

# A STABLE CARDINALITY DISTANCE FOR TOPOLOGICAL CLASSIFICATION

VASILEIOS MAROULAS, CASSIE PUTMAN MICUCCI, AND ADAM SPANNAUS

**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.

## 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 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 [21] and [31], 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 [8, 18, 19, 22, 32, 37] 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 [15] and [32]. 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 [40] 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 [40] is not a realistic representation of experimental APT data, where about 65% of the data is missing [35] and is corrupted by

---

*Key words and phrases.* Stability, Classification, Persistent Homology, Persistence Diagrams, Crystal Structure of Materials.

This work has been partially supported by the ARO Grant # W911NF-17-1-0313, and the NSF DMS-1821241.



FIGURE 1. Example of body-centered cubic, (BCC), (A) and face-centered cubic, (FCC), (B) 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.

more observational noise [31]. 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 [39] 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 [4, 12, 24, 26, 27, 28, 38, 41] 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 [5, 25] or vectorized and input into standard classification algorithms, such as support vector machines [1, 3].

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 [9]. Motivated by [25], 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 [7]; 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 is:

- (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 Section 2. Stability results of the  $d_p^c$  distance are in Section 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 Section 5.

## 2. PERSISTENT HOMOLOGY BACKGROUND

This section succinctly explains the construction of persistence diagrams, which are topological summaries of the underlying 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 [11].

**Definition 1.** A  $v$ -simplex is the convex hull of an affinely independent point set of size  $v + 1$ .

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

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

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  $\nu$ -simplex at the points corresponding to the intersection of  $\nu$  balls. See Fig. 2 for an illustration. For the Vietoris-Rips complex corresponding to  $\epsilon$ , denoted by  $VR_\epsilon$ , it is clear that  $VR_\epsilon \subset VR_{\epsilon'}$  for  $\epsilon < \epsilon'$ . Thus we need only examine specific  $\epsilon$  values corresponding to the emergence and disappearance of homological features. These  $\epsilon$  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 [38] 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 [14] and the distance to measure function as in [6]. 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  $W_p(X, Y) = (\inf_{\eta: X \rightarrow Y} \sum_{x \in X} \|x - \eta(x)\|_\infty^p)^{\frac{1}{p}}$ , where the infimum is taken over all bijections  $\eta$ , 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. 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, [25] 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

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

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

**Remark 2.** Note that this distance can be applied to arbitrary point clouds with finite cardinality as well. As shown in [25], 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  $d_p^c(X, Y) \rightarrow c$ .

