# Unsupervised topological learning for identification of atomic structures

Sébastien Becker [1,2] Emilie Devijver,[2] Rémi Molinier [3] and Noël Jakse [1]

[1]*University of Grenoble Alpes, CNRS, Grenoble INP, SIMaP, F-38000 Grenoble, France*
[2]*University of Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France*
[3]*University of Grenoble Alpes, CNRS, IF, F-38000 Grenoble, France*

We propose an unsupervised learning methodology with descriptors based on topological data analysis (TDA) concepts to describe the local structural properties of materials at the atomic scale. Based only on atomic positions and without *a priori* knowledge, our method allows for an autonomous identification of clusters of atomic structures through a Gaussian mixture model. We apply successfully this approach to the analysis of elemental Zr in the crystalline and liquid states as well as homogeneous nucleation events under deep undercooling conditions. This opens the way to deeper and autonomous study of complex phenomena in materials at the atomic scale.

## I. INTRODUCTION

Knowledge of microscopic events driving properties of matter is of fundamental and technological interest. More precisely, understanding the structure at the atomic scale is crucial for the design of new materials, which can play a role in solving contemporary economical and social issues such as the reduction of energy consumption or new drug design [1]. However, detailed information on such a small spatial scale is still out of reach experimentally.

To access physical and chemical properties at the atomic scale, simulation tools such as molecular dynamics (MD) represent dedicated in-depth means [2]. Thanks to the increase of computational power, such simulations can nowadays easily reach millions of atoms evolving through several nanoseconds. Although production of large amounts of valuable data has been streamlined, the processing of the "big data" is far from being obvious and has led to the fourth paradigm [3] of science in many scientific fields since the early 2000s. Likewise, in condensed matter physics and materials science, new protocols and tools based on supervised learning and neural networks [4–7] as well as unsupervised learning [8–12], or even combinations of these approaches [13], are currently built at an accelerated rate.

In this paper, we propose an unsupervised method to analyze structural information, where descriptors from atomic positions are constructed using persistent homology (PH), a classical topological data analysis (TDA) tool [14]. Persistent homology has been successfully applied to MD simulations to analyze the medium-range structural environments [15] in amorphous solids [16–18], ice [19], and complex molecular liquids [20]. However, to the best of our knowledge, PH has never been used as a descriptor to encode local atomic structures, whereas it provides topological features at different resolutions and for different levels of homology. Similar local atomic structures, with respect to those topological

features, are clustered using a Gaussian mixture model (GMM). To illustrate the potential of this approach, we analyze here previous simulations of homogeneous crystal nucleation of elemental zirconium (Zr) [21].

Section II is dedicated to a presentation of our method illustrated on pure Zr simulations, while Sec. III presents and discusses the physical results highlighted on these simulations. Finally, Sec. IV draw the conclusions.

## II. UNSUPERVISED LEARNING BASED ON TOPOLOGICAL DESCRIPTORS

We present in this section the complete methodology behind our unsupervised protocol, giving all the steps necessary for its implementation.

### A. Data production

Our unsupervised learning approach is illustrated on pure Zr simulations. All the simulations mentioned in this paper have been performed with the LAMMPS code [22] in the isobaric-isothermal ensemble with the Nosé-Hoover thermostat and barostat [23] and interatomic interaction described by the MEAM potential [24] developed in Ref. [21]. For a detailed description of the procedure of the simulation, we refer the reader to Ref. [21], and only the essential characteristics are recalled here. The phase space trajectory is obtained by integrating Newton's equations numerically using the Verlet's algorithm in its velocity form, with a time step of 2 fs. A simulation box with $N = 1\,024\,000$ atoms is considered with periodic boundary conditions (PBC) in the three directions of space. This system is first equilibrated in the liquid state above the melting points at $T = 2500$ K and quenched with a cooling rate of $10^{12}$ K/s at ambient pressure down to the nose of the time-temperature-transformation (TTT) curve at
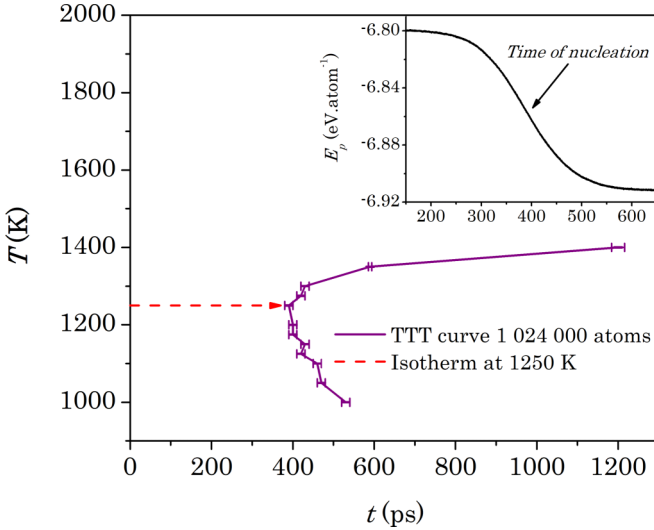
FIG. 1. Time-temperature-transformation curve with 1 024 000 atoms in a temperature range near the nose, which was obtained over five independent runs for each temperature. Inset: Evolution of the potential energy as a function of time along a nucleation process. See more details in Ref. [21].
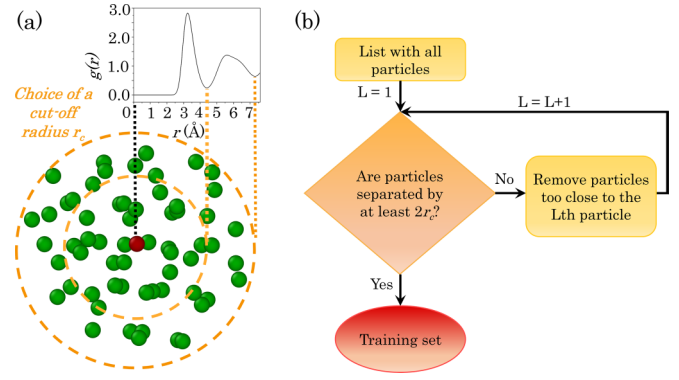


FIG. 2. (a) Schematic illustration of the first and second neighbor shells of a central particle (in red) with cut-off radii determined by the radial distribution function $g(r)$ of undercooled Zr at 1250 K. (b) Flowchart illustrating how to define a training set with independent structures. First, all particle positions in the simulation box are randomly ordered and set in a list. Then, the first particle of this list is set as a reference, and all particles which are closer than twice the selected cut-off radius $r_c$ from this reference particle are removed. Changing the reference particle, this process is iteratively repeated until all central particles meet the distance criterion based on $r_c$.

$T = 1250$ K as shown in Fig. 1. The homogeneous nucleation process is then observed along this isotherm.

When the onset of nucleation is observed, typically at the nucleation time, a $N$-atom configuration is extracted and its inherent structure is used to build a training set. The inherent structures [25] are obtained by minimizing the energy by means of a conjugate gradient algorithm, implemented in LAMMPS, to bring the system to its local minimum in the potential energy surface to suppress the thermal noise.

### B. Data preparation

Local atomic structures in the configuration are defined through a coordination sphere, consisting of a central particle and its surrounding environment up to a chosen cut-off radius. The cut-off radius is defined thanks to the radial distribution function $g(r)$, which gives the probability to find a particle at a distance $r$ from a particle at the origin and is defined by

$$g(r) = \frac{N}{V} \frac{n(r)}{4\pi r^2 \Delta r},\tag{1}$$

where $n(r)$ stands for the mean number of particles in a spherical shell of radius $r$ and thickness $\Delta r$ centered on the origin. A typical representation of $g(r)$ for the liquid state of Zr is depicted in Fig. 2(a). Each minimum of the radial distribution function can reasonably be used as a cut-off radius to obtain local environments that represent consecutive neighbor shells of each central particle. The first minimum (beyond the first maximum) defines the end of the first neighbor shell (4.43 Å from the central particle in Fig. 2(a)), the second minimum defines the end of the second neighbor shell (7.28 Å from the central particle), and so on.

Among the million particles in the configuration, local atomic structures are subsampled to construct a training set for the learning. To be representative of the configuration while satisfying statistical independence, the subsampling should pave the whole simulation box with respect to the PBC. For a given cut-off radius, the subsample is one among the (almost) maximal sets such that the central particles of the local atomic structures are separated by at least twice this cut-off radius. More precisely, such sets are computed through an iterative process summarized on the flowchart Fig. 2(b). The choice of the cut-off radius (between first shell, second shell, etc.) as well as the number of structures in the training set will be justified in Sec. II C.

Once the central particles are identified, the Python package pyscal [26] is used on the full configuration to efficiently extract the particles coordinates of the local atomic structures.

### C. Persistent homology as a local atomic structure descriptor

To encode topological information of the local atomic structures, we use PH, a popular TDA tool [27,28] that detects relevant topological features from a point cloud.

Given a set of points $X$, one can construct a collection of simplicial complexes [29], called *Vietoris-Rips complexes* as follows. Given $t \geqslant 0$, we consider the simplicial complex $\mathcal{X}_t$ which is the union of $k$-simplices for $k \in \mathbb{N}$, where 0-simplices are all the points of $X$, 1-simplices are segments with extremities in $X$ of length smaller than $t$, 2-simplices are full triangles with vertices in $X$ and distant of at most $t$ from one another, etc. In the end, we get an increasing sequence (called filtration) $\mathcal{X} = (\mathcal{X}_t)_{t \geqslant 0}$ of simplicial complexes. A schematic representation of this filtration process is depicted in Fig. 3. Actually, since the number of points is finite, changes in these complexes only appear at a finite sequence of steps $t_1 < \cdots < t_n$ which gives a finite filtration of simplicial complexes $\mathcal{X}_0 = \mathcal{X}_{t_0} \subset \mathcal{X}_{t_1} \subset \cdots \subset \mathcal{X}_{t_n}$. For each space $\mathcal{X}_t$ of the filtration, we compute the simplicial homology, which is the sequence of their homology groups $[H_k(\mathcal{X}_t)]_{k \geqslant 0}$, further denoted $(H_k)_{k \geqslant 0}$ when the space is understood or not specified for general consideration. The dimension of these homology
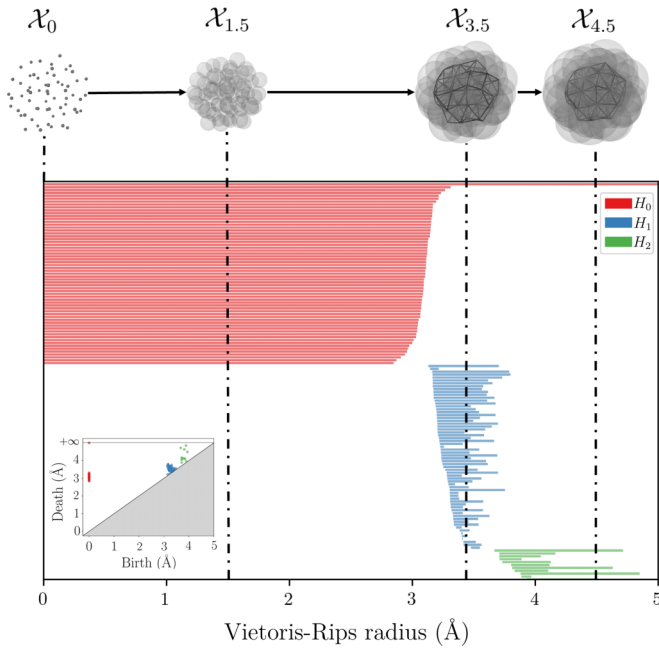
FIG. 3. Schematic snapshots of the *Vietoris-Rips complexes* filtration $(\mathcal{X}_t)_{t \geqslant 0}$ applied to a local structure with two neighbor shells. The *barcode* below encodes the persistence of each topological feature by a line starting at its birth (the radius where its appear) and finishing at its death (the radius where it disappears). The colors correspond to the homological dimension (red, blue, and green for $H_0$, $H_1$, and $H_2$ respectively) of each topological feature. The insert represents the corresponding PD where each point corresponds to a line and thus its associated topological feature with coordinates (birth, death).

groups gives insights on the dataset. For instance, the dimension of $H_0$ gives the number of connected components, the dimension of $H_1$ the number of holes, and the dimension of $H_2$ the number of cavities inside the simplicial complex. Given a filtration of simplicial complexes $(\mathcal{X}_t)_{t \geqslant 0}$, one can keep track of topological features (encoded by elements in the homology groups) and how "persistent" they are, i.e., at which $t$ they appear (the birth) and at which $t$ they disappear (the death).

A persistence diagram (PD) summarizes this information by plotting the pair (birth, death) of each topological feature of the filtration on a graph, as illustrated in Fig. 3 or Fig. 4.

Using Python packages gudhi [30] and ripser.py [31], the individual PDs of the local atomic structures from the previously extracted training set are computed for each homological dimensions $H_0$, $H_1$, and $H_2$. To remove topological noise when considering $H_1$ and $H_2$, we use a subsampling approach as introduced in Ref. [32].

While PDs are usually compared using the Bottleneck distance, their space equipped with this metric cannot be embedded into a Euclidean space nor even a normed vector space [33]. To tackle this problem, several mapping into vector spaces have been proposed in the literature. In this paper, we use a method developed in Ref. [34] and classically used to study three-dimensional shapes. Each coordinate of the topological vector is associated to a pair of points $(x, y)$ in a PD $D$ for a fixed level of homology, except for the infinite point, and is calculated by

$$m_D(x, y) = \min\{\|x - y\|_\infty, d_\Delta(x), d_\Delta(y)\}, \qquad (2)$$

where $d_\Delta(\cdot)$ denotes the $\ell^\infty$ distance to the diagonal, and those coordinates are sorted by decreasing order. To reduce the dimension, a number of coordinates equal to the number of points in the PD for the current homological dimension is kept, as proposed in the gudhi's package. Remark that the dimension of each topological vector depends on the local atomic structure, so we fill in each vector with a noninformative value to reach the maximal dimension of the descriptor space. Here we decided to fill the vectors with the value $-1$ instead of 0 as proposed in Ref. [34], as distances between points in PDs are always nonnegative, and a zero, i.e., two points of the PD with the same birth and death, corresponds to relevant information in our context.

To highlight the importance of these zeros in our topological vectors, let us consider Zr atoms on a body-centered cubic (bcc) lattice up to the second neighbor shell (7.16 Å from the central particle) and Zr atoms on a face-centered cubic (fcc) lattice in the same shell as shown in Fig. 4. Even though these structures are fundamentally different, it can be noted that the only thing that differs between the topological vectors
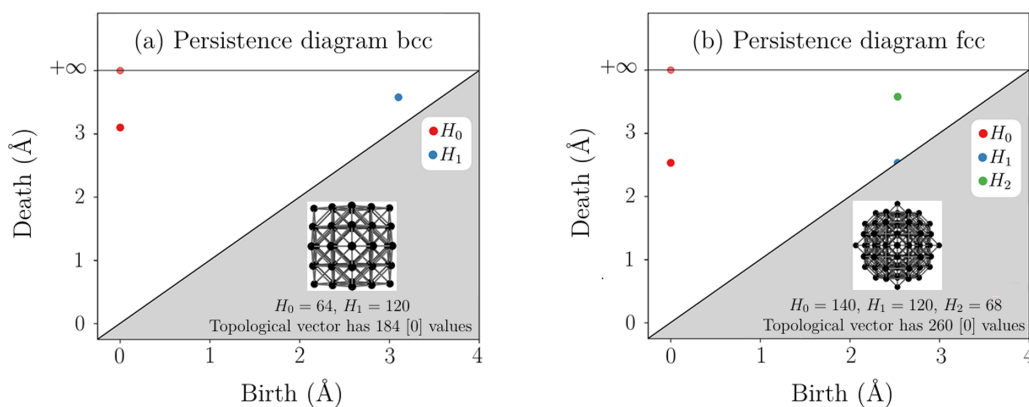


FIG. 4. (a) Zr atoms on a bcc lattice and the corresponding PD. (b) Zr atoms on a fcc lattice and the corresponding PD. The homological dimensions $H_0$, $H_1$, and $H_2$ are depicted in red, blue, and green respectively. As these structures are on periodic lattices, all topological features of the filtration from a specific homological dimension have the same pair (birth, death).
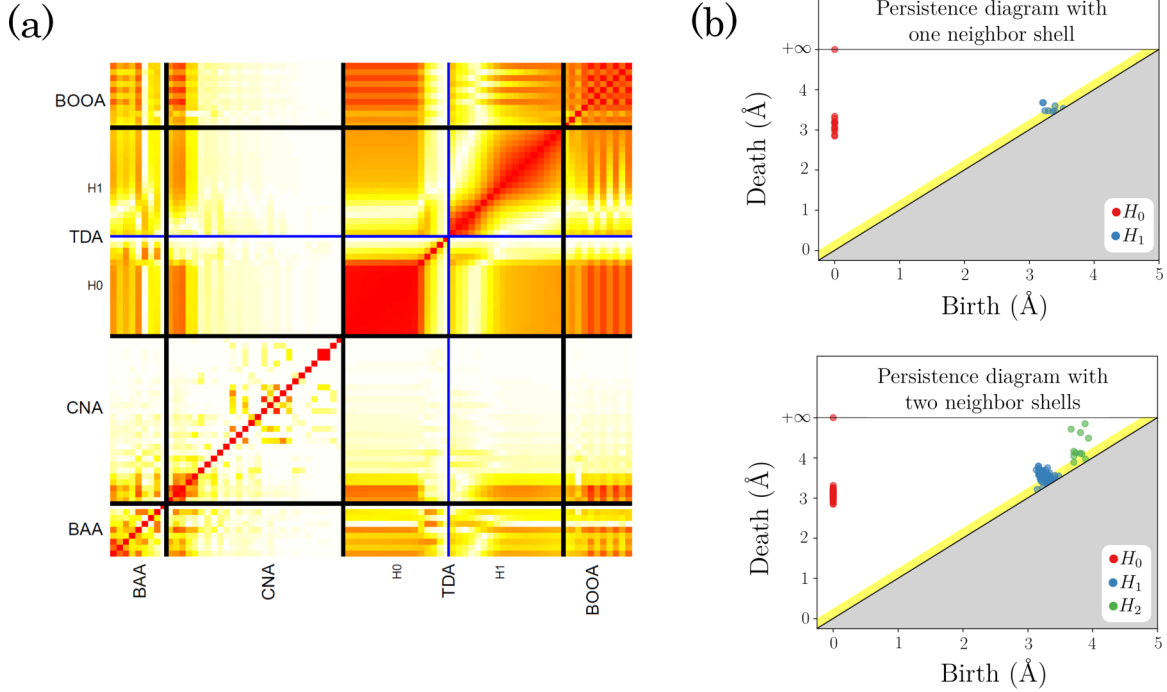
FIG. 5. (a) Empirical correlation matrix (absolute value) between several descriptor families: BAA, CNA, TDA, and BOOA. Low values are white, high values are red. We distinguish between the descriptors by black lines, and between levels of homology with a blue line ($H_1$ and $H_1$). (b) PDs of a local atomic structure with one and two neighbor shells. The yellow line corresponds to a threshold computed with the subsampling approach to remove the noise.

obtained from each structure is the dimension of the vector (64+120 [0] values for the bcc lattice, 140+52+68 [0] values for the fcc lattice). Thus, if we fill the smaller with zeros, then we will obtain two identical vectors, i.e., the same point in the descriptor space.

Finally, to illustrate the relevance of our topological signature to study the local atomic structures, we confront it on our Zr simulation to widely used classical physical descriptors such as the bond angle analysis (BAA) [35], the common neighbor analysis (CNA) [36], and the bond-orientational order analysis (BOOA) [37], in its averaged definition [38]. The latter computes the order parameters $\bar{q}_l$ given by

$$\bar{q}_l(i) = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} \left| \frac{1}{\tilde{N}_b(i)} \sum_{k=0}^{\tilde{N}_b(i)} q_{lm}(k) \right|^2} \quad (3)$$

with the complex vector

$$q_{lm}(i) = \frac{1}{N_b(i)} \sum_{j=1}^{N_b(i)} Y_{lm}(\mathbf{r}_{ij}), \quad (4)$$

in which $\tilde{N}_b(i)$ is accounting for all the neighbors of a central particle $i$ plus $i$ itself, $N_b(i)$ for the sole number of nearest neighbors and $Y_{lm}(\mathbf{r}_{ij})$ are the spherical harmonics functions of the vector $\mathbf{r}_{ij}$ from particle $i$ to $j$, with $l$ an integer ranging from 2 to 12 and $m \in [-l, +l]$. A direct comparison of the CNA and BOOA, both being mostly used, can be found in Ref. [39]. These classical descriptors characterize the radial and/or angular distribution of bonded pairs of atoms in the coordination sphere. The averaged version of the BOOA is

preferred over the regular definition as a considerable improvement of the accuracy by taking into account information from the neighboring particles beyond the first neighbor shell. Figure 5(a) gives an empirical correlation matrix between the coordinates of our TDA descriptor, BAA, BOOA, and CNA. This matrix is constructed on a sample of approximately 25 000 local structures extracted as described in Sec. II B with a cut-off radius corresponding to the first minimum of the radial distribution function. The latter choice was made as a descriptor, as the CNA cannot be applied beyond the first shell of neighbors. The empirical correlation matrix shows that the topological signature is highly correlated with all the other descriptors. Some specificities of the data are highlighted: As all local neighborhood contain at least 10 particles, the first 10 components of $H_0$ are highly correlated; they are correlated with high values of $H_1$, too, because most of the local structures have few $H_1$ components (50% of the population has fewer than five components); $H_2$ components are not depicted here as there is none in local atomic structures with only the first neighbor shell; the correlation matrix associated to the CNA is sparse, because only a few bonds, which are the first ones in the matrix (using the indexing of Faken and Jónsson [40]: [666] and [444] for the bcc ordering and [555], [544], and [433] for the icosahedral ordering), are present in most of the structures. It also explains the low correlation between the last bonds in the matrix of CNA with the other descriptors.

This example on the first neighbor shell illustrates how relevant signatures from $H_0$ and $H_1$ are. To increase the number of $H_0$ and $H_1$ components and also to capture information on $H_2$ (which appears when considering more than just one

neighbor shell), we built a training set of 5314 local atomic structures with two neighbor shells to construct our model, approaching the intermediate order. This leads here to 68 $H_0$ components, 100 $H_1$ components, and 22 $H_2$ components, instead of only 17 $H_0$ components and 18 $H_1$ components with one neighbor shell. Figure 5(b) shows a representative example of the differences in the PDs of a structure with one neighbor shell and the same structure with two neighbor shells. The second neighbor shell plays also a crucial role in the structural description [41,42].

Albeit considering higher-order shells is possible with our persistent homological description of local structures, the advantage we gain in topological information is balanced by a too-high spatial resolution of the local structures, leading to a loss of information in the GMM clustering. Thus, we have restricted ourselves to the second neighbor shell, through a trade-off between the structural information and a coarsening of the spatial resolution, already observed for the averaged BOOA [38]. Note that the reason for using the radial distribution function $g(r)$ to determine the second shell of neighbors is to fulfill this compromise; adaptive methods which are more general, such as SANN [43] or RAD [44] algorithms being limited to the first shell of neighbors. It is worth mentioning that the clustering remains essentially unchanged by varying the cut-off radius by $\pm5\%$ about the second minimum, which is compatible with the volume change during the crystal nucleation.

### D. Gaussian mixture model clustering

A GMM groups data points into clusters within an unsupervised way through a mixture of $M$ Gaussian distributions $[\phi(\,\cdot\,;\boldsymbol{\mu}_m, \Sigma_m)]_{1\leqslant m\leqslant M}$ of weights $(\alpha_m)_{1\leqslant m\leqslant M}$ as

$$\sum_{m=1}^{M}\alpha_m\phi(\,\cdot\,;\boldsymbol{\mu}_m, \Sigma_m), \tag{5}$$

where $\boldsymbol{\mu}_m$ is the mean and $\Sigma_m$ the covariance matrix of the $m$th Gaussian distribution. After standardizing the descriptor space, an expectation-maximization algorithm [45] is used as an iterative method to estimate the unknown parameters $(\alpha_m, \mu_m, \Sigma_m)_{1\leqslant m\leqslant M}$. We use the implementation in the Python package `scikit-learn` [46] with full covariance matrices and 3000 $k$-means [47] initializations.

To select the number of Gaussian components, i.e., the number of clusters, the integrated completed likelihood (ICL) [48] criterion has been used:

$$\mathrm{ICL} = -2\ln(\hat{L}) + D\ln(n) - 2\sum_{k=1}^{n}\sum_{m=1}^{M}\hat{\tau}_{m,k}ln(\hat{\tau}_{m,k})$$

$$= \mathrm{BIC} - 2\sum_{k=1}^{n}\sum_{m=1}^{M}\hat{\tau}_{m,k}ln(\hat{\tau}_{m,k}), \tag{6}$$

where $\hat{L}$ denotes the likelihood evaluated on the estimator, $D$ the number of parameters to be estimated [$D = M(d + d^2 + 1)$, with $d$ the dimension of the topological vectors], and $\hat{\tau}_{m,k}$ denotes the probability to belong to the $m$th component conditionally to the observation $x_k$. The ICL generalizes the widely used Bayesian information criterion (BIC) [49] to clus-
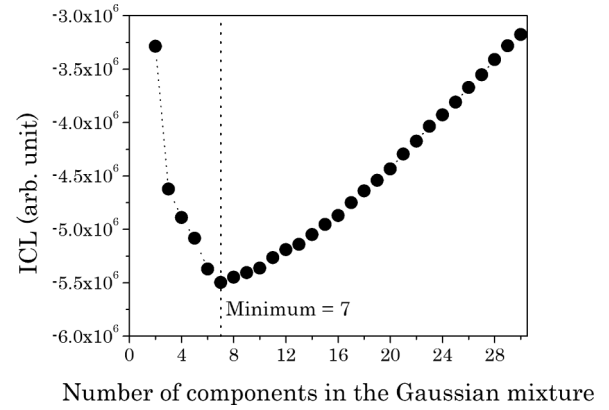


FIG. 6. ICL criterion on the training set. Its minimum corresponds to the maximum likelihood, which leads to the optimal number of clusters to build the model.

tering methods by adding an entropic penalty computed from the posterior probabilities of the data to be assigned to each Gaussian component. The number of clusters in the model is set to the number achieving the minimum of ICL.

## III. APPLICATION OF THE UNSUPERVISED APPROACH ON ELEMENTAL ZIRCONIUM SIMULATIONS

We focus in this section on the application of our method, called TDA-GMM, to elemental Zr simulations. It demonstrates that this protocol provides relevant structural information in a context as challenging as homogeneous nucleation.

### A. Learning the model

As mentioned in Sec. II B, a configuration is chosen in course of the nucleation process, where the supercooled liquid coexists with crystalline nuclei, to capture all structural atomic events of interest. Figure 6 depicts the ICL computed on the extracted training set as described in Sec. II C, leading to select seven clusters [39]. Each cluster $C_i$ is represented by the local structure that is the closest to the mean of the corresponding Gaussian component, as illustrated in Fig. 7(a). Meanwhile classical embeddings [39,50,51] do not allow to visually interpret the width and shape of the different Gaussians, an analysis of the eigenvalues of the covariance matrices highlights the different elliptical shapes. This proves the necessity of GMM over simpler unsupervised algorithms such as $k$-means which would only be suitable to fit hyperspheres.

Once the model is learned, it can be applied to new uncategorized local atomic structures through their topological signatures. The robustness of our model is checked by considering a test set consisting of all the structures in the configuration but the one in the training set. Most of the structures in this test set (99.998%) have probabilities beyond 0.999 to be in one cluster of the model, and the remaining ones still have probabilities higher than 0.5. Further analysis of the fit of the model to the test set is performed using the Mahalanobis distance (MaD) [52], which computes the distance between an empirical Gaussian distribution and a dataset. Using hard-assignment to affect each structure to a
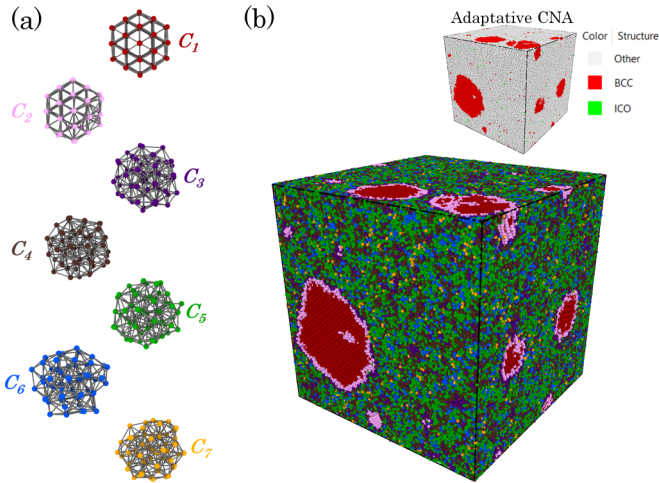
FIG. 7. (a) Representation in the real space of the local structures assigned to each cluster $C_i$. (b) Snapshot of the one million atoms in the simulation box after the clustering in the descriptor space. In the inset, the classification obtained from an adaptive CNA method is depicted.
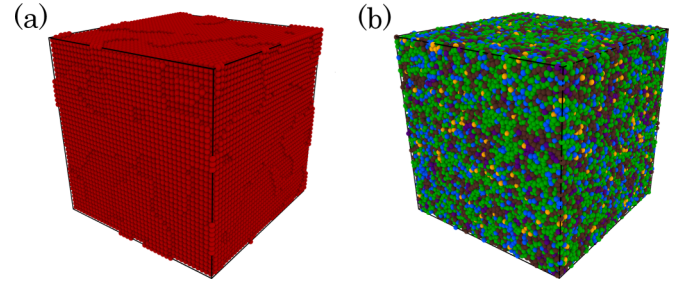


FIG. 8. Snapshots at $T = 1250$ K of Zr bulk crystal (a) and liquid (b) brought in their inherent local structures. Atoms are colored according to the cluster they belong to (see Fig. 7).

cluster, we compute the MaD for each cluster between the test set and the multivariate Gaussian distribution. Outliers are then considered based on the interquartile range rule. Table I presents the percentage of structures in each cluster viewed as outliers according to this criterion: more than 96% of the structures in the test set have a MaD in agreement with their assigned Gaussian distribution of the model.

The final clustering is shown in Fig. 7(b) on the entire configuration depicted in the real space of the simulation box with the help of OVITO [53]. In the inset, the classification found by a classical adaptive CNA algorithm as implemented in OVITO is represented for comparison. In contrast to TDA-GMM, only two types of structures that correspond to strictly crystalline or icosahedral structures were successfully identified.

### B. Description of the bulk crystal and liquid

We have first applied our model to a configuration with $N = 128\,000$ atoms in the bulk crystal and liquid at $T = 1250$ K. Figure 8 and Table II shows the results obtained in both cases through the central atoms of the local structures associated with each cluster.

The structures of the bulk crystal are classified as local structures in cluster $C_1$. Referring to Fig. 7(a), the atoms of the mean local structure associated with $C_1$ are indeed on a periodic lattice. In the bulk liquid, more than 99.96% of the

identified local structures belong to clusters $C_3, C_4, C_5, C_6$, and $C_7$. From this simple application, and without any information about the physical nature of the atomic structures, the trained TDA-GMM model distinguishes solidlike from liquidlike clusters.

### C. Description of the homogeneous crystal nucleation

During the process of homogeneous nucleation along an isotherm, the internal energy of the system undergoes a sharp drop. It follows the growth of the nuclei that carry the crystalline periodicity, until the state of least energy is reached. To autonomously detect these nuclei among the clusters obtained by the TDA-GMM method, five configurations along an isotherm at the nose of the TTT at $T = 1250$ K with $N = 1\,024\,000$ atoms are selected. They range from the onset of nucleation at 160 ps from the quench, up to a configuration in the polycrystalline bulk at 1560 ps. The model built in Sec. III A corresponds to a configuration in the course of nucleation at 360 ps. Table III shows the evolution of the proportion of each cluster in these configurations.

Local structures from clusters $C_1$ and $C_2$ follow a fast growth until they become the majority in the final bulk. Conversely, the other clusters undergo a more or less sharp decrease in their proportions over time. As mentioned in Sec. III B, local structures from $C_1$ correspond to a crystalline ordering whereas atoms of the mean local structure associated with $C_2$ are on a partial periodic lattice. Figure 9 represents the snapshots of these clusters in these different configurations. The other clusters, classified as liquid, show that apart from the nuclei, there exist complex structural

TABLE I. Proportions of structures and percentage of outliers in each cluster of the test set based on the interquartile range rule on the MaD between the test set and their assigned multivariate Gaussian distribution of the model in the descriptor space.

| Clusters | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | Total |
|---|---|---|---|---|---|---|---|---|
| Proportion (%) | 11.70 | 7.22 | 9.28 | 24.45 | 33.46 | 11.03 | 2.86 | 100 |
| Outliers (%) | 2.47 | 2.12 | 6.29 | 3.25 | 3.67 | 3.88 | 3.06 | 3.56 |

TABLE II. Proportion of each cluster $C_i$ in the bulk crystal and liquid.

| Bulk | Crystal | Liquid |
|---|---|---|
| $C_1$ (%) | 100 | 0.00 |
| $C_2$ (%) | 0 | 0.03 |
| $C_3$ (%) | 0 | 6.49 |
| $C_4$ (%) | 0 | 29.83 |
| $C_5$ (%) | 0 | 45.22 |
| $C_6$ (%) | 0 | 14.70 |
| $C_7$ (%) | 0 | 3.72 |

TABLE III. Proportion of each cluster $C_i$ at different times of the nucleation process. Superscripts (M) and (S) correspond, respectively, to the configuration used to train the model and the solidified configuration.

| Time (ps) | 160 | 200 | 240 | 280 | 310$^{(M)}$ | 1560$^{(S)}$ |
|---|---|---|---|---|---|---|
| $C_1$ (%) | 0.25 | 0.95 | 2.85 | 6.83 | 12.66 | 67.21 |
| $C_2$ (%) | 0.67 | 1.37 | 2.76 | 4.96 | 6.85 | 22.65 |
| $C_3$ (%) | 7.37 | 7.68 | 8.18 | 8.75 | 8.51 | 6.99 |
| $C_4$ (%) | 29.99 | 29.47 | 28.41 | 26.52 | 24.22 | 1.58 |
| $C_5$ (%) | 43.69 | 42.87 | 40.99 | 37.42 | 33.78 | 0.90 |
| $C_6$ (%) | 14.44 | 14.14 | 13.48 | 12.36 | 11.16 | 0.35 |
| $C_7$ (%) | 3.59 | 3.53 | 3.32 | 3.15 | 2.82 | 0.31 |

heterogeneities in the supercooled liquid, which we describe in Sec. III E, and which are successfully highlighted by the TDA-GMM method. Such an assessment has recently been brought out as well in supercooled liquid mixtures by another unsupervised method based on an information-theoretic approach [12].

The independent nuclei formed by local structures from $C_1$ and/or $C_2$ are extracted using a distance-based neighboring criterion as implemented in OVITO. This criterion is set as the first minimum of $g(r)$ in such a manner that two particles are connected if they are at a distance less than or equal to this threshold. The size distributions of these independent nuclei are then obtained by accounting for all atoms of the overlapping local atomic structures of two neighbor shells defined by each central particle belonging to $C_1$ and/or $C_2$. The critical size is estimated by the smallest cluster that persists over the several configurations of nucleation, with at least the same number of atoms. In such a high undercooling regime, this leads to a critical nucleus size of 70–80 atoms.

### D. Translational and orientational orderings

Two physical orderings which have been shown to be important parameters driving nucleation are the translational and orientational orderings [54–56].

With the nuclei previously identified, their centers of mass are easily inferred. Based on these positions, the spatial evolution of the translational ordering from the nuclei to the liquid can be quantified through density. The density is indeed the dedicated indicator to measure the fluctuations in the relative positions of neighboring atoms: for each cluster $C_i$ and its relative number of atoms $K_i(r_n)$ in a spherical shell of volume $V_s$ extending from the center of a nucleus to a distance $r_n$

to $(r_n + \epsilon)$ with $\epsilon = 1$ Å, the radial partial atomic density profiles $\rho_i(r_n)$ are computed as:

$$\rho_i(r_n) = \frac{K_i(r_n)}{V_s}, \quad \text{with } V_s = \frac{4}{3}\pi\left[(r_n + \epsilon)^3 - r_n^3\right]. \quad (7)$$

On the other hand, the analysis of the orientational ordering provides information on the fluctuations in relative geometric bonding between neighboring atoms. Such a characterization can be performed with help of the classical methods mentioned in Sec. II C. For instance, the BOOA has shown to be successful to analyze the crystallization of colloidal suspensions [57,58]. Here, following the indexing of Faken and Jónsson, the CNA was performed in each cluster as a representative measure [59]. We checked, as indicated by the high correlations in Fig. 5(a), that these two methods give similar information about the orientational ordering [39].

This translational and orientational orderings analysis shows us a general behavior, even for the precritical nuclei, which is depicted in Fig. 10. Figure 10(a) reveals that all the nuclei emerge with a translational ordering which corresponds to the density of the bulk crystal. Despite the small density window between the bulk crystal and liquid at $T = 1250$ K, it is clearly spotted that the sum of the respective densities $\rho_i$ of the clusters $C_i$ associated with the nuclei or liquid leads to the density values of the bulk crystal and liquid, respectively. Fig. 10(b) confirms the crystalline structures in $C_1$ and $C_2$ which exhibit bcc bonds ([666]: 8, [444]: 6) and thus a concurrent emergence of the orientational ordering in the nucleation process. The slight percentage of icosahedral bonds [555] in $C_2$ tends to explain the partial periodicity observed in Fig. 7(a). On the other hand, the local structures from clusters $C_4$ to $C_7$ share strong fivefold symmetries bonds characteristics of the undercooled liquid, although they also possess nonnegligible relative bcc bonds. About 4% of [666] in these clusters are in excess to form bcc structures and have the potential to combine with [555] bonds to form Z14, Z15, and Z16 Frank-Kasper polyhedra [60]. This can lead to a geometrical frustration able to slow down the onset of the nucleation in the liquid part [61,62]. Lastly, $C_3$ is a more peculiar case, with a strong distribution at the border of nuclei while also being present in the deeper liquid. Its presence in the frontier region is explained by the spatial resolution already mentioned in Sec. II C, while its presence in the liquid can be seen as a precursor for the emergence of the embryos. This is in agreement with the latest view of a heterogeneous scenario in which precursors of the crystalline ordering arise from structural heterogeneities in the liquid [63–65] and which we develop in Sec. III E.
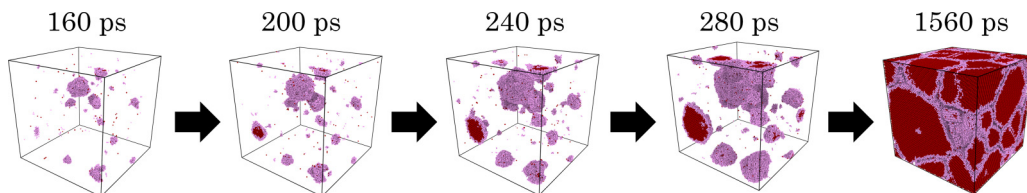


160 ps    200 ps    240 ps    280 ps    1560 ps

FIG. 9. Snapshots of configurations along nucleation at the nose of the TTT of Zr at $T = 1250$ K. The local structures in the clusters $C_1$ and $C_2$ identified as carrying the crystalline periodicity are depicted through their central particle in red and pink, respectively.
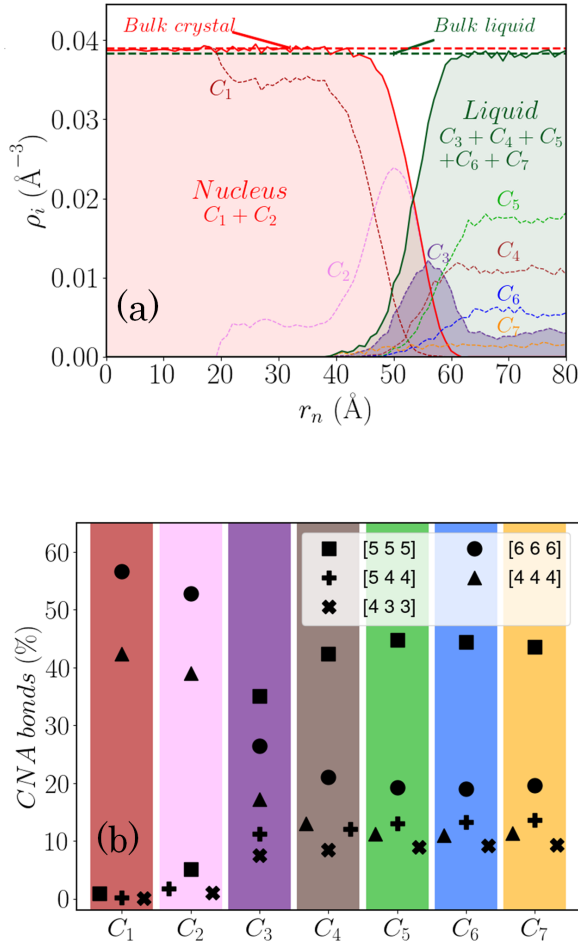
TABLE IV. The $p$ values computed on the projection of atomic positions at an early nucleation stage (160 ps) onto each direction of the simulation box from a KS test against the uniform distribution.

| $p$ value | $x$ | $y$ | $z$ |
|---|---|---|---|
| $C_1$ | 0.000 | 0.000 | 0.000 |
| $C_2$ | 0.000 | 0.000 | 0.000 |
| $C_3$ | 0.000 | 0.000 | 0.000 |
| $C_4$ | 0.000 | 0.000 | 0.007 |
| $C_5$ | 0.001 | 0.000 | 0.000 |
| $C_6$ | 0.007 | 0.331 | 0.001 |
| $C_7$ | 0.192 | 0.156 | 0.005 |

empirical cumulative distribution function is not defined, a Kolmogorov-Smirnov (KS) test [66] was performed on the projection of atomic positions in the three directions of space. For a level 0.01, if the uniform distribution is rejected on at least one projection, it is rejected for the entire multivariate dataset. Table IV depicts the results through the $p$ values of the test obtained from the configuration at 160 ps. For all the clusters the test is always rejected, meaning that their distributions are not uniform, in other words, indicating structural heterogeneities in the simulation box.

## IV. CONCLUSIONS

We used PH, a main computational tool in TDA, to build descriptors of local atomic structures. A GMM allowed us to autonomously identify clusters of similar local structures in the space of these topological descriptors. We applied this unsupervised approach to the analysis of several MD configurations of Zr in the challenging context of homogeneous nucleation. We highlighted some interesting properties of the phenomenon such as a concurrent emergence of the translational and orientational orderings and structural heterogeneities in the undercooled liquid. A closer look at the nucleation process of other pure metals has been investigated through this TDA-GMM method in our latest work [67]. An extension of the analysis on multicomponent systems where chemical orderings arise would be especially relevant for further inspection and control of such mechanisms to enhance materials design. More generally, this unsupervised learning methodology opens the door to further study of structure-dependent phenomena at the atomic scale.

FIG. 10. (a) Representation on a nucleus of the typical translational ordering of the nuclei and their surrounding liquid through the radial density profile of the $C_i$ clusters. The corresponding densities of the bulk crystal and undercooled liquid have been determined at $T = 1250$ K and constant ambient pressure. This behavior of the translational ordering is representative of all nuclei in the configurations, even the precritical ones. (b) Typical orientational ordering of the $C_i$ clusters in the configurations through CNA. [555], [544], and [433] stand for fivefold symmetries bonds and, [666] and [444] for bonds relative to a bcc crystal ordering.

There is no consensus in the literature for a general pathway of the nucleation process of materials, which seems to be system dependent: Hard spheres systems undergo concurrent translational and orientational ordering [56], whereas a decoupling of these two symmetries has been observed in colloidal systems [54]. For a pure metal like Zr, our unsupervised analysis brings us to the conclusion that the translational and orientational orderings arise synchronously.

### E. Liquid heterogeneities

To assess the scenario mentioned in Sec. III D, heterogeneity at an early nucleation stage is evaluated by a comparison of the distribution of atoms belonging to each cluster against the uniform distribution. As the comparison of distributions is particularly complex in three dimensions since the

[1] G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen, and A. Michaelides, Chem. Rev. **116**, 7078 (2016).

[2] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed. (Academic, San Diego, 2002).

[3] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft Research, Redmond, 2009).

[4] P. Geiger and C. Dellago, J. Chem. Phys. **139**, 164105 (2013).

[5] C. Dietz, T. Kretz, and M. H. Thoma, Phys. Rev. E **96**, 011301(R) (2017).

[6] E. Boattini, M. Ram, F. Smallenburg, and L. Filion, Mol. Phys. **116**, 3066 (2018).

[7] R. S. DeFever, C. Targonski, S. W. Hall, M. C. Smith, and S. Sarupria, Chem. Sci. **10**, 7503 (2019).

[8] W. F. Reinhart, A. W. Long, M. P. Howard, A. L. Ferguson, and A. Z. Panagiotopoulos, Soft Matter **13**, 4733 (2017).

[9] W. F. Reinhart and A. Z. Panagiotopoulos, Soft Matter **14**, 6083 (2018).

[10] M. Ceriotti, J. Chem. Phys. **150**, 150901 (2019).

[11] E. Boattini, M. Dijkstra, and L. Filion, J. Chem. Phys. **151**, 154901 (2019).

[12] J. Paret, R. L. Jack, and D. Coslovich, J. Chem. Phys. **152**, 144502 (2020).

[13] M. Spellings and S. C. Glotzer, AIChE J. **64**, 2198 (2018).

[14] F. C. Motta, in *Advances in Nonlinear Geosciences*, edited by A. A. Tsonis (Springer International Publishing, Cham, 2018), pp. 369–391.

[15] H. Tanaka, H. Tong, R. Shi, and J. Russo, Nat. Rev. Phys. **1**, 333 (2019).

[16] T. Nakamura, Y. Hiraoka, A. Hirata, E. G. Escolar, and Y. Nishiura, Nanotechnology **26**, 304001 (2015).

[17] Y. Hiraoka, T. Nakamura, A. Hirata, E. G. Escolar, K. Matsue, and Y. Nishiura, Proc. Natl. Acad. Sci. U.S.A. **113**, 7035 (2016).

[18] A. Hirata, T. Wada, I. Obayashi, and Y. Hiraoka, Commun. Mater. **1**, 98 (2020).

[19] S. Hong and D. Kim, J. Phys.: Condens. Matter **31**, 455403 (2019).

[20] K. Sasaki, R. Okajima, and T. Yamashita, *Liquid Structures Characterized by a Combination of the Persistent Homology Analysis and Molecular Dynamics Simulation* (Thessaloniki, Greece, 2018), p. 020015.

[21] S. Becker, E. Devijver, R. Molinier, and N. Jakse, Phys. Rev. B **102**, 104205 (2020).

[22] S. J. Plimpton, J. Comput. Phys. **117**, 1 (1995).

[23] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids: Second Edition* (Oxford University Press, Oxford, 2017).

[24] M. I. Baskes, Phys. Rev. B **46**, 2727 (1992).

[25] F. H. Stillinger and T. A. Weber, Phys. Rev. A **25**, 978 (1982).

[26] S. Menon, G. Leines, and J. Rogal, J. Open Source Softw. **4**, 1824 (2019).

[27] H. Edelsbrunner, D. Letscher, and A. Zomorodian, Discr. Comput. Geom. **28**, 511 (2002).

[28] A. Zomorodian and G. Carlsson, Discr. Comput Geom **33**, 249 (2005).

[29] R. Ghrist, Bull. Am. Math. Soc. **45**, 61 (2007).

[30] C. Maria, J.-D. Boissonnat, M. Glisse, and M. Yvinec, The Gudhi Library: Simplicial Complexes and Persistent Homology, INRIA, RR-8548 (2014).

[31] C. Tralie, N. Saul, and R. Bar-On, J. Open Source Softw. **3**, 925 (2018).

[32] B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh, Ann. Statist. **42**, 2301 (2014).

[33] M. Carrière and U. Bauer, *Proceedings of the 35th International Symposium on Computational Geometry*, edited by G. Barequet and Y. Wang (Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2019), pp. 21:1–21:15.

[34] M. Carrière, S. Y. Oudot, and M. Ovsjanikov, Comput. Graph. Forum **34**, 1 (2015).

[35] G. J. Ackland and A. P. Jones, Phys. Rev. B **73**, 054104 (2006).

[36] J. D. Honeycutt and H. C. Andersen, J. Phys. Chem. **91**, 4950 (1987).

[37] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, Phys. Rev. B **28**, 784 (1983).

[38] W. Lechner and C. Dellago, J. Chem. Phys. **129**, 114707 (2008).

[39] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.105.045304 for a comparison of the CNA and the BOOA; boxplots of the topological vectors in the homological dimensions $H_0$, $H_1$, and $H_2$ for each cluster $C_i$; and a principal component analysis [50] on the first two axes and a t-SNE representation [51] of the trained model.

[40] D. Faken and H. Jónsson, Comput. Mater. Sci. **2**, 279 (1994).

[41] A. K. Soper and M. A. Ricci, Phys. Rev. Lett. **84**, 2881 (2000).

[42] Z. Yan, S. V. Buldyrev, P. Kumar, N. Giovambattista, P. G. Debenedetti, and H. E. Stanley, Phys. Rev. E **76**, 051201 (2007).

[43] J. A. van Meel, L. Filion, C. Valeriani, and D. Frenkel, J. Chem. Phys. **136**, 234107 (2012).

[44] J. Higham and R. H. Henchman, J. Chem. Phys. **145**, 084108 (2016).

[45] A. P. Dempster, N. M. Laird, and D. B. Rubin, J. R. Stat. Soc.: Ser. B (Methodol.) **39**, 1 (1977).

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau, J. Mach. Learn. Res. **12**, 2825 (2011).

[47] S. Lloyd, IEEE Trans. Inform. Theory **28**, 129 (1982).

[48] C. Biernacki, G. Celeux, and G. Govaert, IEEE Trans. Pattern Anal. Mach. Intell. **22**, 719 (2000).

[49] G. Schwarz, Ann. Statist. **6**, 461 (1978).

[50] I. T. Jolliffe, *Principal Component Analysis, Second Edition* (Springer, New York, 2002).

[51] L. van der Maaten and G. Hinton, J. Mach. Learn. Res. **9**, 2579 (2008).

[52] P. C. Mahalanobis, Proc. Natl. Inst. Sci. (Calc.) **2**, 49 (1936).

[53] A. Stukowski, Model. Simul. Mater. Sci. Eng. **18**, 015012 (2010).

[54] J. Russo and H. Tanaka, Sci. Rep. **2**, 505 (2012).

[55] J. Russo and H. