Chapter 5
Persistent Homology and Materials Informatics

Mickaël Buchet, Yasuaki Hiraoka and Ippei Obayashi

Abstract This paper provides an introduction to persistent homology and a survey of its applications to materials science. Mathematical prerequisites are limited to elementary linear algebra. Important concepts in topological data analysis such as persistent homology and persistence diagram are explained in a self-contained manner with several examples. These tools are applied to glass structural analysis, crystallization of granular systems, and craze formation of polymers.

Keywords Persistent homology · Materials informatics · Topological data analysis

5.1 Introduction

Because of the rapid growth of computers, internet, and experimental measurement devices, huge amounts of data are now available and they induce drastic changes in scientific activities. Namely, data-driven science has recently emerged and this new trend also applies to materials science, leading to a new concept of materials informatics. The basic strategy is to try to capture meaningful information embedded in the database using machine learning. The readers can discover results at the frontiers of materials informatics from some papers in this book.

A key to the success of materials informatics is to select compact descriptors of data to appropriately study materials properties. Available data is large and compli-

M. Buchet · Y. Hiraoka (✉) · I. Obayashi
Advanced Institute for Materials Research (WPI-AIMR), Tohoku University, 2 Chome-1-1 Katahira, Aoba Ward, Sendai 980-8577, Japan
e-mail: hiraoka@tohoku.ac.jp

Y. Hiraoka
Center for Materials research by Information Integration (CMI2), Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), 1 Chome-2-1 Sengen, Ibaraki Prefecture, Tsukuba 305-0047, Japan

Y. Hiraoka
Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan

I. Tanaka (ed.), *Nanoinformatics*, https://doi.org/10.1007/978-981-10-7617-6_5

cated. Therefore, good descriptors are required for efficient applications of machine learning, expanding the possibilities beyond conventional descriptors.

This story applies not only to materials science, but also to various communities in science and technology. Topological data analysis (TDA) has emerged in this century [1] and shed a new light on data science. A distinguishing property of TDA is that it provides tools for capturing the *shape of data* in a multi-scale way. They capture topological and geometric features embedded in data and enable the study of relationships of those detected features in different scales. Nowadays, topological data analysis is applied to a wide variety of scientific and industrial areas (e.g., materials science, life science, neuroscience, and social networks).

A particularly important tool in TDA is persistent homology and persistence diagrams. Briefly speaking, these tools describe topological features characterized by holes in data (components, rings, cavities, etc.). Practically, the input to persistent homology is usually given as a finite point set in a Euclidean space or digital images of any dimension. In materials science, atomic (or particle) configurations obtained by molecular dynamics simulations as well as digital images observed by experiments can be studied by these tools. The persistence diagram is a two-dimensional histogram compactly expressing the output of persistent homology. Based on this visualization, we can easily study higher dimensional topological features in a multi-scale way.

The purpose of this paper is to provide a self-contained introduction to persistent homology and survey several applications to materials science [2–5]. We only assume knowledge of elementary linear algebra and show several examples to help the readers' understanding. We hope that this paper will be useful for materials scientists to get used to persistent homology.[1]

5.2 Mathematical Background

First, we review the mathematical background behind topological data analysis. Our goal is to provide both a rigorous mathematical development and easily understandable intuition. The aim of topological data analysis is to provide an understanding of the structure of data. For that, we first need to define what we are looking for and then describe how to extract this information.

5.2.1 Homology

The structure we study is called homology. While homology is not as descriptive as the maybe more classical concept of homotopy, it does present the undeniable

[1]The readers can obtain further information of materials TDA project organized by our group from the website http://www.wpi-aimr.tohoku.ac.jp/hiraoka_labo/index.html.

advantage of being computable. For the sake of simplicity, we will only introduce the concept of simplicial homology.

We will endeavor to present the concept from the algebraic side while maintaining a geometric intuition. We fix a set called the set of indices. In our case, we will only use the set of integers $\mathbb{N}$.

Definition 5.2.1 A k-simplex is a set of $k+1$ indices.

This very simple definition describes an abstract simplex. It can have an intuitive geometric counterpart. Given a set of points numbered by indices, the geometric k-simplex corresponding to a subset of indices is the convex hull of the subset of points corresponding these indices. Within this geometric framework, a 0-simplex is simply a point, a 1-simplex is an edge, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, and so on (see Fig. 5.1).

Definition 5.2.2 A simplicial complex X is a set of simplices such that for any $\sigma \in X$ and any $\sigma' \subset \sigma$, $\sigma' \in X$.

Therefore, a simplicial complex is a set of simplices with a very natural and simple rule ensuring coherence. For example, if a triangle belongs to the simplicial complex X, then the three edges that border it also belong to X as well as the three vertices. Figure 5.2 illustrates this property. While the left object is a simplicial complex, the middle one is not because the edge e is missing while the upper triangle exists. The right one is also incorrect. A consequence of the definition is that the intersection of two simplices is either empty or a simplex belonging to the simplicial complex. Here p is the intersection of two simplices but it does not appear as a simplex. Note that just adding p would not be sufficient to fix the construction.

We now introduce an algebraic notion of orientation to our simplices. Namely, we fix an ordering on the indices.

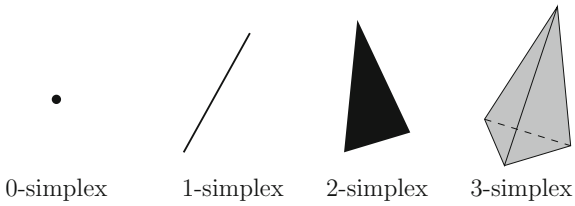

Fig. 5.1 Example of geometric simplices

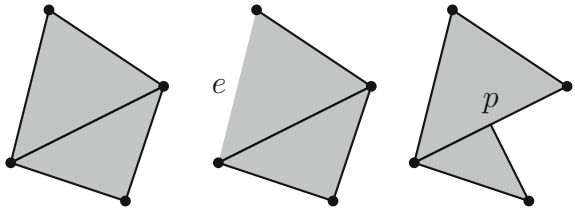

Fig. 5.2 One simplicial complex (left) and some objects that are not simplicial complexes

Definition 5.2.3 Given a set of indices $\{v_1, \dots, v_k\}$, we define the oriented simplex $\sigma = [v_1, \dots, v_k]$ as an ordered set. The opposite simplex is obtained by permuting two indices: $[v_1, \dots v_i, \dots, v_j, \dots, v_k] = -[v_1, \dots, v_j, \dots, v_i, \dots, v_k]$.