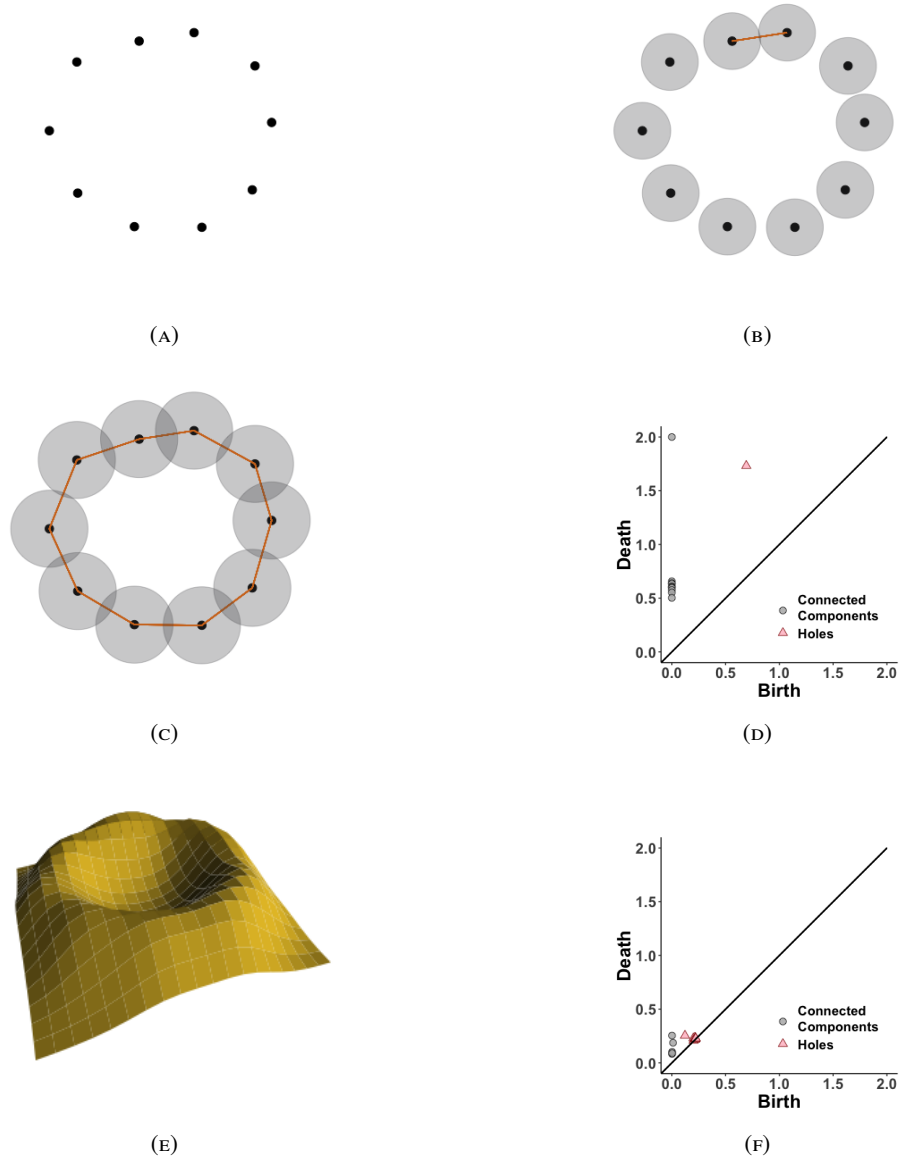


FIGURE 2. Begin with a point cloud (A). After increasing the radius of the balls around the points, a 1-simplex (line segment) forms in the corresponding Vietoris-Rips complex, (B). Eventually, more 1-simplices are added and a 1-dim hole forms (C). In (D), the persistence diagram tracks all the birth 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.

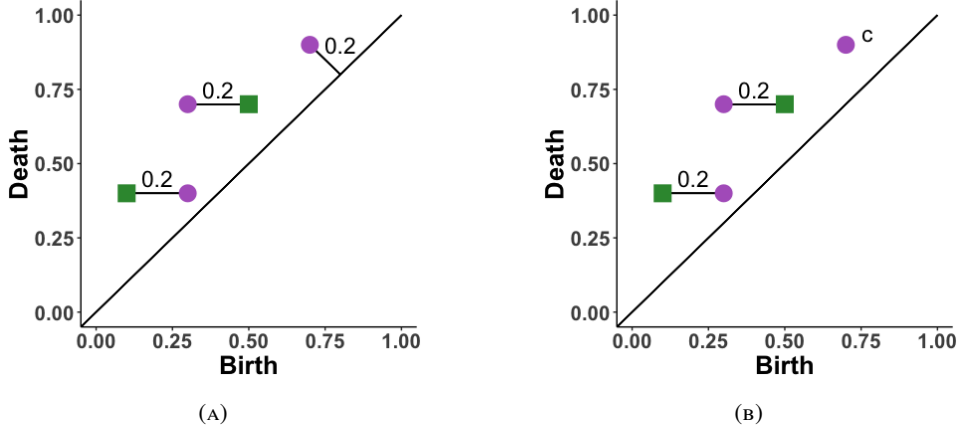


FIGURE 3. Consider two persistence diagrams, one given by the green squares and another by the purple circles. (A) The Wasserstein distance imposes a cost of 0.2 to the extra purple point (the  $\ell^\infty$ -distance to the diagonal). (B) The  $d_p^c$  distance imposes a penalty  $c$  on the point instead.

### 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 [14], 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 [10] and [9] 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 [14], 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 6.** *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 Def. 5

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

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 7.** *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  $\pi_i$  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  $\pi_i$  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 6, 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  $E = D_A - D_{A_i}$ .

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

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  $h_\eta \subset [0, \infty)$  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  $g_\eta$  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  $g_\eta < d_j^A \leq g_{\eta+1}$ . Let  $h_\eta < d_j^A < h_{\eta+1}$ , and suppose that it is not the case that  $h_\eta < d_j^A \leq h_{\eta+1}$ . 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 6, take  $|A_i| = |A|$  without loss of generality. By Lemma 7 (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 7 (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  $\pi_i$  in Lemma 7 (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  $\epsilon$  and a certain number of the distances are greater than  $\epsilon$ . Lemma 7 (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  $\pi_i$ .

