



A stable cardinality distance for topological classification

Vasileios Maroulas¹ · Cassie Putman Micucci¹ · Adam Spannaus¹

Received: 12 December 2018 / Revised: 20 September 2019 / Accepted: 18 November 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

This work incorporates topological features via persistence diagrams to classify point cloud data arising from materials science. Persistence diagrams are multisets summarizing the connectedness and holes of given data. A new distance on the space of persistence diagrams generates relevant input features for a classification algorithm for materials science data. This distance measures the similarity of persistence diagrams using the cost of matching points and a regularization term corresponding to cardinality differences between diagrams. Establishing stability properties of this distance provides theoretical justification for the use of the distance in comparisons of such diagrams. The classification scheme succeeds in determining the crystal structure of materials on noisy and sparse data retrieved from synthetic atom probe tomography experiments.

Keywords Stability · Classification · Persistent homology · Persistence diagrams · Crystal structure of materials

Mathematics Subject Classification 62H30 · 62P30 · 55N99 · 54H99

1 Introduction

A crucial first step in understanding properties of a crystalline material is determining its crystal structure. For highly disordered metallic alloys, such as high entropy

This work has been partially supported by the ARO Grant # W911NF-17-1-0313, the NSF DMS-1821241, and UTK 2019 Research Seed Funding-Interdisciplinary.

✉ Vasileios Maroulas
vmaroula@utk.edu

Cassie Putman Micucci
cputman@vols.utk.edu

Adam Spannaus
aspannaus@gmail.com

¹ Department of Mathematics, University of Tennessee, Knoxville, TN 37996, USA

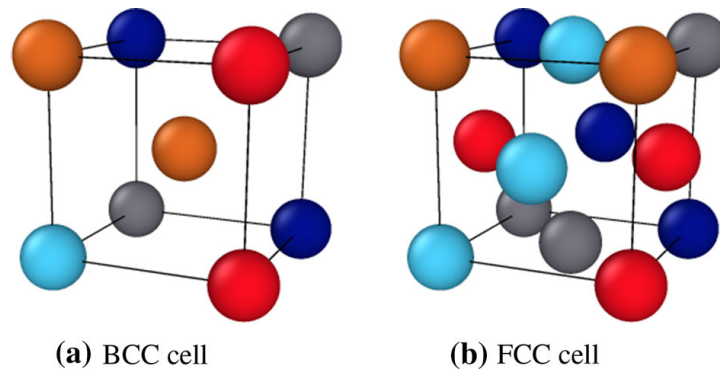


Fig. 1 Example of **a** body-centered cubic, and **b** face-centered cubic unit cells without additive noise or sparsity. Notice there is an essential topological difference between the two structures: The body-centered cubic structure has one atom at its center, whereas the face-centered cubic is hollow in its center, but has one atom in the middle of each of its faces

alloys (HEAs), atom probe tomography (APT) gives a snapshot of the local atomic environment. APT has two main drawbacks: experimental noise and missing data. Approximately 65% of the atoms in a sample are not registered in a typical experiment, and those atoms that are captured have their spatial coordinates corrupted by experimental noise. As noted by Kelly et al. (2013) and Miller et al. (2012), APT has a spatial resolution approximately the length of the unit cell we consider, as seen in Fig. 1. Hence the process is unable to see the finer details of a material, making the determination of a lattice structure a challenging problem. Existing algorithms for detecting the crystal structure (Chisholm and Motherwell 2004; Hicks et al. 2018; Honeycutt and Andersen 1987; Larsen et al. 2016; Moody et al. 2011; Togo and Tanaka 2018) are not able to establish the crystal lattice of an APT dataset, as they rely on symmetry arguments. Consequently, the field of atom probe crystallography, i.e., determining the crystal structure from APT data, has emerged in recent years (Gault et al. 2012 and Moody et al. 2011). These algorithms rely on knowing the global lattice structure *a priori* and aim to determine local small-scale structures within a larger sample. For some materials this information is readily known, for others, such as HEAs, the global structure is unknown and must be inferred. A recent work by Ziletti et al. (2018) proposes a machine-learning approach to classifying crystal structures of a noisy and sparse materials dataset, without knowing the global structure *a priori*. The authors employ a convolutional neural network for classifying the crystal structure by looking at a diffraction image, a computer-generated diffraction pattern. The authors suggest their method could be used to determine the crystal structure of APT data or other noisy and sparse data from materials science. However, the synthetic data considered in Ziletti et al. (2018) is not a realistic representation of experimental APT data, where about 65% of the data is missing (Santodonato et al. 2015) and is corrupted by more observational noise (Miller et al. 2012). Most importantly, their synthetic data is either sparse or noisy, not a combination of both. We consider a combination of noise and sparsity, such as is the case in real APT data.

In this work, we provide a machine learning approach to classify the crystal structure of a noisy and sparse materials dataset. Specifically, we consider materials that are either body-centered cubic (BCC) or face-centered cubic (FCC), as these lattice

structures are the essential building blocks of HEAs (Zhang et al. 2014) and have fundamental differences that set them apart in the case of noise-free, complete materials data. The BCC structure has a single atom in the center of the cube, while the FCC has a void in its center but has atoms on the center of the cubes' faces, see Fig. 1. These two crystal structures are distinct when viewed through the lens of Topological Data Analysis (TDA). Differentiating between the holes and connectedness of these two lattice structures allows us to create an accurate classification rule. This fundamental distinction between BCC and FCC point clouds is captured well by topological methods and explains the high degree of accuracy in the classification scheme presented herein.

TDA provides input features for machine learning algorithms, as well as a useful toolbox for classification. Several authors have used TDA on real-world problems, see Carlsson et al. (2005), Edelsbrunner et al. (2000), Marchese and Maroulas (2016), Marchese et al. (2017), Wasserman (2018), Zomorodian and Carlsson (2005), Maroulas et al. (2019a,b) and the references therein. Persistent homology, which measures changes in topological features over different scales, is the main framework considered by these authors.

Persistent homology is applicable to classification problems as it studies and differentiates holes within data as viewed in different dimensions, e.g., the space enclosed by a loop is a one-dimensional hole. Overall, persistent homology provides a summary of the connectedness and holes (empty space in atomic cells) of data, which indirectly gives information about the shape of the data as well. Indeed, persistent homology records when different homological features emerge and vanish in the data. This analysis quantifies the significance of a homological feature and provides a tool to contend with noisy data. The appearance and disappearance of each homological feature is calculated and recorded in a persistence diagram. Persistence diagrams yield topological summaries of the persistent homology of a dataset and are rich sources of detail about underlying topological features. The diagrams could be used in distance-based classifiers (Marchese and Maroulas 2018; Carriere et al. 2017) or vectorized and input into standard classification algorithms, such as support vector machines (Adams et al. 2017; Bubenik 2015).

Distances on the space of persistence diagrams yield a means of comparison between diagrams. The Wasserstein and bottleneck distances compute the cost of the optimal matching between the points in each persistence diagram, while allowing matching to additional points on the diagonal to allow for cardinality differences and to prove stability properties as in Cohen-Steiner et al. (2007). Motivated by Marchese and Maroulas (2018), we consider here the d_p^c distance, a distance on the space of persistence diagrams. This distance employs the cardinality of the persistence diagrams, as well as distances between points in the diagrams. It calculates the cost of an optimal matching between the persistence diagrams without any points added to the diagonal. A regularization term then considers the cardinality differences between persistence diagrams.

The stability of the d_p^c distance is also verified in this paper. This property guarantees that when the distances between point clouds go to zero, the distances between the associated persistence diagrams go to zero as well. Another formulation of this stability is given in Chazal et al. (2014); using a related approach, we show continuity of the

mapping of point cloud to persistence diagram under the d_p^c distance. This analysis provides insight into how the cardinality of the diagrams changes with the size of the input point clouds. Additionally, using statistics on the diagram's cardinality generates corresponding prediction intervals, which give probabilistic bounds on the d_p^c distances between persistence diagrams. The idea is that point clouds generated from the same process have small variability with respect to cardinality of the persistence diagrams.

The contributions of this work are:

1. The stability of the d_p^c distance in a continuous fashion.
2. Theoretical and statistical bounds on the number of 1-dim holes represented in a persistence diagram based on the cardinality of the underlying point cloud.
3. A d_p^c distance based classification algorithm for the crystal structure of high entropy alloys using synthetic atom probe tomography experiments.

