000 configurations in each experiment. The proportion of BCC structures are given, and varied between experiments. The proportion of FCC configurations is 1-BCC%.

BCC proportion	10%	25%	40%	60%	75%	90%
Std. Dev	Accuracy					
$\sigma = 0.25$	96.72%	92.32%	88.56%	88.48%	90.24%	94.24%
$\sigma = 0.5$	99.96%	99.84%	100%	100%	100%	100%
$\sigma = 0.75$	95.76%	89.86%	82.88%	82.24%	85.84%	95.04%
$\sigma = 1$	94.72%	85.76%	81.44%	83.2%	84.08%	94%

4.2.1. Imbalanced Classification

Continuing our study of the binary classification problem, we investigated the effect of varying the proportion of BCC vs. FCC lattice structures had on the resulting classification accuracy. We considered the same combinations of sparsity and additive noise as in section 4.2, but we varied the proportion of BCC structures in the entire dataset between 10% and 90%. The remaining percentage was composed of FCC structures so that the total number of structures was 5,000. We observe a level of accuracy in this setting similar to those observed in the previous experiment; these accuracy scores are presented in table 2. We observe that the classification scheme is robust against not only the perturbations and missing data expected from an APT experiment, but class imbalance as well.

4.2.2. Multi-class Classification

As a final experiment, to the previous setting of varying the proportion of BCC vs. FCC structures, we add a constant number of HCP structures to the data set. All lattice structures in this experiment are perturbed by Gaussian noise with a standard deviation of 0.25, as the noise was found in a previous study to follow a narrowly peaked distribution, as opposed to a wide Gaussian distribution [29]. From each of these datasets, we removed γ% of the atoms. The results of this experiment are in table 3. In this scenario, the primary challenge is to correctly identify the FCC and HCP lattices. While these two structures are distinct, they have the same density, i.e., the same number of atoms per unit volume, and only have a subtle variation in their identifying geometry. Indeed, there is a non-trivial decrease in accuracy when the HCP lattices are introduced into the dataset. Specifically, the accuracy declines as the proportion of FCC structures increases relative to the number of HCP lattice types and is the dominant class represented in the dataset. When the BCC proportion comprises 10% of the dataset, the proportion of FCC to HCP lattices is approximately 2:1,

Table 3: Mean 10-fold cross validation accuracy, classifying synthetic APT data with $\mathcal{N}(0, 0.25^2)$ added noise and proportion $\gamma \in (0, 1)$ missing. We consider three classes, BCC, FCC, and HCP structures, in this synthetic APT dataset. We varied the proportion of 5,000 configurations between BCC and FCC lattices. The BCC proportion of these structures are given and the fraction of FCC configurations is 1-BCC%. To these 5,000 structures we added a constant 2,500 HCP lattice structures in each instance.

BCC proportion	10%	25%	40%	60%	75%	90%
Proportion Missing	Accuracy					
$\gamma = 0.33$	60.67%	69.84%	84.84%	86.51%	78.88%	88.39%
$\gamma = 0.50$	68.33%	74.76%	85.16%	88.13%	82.40%	89.45%

and the classifier's accuracy is decreased as compared to settings with less class imbalance in the dataset.

4.3. APT Experimental Data

We now turn to our original problem of determining the local lattice structure of an HEA from the experimental APT data. We apply our materials fingerprinting method to the APT experimental data from two HEAs, $Al_{1.3}$CoCrCuFeNi and $Al_{0.3}$CoCrFeNi (FCC). Recalling section 2.2, the former has both BCC and FCC phases, while the was determined to be FCC through XRD experiments [31]. The challenge is to uncover the true atomic-level structure amid the noise and missing data. Using our materials fingerprinting methodology, we are able to classify the lattice structure of 200,000 atomic neighborhoods, split evenly between BCC and FCC lattice types, from these APT datasets at **99.97**% accuracy with 10-fold cross validation.