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 8** ([33]). *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 9** ([17]). *For a finite point cloud with no more than  $\rho$  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*

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

**Proposition 10.** *Consider a point cloud in  $\mathbb{R}^d$  with  $\rho$  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  $\rho$  points that has  $b_1 = 0$ , simply take the  $\rho$  points and arrange them on a line. To form a point cloud with  $\rho$  points that has  $b_1 = \lfloor \frac{\rho}{2} \rfloor - 1$ , arrange the  $\rho$  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,



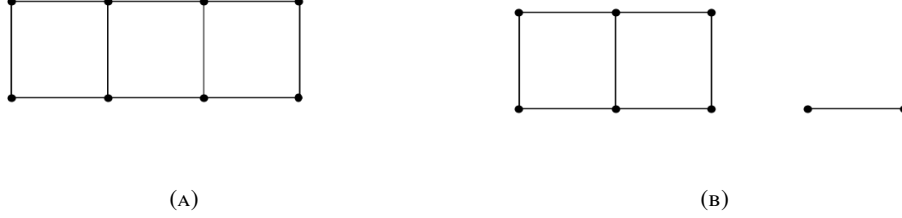


FIGURE 4. An example of 8-point arrangements to visualize the proof of Proposition 10. (A) A 3-hole configuration vs. (B) a 2-hole configuration.

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  $\rho$  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 9, the number of 1-dim holes in the Vietoris-Rips complex for a fixed radius  $\epsilon$  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  $\epsilon$  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 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 10 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 [13]. 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  $\mathbf{b}_1 = \mathbf{b}_0\gamma + \epsilon$ , where  $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ . The approximation is then given by  $\mathbf{b}_0\hat{\gamma} = \mathbf{b}_1$ , with  $\hat{\gamma} = (\mathbf{b}_0^T \mathbf{W} \mathbf{b}_0)^{-1} \mathbf{b}_0^T \mathbf{W} \mathbf{b}_1$ . In turn, Proposition 11 provides bounds from prediction intervals using weighted least squares for the  $d_p^c$  distance.

**Proposition 11.** *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  $\mathbf{b}_1 = \mathbf{b}_0 \gamma + \epsilon$ , 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

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

*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  $\widehat{b}_1^*$  is sought. To calculate this interval for a new response from the mean predicted response  $\widehat{b}_1^* = \widehat{\gamma} b_0^*$ , note that  $\widehat{b}_1^* - b_1^*$  has the distribution  $\frac{\widehat{b}_1^* - b_1^*}{\text{Var}(\widehat{b}_1^* - b_1^*)} \sim t_{N-2}$ . Also,  $\text{Var}(\widehat{b}_1^* - b_1^*) = \text{Var}(\epsilon)[1 \ b_0^*](\mathbf{b}_0^T \mathbf{W} \mathbf{b}_0)^{-1} [1 \ b_0^*]^T + \frac{\text{Var}(\epsilon)}{w^*}$ , where  $w^* = \frac{1}{b_0^*}$ , the weight corresponding to  $b_0^*$ . Prediction intervals for  $b_1^*$  are thus  $\widehat{b}_1^* \pm t_{1-\alpha/2, N-2} s \sqrt{[1 \ b_0^*](\mathbf{b}_0^T \mathbf{b}_0)^{-1} [1 \ b_0^*]^T + b_0^*}$ , where  $s^2 = \frac{\widehat{\epsilon}^T \mathbf{W} \widehat{\epsilon}}{N-2}$ , the unbiased estimator for  $\text{Var}(\epsilon)$ , using the residuals  $\widehat{\epsilon}$ . 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. 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 Figs. 1A and 1B). 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 Figs. 1A and 1B) one may see that a BCC unit cell has two atoms, one at the center and  $1/8^{\text{th}}$  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  $1/8^{\text{th}}$  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. Figs. 7c and 7d 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 homological 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 11. Given the salient topological