The work is organized as follows. Relevant definitions and concepts necessary for persistent homology are presented in Sect. 2. Stability results of the d_p^c distance are in Sect. 3, as well as prediction interval bounds. Section 4 demonstrates a classification scheme for materials science data retrieved from synthetic APT experiments. We conclude and provide future directions in Sect. 5.

2 Persistent homology background

This section succinctly explains the construction of persistence diagrams, which are topological summaries of the underlying point cloud space. The Vietoris–Rips complex provides the necessary computational link between the point cloud, a subset of $\mathbb{R}^d$ under the Euclidean distance, and its persistence diagram. Below we give a brief summary of the necessary background. For a detailed treatment, see Edelsbrunner and Harer (2010).

Definition 1 A ν -simplex is the convex hull of an affinely independent point set of size $\nu + 1$.

Definition 2 For a set of points $\mathcal{P}$, an abstract simplicial complex σ is a collection of finite subsets of $\mathcal{P}$ such that for every set A in σ and every nonempty set $B \subset A$, we have that B is in σ . The elements of σ are called abstract simplices and are the combinatorial analogues of the geometric simplices in Definition 1.

Definition 3 For a given threshold ϵ , the Vietoris–Rips complex is a simplicial complex formed from a set such that corresponding to each subset of ν points of the set, an ν -simplex is included in the Vietoris–Rips complex each time the subsets have pairwise distances at most ϵ .

The Vietoris–Rips complex can be visualized by placing a ball of radius $\epsilon/2$ at each point in the set and then adding a ν -simplex at the points corresponding to the intersection of ν balls. See Fig. 2 for an illustration. For the Vietoris–Rips complex corresponding to ϵ , denoted by VR_ϵ , it is clear that $VR_\epsilon \subset VR_{\epsilon'}$ for $\epsilon < \epsilon'$. Thus we need only examine specific ϵ values corresponding to the emergence and disappearance

of homological features. These ϵ values are recorded as ordered pairs (b, d) in a persistence diagram, where b denotes the birth of a feature and d its death.

As can be seen in Fig. 2, a 0-dim homological feature is a connected component of a simplex, a 1-dim homological feature is a hole, such as those created by a loop or the circle S^1 , and a 2-dim homological feature describes voids, e.g. the inside of a sphere; see Wasserman (2018) for details. Higher dimensional data analogously yields higher dimensional holes.

Remark 1 Persistence diagrams can also be computed using a pertinent function g from a topological space to $\mathbb{R}$. Such a function can act as an approximation to a point cloud; typical functions used are kernel density estimators as in Fasy et al. (2014) and the distance to measure function as in Chazal et al. (2011). Homological features are born and die within the sublevel sets $g^{-1}(-\infty, t]$ as t increases. These birth and death times create another persistence diagram, see Fig. 2f.

To calculate the similarity between diagrams for classification problems, a distance on the space of persistence diagrams is needed. A typical distance is the Wasserstein distance.

Definition 4 The p -Wasserstein distance between two persistence diagrams X and Y is given by $W_p(X, Y) = \left(\inf_{\eta: X \rightarrow Y} \sum_{x \in X} \|x - \eta(x)\|_\infty^p \right)^{\frac{1}{p}}$, where the infimum is taken over all bijections η , and the points of the diagonal are added with infinite multiplicity to each diagram. If $p \rightarrow \infty$, then $W_\infty(X, Y) = \inf_{\eta: X \rightarrow Y} \sup_{x \in X} \|x - \eta(x)\|_\infty$ is the bottleneck distance between diagrams X and Y .

The Wasserstein distance yields the penalty of matched points under the optimal bijection. Points can be matched to the diagonal of each persistence diagram, which is assumed to have infinitely many points with infinite multiplicity; this ensures that a bijection between X and Y actually exists, since X and Y may not have the same cardinality. In other words, the Wasserstein distance gives no explicit penalty for differences in cardinality between two diagrams (Fig. 3). Instead, the Wasserstein distance penalizes unmatched points by using their distance to the diagonal. However, cardinality differences may play a key role in machine learning problems, and to that end, Marchese and Maroulas (2018) proposed the d_p^c distance given below.

Definition 5 Let X and Y be two persistence diagrams with cardinalities n and m respectively such that $n \leq m$ and denoted $X = \{x_1, \dots, x_n\}$, $Y = \{y_1, \dots, y_m\}$. Let $c > 0$ and $1 \leq p < \infty$ be fixed parameters. The d_p^c distance between two persistence diagrams X and Y is

$$d_p^c(X, Y) = \left(\frac{1}{m} \left(\min_{\pi \in \Pi_m} \sum_{\ell=1}^n \min(c, \|x_\ell - y_{\pi(\ell)}\|_\infty)^p + c^p |m - n| \right) \right)^{\frac{1}{p}}, \quad (1)$$

where Π_m is the set of permutations of $(1, \dots, m)$. If $m < n$, define $d_p^c(X, Y) := d_p^c(Y, X)$.

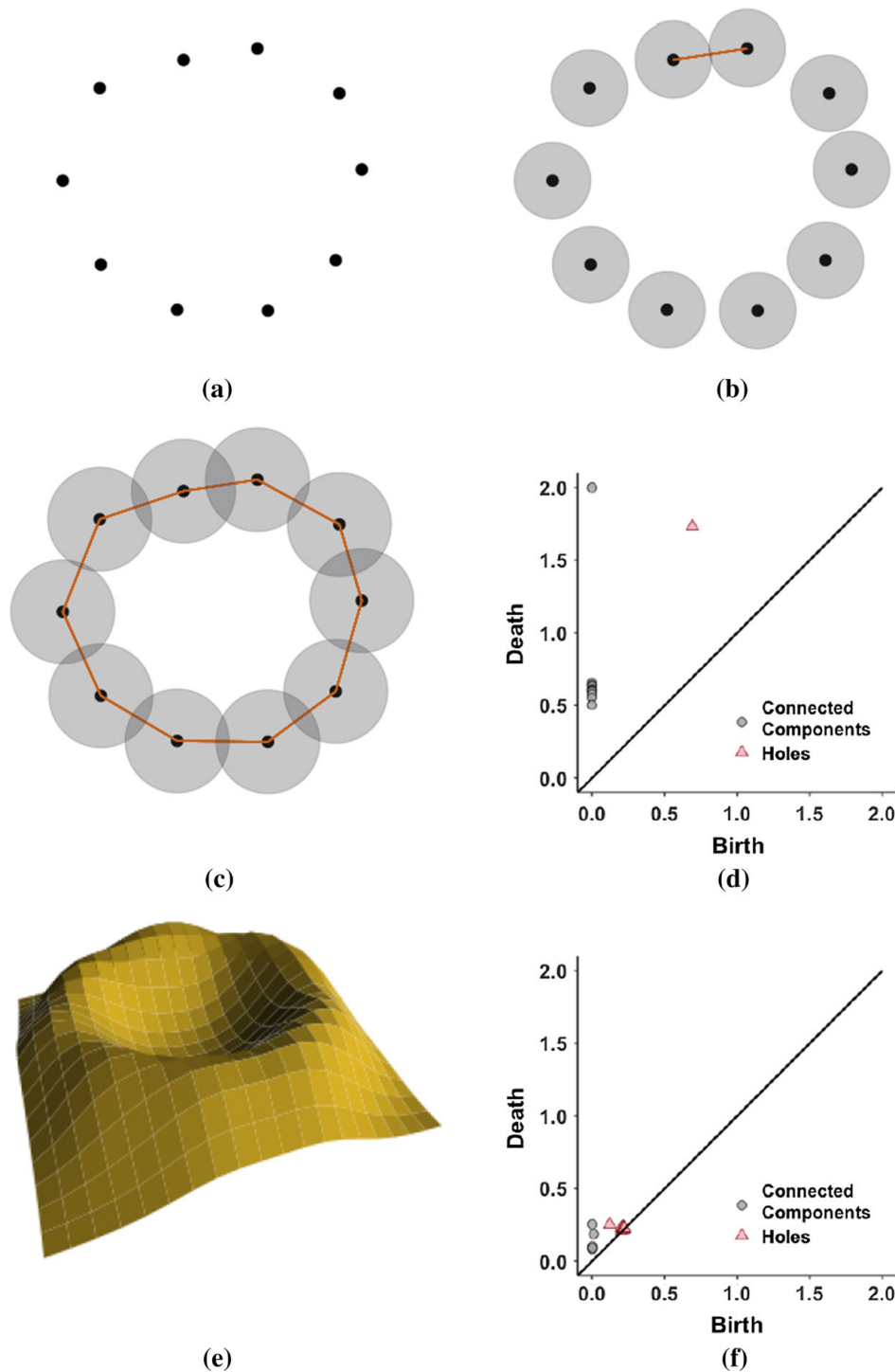


Fig. 2 Begin with a point cloud (a). After increasing the radius of the balls around the points, a 1-simplex (line segment) forms in the corresponding Vietoris–Rips complex, (b). Eventually, more 1-simplices are added and a 1-dim hole forms (c). In (d), the persistence diagram tracks all the birth and death times, with respect to the radius $\epsilon/2$, of the homological features for each dimension. Using the same points as in (a), the kernel density estimator function for this point cloud is plotted in (e). A corresponding persistence diagram is created using sublevel sets in (f). Note the difference between the persistence diagrams in (d) and (f). The persistence diagram created in (f) has noisy 1-dim features that are not present in the persistence diagram created directly from the data points

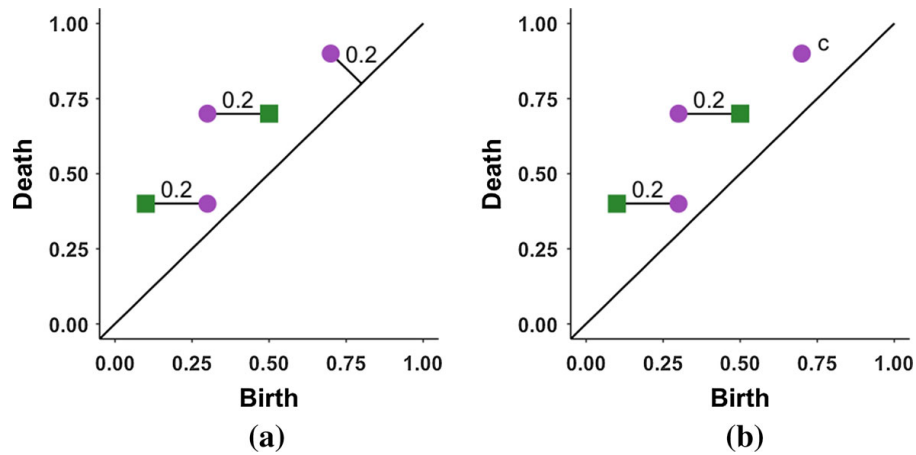


Fig. 3 Consider two persistence diagrams, one given by the green squares and another by the purple circles. **a** The Wasserstein distance imposes a cost of 0.2 to the extra purple point (the ℓ^∞ distance to the diagonal). **b** The d_p^c distance imposes a penalty c on the point instead (color figure online)

Remark 2 Note that this distance can be applied to arbitrary point clouds with finite cardinality as well. As shown in Marchese and Maroulas (2018), a smaller c in Eq. (1) accounts for local geometric differences, while a larger c focuses on global geometry. It is precisely by considering differences in cardinality that the d_p^c distance can distinguish between features of the point cloud that other distances may miss. Also in Eq. (1), if X is fixed and $m \rightarrow \infty$, then $d_p^c(X, Y) \rightarrow c$.

3 Stability properties for d_p^c distance

The stability of the d_p^c distance is proved in this section. Stability of the distance under investigation means that small perturbations in the underlying space result in small perturbations of the generated persistence diagrams. Adopting the approach of estimating a point cloud via a pertinent function, e.g. a kernel density estimator as in Fasy et al. (2014), persistence diagrams may be constructed using sublevel sets as in Fig. 2f and Remark 1. Their differences can be computed using the Wasserstein and bottleneck distances. Using this functional representation, stability of the Wasserstein and bottleneck distances has been shown in Cohen-Steiner et al. (2010) and Cohen-Steiner et al. (2007) respectively, by verifying Lipschitz (and respectively Hölder) continuity of the mapping from the underlying function of the data to its persistence diagram in the bottleneck and Wasserstein distances. Considering discrete point clouds whose distances shrink to zero, Theorem 1 shows that the distance between persistence diagrams goes to zero as well.

Theorem 1 (Stability Theorem) *Consider $c > 0$ and $1 \leq p < \infty$. Let A be a finite nonempty point cloud in $\mathbb{R}^d$. Suppose that $\{A_i\}_{i \in \mathbb{N}}$ is a sequence of finite nonempty point clouds such that $d_p^c(A, A_i) \rightarrow 0$ as $i \rightarrow \infty$. Let X^k and X_i^k be the k -dim persistence diagrams created from the Vietoris–Rips complex for A and A_i respectively. Then $d_p^c(X^k, X_i^k) \rightarrow 0$ as $i \rightarrow \infty$.*

Note that Theorem 1 does not depend on a function created from the points such as a kernel density estimator as in Fasy et al. (2014), but simply on the points themselves and the Vietoris–Rips complex generated from these points. In fact, Theorem 1 shows that the mapping from a point cloud to the persistence diagram of its Vietoris–Rips complex is continuous under the d_p^c distance. This continuous-type stability result is weaker than Lipschitz stability. In order to prove Theorem 1, we first show that if the d_p^c distance between the underlying point clouds goes to 0, then eventually the size of the point clouds must be the same.

Lemma 1 *Let A and A_i be as in Theorem 1 such that $d_p^c(A, A_i) \rightarrow 0$ as $i \rightarrow \infty$. Then A_i and A have the same number of points for $i \geq N_0$ for some $N_0 \in \mathbb{N}$.*

Proof Denote by $|A|$ the number of points in the point cloud A . Suppose that $|A_i| \neq |A|$ infinitely often. Since $d_p^c(A, A_i) \rightarrow 0$, for every $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $i \geq N$ implies that $d_p^c(A, A_i) < \epsilon$. Let $\epsilon = \frac{c}{|A|+1}$, noting that $|A|$ is fixed. By assumption $|A_i| < |A|$, $|A_i| > |A|$, or both, infinitely often. If $|A| < |A_i|$, then by Definition 5

$$d_p^c(A, A_i) \geq \left(c^p \frac{|A_i| - |A|}{|A_i|} \right)^{\frac{1}{p}} \geq c \frac{|A_i| - |A|}{|A_i|}. \quad (2)$$

The function $h : \mathbb{N} \rightarrow \mathbb{R}$ given by $h(z) = \frac{z-|A|}{z}$ is strictly increasing. Whenever $|A| < |A_i|$, we have $|A_i| \geq |A|+1$. The restriction of h to $\{|A|+1, |A|+2, |A|+3, \dots\}$ achieves its minimum at $|A|+1$. This shows that the RHS of Eq. (2) is greater than or equal to $\frac{c}{|A|+1}$, whenever $|A| < |A_i|$, which by assumption happens infinitely often. This contradicts $d_p^c(A, A_i) < \epsilon$ for all $i \geq N$. The case where $|A| > |A_i|$ follows similarly. $\square$

Lemma 2 *Let A and A_i be as in Theorem 1. Suppose the points of each point cloud A_i are ordered so that $A_i = \{a_{\pi_i(1)}, a_{\pi_i(2)}, \dots, a_{\pi_i(|A|)}\}$, where π_i is the permutation used to calculate the d_p^c distance between A_i and A as in Eq. (1). Let D_A and D_{A_i} be the distance matrices for the points of A and A_i respectively, i.e., the kl -th entry of D_A is $\|a_k - a_l\|_d$. Then,*

- (i) $\|D_A - D_{A_i}\|_\infty \rightarrow 0$ as $i \rightarrow \infty$, and
- (ii) for some $N_1 \in \mathbb{N}$, the order of the entries of the upper triangular portion of D_A and D_{A_i} is the same for $i \geq N_1$, up to permutation when either D_A or D_{A_i} have duplicate entries.

Proof (i) Let $A = \{a_1, \dots, a_k\}$, $A_i = \{a_1^i, \dots, a_k^i\}$, and $\lambda_\alpha^i = \|a_\alpha - a_{\pi_i(\alpha)}^i\|_d$ for the permutation π_i in the d_p^c distance between A_i and A . Suppose that $d_p^c(A, A_i) \rightarrow 0$. Note that since c is fixed, then by Lemma 1, there is some N_c such that eventually $d_p^c(A_i, A) = \left(\frac{1}{|A|} \min_{\pi_i \in \Pi_{|A|}} \sum_{\ell=1}^{|A|} \|a_\ell - a_{\pi_i(\ell)}^i\|_d^p \right)^{\frac{1}{p}}$ for $i \geq N_c$. By assumption $d_p^c(A, A_i) \rightarrow 0$, which shows that $|A|^{-\frac{1}{p}} \|\lambda\|_p \rightarrow 0$ as $i \rightarrow \infty$. Thus $\|\lambda^i\|_p \rightarrow 0$ as $i \rightarrow \infty$.

Now, let $E = D_A - D_{A_i}$.

$$\begin{aligned} \|E\|_\infty &= \max_{k,l} |\|a_k - a_l\|_d - \|a_k^i - a_l^i\|_d| \\ &= \max_{k,l} |\|a_k - a_l\|_d + \|a_l - a_k^i\|_d - \|a_l - a_k^i\|_d - \|a_k^i - a_l^i\|_d| \\ &\leq |\|a_k - a_l\|_d - \|a_l - a_k^i\|_d| + |\|a_k^i - a_l^i\|_d - \|a_l - a_k^i\|_d| \\ &\leq \|a_k - a_k^i\|_d + \|a_l - a_l^i\|_d \end{aligned} \quad (3)$$

The last term in Eq. (3) goes to 0 as $i \rightarrow \infty$, proving (i).

(ii) Suppose that the m distinct upper triangular entries of D_A are ordered from smallest to largest, say $d_1^A < d_2^A < \dots < d_m^A$, where $m \leq |A|(|A| - 1)/2$. For $\eta \in \{1, \dots, m+1\}$ let $h_\eta \in [0, \infty)$ be a sequence such that $h_1 < d_1^A < h_2 < d_2^A < \dots < h_m < d_m^A < h_{m+1}$. Let $\|D_A - D_{A_i}\|_\infty < \frac{h}{2}$, where $h = \min_{\eta_1, \eta_2 \in \{1, \dots, m+1\}} \{|h_{\eta_1} - h_{\eta_2}|\}$. We show that there exists a sequence g_η such that $|h_\eta - g_\eta| < 2h$ for each $\eta \in \{1, \dots, m+1\}$ and $h_\eta < d_j^A < h_{\eta+1}$ implies $g_\eta < d_j^{A_i} \leq g_{\eta+1}$. Let $h_\eta < d_j^A < h_{\eta+1}$, and suppose that it is not the case that $h_\eta < d_j^{A_i} \leq h_{\eta+1}$. Since $\|D_A - D_{A_i}\|_\infty < \frac{h}{2}$, then either $d_j^{A_i} \in (h_{\eta-1}, h_\eta]$ or $d_j^{A_i} \in (h_{\eta+1}, h_{\eta+2}]$. If the first case is true, then take $g_\eta = d_j^A - \frac{h}{2}$. If the second, then take $g_\eta = d_j^A + \frac{h}{2}$. This proves the existence of the sequence. Now proceeding by contradiction, if the lemma does not hold for some entries $d_j^A \in D_A$ and $d_j^{A_i} \in D_{A_i}$, then take $\|D_A - D_{A_i}\|_\infty < \frac{1}{2}|d_j^A - d_j^{A_i}|$. $\square$

Proof of Theorem 1 By Lemma 1, take $|A_i| = |A|$ without loss of generality. By Lemma 2 (i), $\|D_A - D_{A_i}\|_\infty \rightarrow 0$ as $i \rightarrow \infty$. If the Vietoris–Rips complex were computed at every threshold value in $[0, \infty)$, then the birth and death times of all features of all dimensions would be distances between points in the underlying point cloud (including the birth time of 0 in the 0-dim diagram). Since the order of the entries of D_A and D_{A_i} may be taken to be the same from Lemma 2 (ii), the same number of simplices are formed in the complexes for A and A_i for each dimension of simplex. Also, the labels of the simplices according to the points of A and A_i are given from the permutation π_i in Lemma 2 (i).

Now, for 0-dim it is clear that for the cardinalities of the persistence diagrams, $|X^0| = |X_i^0|$ since for the sizes of their associated point clouds, $|A_i| = |A|$. For a higher dimensional feature ($k \geq 1$) to appear in the complex, we must have that a certain number of the distances are less than or equal to the threshold ϵ and a certain number of the distances are greater than ϵ . Lemma 2 (ii) shows that although the thresholds where the features are created may be different, the same number of features are formed in the Vietoris–Rips complexes of A and A_i , and these features are formed in the same order and with the points that correspond under π_i .

If $X^k = \{x_1, x_2, \dots, x_{|X^k|}\}$ and $X_i^k = \{x_1, x_2, \dots, x_{|X_i^k|}\}$, then we have that $|X^k| = |X_i^k|$ and that $d_p^c(X^k, X_i^k) < 2h$. Thus $d_p^c(X^k, X_i^k) \rightarrow 0$ as $i \rightarrow \infty$. $\square$

To provide a practical way to control c in computing the d_p^c distance of Eq. (1) and consequently compute the possible fluctuations of the d_p^c distance, a probabilistic

upper bound, which relies on least squares, is provided. Specifically, the following analysis gives predictions on the number of 1-dim holes represented in the persistence diagram, which we denote by b_1 . The parameter b_1 relies on the number of connected components (or equivalently the number of points in the point cloud) represented in the persistence diagram, denoted by b_0 .

Definition 6 (Pfender and Ziegler 2004) The kissing number in $\mathbb{R}^d$ is the maximum number of nonoverlapping unit spheres that can be arranged so that each touches another common central unit sphere.

Lemma 3 (Goff 2011) For a finite point cloud with no more than ρ points in $\mathbb{R}^d$ under the Euclidean distance, let $M_d(\rho)$ denote the maximum possible number of 1-dim holes in the Vietoris–Rips complex for the point cloud for a given threshold. Then

$$M_d(\rho) \leq (K_d - 1)\rho. \quad (4)$$

Proposition 1 Consider a point cloud in $\mathbb{R}^d$ with ρ points and its associated persistence diagram. Let B_1 denote the possible range of the number of 1-dim holes b_1 . Then B_1 is such that $\{0, 1, \dots, \lfloor \frac{\rho}{2} \rfloor - 1\} \subseteq B_1 \subseteq \{0, 1, \dots, \frac{1}{2}(K_d - 1)\rho^2(\rho - 1)\}$, i.e. the possible range of b_1 is expanding as the number of points, b_0 , in the point cloud increases.

Proof We first show the inclusion $\{0, 1, \dots, \lfloor \frac{\rho}{2} \rfloor - 1\} \subseteq B_1$. To form a point cloud with ρ points that has $b_1 = 0$, simply take the ρ points and arrange them on a line. To form a point cloud with ρ points that has $b_1 = \lfloor \frac{\rho}{2} \rfloor - 1$, arrange the ρ points in two rows each with $\lfloor \frac{\rho}{2} \rfloor$ points. Set the spacing between adjacent points in each of the rows to be 1 and then place the two rows directly beside each other so that for each point in the first row, there is exactly one point in the second row at a distance of 1. Adding edges appropriately creates $b_1 = \lfloor \frac{\rho}{2} \rfloor - 1$ squares with side length 1. Thus, creating the Vietoris–Rips complex and corresponding diagram gives $b_1 = \lfloor \frac{\rho}{2} \rfloor - 1$. For an illustration of the arrangement, see Fig. 4a.

To form a point cloud with ρ points that has $b_1 \in \{1, 2, \dots, \lfloor \frac{\rho}{2} \rfloor - 2\}$, arrange $2(b_1 + 1)$ points in two rows as in Fig. 4a. Arrange the other $\rho - 2(b_1 + 1)$ points in a line with the minimum distance from any points in the line to points of the two rows such that it is greater than or equal to 1. Then exactly b_1 holes are formed from the two rows, with no holes formed by the line. For an illustration, see Fig. 4b.