We choose a field k in order to study the topology of simplicial complexes with the use of homology. Given a simplicial complex X, let $X^{(n)}$ be the set of all n-simplices of X. We use this set as the generating elements of the k-vector space $\Delta_n(X)$. This means that an element of $\Delta_n(X)$ is of the form $\sum_{\sigma \in X^{(n)}} \alpha_\sigma \sigma$ where $\{\alpha_\sigma\}$ are coefficient in k. The addition operation is naturally $\sum_{\sigma \in X^{(n)}} \alpha_\sigma \sigma + \sum_{\sigma \in X^{(n)}} \alpha'_\sigma \sigma = \sum_{\sigma \in X^{(n)}} (\alpha_\sigma + \alpha'_\sigma)\sigma$.

The next tool we need is to describe faces of a given simplex σ. We do so by indicating which vertex is opposite to it.

Definition 5.2.4 Given an ordered n-simplex $\sigma = [v_0, \dots, v_n]$, we write $[v_0, \dots \hat{v}_i, \dots v_n]$ the $(n-1)$-simplex obtained by removing the index v_i.

Note that if an n-simplex σ belongs to a simplicial complex X, any one of its faces is a $(n-1)$-simplex and also belongs to X. We can hence define the following map.

Definition 5.2.5 Given a simplicial complex X, the boundary map $\partial_n : \Delta_n(X) \to \Delta_{n-1}(X)$ is defined on the basis elements by:

$$\partial_n([v_0, \dots, v_n]) = \sum_{i=0}^{n} (-1)^i [v_0, \dots \hat{v}_i, \dots v_n].$$

For example, the definition for $n = 1, 2$ is given by $\partial_1([v_0, v_1]) = [v_1] - [v_0]$ and $\partial_2([v_0, v_1, v_2]) = [v_1, v_2] - [v_0, v_2] + [v_0, v_1]$. By extending this operator to all elements of $\Delta_n(X)$, we obtain a linear map. Geometrically, the boundary operator extracts the boundary of a chain while respecting the orientation (see Fig. 5.3).

By combining these operations for each dimension, we obtain the chain complex:

$$\cdots \longrightarrow \Delta_{n+1}(X) \xrightarrow{\partial_{n+1}} \Delta_n(X) \xrightarrow{\partial_n} \cdots \longrightarrow \Delta_1(X) \xrightarrow{\partial_1} \Delta_0(X) \xrightarrow{\partial_0} 0$$

Note that the composition of two consecutive boundary operators is zero. In other words, for any n, $\partial_{n-1}\partial_n = 0$. This property expresses the geometric fact that the boundary of the boundary of an object is empty.

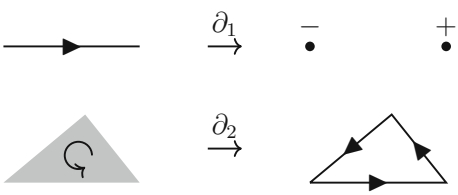

Fig. 5.3 Examples of boundaries

Let Ker $\partial_n = \{c \in \Delta_n(X) : \partial_n c = 0\}$ and Im $\partial_{n+1} = \{c \in \Delta_n(X) : c = \partial_{n+1} c',\ c' \in \Delta_{n+1}(X)\}$, be the kernel and the image of the boundary maps. From the above property, we have Im $\partial_{n+1} \subset$ Ker ∂_n. We can thus define homology by quotienting subspaces.

Definition 5.2.6 The n-dimensional homology of X is defined as $H_n(X) =$ Ker ∂_n/Im ∂_{n+1}.

Intuitively, homology describes holes of the structure. By counting generators of homology, we obtain the Betti numbers which count topological features. The Betti number in dimension 0 gives the number of connected components. In dimension 1, it corresponds to the number of holes and in dimension 2 to the number of cavities, and then generalizes to higher dimensions.

We now give an example of a simplicial complex with five vertices in Fig. 5.4 and compute its homology.

In this simplicial complex, the simplex of highest dimension is the 2-simplex, a.k.a. triangle, $[1, 2, 3]$. Therefore, $\Delta_2(X) = k[1, 2, 3]$. Looking at dimension 1 simplices, we can see five different edges. Therefore, $\Delta_1(X) = k[1, 4] \oplus k[4, 2] \oplus k[1, 2] \oplus k[2, 3] \oplus k[1, 3]$. Finally, we have 5 points and, therefore, $\Delta_0(X) = k[1] \oplus k[2] \oplus k[3] \oplus k[4] \oplus k[5]$.

First, remark that for any dimension $n \geq 3$, the boundary map $\partial_n = 0$ and, therefore, Ker $\partial_n = 0$ and $H_n(X) = 0$. We first need to consider the matrix associated with ∂_2. Writing the matrix M_2 associated with the boundary map ∂_2, we obtain,

$$M_2 = \begin{array}{c} \begin{matrix} [1,2,3] \end{matrix} \\ \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ -1 \end{pmatrix} \end{array} \begin{array}{l} [1,4] \\ [4,2] \\ [1,2] \\ [2,3] \\ [1,3] \end{array}$$

We can immediately deduce that Ker $\partial_2 = 0$ and Im $\partial_2 = k([1, 2] + [2, 3] - [1, 3])$. Hence $H_2(X) =$ Ker ∂_2/Im $\partial_3 = 0$. To compute $H_1(X)$, we also need to consider the matrix M_1 associated with ∂_1.

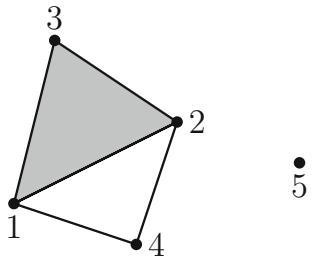

Fig. 5.4 Example of a simplicial complex

$$M_1 = \begin{array}{c} \begin{matrix} [1,4] & [4,2] & [1,2] & [2,3] & [1,3] \end{matrix} \\ \begin{pmatrix} -1 & 0 & -1 & 0 & -1 \\ 0 & 1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} [1] \\ [2] \\ [3] \\ [4] \\ [5] \end{matrix}$$

A simple computation yields that $\text{Ker}\ \partial_1 = k([1,4] + [4,2] - [1,2]) + k([1,2] + [2,3] - [1,3])$. Therefore, the homology $H_1(X) = \text{Ker}\ \partial_1 / \text{Im}\ \partial_2 = k([1,4] + [4,2] - [1,2] + \text{Im}\ \partial_2) \cong k$. In other words, the one-dimensional homology is isomorphic to k and, therefore, has dimension 1. It means that there exists one hole. Moreover, one possible representative of the class is the cycle $[1,4] + [4,2] - [1,2]$. Note that this representative is not unique as $[1,4] + [4,2] + [2,3] - [1,3]$ is also a representative of the same class. Intuitively, the quotient operation means that given a cycle in dimension d, we can add or remove the boundary of simplices of dimension $d+1$ without changing the equivalence class. In our example, the cycle corresponding to the hole is equivalent to the one obtained by adding the boundary of the triangle $[1,2,3]$ to it.