5. Discussion

We have described materials fingerprinting, a topologically-based methodology for classifying the crystal structure of the HEA APT data with near-perfect accuracy especially in the binary case. Starting from a collection of atomic neighborhoods generated by an APT experiment, we extract the fundamental topology of the structure and record the information in a persistence diagram. These diagrams succinctly encode the essential topology of an atomic neighborhood over different length scales in various dimensions. It is by computing the persistent homology of the data that we are able to see through the noise and fill in the sparsity to see where these lattice structures are connected and where they are not. Our materials fingerprinting methodology uses the mean and variance of the d_p^c distance between persistence diagrams to create input for a machine learning algorithm. This distance not only measures differences in the diagrams but accounts for different numbers of points between diagrams being compared. This latter point is salient, as BCC and FCC unit cells each contain a different number of atoms, and this distinction must be taken into account. Basing our materials fingerprint on topological features in conjunction with the number of atoms in each neighborhood, we represent the necessary topological and numeric information required to differentiate between the lattice structures considered here, with the appropriate choice of metric. Indeed, by adopting this point of view, we are able to qualitatively retain the essential geometric information of these crystal structures and use this information to predict with greater than 99% accuracy the crystal structure of real APT data.

The impact of the present work is two-fold. First, the input data to our algorithm is point clouds generated by HEAs resulting from APT experiments. The process can be generalized to other lattice types by incorporating additional crystal structures into the materials fingerprint training set. Indeed, the methodology described herein does not depend on the labels of the data. It takes in the materials data and creates the information-rich persistence diagrams, from which we examine homological differences

between the diagrams in various dimensions. The data analysis can be performed on multiphase samples, although the characterization of individual configurations may need be to be first preceded by classification of domains based on compositional differences, for example. An alternative for comparisons between a multitude of structures is outlined in [46], in which different topological descriptors are invoked that consider the electronegativity of the atoms as a feature when creating the persistence diagrams. Such a methodology may be used in conjunction with a previous work [47] that identifies a mapping between the APT data and a known crystal structure, to aid researchers in understanding the local structure of materials characterized through the APT process.

Acknowledgments

The authors are grateful to two anonymous referees for helpful comments and suggestions that substantially improved the manuscript. The APT experiments were conducted at the Oak Ridge National Laboratory's Center for Nanophase Materials Sciences (CNMS), which is a U.S. DOE Office of Science User Facility. The authors would like to thank Jonathan Poplawsky for insightful discussions about the APT method. V. M. is grateful for support from ARO Grant # W911NF-17-1-0313 and the NSF DMS-1821241. D.K. and V.M. are grateful for support from a UTK Seed grant. A.S., C.M., and F.N, acknowledge the Mathematics Department of the University of Tennessee, where A.S. and C.M. conducted this research as part of their Ph.D studies and F.N. was a Post-Doctoral research associate. This research used resources of the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable,world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

References

[1] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature 559 (2018) 547.

[2] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, O. Levy, The high-throughput highway to computational materials design, Nature materials 12 (2013) 191.

[3] N. Islam, W. Huang, H. L. Zhuang, Machine learning for phase selection in multi-principal element alloys, Computational Materials Science 150 (2018) 230–235.

[4] Q. Zhou, P. Tang, S. Liu, J. Pan, Q. Yan, S. Zhang, Learning atoms for materials discovery, Proceedings of the National Academy of Sciences 115 (2018) E6411–E6417.

[5] A. Ziletti, D. Kumar, M. Scheffler, L. M. Ghiringhelli, Insightful classification of crystal structures using deep learning, Nature communications 9 (2018) 2775.

[6] A. Agrawal, A. Choudhary, Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science, Apl Materials 4 (2016) 053208.

[7] N. Laboratory, Nomad, 2015. URL: http://nomad-coe.eu.

[8] C. M. Rost, E. Sachet, T. Borman, A. Moballegh, E. C. Dickey, D. Hou, J. L. Jones, S. Curtarolo, J.-P. Maria, Entropy-stabilized oxides, Nature communications 6 (2015) 8485.

[9] Y. Zhang, T. T. Zuo, Z. Tang, M. C. Gao, K. A. Dahmen, P. K. Liaw, Z. P. Lu, Microstructures and properties of high-entropy alloys, Progress in Materials Science 61 (2014) 1–93.

[10] Y. Zhang, Y. J. Zhou, J. P. Lin, G. L. Chen, P. K. Liaw, Solid-solution phase formation rules for multi-component alloys, Advanced Engineering Materials 10 (2008) 534–538.

[11] J.-W. Yeh, S.-K. Chen, S.-J. Lin, J.-Y. Gan, T.-S. Chin, T.-T. Shun, C.-H. Tsau, S.-Y. Chang, Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes, Advanced Engineering Materials 6 (2004) 299–303.

[12] L. J. Santodonato, Y. Zhang, M. Feygenson, C. M. Parish, M. C. Gao, R. J. Weber, J. C. Neuefeind, Z. Tang, P. K. Liaw, Deviation from high-entropy configurations in the atomic distributions of a multi-principal-element alloy, Nature communications 6 (2015) 5964.

[13] J.-W. Yeh, Physical metallurgy of high-entropy alloys, Jom 67 (2015) 2254–2261.

[14] Y. Shi, B. Yang, P. K. Liaw, Corrosion-resistant high-entropy alloys: A review, Metals 7 (2017) 43.

[15] B. Gludovatz, A. Hohenwarter, D. Catoor, E. H. Chang, E. P. George, R. O. Ritchie, A fracture-resistant high-entropy alloy for cryogenic applications, Science 345 (2014) 1153–1158.

[16] Z. Lei, X. Liu, Y. Wu, H. Wang, S. Jiang, S. Wang, X. Hui, Y. Wu, B. Gault, P. Kontis, Enhanced strength and ductility in a high-entropy alloy via ordered oxygen complexes, Nature 563 (2018) 546.

[17] Z. Li, K. G. Pradeep, Y. Deng, D. Raabe, C. C. Tasan, Metastable high-entropy dual-phase alloys overcome the strength–ductility trade-off, Nature 534 (2016) 227.

[18] M.-H. Tsai, J.-W. Yeh, High-entropy alloys: a critical review, Materials Research Letters 2 (2014) 107–123.

[19] M. A. Hemphill, T. Yuan, G. Wang, J. Yeh, C. Tsai, A. Chuang, P. K. Liaw, Fatigue behavior of $Al_{0.5}$CoCrCuFeNi high entropy alloys, Acta Materialia 60 (2012) 5723–5734.

[20] Z. Tang, T. Yuan, C. Tsai, J. Yeh, C. D. Lundin, P. K. Liaw, Fatigue behavior of a wrought $Al_{0.5}$CoCrCuFeNi two-phase high-entropy alloy, Acta Materialia 99 (2015) 247–258.

[21] J. Guo, H. Wang, F. von Rohr, Z. Wang, S. Cai, Y. Zhou, K. Yang, A. Li, S. Jiang, Q. Wu, Robust zero resistance in a superconducting high-entropy alloy at pressures up to 190 gpa, Proceedings of the National Academy of Sciences 114 (2017) 13144–13147.

[22] P. Koželj, S. Vrtnik, A. Jelen, S. Jazbec, Z. Jagličić, S. Maiti, M. Feuerbacher, W. Steurer, J. Dolinšek, Discovery of a superconducting high-entropy alloy, Physical Review Letters 113 (2014) 107001.

[23] M. P. Moody, B. Gault, L. T. Stephenson, R. K. Marceau, R. C. Powles, A. V. Ceguerra, A. J. Breen, S. P. Ringer, Lattice rectification in atom probe tomography: Toward true three-dimensional atomic microscopy, Microscopy and Microanalysis 17 (2011) 226–239.

[24] T. F. Kelly, M. K. Miller, K. Rajan, S. P. Ringer, Atomic-scale tomography: A 2020 vision, Microscopy and Microanalysis 19 (2013) 652–664.

[25] M. K. Miller, T. F. Kelly, K. Rajan, S. P. Ringer, The future of atom probe tomography, Materials Today 15 (2012) 158–165.

[26] D. Hicks, C. Oses, E. Gossett, G. Gomez, R. H. Taylor, C. Toher, M. J. Mehl, O. Levy, S. Curtarolo, Aflow-sym: platform for the complete, automatic and self-consistent symmetry analysis of crystals, Acta Crystallographica Section A: Foundations and Advances 74 (2018) 184–203.

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

[28] P. M. Larsen, S. Schmidt, J. Schiøtz, Robust structural identification via polyhedral template matching, Modelling and Simulation in Materials Science and Engineering 24 (2016) 055007.

[29] B. Gault, M. P. Moody, J. M. Cairney, S. P. Ringer, Atom probe crystallography, Materials Today 15 (2012) 378–386.

[30] A. Stukowski, Visualization and analysis of atomistic simulation data with OVITO-the Open Visualization Tool, MODELLING AND SIMULATION IN MATERIALS SCIENCE AND ENGINEERING 18 (2010). doi:{10.1088/0965-0393/18/1/015012}.

[31] H. Diao, D. Ma, R. Feng, T. Liu, C. Pu, C. Zhang, W. Guo, J. D. Poplawsky, Y. Gao, P. K. Liaw, Novel nial-strengthened high entropy alloys with balanced tensile strength and ductility, Materials Science and Engineering: A 742 (2019) 636–647.

[32] A. Marchese, V. Maroulas, Signal classification with a point process distance on the space of persistence diagrams, Advances in Data Analysis and Classification 12 (2018) 657–682.

[33] A. Marchese, V. Maroulas, J. Mike, K-means clustering on the space of persistence diagrams, in: Wavelets and Sparsity XVII, volume 10394, International Society for Optics and Photonics, 2017, p. 103940W.

[34] F. Nasrin, C. Oballe, D. Boothe, V. Maroulas, Bayesian topological learning for brain state classification, in: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), IEEE, 2019, pp. 1247–1252.

[35] Y. Hiraoka, T. Nakamura, A. Hirata, E. G. Escolar, K. Matsue, Y. Nishiura, Hierarchical structures of amorphous solids characterized by persistent homology, Proceedings of the National Academy of Sciences (2016) 201520877.

[36] I. Donato, M. Gori, M. Pettini, G. Petri, S. De Nigris, R. Franzosi, F. Vaccarino, Persistent homology analysis of phase transitions, Physical Review E 93 (2016) 052138.

[37] Y. Lee, S. D. Barthel, P. Dłotko, S. M. Moosavi, K. Hess, B. Smit, Quantifying similarity of pore-geometry in nanoporous materials, Nature Communications 8 (2017) 15396.

[38] W. Guo, D. A. Garfinkel, J. D. Tucker, D. Haley, G. A. Young, J. D. Poplawsky, An atom probe perspective on phase separation and precipitation in duplex stainless steels, Nanotechnology 27 (2016) 254004.

[39] H. Edelsbrunner, J. Harer, Persistent homology-a survey, Contemporary Mathematics 453 (2008) 257–282.

[40] H. Edelsbrunner, J. Harer, Computational Topology: An Introduction, American Mathematical Society, Providence, RI, 2010.

[41] T. Kaczynski, K. Mischaikow, M. Mrozek, Computational homology, volume 157, Springer Science & Business Media, 2006.

[42] V. Maroulas, C. P. Micucci, A. Spannaus, A stable cardinality distance for topological classification, Advances in Data Analysis and Classification (2019) 1–18. doi:10.1007/s11634-019-00378-3.

[43] J. Friedman, T. Hastie, R. Tibshirani, et al., Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors), The Annals of Statistics 28 (2000) 337–407.

[44] T. Hastie, Generalized additive models, 1990.

[45] T. F. Chan, G. H. Golub, R. J. LeVeque, Algorithms for computing the sample variance: Analysis and recommendations, The American Statistician 37 (1983) 242–247.

[46] J. Townsend, C. P. Micucci, J. H. Hymel, V. Maroulas, K. D. Vogiatzis, Representation of molecular structures with persistent homology for machine learning applications in chemistry, Nature communications 11 (2020) 1–9.

[47] A. Spannaus, V. Maroulas, D. J. Keffer, K. J. H. Law, Bayesian point set registration, in: 2017 MATRIX Annals, Springer, 2019, pp. 99–120.