Next, we verify the inclusion $B_1 \subseteq \{0, 1, \dots, \frac{1}{2}(K_d - 1)\rho^2(\rho - 1)\}$. By Lemma 3, the number of 1-dim holes in the Vietoris–Rips complex for a fixed radius ϵ for the point cloud is bounded above by $(K_d - 1)\rho$. The homology of the Vietoris–Rips complex changes at most $\binom{\rho}{2}$ times as the radius ϵ increases due to the maximum of $\binom{\rho}{2}$ distinct distances between points in the point cloud. Therefore, there can be at most $\frac{1}{2}(K_d - 1)\rho^2(\rho - 1)$ 1-dim holes formed over the entire evolution of the Vietoris–Rips complex. This gives the desired bound of $b_1 \leq \frac{1}{2}(K_d - 1)\rho^2(\rho - 1)$. $\square$

Now, let N point clouds be generated from some process, and N corresponding persistence diagrams be created. For each persistence diagram $X_i^k, k \in \{0, 1\}, i = 1, \dots, N$, record the cardinality b_0^i of the 0-dim diagram and the cardinality b_1^i of

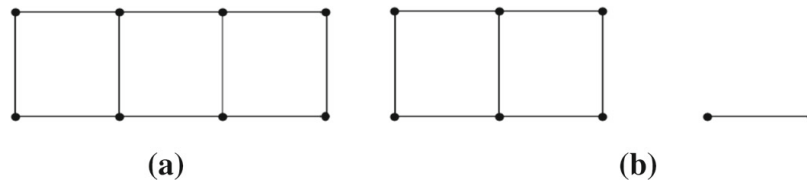


Fig. 4 An example of 8-point arrangements to visualize the proof of Proposition 1. **a** A 3-hole configuration versus **b** a 2-hole configuration

the 1-dim diagram. Let $\mathbf{b}_0 \in \mathbb{R}^{N \times 2}$ be the predictor matrix whose rows are $[1 \ b_0^i]$ and $\mathbf{b}_1 \in \mathbb{R}^N$ be the vector of responses with entries b_1^i . Proposition 1 gives that the possible range of $\mathbf{b}_1$ is increasing as $\mathbf{b}_0$ grows, which yields that an increase in variance as $\mathbf{b}_0$ grows may be present, i.e., heteroscedasticity exists. Thus the analysis of the change in number of 1-dim holes as the size of the point cloud changes needs to account for heteroscedasticity in order to capture the non-constant variance behavior. Therefore to estimate the number of 1-dim holes, we use weighted least squares as in Efron and Hastie (2016). If $\mathbf{W} \in \mathbb{R}^{N \times N}$ is the weight matrix $\mathbf{W} = \text{diag}(a_1, \dots, a_N)$, then a weighted least-squares regression can be found for $\mathbf{b}_1 = \mathbf{b}_0 \boldsymbol{\gamma} + \boldsymbol{\epsilon}$, where $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. The approximation is then given by $\mathbf{b}_0 \hat{\boldsymbol{\gamma}} = \mathbf{b}_1$, with $\hat{\boldsymbol{\gamma}} = (\mathbf{b}_0^T \mathbf{W} \mathbf{b}_0)^{-1} \mathbf{b}_0^T \mathbf{W} \mathbf{b}_1$. In turn, Proposition 2 provides bounds from prediction intervals using weighted least squares for the d_p^c distance.

Proposition 2 Suppose N point clouds are generated from a process, and N corresponding persistence diagrams are created. For each persistence diagram $X_i^k, k \in \{0, 1\}$, record the cardinality of the 0-dim diagram b_0^i and of the 1-dim diagram b_1^i . Let $\mathbf{b}_0 \in \mathbb{R}^{N \times 2}$ be the predictor matrix whose rows are $[1 \ b_0^i]$ and $\mathbf{b}_1 \in \mathbb{R}^N$ be the vector of responses of b_1^i . Assume the model $\mathbf{b}_1 = \mathbf{b}_0 \boldsymbol{\gamma} + \boldsymbol{\epsilon}$, where $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ depends on the value of the input b_0^i . Let X^1 and Y^1 be persistence diagrams generated from the same process as $\mathbf{b}_0$ with $|X^0| = \mu$. Considering the $(1 - \alpha) \cdot 100\%$ -level prediction interval for $\mathbf{b}_1$, the distance $d_p^c(X^1, Y^1)$ is bounded above by

$$\left(\min_{\pi \in \Pi_m} \sum_{\ell=1}^n \min(c, \|x_{\ell}^1 - y_{\pi(\ell)}^1\|_{\infty})^p + c^p 2t_{1-\alpha, N-2s} \sqrt{[1 \ \mu](\mathbf{b}_0^T \mathbf{W} \mathbf{b}_0)^{-1} [1 \ \mu]^T + \mu} \right)^{\frac{1}{p}}.$$

Proof Prediction intervals can be constructed for the cardinality of a 1-dim diagram for an instance of point cloud size b_0^* using standard results on weighted least squares. Specifically, for level $(1 - \alpha) \cdot 100\%$ a prediction interval for the new response $\hat{b}_1^*$ is sought. To calculate this interval for a new response from the mean predicted response $\hat{b}_1^* = \hat{\boldsymbol{\gamma}} b_0^*$, note that $\hat{b}_1^* - b_1^*$ has the distribution $\frac{\hat{b}_1^* - b_1^*}{\text{Var}(\hat{b}_1^* - b_1^*)} \sim t_{N-2}$. Also, $\text{Var}(\hat{b}_1^* - b_1^*) = \text{Var}(\boldsymbol{\epsilon}) [1 \ b_0^*] (\mathbf{b}_0^T \mathbf{W} \mathbf{b}_0)^{-1} [1 \ b_0^*]^T + \frac{\text{Var}(\boldsymbol{\epsilon})}{w^*}$, where $w^* = \frac{1}{b_0^*}$, the weight corresponding to b_0^* . Prediction intervals for b_1^* are thus $\hat{b}_1^* \pm t_{1-\alpha/2, N-2s} \sqrt{[1 \ b_0^*] (\mathbf{b}_0^T \mathbf{W} \mathbf{b}_0)^{-1} [1 \ b_0^*]^T + \frac{\hat{\boldsymbol{\epsilon}}^T \mathbf{W} \hat{\boldsymbol{\epsilon}}}{N-2}}$, where $s^2 = \frac{\hat{\boldsymbol{\epsilon}}^T \mathbf{W} \hat{\boldsymbol{\epsilon}}}{N-2}$, the unbiased estimator for $\text{Var}(\boldsymbol{\epsilon})$, using the residuals $\hat{\boldsymbol{\epsilon}}$. Thus the cardinality difference term

in the calculation of the d_p^c distance as in Eq. (1) is bounded above by the length of the prediction interval with $(1 - \alpha) \cdot 100\%$ -level confidence. Substituting this length into Eq. (1) gives the result. $\square$

4 Classification of materials data

Here we describe the d_p^c -distance based classification of crystal structures of high entropy alloys (HEAs) using atom probe tomography (APT) experiments (Fig. 5). Recall that the building blocks of HEAs are either body-centered cubic (BCC) or face-centered cubic (FCC). Topological considerations are a natural fit for this problem since BCC and FCC crystal structures enjoy a different atomic configuration within a unit cell. Indeed, the BCC structure has one atom at its center, but the FCC contains a void (recall Fig. 1a, b). This distinction is important from the viewpoint of persistent homology.

However, topology alone is insufficient to distinguish between noisy and sparse BCC and FCC lattice structures accurately. If we count the number of atoms in a unit cell (see Fig. 1a, b) one may see that a BCC unit cell has two atoms, one at the center and $1/8$ th of an atom at the unit cell's corners, as it shares part of these corner atoms with its neighboring cells. Similarly, an FCC unit cell has four atoms; the same $1/8$ th of the corner atoms plus one-half of each of the six atoms on the cell's faces. In both cases, the atoms on the faces and lattice points are shared with the cell's neighbors and are only counted as a proportion contributing to the unit cell.

Another way to see this difference in cardinality is by plotting the number of connected components against the number of holes for both BCC and FCC crystal structures. Figures 6, 7c, d depict that FCC structures have larger point clouds, and consequently, a greater number of connected components. Observe in Fig. 6 that the number of connected components and 1-dim holes are greater in the FCC diagrams than the BCC diagrams. Consequently, we must account for more than just homo-

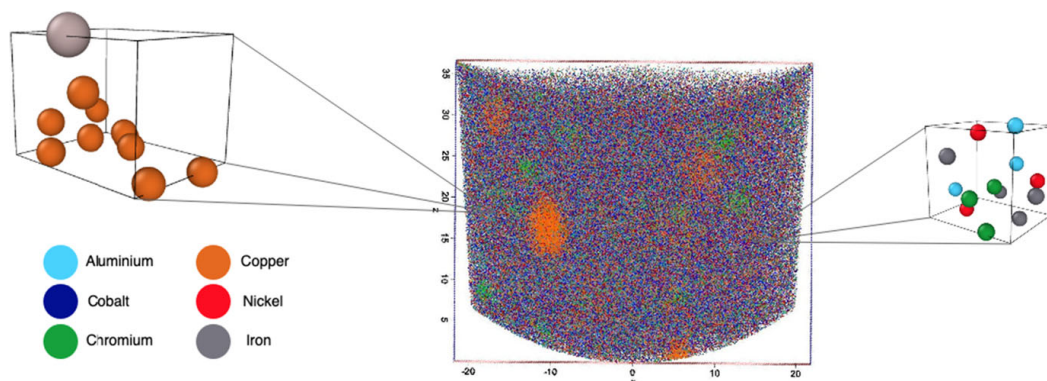


Fig. 5 Image of APT data with atomic neighborhoods shown in detail on the left and right. Each pixel represents a different atom, the neighborhood of which is considered. Certain patterns with distinct crystal structures exist, e.g., the orange region is copper-rich (left), but overall no pattern is identified. Putting a single atomic neighborhood under a microscope, the true crystal structure of the material, which could be either BCC (Fig. 1a) or FCC (Fig. 1b), is not revealed. This distinction is obscured due to experimental noise (color figure online)

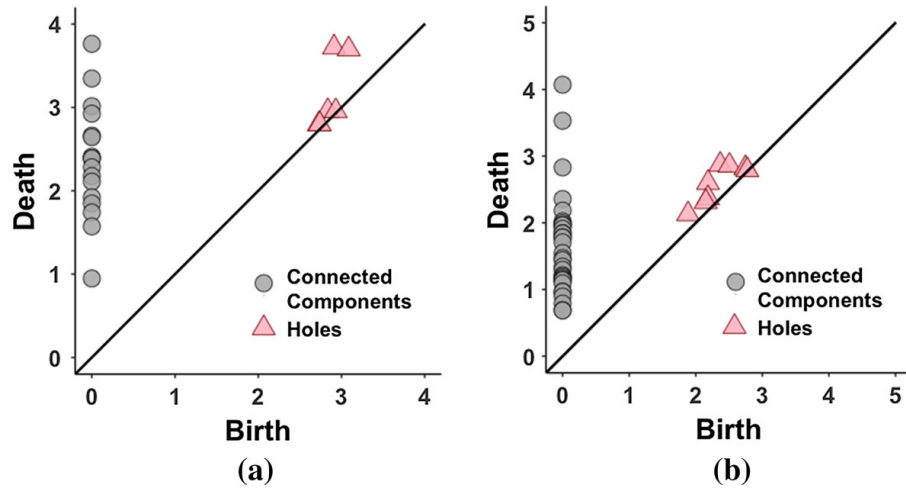


Fig. 6 **a** A persistence diagram generated by a BCC lattice. **b** A persistence diagram generated by an FCC lattice. The data has a noise standard deviation of $\tau = 0.75$ and 67% of the atoms are missing. Note that the BCC diagram has two prominent (far from the diagonal) points representing 1-dim holes and fewer connected components and 1-dim holes than does the FCC diagram

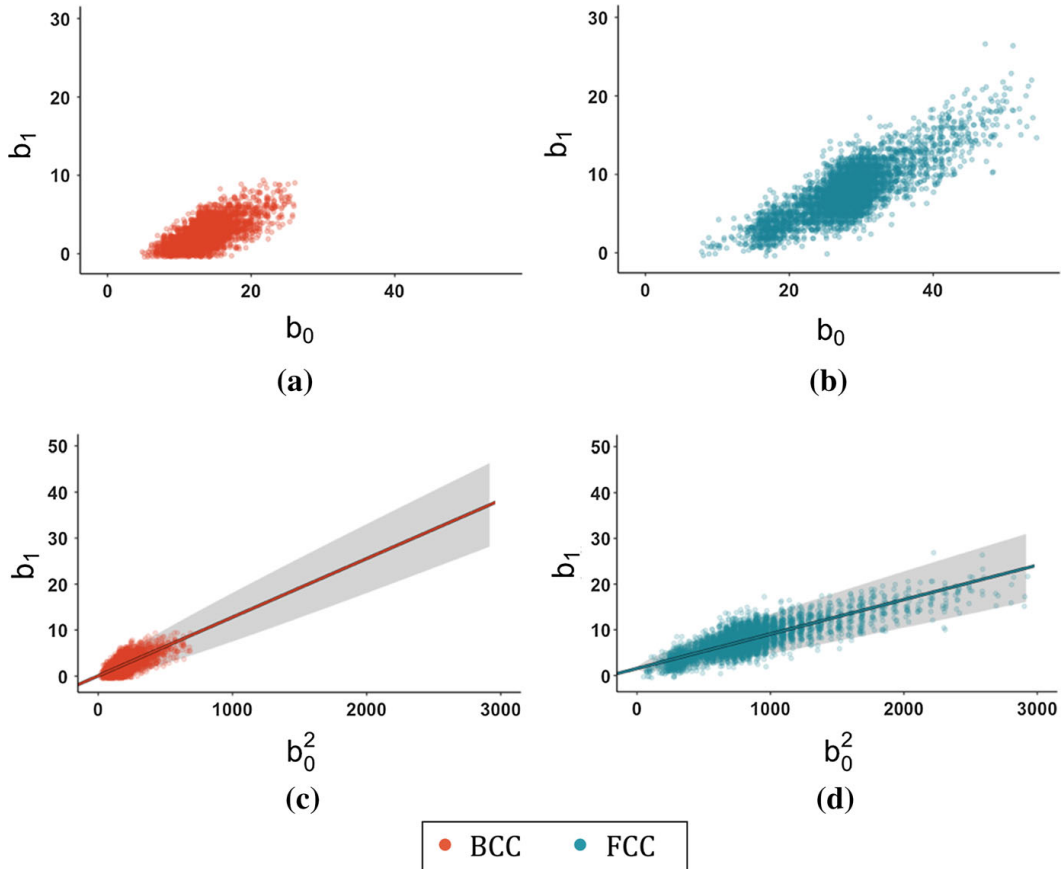


Fig. 7 *Top*: Number of connected components (in this case atoms), b_0 , against the number of 1-dim homological features, b_1 , of the persistence diagrams. One can see the presence of heteroscedasticity since the variance of b_1 increases as b_0 increases. *Bottom*: Same as in top but using a quadratic transformation of the predictor variable, along with the weighted least squares fit line and 95% prediction intervals provided by Proposition 2

logical differences when considering persistence diagrams derived from these atomic neighborhoods. Variability in the size of the underlying point clouds must be considered, as verified in Proposition 2. Given the salient topological and cardinality differences between these two crystal structures, we seek to classify their associated persistence diagrams via these essential differences. To that end, we consider the d_p^c distance given in Eq. (1).

In the numerical experiments, the point clouds (atomic neighborhoods) are extracted from a sample containing approximately 10,000 atoms. We remove atoms, to create sparsity, and add Gaussian noise to the larger sample mirroring those levels found in true APT experimental data. To create these neighborhoods, we consider a fixed volume around each atom in the perturbed sample and those atoms within the volume are recorded for our classification methodology. Here we consider $N = 1,000$ synthetic atomic neighborhoods ($N_{BCC} = 500$ BCC structures and $N_{FCC} = 500$ FCC structures) with noise and sparsity levels similar to those found in true APT experiments. Let $\mathbf{q} = (q_1, \dots, q_M)^T$ be the atoms' positions within an atomic neighborhood. Applying the persistent homology machinery of Sect. 2, one obtains the associated persistence diagram denoted by $X_{\mathbf{q}}$, see Fig. 6. For our classification problem, we are interested in the conditional probability, $\tilde{\pi}_j = \mathbb{P}(Y_i = j \mid X_i)$, of the persistence diagram X_i being in class Y_j , for $j = 0$ (BCC) or $j = 1$ (FCC). To that end, we consider a logistic regression model,

$$\log \left(\frac{\tilde{\pi}_j}{1 - \tilde{\pi}_j} \right) = \alpha + \sum_{i=1}^L \varphi_i(\mathbf{\Sigma}_i), \quad (5)$$

where φ_i is some pertinent smooth function, and $\mathbf{\Sigma} \in \mathbb{R}^{N \times 8}$ is the feature matrix whose i^{th} row is

$$\mathbf{\Sigma}_i = (\mathbb{E}_{i,B}^0, \mathbb{E}_{i,B}^1, \text{Var}_{i,B}^0, \text{Var}_{i,B}^1, \mathbb{E}_{i,F}^0, \mathbb{E}_{i,F}^1, \text{Var}_{i,F}^0, \text{Var}_{i,F}^1). \quad (6)$$

For any persistence diagram X_i^k with k -dimensional homology ($k = 0, 1$), $\mathbb{E}_{i,B}^k = \frac{1}{N_{BCC}} \sum_{j=1}^{N_{BCC}} d_p^c(X_i^k, X_j^k)$ and $\text{Var}_{i,B}^k = \frac{1}{N_{BCC}-1} \sum_{j=1}^{N_{BCC}} (d_p^c(X_i^k, X_j^k) - \mathbb{E}_{i,B}^k)^2$ respectively yield the average and variance of the distance between X_i^k and the collection of all BCC persistence diagrams. Similarly, $\mathbb{E}_{i,F}^k$ and $\text{Var}_{i,F}^k$ are the average and variance of the distance between X_i^k and the collection of all FCC persistence diagrams.

We perform 10-fold cross validation on the 1,000 synthetic crystal structures. In other words, the data is divided randomly into 10 folds, and 9 folds of the data are used as a training set. For any unknown crystal structure in the remaining fold, the feature vector of the unknown crystal structure is computed according to Eq. (6) and used as input for the decision tree classifier. Similarly, the other 9 folds are each used once as test sets employing the same procedure. The tree finds the best fit for the features from the additive model in Eq. (5) and returns the class of the unknown structure.

For our numerical experiments, the persistence diagrams are constructed using the C++ Ripser software, and the scikit-learn decision tree implementation. The studies (Miller et al. 2012; Santodonato et al. 2015) estimate that approximately 65% of the data is missing. However, an estimate of the experimental noise is not provided. In fact,

Table 1 The atomic positions in the APT data are $\mathcal{N}(0, \tau^2)$ distributed with 67% of the atoms missing

τ	c -value	Accuracy (%)
0.0	0.01	99
0.25	0.05	99.4
0.75	0.03	96.5
1.0	0.13	96.4

We employ the d_p^c classifier, where c has been optimized in each noise level case. The accuracy in the 10-fold cross validation is listed in the third column

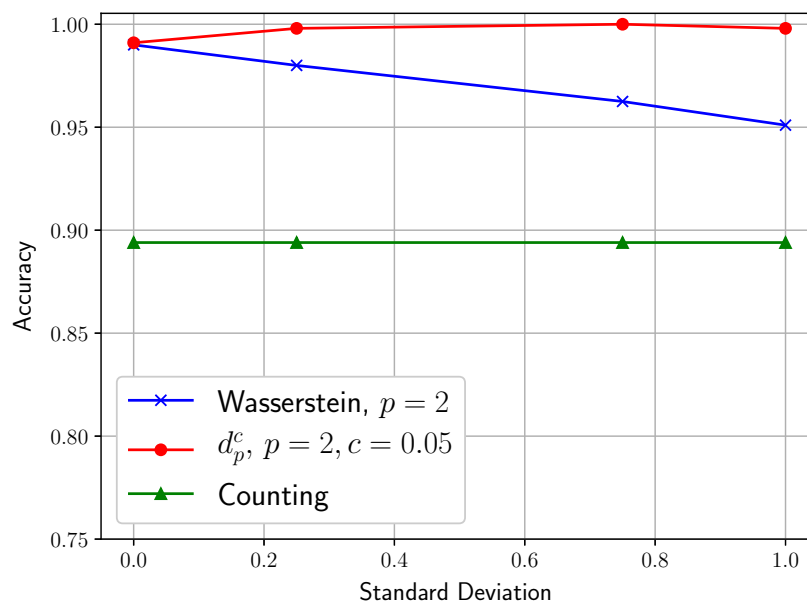


Fig. 8 10-fold cross validation accuracy scores for d_p^c (red), Wasserstein (blue), and counting (green) classifiers, plotted against different standard deviations, τ , (see Table 1) of the normally distributed noise of the atomic positions. In each instance, the sparsity has been fixed at 67% of the atoms missing, as in a true APT experiment (color figure online)

as noted by Larson (2013), Miller (2014), the noise varies between experiments and specimens. Our synthetic data replicates this resolution by drawing from a Gaussian, (Gault et al. 2010; McNutt et al. 2017; Moody et al. 2011), $\mathcal{N}(0, \tau^2)$, with four different levels of variance to give a more representative approximation of true APT datasets. Computing the d_p^c distances with $p = 2$ to imitate typical Euclidean distance, we find different values of c via a grid search for these four different levels of variance, τ^2 , in both 0- and 1-dim homology, employing a different dataset than is used for the classification. In each case, a geometric sequence of 10 values between 0.01 and 1 is taken into account. The results and the associated algorithmic accuracy are presented in Table 1.

As a comparison the feature matrix in Eq. (6) is also calculated using the Wasserstein distance, choosing $p = 2$. Moreover, we adopt a counting classifier which takes into account only the number of points in an atomic neighborhood as the input feature in the tree classifier. Our d_p^c classifier successfully dichotomizes these 1000 persistence diagrams generated by BCC and FCC lattice structures at better than 96%

accuracy, where accuracy is measured as $(1 - \text{Misclassification rate})$. The d_p^c classifier outperforms both the Wasserstein and the counting classifier, see Fig. 8. These results demonstrate that using just the differences in cardinality between the two classes of crystal structures is insufficient to distinguish between them.

As demonstrated in Proposition 2, there is a relationship between the number of connected components, $\mathbf{b}_0$, (number of atoms in this case) and the number of 1-dim homological features, $\mathbf{b}_1$, in the persistence diagrams. Fig. 7a, b demonstrate this relationship, as well as the presence of heteroscedasticity between $\mathbf{b}_0$ and $\mathbf{b}_1$, also verified by the Breusch-Pagan test (Breusch and Pagan 1979) with a p value of 9.3×10^{-54} for FCC cells and a p value of 2.01×10^{-47} for BCC cells. Figs. 7a, b also provide 95% prediction intervals for $\mathbf{b}_1$ based on the weighted least squares regression analysis of Proposition 2. To that end, this exact fine balance between the number of atoms in a neighborhood and the associated topology created by the positions of these atoms in the cubic cell is captured by the d_p^c distance.

5 Conclusions

This work combined statistical learning and topology to classify the crystal structure of high entropy alloys using atom probe tomography (APT) experiments. These APT experiments produce a noisy and sparse dataset, from which we extract atomic neighborhoods, i.e., atoms within a fixed volume forming a point cloud, and apply the machinery of Topological Data Analysis (TDA) to these point clouds. Viewed through the lens of TDA, these point clouds are a rich source of topological information. Indeed, employing persistent homology, we summarized the shape of these atomic neighborhoods and classified their crystal structures as either BCC or FCC. The classifier was based on features derived from the new distance on persistence diagrams, denoted herein by d_p^c . This distance is different from all other existing distances on persistence diagrams in that it explicitly penalizes differences in cardinality between diagrams.

We proved a stability result for the d_p^c distance, demonstrating that small perturbations of the underlying point clouds resulted in small changes to the d_p^c distance. We also provided guidance for the choice of the c parameter by looking at confidence bounds using a function of the cardinalities of the persistence diagrams.

The classification results presented herein could aid materials science researchers by providing a previously unavailable representation of the local atomic environment of high entropy alloys from APT data. The methodology need not be limited to a binary choice between BCC and FCC, e.g. entropy-stabilized oxides (Rost et al. 2015) are amenable to APT characterizations and our process could be generalized to those materials as well. Moreover, as APT experiments produce datasets on the order of 10 million atoms, materials science research has moved into the realm of big data, and the necessary computational and modelling tools have yet to be developed for this regime according to Katsoulakis and Zabaras (2017). The d_p^c classifier, coupled with our ongoing research of quantifying local atomic distributions as in Spannaus et al. (2019), aims to recover global atomic structure of high entropy alloys.

Acknowledgements The authors would like to thank the anonymous associate editor and two anonymous reviewers for their insightful comments which substantially improved the manuscript. Moreover, the authors would like to thank Professor David J. Keffer (Department of Materials Science and Engineering at The University of Tennessee) for providing the codes which create the realistic APT datasets and for useful discussions, as well as Professor Kody J.H. Law (School of Mathematics at the University of Manchester) for insightful discussions.

References

- Adams H, Emerson T, Kirby M, Neville R, Peterson C, Shipman P, Chepushtanova S, Hanson E, Motta F, Ziegelmeier L (2017) Persistence images: a stable vector representation of persistent homology. *J Mach Learn Res* 18(1):218–252
- Breusch TS, Pagan AR (1979) A simple test for heteroscedasticity and random coefficient variation. *Econ J Econ Soc* 47(5):1287–1294
- Bubenik P (2015) Statistical topological data analysis using persistence landscapes. *J Mach Learn Res* 16(1):77–102
- Carlsson G, Zomorodian A, Collins A, Guibas LJ (2005) Persistence barcodes for shapes. *Int J Shape Model* 11(02):149–187
- Carriere M, Cuturi M, Oudot S (2017) Sliced Wasserstein kernel for persistence diagrams. In: *Proceedings of the 34th international conference on machine learning-volume 70, JMLR. org*, pp 664–673
- Chazal F, Cohen-Steiner D, Mérigot Q (2011) Geometric inference for probability measures. *Found Comput Math* 11(6):733–751
- Chazal F, de Silva V, Oudot S (2014) Persistence stability for geometric complexes. *Geom Dedic* 173(1):193–214
- Chisholm JA, Motherwell S (2004) A new algorithm for performing three-dimensional searches of the cambridge structural database. *J Appl Crystallogr* 37(2):331–334
- Cohen-Steiner D, Edelsbrunner H, Harer J (2007) Stability of persistence diagrams. *Discrete Comput Geom* 37(1):103–120
- Cohen-Steiner D, Edelsbrunner H, Harer J, Mileyko Y (2010) Lipschitz functions have L_p -stable persistence. *Found Comput Math* 10(2):127–139
- Edelsbrunner H, Harer J (2010) *Computational topology: an introduction*. American Mathematical Society, Providence
- Edelsbrunner H, Letscher D, Zomorodian A (2000) Topological persistence and simplification. In: *Proceedings 41st annual symposium on foundations of computer science, 2000, IEEE*, pp 454–463
- Efron B, Hastie T (2016) *Computer age statistical inference*, vol 5. Cambridge University Press, Cambridge
- Fasy BT, Lecci F, Rinaldo A, Wasserman L, Balakrishnan S, Singh A et al (2014) Confidence sets for persistence diagrams. *Ann Stat* 42(6):2301–2339
- Gault B, Moody MP, De Geuser F, La Fontaine A, Stephenson LT, Haley D, Ringer SP (2010) Spatial resolution in atom probe tomography. *Microsc Microanal* 16(1):99–110
- Gault B, Moody MP, Cairney JM, Ringer SP (2012) Atom probe crystallography. *Materials Today* 15(9):378–386
- Goff M (2011) Extremal Betti numbers of Vietoris–Rips complexes. *Discrete Comput Geom* 46(1):132–155
- Hicks D, Oses C, Gossett E, Gomez G, Taylor RH, Toher C, Mehl MJ, Levy O, Curtarolo S (2018) Aflow-sym: platform for the complete, automatic and self-consistent symmetry analysis of crystals. *Acta Crystallogr Sect A Found Adv* 74(3):184–203
- Honeycutt JD, Andersen HC (1987) Molecular dynamics study of melting and freezing of small lennard-jones clusters. *J Phys Chem* 91(19):4950–4963
- Katsoulakis MA, Zabaras N (2017) Special issue: predictive multiscale materials modeling. *J Comput Phys* 338(1):18–20
- Kelly TF, Miller MK, Rajan K, Ringer SP (2013) Atomic-scale tomography: a 2020 vision. *Microsc Microanal* 19(3):652–664
- Larsen PM, Schmidt S, Schiøtz J (2016) Robust structural identification via polyhedral template matching. *Modell Simul Mater Sci Eng* 24(5):055007
- Larson DJ (2013) *Local electrode atom probe tomography: a user's guide*. Springer, Berlin

- Marchese A, Maroulas V (2016) Topological learning for acoustic signal identification. In: 2016 19th International conference on information fusion (FUSION), IEEE, pp 1377–1381
- Marchese A, Maroulas V (2018) Signal classification with a point process distance on the space of persistence diagrams. *Adv Data Anal Classif* 12(3):657–682
- Marchese A, Maroulas V, Mike J (2017) K-means clustering on the space of persistence diagrams. In: Wavelets and sparsity XVII, International Society for Optics and Photonics, vol 10394, p 103940W
- Maroulas V, Mike JL, Oballe C (2019) Nonparametric estimation of probability density functions of random persistence diagrams. *J Mach Learn Res* 20(151):1–49
- Maroulas V, Nasrin F, Oballe C (2019) A bayesian framework for persistent homology. *SIAM J Math Data Sci* Appear. [arXiv:1901.02034](https://arxiv.org/abs/1901.02034)
- McNutt NW, Rios O, Maroulas V, Keffer DJ (2017) Interfacial Li-ion localization in hierarchical carbon anodes. *Carbon* 111:828–834
- Miller MK (2014) Atom-probe tomography: the local electrode atom probe. Springer, Berlin
- Miller MK, Kelly TF, Rajan K, Ringer SP (2012) The future of atom probe tomography. *Mater Today* 15(4):158–165
- Moody MP, Gault B, Stephenson LT, Marceau RK, Powles RC, Ceguerra AV, Breen AJ, Ringer SP (2011) Lattice rectification in atom probe tomography: toward true three-dimensional atomic microscopy. *Microsc Microanal* 17(2):226–239
- Pfender F, Ziegler GM (2004) Kissing numbers, sphere packings, and some unexpected proofs. *Not Am Math Soc* 51:873–883
- Rost CM, Sachet E, Borman T, Moballeghe A, Dickey EC, Hou D, Jones JL, Curtarolo S, Maria JP (2015) Entropy-stabilized oxides. *Nat Commun* 6:8485
- Santodonato LJ, Zhang Y, Feygenson M, Parish CM, Gao MC, Weber RJ, Neuefeind JC, Tang Z, Liaw PK (2015) Deviation from high-entropy configurations in the atomic distributions of a multi-principal-element alloy. *Nat Commun* 6:5964
- Spannaus A, Maroulas V, Keffer DJ, Law KJH (2019) Bayesian point set registration. In: 2017 MATRIX Annals, Springer, pp 99–120
- Togo A, Tanaka I (2018) Spglib : a software library for crystal symmetry search. [arXiv:1808.01590](https://arxiv.org/abs/1808.01590)
- Wasserman L (2018) Topological data analysis. *Annu Rev Stat Appl* 5:501–532
- Zhang Y, Zuo TT, Tang Z, Gao MC, Dahmen KA, Liaw PK, Lu ZP (2014) Microstructures and properties of high-entropy alloys. *Prog Mater Sci* 61:1–93
- Ziletti A, Kumar D, Scheffler M, Ghiringhelli LM (2018) Insightful classification of crystal structures using deep learning. *Nat Commun* 9(1):2775
- Zomorodian A, Carlsson G (2005) Computing persistent homology. *Discrete Comput Geom* 33(2):249–274

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