To finish, remark that $\text{Im}\ \partial_1 = k([4] - [1]) + k([2] - [4]) + k([3] - [2])$ and that ∂_0 is a zero map. Therefore, $\text{Ker}\ \partial_0 = k[1] + k[2] + k[3] + k[4] + k[5]$ and $H_0(X) \cong k^2$ which indicates the presence of two connected components.

5.2.2 From Point Sets to Simplicial Complexes

The construction of simplicial homology relies on simplicial complexes. The first task is to build such a simplicial complex from our data. We consider here an input given as a set of points $P \subset \mathbb{R}^d$ in a Euclidean space. We want to build a geometric simplicial complex, id est a continuous space, from the point set P which is a discrete space. To do so, we consider balls around these points.

Given a radius r and a point x, we denote $B(x,r)$ the ball centered at x and of radius r. We consider the union $\cup_{x \in P} B(x,r)$ of all balls of radius r centered at points of P. We define the nerve of the union of balls also called the Čech complex, which is a geometric simplicial complex whose vertices are the points of P.

Definition 5.2.7 The Čech complex is defined as $C_r(P) = \{\sigma | \cap_{p \in \sigma} B(p,r) \neq \emptyset\}$.

Each point is associated with a ball. Note that all the balls are non-empty if $r > 0$ and, therefore, all points of P belong to the Čech complex. An edge belongs to the complex if and only if the two balls corresponding to its extremities intersect. Similarly, a triangle requires the common intersection of its three vertices' balls to be non-empty to belong to the Čech complex.

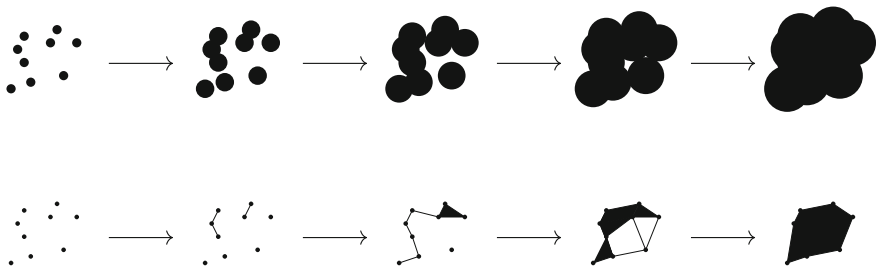

Fig. 5.5 Example of Čech complex

Considering the Čech complex is enough to study the topology of the union of balls as the Nerve Theorem [6, 4G.3] implies:

Proposition 5.2.8 *Given a set of points P in a Euclidean space and a radius r, the union of balls $\cup_{p\in P} B(p,r)$ and the Čech complex $C_r(P)$ are homotopy equivalent.*

Intuitively, two spaces are homotopy equivalent if we can deform continuously one into the other. Therefore, they have the same topological structure and studying the homology of one is equivalent to studying the homology of the other one. The construction is illustrated in Fig. 5.5.

It is important to note that the construction can be made with any union of balls. The Nerve Theorem is not limited to Čech complexes. From an applicative standpoint in material science, the notion of weighted Čech complexes is especially useful. When the input is a set of atomic positions with different type of atoms, we can reflect the size of each particular atom by modifying the radius accordingly. We obtain a union of balls with different radii, bigger atoms having larger balls.

5.2.3 Persistent Homology

A major problem that arises is the choice of the radius r. Choosing a radius gives a snapshot of the topology at the corresponding scale but does not encapture the whole topological structure. Persistent homology is a tool that allows multi-scale analysis. Instead of looking at one given radius, we can look at the evolution of topological features across scales.

In the context of material science, this allows to not only detect topological features but also to classify them depending on their scale. This is related to the diameter and the geometry of holes and cavities.

First, notice that the union of balls we considered previously possesses a natural inclusion when the radius increases. Given some radii $r_1 \leq \cdots \leq r_i \leq \cdots \leq r_l$, we have:

$$\cup_p B(p,r_1) \hookrightarrow \cup_p B(p,r_2) \hookrightarrow \cdots \hookrightarrow \cup_p B(p,r_i) \hookrightarrow \cdots \hookrightarrow \cup_p B(p,r_l).$$

This sequence can be transformed in a sequence of inclusions between simplicial complexes by taking the nerve of each union of balls. We obtain the following Čech filtration.

$$C_{r_1}(P) \hookrightarrow C_{r_2}(P) \hookrightarrow \cdots \hookrightarrow C_{r_i}(P) \hookrightarrow \cdots \hookrightarrow C_{r_l}(P).$$

We then use the homological construction for each of these spaces to obtain a sequence of vector spaces linked by linear maps. We denote $H_n(C_r(P))$ the homology vector space built using $C_r(P)$ for a given dimension n. Since the choice of the working dimension does not have an influence on the theoretical results, we indicate it by writing $H_*(C_r(P))$.

Definition 5.2.9 Given an ordered index set I and a field k, a persistence module H is a sequence $(\Phi_i)_{i\in I}$ of vector spaces and linear maps $(\phi_i^j)_{i\leq j}$ where $\phi_i^j : \Phi_i \to \Phi_j$ and for all $i \leq j \leq k$, $\phi_i^k = \phi_j^k \circ \phi_i^j$.

A persistence module is a sequence of vector spaces linked by linear maps. The condition on the linear maps is that they commute. Intuitively, this means that we can decompose and recompose them. Working on the previous chain sequence, we build at homology level the following persistence module.

$$H_*(C_{r_1}(P)) \to H_*(C_{r_2}(P)) \to \cdots \to H_*(C_{r_i}(P)) \to \cdots \to H_*(C_{r_l}(P))$$

The Persistent Nerve Lemma [7] guarantees that this persistent module is isomorphic to the one we can build using the union of balls. Therefore, studying the Čech filtered complex is equivalent to studying the filtered union of balls.

The critical property of the persistence module is its decomposability. Indecomposables, in other words, the building blocks, are called interval modules. They consist of a sequence of one-dimensional vector spaces linked by identity maps.

$$0 \to k \to k \to k \to 0 \to 0$$

In this example of an interval module, we have six values of indices we name $\{1, \ldots, 6\}$. The interval spans from the second to the fourth so we denote it $I[2, 4]$. All maps between the nonzero vector spaces are identity maps.

The following property ensures that the persistence modules we consider are uniquely decomposables into a direct sum of intervals.

Proposition 5.2.10 *A persistence module whose every vector space is finite dimensional is uniquely decomposable into a direct sum of interval modules.*

Note that in our setting, we build finite simplicial complexes from finite point sets. Therefore, everything is finite, especially the dimension of the vector spaces. Thus the Proposition applies. There exist various more general variants [8, 9] of this result but we limit ourselves to this one for the sake of simplicity.

Intuitively, intervals have a birth, the first index where the vector space is nonzero, and a death, the first index where the vector space is zero after having been nonzero. The first index for which a simplex σ belongs to the complex is called the apparition time of σ. Intervals correspond to the existence of topological features. In the case of a one-dimensional cycle, for example, the birth corresponds to the apparition time of the edge forming the cycle and the death corresponds to the apparition time of the triangle that fills it completely.

Formally, a persistence module H can be associated with a set of pairs (b_i, d_i) such that:

$$H = \bigoplus I[b_i, d_i]$$

We can represent each of the interval $I[b, d]$ as a bar starting at b and ending at d. We thus obtain a figure called barcode that describes the decomposition of the persistence module. Figure 5.6 shows an example of barcode.

There exists a natural bijection from barcodes to multi-sets of $\mathbb{R}^2$ denoted $D = \{(b, d)\}$. This multi-set is called a persistence diagram (PD for short) and is often represented as in Fig. 5.7.

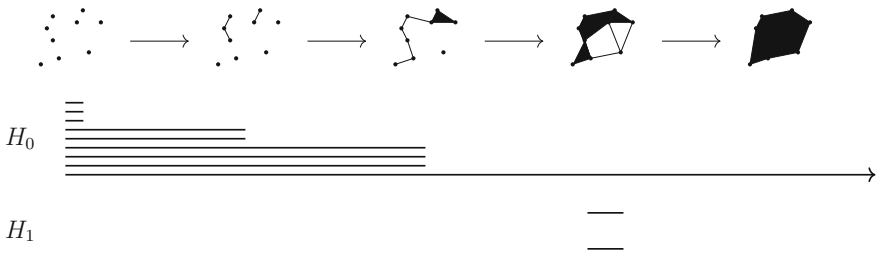

Fig. 5.6 Simplicial complex, topological features, and barcode for zero and one-dimensional homology

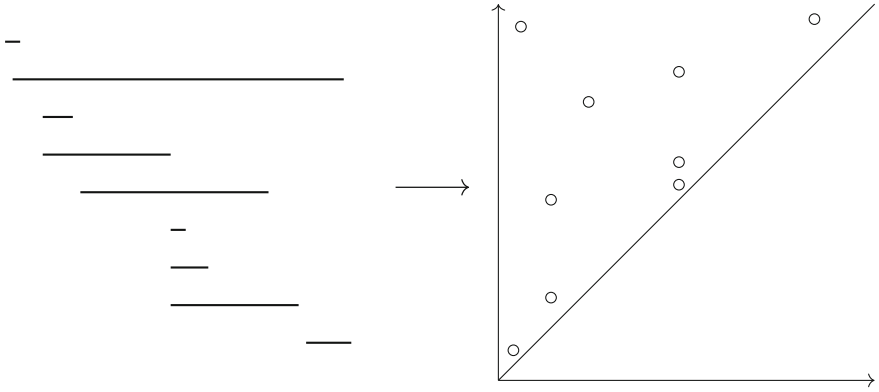

Fig. 5.7 From barcode to persistence diagram

Interpretation of persistence diagrams reveals two different kinds of information. First, it indicates, which features are probably relevant as they are those far away from the diagonal. Second, it can separate features according to a combination of size and shape that contribute to their lifespan.

5.2.4 Computation

From a computational point of view, persistent homology is very intuitive. Considering that we build the simplicial complex from scratch, we add one simplex at a time according to their apparition time. If multiple simplices are added at the same time, we can arbitrarily choose the order in which we insert them. This allows us to maintain a simplicial complex at all steps.

When a d-simplex is inserted, there are two possible cases. Either the simplex is *negative* which means that it destroys a $(d-1)$-dimensional feature, or it is *positive* and creates a d-dimensional feature. Figure 5.8 shows the two kinds of 1-simplices. Note that the object on the left has two connected components and no cycle. The first edge we introduce kills one of the connected component and, therefore, is negative. The second one has its two extremities in the same connected component and, therefore, is positive, creating a cycle.

To compute the barcode, a positive simplex is trivial to handle. We just need to create a new bar. However, a negative simplex is more complicated. We need to find which feature is being killed and that is nontrivial. In our example, we do not know which of the two connected component should be considered as dead and which one is still alive. Persistent homology follows the rule that the oldest one survives. Therefore, we kill the one that appeared last.

This very intuitive algorithm has an algebraic counterpart. We build a boundary matrix that contains the incidence information of all simplices. Each column and row represent a simplex and they are ordered by apparition time. Rows are the boundaries of columns.

Computing persistent homology is equivalent to reducing that matrix with the following rules. Every time we introduce a new simplex, id est a new column, we are free to use the columns on the left and add multiple of them to the new column. Any

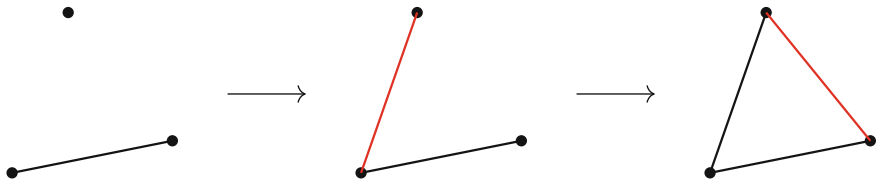

Fig. 5.8 Insertion of the two types of edges

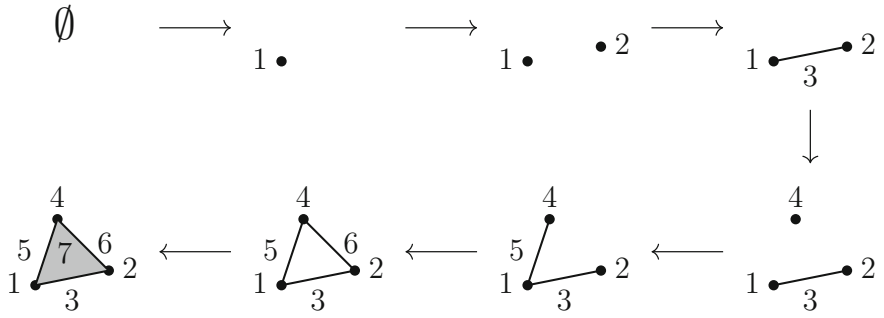

Fig. 5.9 Filtration on a triangle

zero column corresponds to a topological feature. A nonzero column corresponds to the death of the feature created at the time of the lowest nonzero index.

We now provide a simple example and do the whole computation. We build a complex containing a triangle, its edges and vertices filtered in the order shown in Fig. 5.9.

We fix an arbitrary orientation on every simplex by sorting indices in increasing order of apparition. Therefore, we consider the boundary of edge [3] to be [2] − [1]. We then obtain the following boundary matrix.

$$\begin{array}{c} \\ [1] \\ [2] \\ [3] \\ [4] \\ [5] \\ [6] \\ [7] \end{array} \begin{array}{c} \begin{array}{ccccccc} [1] & [2] & [3] & [4] & [5] & [6] & [7] \end{array} \\ \begin{pmatrix} 0 & 0 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

First, note that this matrix is upper triangular. This is a direct consequence of having a filtered complex. A simplex cannot appear before one of its faces.

We now do the computation for this example. First, we introduce columns [1] and [2] which are zero and corresponds to 0-simplices. Therefore, it creates two connected components. Then we add [3] which cannot be reduced by elements on its left and, therefore, kills a feature. The lowest nonzero entry corresponds to line [2] so [3] kills the feature created by [2]. In the same way, [4] creates a new connected component killed by [5].

The insertion of [6], however, introduces a column that can be reduced using columns located on its left. More precisely [6] = [5] − [3]. Note that it is easy to detect such a case as it suffices to look at the lowest nonzero entry, cancel it and then recurse. Hence [6] creates a cycle, id est a one-dimensional feature, which is then killed by the insertion of [7].

The resulting matrix can be expressed as:

$$\begin{array}{c} \\ [1] \\ [2] \\ [3] \\ [4] \\ [5] \\ [6] \\ [7] \end{array} \begin{array}{c} \begin{array}{ccccccc} [1] & [2] & [3] & [4] & [5] & [6]+[3]-[5] & [7] \end{array} \\ \begin{pmatrix} 0 & 0 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Note that the algorithm provides a few extra information for free. We obtain matches between positive simplices and negative ones. Moreover, we get a representant of each homology class being created. Here, the cycle can be represented by $[6] + [3] - [5]$. Beware that this representant is not necessarily the unique representant in its class nor looks good from a geometric point of view. Its structure is disconnected from the geometry.

This algorithm has a worst case running time that is cubic in the number of simplices. In practice, however, implementations work much faster, mostly because of the sparsity of the boundary matrix. There are numerous libraries that compute persistent homology and that are aimed at different public. Some of the most recent ones are the TDA package in R [10] intended for statisticians, DIPHA [11] and GUDHI [12] that are state-of-the-art approaches from the computational topology community or HomCloud [13] which aims at a more experimentalist public with additional tools and graphical output. This list is non-exhaustive and many more exist.

5.2.5 *Digital Images*

Until now we focused on point sets. We now look into what is different when we want to work with digital images.

By digital images, we mean a multidimensional array of value that can be either 0 or 1. For example, a two-dimensional array is a black and white image. The tabular structure is particular and our previous geometric construction using the Čech complex is not the most suitable here. We replace simplicial complexes by cubical complexes. The idea is similar but we use squares instead of triangles and cubes instead of tetrahedron and so on.

Taking the example of an image, we build the complex with the following rule. Every pixel is given a value α and the cubic complex at time α contains all pixels whose value is less than α. Moreover, two adjacent pixels are linked if both of them have values below α. Four pixels in a square shape corresponds to a square in the complex if all of them have value less than α. The construction extends naturally to

higher dimensions. Note that the resulting object is indeed a complex in the sense that any element belonging to it has faces that also belong to it.

The next question is how to choose the value α for each pixel. We want to give a description of the topology of the areas, taking geometry into consideration. Note that if we just keep 0 and 1, we do only compare black and white areas. We thus put new values on each pixel depending on the distance to the other color. A black pixel adjacent to a white pixel is valued 0 and then the next black pixel is valued -1 and so on. Conversely, white pixels are valued increasingly depending on the distance to the nearest black pixel. Figure 5.10 shows the example of how to choose α and Fig. 5.11 shows the filtration by those α.

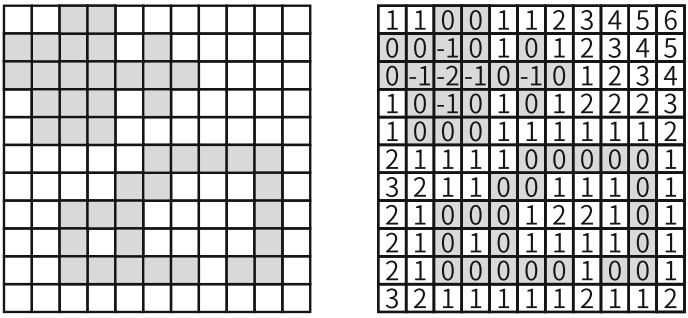

Fig. 5.10 Example of choosing α. The left figure shows an input digital image and the right figure shows the assignment of α on each pixel

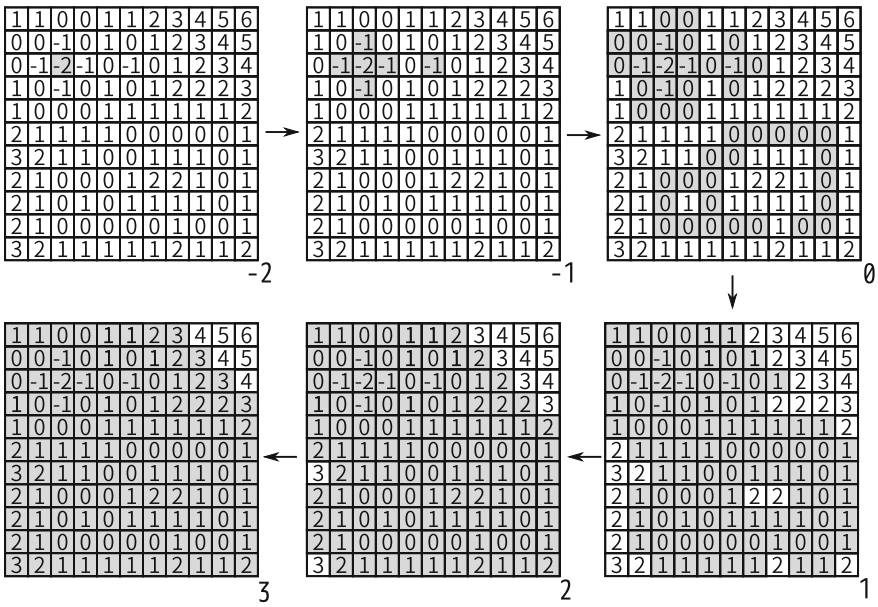

Fig. 5.11 Filtration for a digital image

Our construction provides a way to analyze digital images through the lense of persistent homology. It provides good insight into the structure of objects. Moreover, this simple approach to topological data analysis can be combined to machine learning to obtain interesting results [14].

5.3 Materials TDA

In this section, we briefly explain some applications of persistent homology to materials science. For details of each subject, we refer the readers to the original papers listed therein.

5.3.1 Silica Glass

Our first application is the structural analysis of silica glasses by using persistent homology [2]. There is a long history of trying to understand geometric structures of glass materials. From the experimental side, Xray/neutron scattering diffractions and the transmission electron microscopy (TEM) are often used to study the geometric structures of atomic configurations. On the other hand, from the computational side, molecular dynamics simulations, reverse Monte Carlo, and first-principles calculation based on density function theory are used to simulate atomic configurations. Although our understanding of glass structures is becoming richer, we have not yet reached a sufficient level.

One of the problems we are facing is the lack of appropriate descriptors to compactly and quantitatively express the geometry of glass atomic configurations. In the computational studies, we usually apply radial distribution functions, ring statistics, and Voronoi polyhedron analysis as conventional descriptors to the atomic configurations. However, those tools are restricted to the study of either the zero-dimensional topology (connected components) or single scale properties. As we have seen so far, persistence diagrams provide a tool for multi-scale analysis of higher topological features. This is presumably the most desired function for deeper analysis of amorphous structures.

Our idea is that, given an atomic configuration of silica (SiO_2), we regard it as a point cloud and characterize its geometric and topological structures by using persistent homology. Namely, we put balls with radius r_{Si} and r_{O} on silicon atoms and oxygen atoms, respectively, and gradually increase those radii to study birth and death events of holes in the atomic ball models in a multi-scale way. Technically, the initial radii r_{Si} and r_{O} are determined from the first peak positions of the partial radial distribution functions.

Figure 5.12 shows the one-dimensional persistence diagrams computed in the liquid, glass, and crystal states of silica, respectively. We denote them by $D_1(\mathcal{A}_{\mathrm{liq}})$, $D_1(\mathcal{A}_{\mathrm{amo}})$, and $D_1(\mathcal{A}_{\mathrm{cry}})$, respectively. Recall that the one-dimensional persistence diagram studies ring structures embedded in the atomic configurations. Here, the color bar is plotted on the logarithmic scale. The atomic configurations, consisting

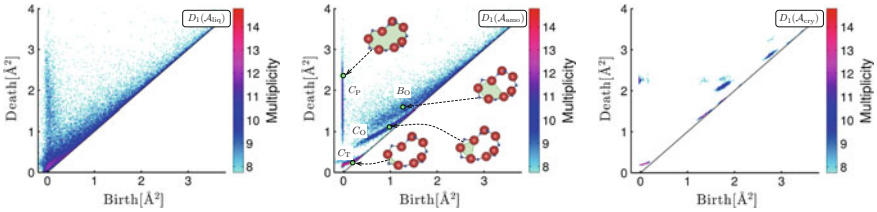

Fig. 5.12 Persistence diagrams of silica in liquid (left), glass (middle), and crystal (right) states (Reproduced from [2])

of 2,700 silicon atoms and 5,400 oxygen atoms, are prepared via the Beest-Kramer-Santen (BKS) model. We refer the readers to the original paper for details on preparing those atomic configurations by molecular dynamics simulations.

As we observe from Fig. 5.12, the persistence diagrams clearly distinguish these three states. Namely, the liquid, glass, and crystal states are characterized by planar (2-dim), curvilinear (1-dim), and island (0-dim) regions of the distributions, respectively. Here, the 0 and 2 dimensionality of the PDs result from the periodic and random atomic configurations of the crystal and liquid states, respectively. In particular, we emphasize that the presence of the curves in $D_1(\mathcal{A}_{\mathrm{amo}})$ clearly distinguishes the glass state from the others. This implies that specific geometric features of the rings generating these curves in $D_1(\mathcal{A}_{\mathrm{amo}})$ play a significant role to elucidate glass states.

Let us consider the meaning of curves. We first remark that, since our system consists of a large enough amount of atoms (8,100 atoms), statistical information is also embedded in each persistence diagram. From this respect, the presence of curve means that generators on each curve are restricted to that curve. Namely, each generator is not allowed to move in the normal direction of the curve, but possibly move to the tangential direction. We recall that generators in the persistence diagram are characterized by ring configurations of atoms. Hence, by pulling back these normal directions of curves, we obtain geometric constraints of local deformations to which atomic configurations are prohibited. In other words, a rigidity information with respect to small deformation of the atomic configuration is embedded in the persistence diagram. Actually, in the original paper, the relationship between persistence diagrams and rigidity based on the small deformation of atomic configurations induced by isotropic pressurization is studied in detail. From the same observation, we also remark that the persistence diagram of crystal state shows further geometric constraints.

The silica is a typical glass material classified as network forming glasses. In [2], we also studied another type of glass materials based on random packing structures. For instance, Fig. 5.13 shows the one-dimensional and two-dimensional persistence diagrams of the Lennard-Jones (LJ) system in crystal and glass states, denoted by $D_k(\mathcal{A}^{\mathrm{LJ}}_{\mathrm{cry}})$ and $D_k(\mathcal{A}^{\mathrm{LJ}}_{\mathrm{amo}})$ $(k = 1, 2)$. In this case, not only the one-dimensional persistence diagrams but also the two-dimensional persistence diagrams show characteristic features. Similar to the silica case, a deviation of the persistence diagrams of the

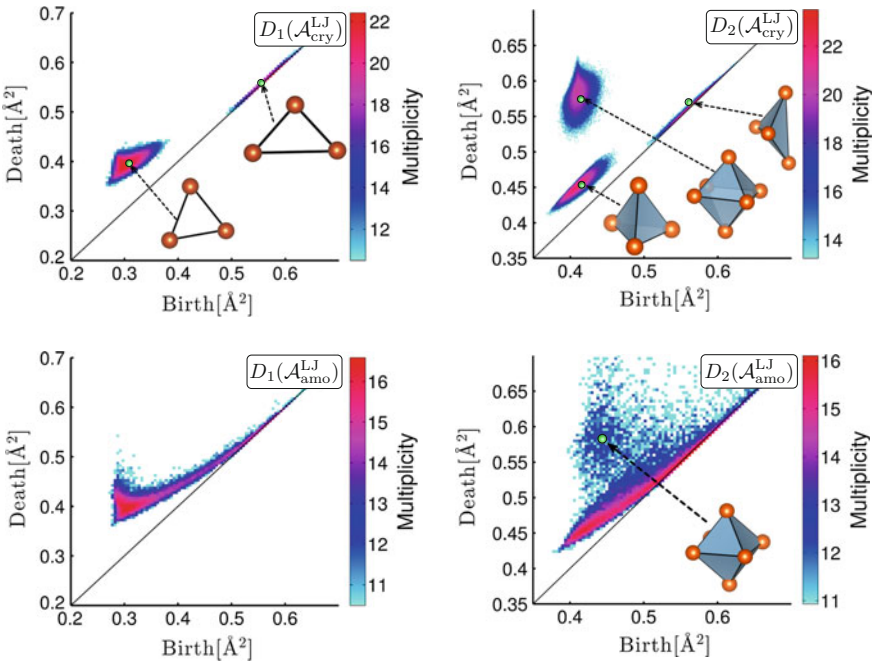

Fig. 5.13 Persistence diagrams of the Lennard-Jones system in crystal and glass states (Reproduced from [2])

glass state from those of the crystal state is observed. In particular, $D_2(\mathcal{A}^{\mathrm{LJ}}_{\mathrm{amo}})$ shows a peak corresponding to octahedral configurations.