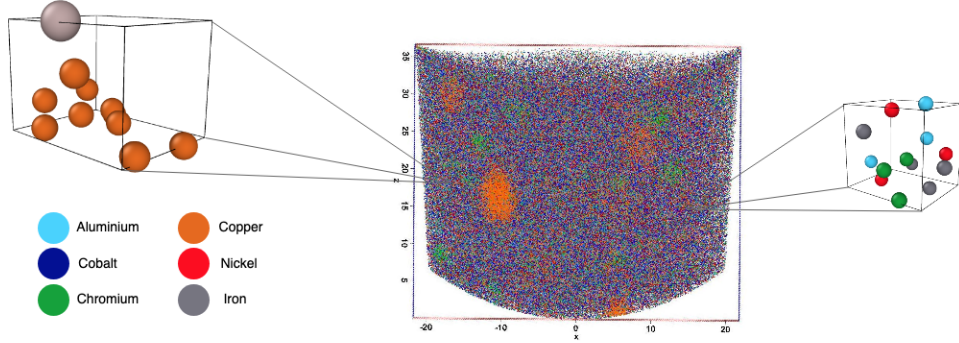


FIGURE 5. Image of APT data with atomic neighborhoods shown in detail on the left and right. Each pixel represents a different atom, the neighborhood of which is considered. Certain patterns with distinct crystal structures exist, e.g., the orange region is copper-rich (left), but overall no pattern 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.

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 Section 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,

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

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

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

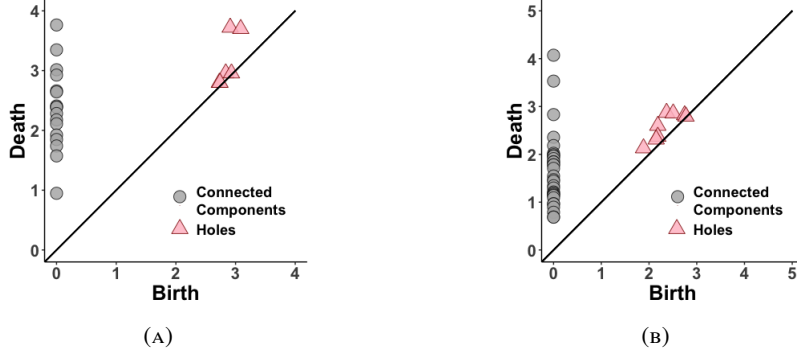


FIGURE 6. Example of persistence diagrams generated by (A) a BCC lattice, and (B) FCC lattice. The data has a noise standard deviation of  $\tau = 0.75$  and 67% of the atoms are missing. Note that the BCC diagram has two prominent (far from the diagonal) points representing 1-dim holes and fewer connected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

$\tau$	$c$ -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%

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

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 [31, 35] estimate that approximately 65% of the data is missing. However, an estimate of the experimental noise is not provided. In fact, as noted by [23, 30], the noise varies between

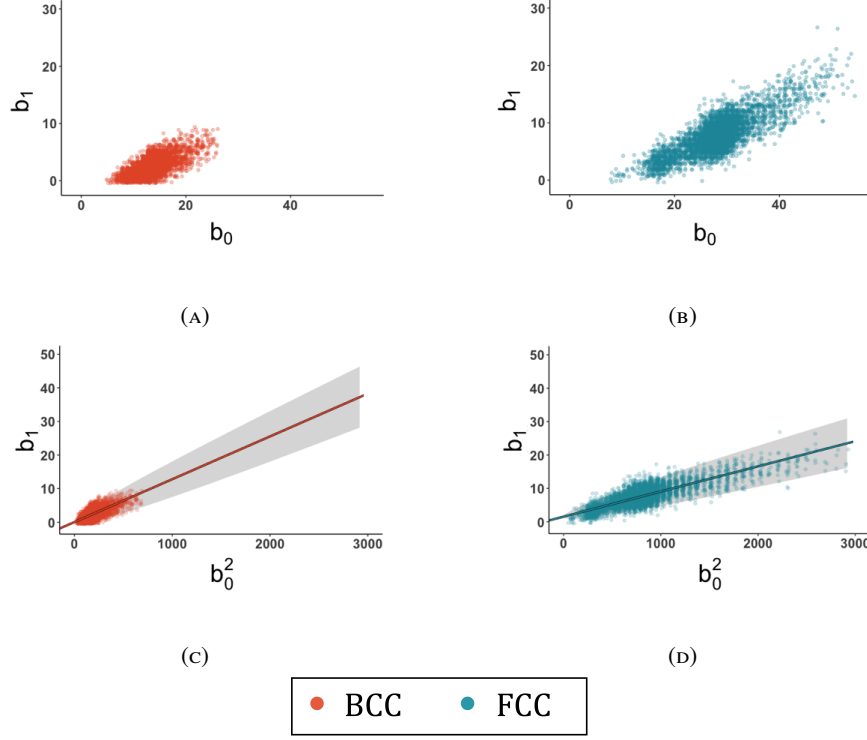


FIGURE 7. *Top*: Number of connected components (in this case atoms),  $\mathbf{b}_0$ , against the number of 1-dim homological features,  $\mathbf{b}_1$ , of the persistence diagrams. One can see the presence of heteroscedasticity since the variance of  $\mathbf{b}_1$  increases as  $\mathbf{b}_0$  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 11.

experiments and specimens. Our synthetic data replicates this resolution by drawing from a Gaussian [16, 29, 32],  $\mathcal{N}(0, \tau^2)$ , 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,  $\tau^2$ , 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 1,000 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

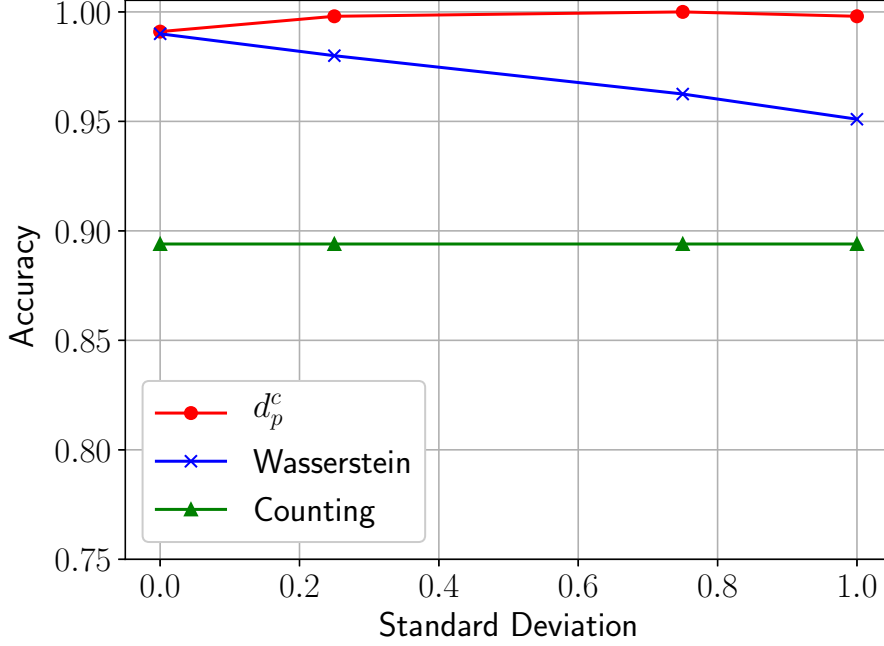


FIGURE 8. 10-fold cross validation accuracy scores for  $d_p^c$  (red), Wasserstein (blue), and counting (green) classifiers, plotted against different standard deviations,  $\tau$ , (see Table 1) of the normally distributed noise of the atomic positions. In each instance, the sparsity has been fixed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

in cardinality between the two classes of crystal structures is insufficient to distinguish between them.

As demonstrated in Proposition 11, 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 Figs. 7A and 7B 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 [2] with a  $p$ -value of  $9.3 \times 10^{-54}$  for FCC cells and a  $p$ -value of  $2.01 \times 10^{-47}$  for BCC cells. Figs. 7A and 7B also provide 95% prediction intervals for  $\mathbf{b}_1$  based on the weighted least squares regression analysis of Proposition 11. 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 [34] 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 [20]. The  $d_p^c$  classifier, coupled with our ongoing research of quantifying local atomic distributions as in [36], aims to recover global atomic structure of high entropy alloys.

#### ACKNOWLEDGMENTS

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

- [1] ADAMS, H., EMERSON, T., KIRBY, M., NEVILLE, R., PETERSON, C., SHIPMAN, P., CHEPUSHTANOVA, S., HANSON, E., MOTTA, F., AND ZIEGELMEIER, L. Persistence images: A stable vector representation of persistent homology. *The Journal of Machine Learning Research* 18, 1 (2017), 218–252.
- [2] BREUSCH, T. S., AND PAGAN, A. R. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society* (1979), 1287–1294.
- [3] BUBENIK, P. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research* 16, 1 (2015), 77–102.
- [4] CARLSSON, G., ZOMORODIAN, A., COLLINS, A., AND GUIBAS, L. J. Persistence barcodes for shapes. *International Journal of Shape Modeling* 11, 02 (2005), 149–187.
- [5] CARRIERE, M., CUTURI, M., AND OUDOT, S. Sliced wasserstein kernel for persistence diagrams. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (2017), JMLR. org, pp. 664–673.
- [6] CHAZAL, F., COHEN-STEINER, D., AND MÉRIGOT, Q. Geometric inference for probability measures. *Foundations of Computational Mathematics* 11, 6 (2011), 733–751.
- [7] CHAZAL, F., DE SILVA, V., AND OUDOT, S. Persistence stability for geometric complexes. *Geometriae Dedicata* 173, 1 (Dec 2014), 193–214.
- [8] CHISHOLM, J. A., AND MOTHERWELL, S. A new algorithm for performing three-dimensional searches of the cambridge structural database. *Journal of applied crystallography* 37, 2 (2004), 331–334.
- [9] COHEN-STEINER, D., EDELSBRUNNER, H., AND HARER, J. Stability of persistence diagrams. *Discrete & Computational Geometry* 37, 1 (2007), 103–120.
- [10] COHEN-STEINER, D., EDELSBRUNNER, H., HARER, J., AND MILEYKO, Y. Lipschitz functions have lp-stable persistence. *Foundations of Computational Mathematics* 10, 2 (Apr 2010).

- [11] EDELSBRUNNER, H., AND HARER, J. *Computational Topology: An Introduction*. American Mathematical Society, Providence, RI, 2010.
- [12] EDELSBRUNNER, H., LETSCHER, D., AND ZOMORODIAN, A. Topological persistence and simplification. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on* (2000), IEEE, pp. 454–463.
- [13] EFRON, B., AND HASTIE, T. *Computer age statistical inference*, vol. 5. Cambridge University Press, 2016.
- [14] FASY, B. T., LECCI, F., RINALDO, A., WASSERMAN, L., BALAKRISHNAN, S., SINGH, A., ET AL. Confidence sets for persistence diagrams. *The Annals of Statistics* 42, 6 (2014), 2301–2339.
- [15] GAULT, B., MOODY, M. P., CAIRNEY, J. M., AND RINGER, S. P. Atom probe crystallography. *Materials Today* 15, 9 (2012), 378–386.
- [16] GAULT, B., MOODY, M. P., DE GEUSER, F., LA FONTAINE, A., STEPHENSON, L. T., HALEY, D., AND RINGER, S. P. Spatial resolution in atom probe tomography. *Microscopy and Microanalysis* 16, 1 (2010), 99–110.
- [17] GOFF, M. Extremal betti numbers of victoris–rips complexes. *Discrete & Computational Geometry* 46, 1 (2011), 132–155.
- [18] HICKS, D., OSES, C., GOSSETT, E., GOMEZ, G., TAYLOR, R. H., TOHER, C., MEHL, M. J., LEVY, O., AND CURTAROLO, S. Aflow-sym: platform for the complete, automatic and self-consistent symmetry analysis of crystals. *Acta Crystallographica Section A: Foundations and Advances* 74, 3 (2018), 184–203.
- [19] HONEYCUTT, J. D., AND ANDERSEN, H. C. Molecular dynamics study of melting and freezing of small lennard-jones clusters. *Journal of Physical Chemistry* 91, 19 (1987), 4950–4963.
- [20] KATSOLAKIS, M. A., AND ZABARAS, N. Special issue: Predictive multiscale materials modeling. *Journal of Computational Physics* 338, 1 (2017).
- [21] KELLY, T. F., MILLER, M. K., RAJAN, K., AND RINGER, S. P. Atomic-scale tomography: A 2020 vision. *Microscopy and Microanalysis* 19, 3 (2013), 652–664.
- [22] LARSEN, P. M., SCHMIDT, S., AND SCHIØTZ, J. Robust structural identification via polyhedral template matching. *Modelling and Simulation in Materials Science and Engineering* 24, 5 (2016), 055007.
- [23] LARSON, D. J. *Local Electrode Atom Probe Tomography : A User's Guide*. Springer, 2013.
- [24] MARCHESE, A., AND MAROULAS, V. Topological learning for acoustic signal identification. In *Information Fusion (FUSION), 2016 19th International Conference on* (2016), IEEE, pp. 1377–1381.
- [25] MARCHESE, A., AND MAROULAS, V. Signal classification with a point process distance on the space of persistence diagrams. *Advances in Data Analysis and Classification* 12, 3 (2018), 657–682.
- [26] MARCHESE, A., MAROULAS, V., AND MIKE, J. K-means clustering on the space of persistence diagrams. In *Wavelets and Sparsity XVII* (2017), vol. 10394, International Society for Optics and Photonics, p. 103940W.
- [27] MAROULAS, V., MIKE, J. L., AND OBALLE, C. Nonparametric Estimation of Probability Density Functions of Random Persistence Diagrams. *Journal of Machine Learning Research* (2019), arXiv:1803.02739.
- [28] MAROULAS, V., NASRIN, F., AND OBALLE, C. Bayesian inference for persistent homology. *arXiv preprint arXiv:1901.02034* (2019).
- [29] MCNUTT, N. W., RIOS, O., MAROULAS, V., AND KEFFER, D. J. Interfacial Li-ion localization in hierarchical carbon anodes. *Carbon* 111 (2017), 828–834.
- [30] MILLER, M. K. *Atom-Probe Tomography : The Local Electrode Atom Probe*. Springer, 2014.
- [31] MILLER, M. K., KELLY, T. F., RAJAN, K., AND RINGER, S. P. The future of atom probe tomography. *Materials Today* 15, 4 (2012), 158–165.
- [32] MOODY, M. P., GAULT, B., STEPHENSON, L. T., MARCEAU, R. K., POWLES, R. C., CEGUERRA, A. V., BREEN, A. J., AND RINGER, S. P. Lattice rectification in atom probe tomography: Toward true three-dimensional atomic microscopy. *Microscopy and Microanalysis* 17, 2 (2011), 226–239.
- [33] PFENDER, F., AND ZIEGLER, G. M. Kissing numbers, sphere packings, and some unexpected proofs. *Notices-American Mathematical Society* 51 (2004), 873–883.
- [34] ROST, C. M., SACHET, E., BORMAN, T., MOBALLEGH, A., DICKEY, E. C., HOU, D., JONES, J. L., CURTAROLO, S., AND MARIA, J.-P. Entropy-stabilized oxides. *Nature communications* 6 (2015), 8485.
- [35] SANTODONATO, L. J., ZHANG, Y., FEYGENSON, M., PARISH, C. M., GAO, M. C., WEBER, R. J., NEUEFEIND, J. C., TANG, Z., AND LIAW, P. K. Deviation from high-entropy configurations in the atomic distributions of a multi-principal-element alloy. *Nature communications* 6 (2015), 5964.
- [36] SPANNAUS, A., MAROULAS, V., KEFFER, D. J., AND LAW, K. J. H. Bayesian point set registration. In *2017 MATRIX Annals*. Springer, 2019, pp. 99–120.
- [37] TOGO, A., AND TANAKA, I. Spglib : a software library for crystal symmetry search. *arXiv preprint arXiv:1808.01590* (2018).
- [38] WASSERMAN, L. Topological data analysis. *Annual Review of Statistics and Its Application* 5 (2018), 501–532.
- [39] ZHANG, Y., ZUO, T. T., TANG, Z., GAO, M. C., DAHMEN, K. A., LIAW, P. K., AND LU, Z. P. Microstructures and properties of high-entropy alloys. *Progress in Materials Science* 61 (2014).



- [40] ZILETTI, A., KUMAR, D., SCHEFFLER, M., AND GHIRINGHELLI, L. M. Insightful classification of crystal structures using deep learning. *Nature communications* 9, 1 (2018), 2775.
- [41] ZOMORODIAN, A., AND CARLSSON, G. Computing persistent homology. *Discrete & Computational Geometry* 33, 2 (2005), 249–274.

DEPARTMENT OF MATHEMATICS - UNIVERSITY OF TENNESSEE, KNOXVILLE, TN 37996  
*E-mail address*, Corresponding author: vmaroula@utk.edu

DEPARTMENT OF MATHEMATICS - UNIVERSITY OF TENNESSEE, KNOXVILLE, TN 37996  
*E-mail address*: cputman@vols.utk.edu

DEPARTMENT OF MATHEMATICS - UNIVERSITY OF TENNESSEE, KNOXVILLE, TN 37996  
*E-mail address*: aspannaus@utk.edu