As we see, the persistence diagrams clarify topological and geometric features embedded in atomic configurations, which cannot be characterized by other conventional methods. Note that those persistence diagrams are computed on atomic configurations given in a fixed system size. Therefore, we need to be careful about the dependence of the system sizes. The scaling properties of PDs with respect to the system size are computationally studied in [4]. Recently, the existence and uniqueness of limiting persistence diagram is mathematically solved in [15].

Starting from the research explained in this subsection, persistence diagrams are nowadays applied to a wide variety of structural analysis of materials.

5.3.2 Grain Packing

In the paper [5], crystallization mechanism of three-dimensional granular packings of frictional spheres is studied at the grain-scale using Xray tomography and persistent homology. Here, we briefly review some of the results.

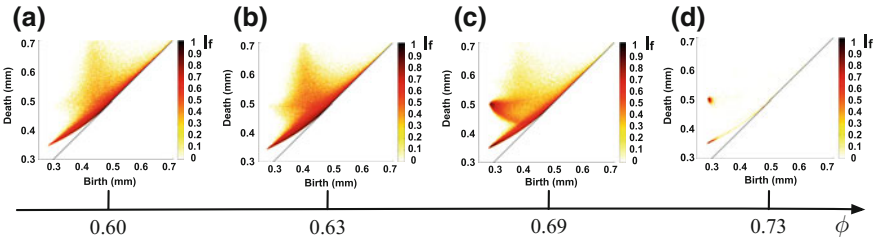

Fig. 5.14 Persistence diagrams of grain configurations for different packing ratios (Reproduced from [5])

In this study, three-dimensional images of granular packings with several packing ratio ϕ are obtained by using XCT, and these images provide precise positional coordinates of grains. Our interest is to characterize the skeleton deformation structures of grain configurations during the crystallization process. For experimental details, please see the original paper.

Figure 5.14 shows the two-dimensional persistence diagrams computed on the grain configurations for four packing ratios $\phi = 0.6, 0.63, 0.69$, and 0.73. Here, we note that the packing ratio $\phi = 0.64$ is known as the Bernal's density at which sharp structural transition to jamming is observed. As we observe from the figure, the persistence diagram (d) at the crystallized state consists of two strong peaks at $(0.288, 0.353)$ and $(0.288, 0.5)$, and they correspond to the regular tetrahedral and the regular octahedral configurations, respectively. We note that the persistence diagram (c) is similar to $D_2(\mathcal{A}^{\mathrm{LJ}}_{\mathrm{amo}})$ in Fig. 5.13 (the Lennard-Jones system), since both are classified as random packing systems.

The tetrahedral peaks are well preserved for all packing ratios, while the octahedral peaks only exist at (c) and (d). Actually, further studies show that the octahedral peaks are only observable for packing ratios $\phi > 0.64$.

Next, let us study the persistence diagram (c) at $\phi = 0.69$ in detail. Figure 5.15a is the same persistence diagram at $\phi = 0.69$, in which four curves (D1, D2, D3, and D4) corresponding to the boundaries are drawn. In the paper, we found the analytical expressions of the actual deformations of grain configurations corresponding to these curves. Figure 5.15b and c show those deformations. It follows from a discussion similiar to the silica glass case that distorted tetrahedra and octahedra are confined in the region bounded by D1-D4 and those deformations give geometric constraints during the crystallization process.

5.3.3 *Craze Formation of Polymer*

Craze formation has been intensively investigated by experiments such as electron microscopy, optical microscopy, atomic force microscopy, and so on. From these

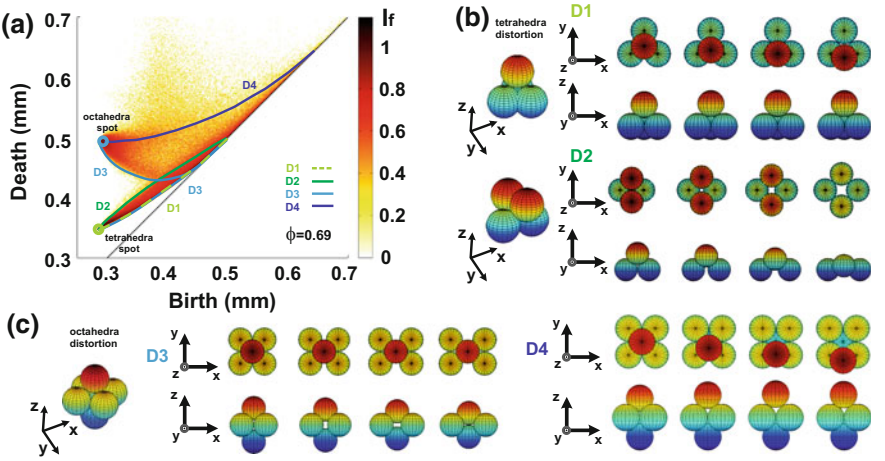

Fig. 5.15 Persistence diagram at $\phi = 0.69$ and the deformations of tetrahedra and octahedra generating the boundary curves D1–D4 (Reproduced from [5])

experimental observations, several kinetic models of craze formation have been proposed so far.

On the other hand, molecular dynamics (MD) simulations have also been applied to understand atomic-scale craze formation mechanisms, which are difficult to observe by experiments. However, the relation between the kinetic models and the MD simulations still remains unclear. This is partially due to the lack of definition of voids in the MD simulations. We note that, since MD simulations are based on the discretized systems, the definition of voids which are consistent with multi-scalability is not trivial. However, such a multi-scalable definition of voids is unavoidable to study the growing process of voids as continuum phenomena, where the kinetic models are discussed. As we now know, persistence diagrams provide an appropriate tool for this purpose.

In the paper [3], a persistent homology analysis is applied to investigate the behavior of nanovoids during the crazing process of glassy polymers. We carry out a coarse-grained molecular dynamics simulation of the uniaxial deformation of an amorphous polymer and analyze the results with persistent homology.

We first compute persistence diagrams of simulation results at each time snapshot. After yielding, several large voids appear, and we detect them from persistence diagrams as generators with large death values as these values measure the size of voids. Then, we reverse the time evolution of the simulation to investigate the initial configurations of those large voids. Then, we revealed that those large voids are created by the coalescences of small voids during craze formation. Figure 5.16 shows some

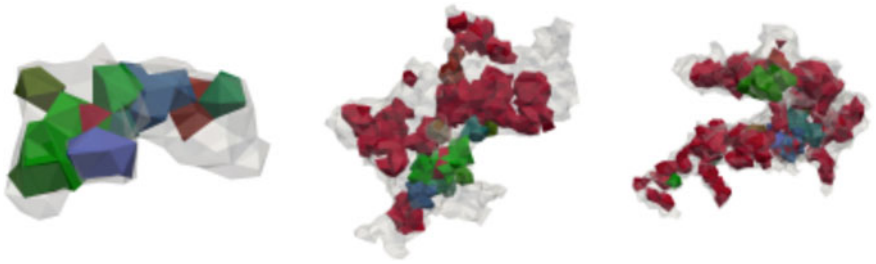

Fig. 5.16 Void percolation (Reproduced from [3])

of those coalescences during crazing, where gray voids correspond to large voids observed after yielding and other colored small voids coalesce to those gray voids. The results suggest that the yielding process should be regarded as the percolation of nanovoids created by deformation.

5.4 Discussions

In this paper, we summarized persistent homology and its applications to materials science. From these applications, we observed that persistence diagrams are significant descriptors for characterizing multi-scale disordered structures in materials. The next stage toward materials informatics is to combine TDA with machine learning.

Machine learning enables us to capture characteristic patterns from a large amount of data, and TDA enables us to summarize the shape of data quantitatively. Therefore, by combining these two data analysis methods, we can effectively capture the characteristic geometric patterns of the data. Since many machine learning methods accept vectors as input data, we need to convert a persistence diagram into a vector. Some vectorization methods are proposed, and here we introduce two methods with some applications.

One method is the persistence image (PI) [16], which uses a histogram on a finite mesh with smoothing and weighting applied. The histogram values are ordered consistently and we treat it as a finite dimensional vector. In [14], PI is used with logistic regression and linear regression to find a hidden relationship between a persistence diagram obtained from data and a parameter bound to data. In that paper, inverse analysis is effectively used to clarify the geometric origins of birth-death pairs important for the relationship. For materials informatics, we can apply the method to find the characteristic geometric patterns of materials data related to their physical properties such as Young's modulus and conductivity.

Another method is the persistence weighted Gaussian kernel (PWGK) [17, 18], a kind of kernel methods. PWGK maps a persistence diagram into a vector in an infinite dimensional Hilbert space. It is impossible to directly treat infinite dimensional vectors on a computer, but using the kernel trick technique, we can indirectly treat

the vectors to apply various kinds of machine learning methods. This method shows good performance in some examples in [17] and is applied to practical problems in [17, 18], e.g., estimating the liquid-glass transition point by using changing point analysis and classifying proteins by using support vector machine.

Acknowledgements The authors appreciate all the collaborators relating materials TDA projects. This work is partially supported by JST CREST Mathematics15656429, JST Materials research by Information Integration Initiative (MI2I) project of the Support Program for Starting Up Innovation Hub, Structural Materials for Innovation Strategic Innovation Promotion Program D72, and New Energy and Industrial Technology Development Organization (NEDO).

References

1. G. Carlsson, Topology and data. Bull. Am. Math. Soc. **46**(2), 255–308 (2009)
2. Y. Hiraoka, T. Nakamura, A. Hirata, E.G. Escolar, K. Matsue, Y. Nishiura, Hierarchical structures of amorphous solids characterized by persistent homology. Proc. Nat. Acad. Sci. **113**(26), 7035–7040 (2016)
3. T. Ichinomiya, I. Obayashi, Y. Hiraoka, Persistent homology analysis of craze formation. Phys. Rev. E **95**(1), 012504 (2017)
4. T. Nakamura, Y. Hiraoka, A. Hirata, E.G. Escolar, Y. Nishiura, Persistent homology and many-body atomic structure for medium-range order in the glass. Nanotechnol. **26**(30), 304001 (2015)
5. M. Saadatfar, H. Takeuchi, V. Robins, N. Francois, Y. Hiraoka, Pore configuration landscape of granular crystallization. Nat. Commun. **8** (2017)
6. A. Hatcher, *Algebraic Topology*. (Cambridge University Press, 2002)
7. F. Chazal S.Y. Oudot, Towards persistence-based reconstruction in euclidean spaces, in *Proceedings of the 24th Annual Symposium on Computational Geometry*, (ACM, 2008), pp. 232–241
8. F. Chazal, V. De Silva, M. Glisse, S. Oudot, The structure and stability of persistence modules (2012), arXiv:1207.3674
9. S.Y. Oudot, *Persistence Theory: From Quiver Representations To Data Analysis*, vol. 209, (American Mathematical Society, 2015)
10. B.T. Fasy, J. Kim, F. Lecci, C. Maria, V. Rouvreau, *TDA: Statistical Tools For Topological Data Analysis* (2014)
11. U. Bauer, M. Kerber, J. Reininghaus, Distributed computation of persistent homology, in *2014 Proceedings of the 16th Workshop on Algorithm Engineering and Experiments (ALENEX)*, (SIAM, 2014), pp. 31–38
12. C. Maria, J.-D. Boissonnat, M. Glisse, M. Yvinec, The gudhi library: simplicial complexes and persistent homology, in *International Congress on Mathematical Software*, (Springer, 2014), pp. 167–174
13. http://www.wpi-aimr.tohoku.ac.jp/hiraoka_labo/homcloud-english.html
14. I. Obayashi, Y. Hiraoka, Persistence diagrams with linear machine learning models (2017), arXiv:1706.10082
15. T.K. Duy, Y. Hiraoka, T. Shirai, Limit theorems for persistence diagrams (2016), arXiv:1612.08371
16. H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, L. Ziegelmeier, Persistence images: a stable vector representation of persistent homology. J. Mach. Learn. Res. **18**(8), 1–35 (2017)
17. G. Kusano, K. Fukumizu, Y. Hiraoka, Kernel method for persistence diagrams via kernel embedding and weight factor (2017), arXiv:1706.03472

18. G. Kusano, Y. Hiraoka, K. Fukumizu, Persistence weighted gaussian kernel for topological data analysis, in *International Conference on Machine Learning* (2016), pp. 2004–2013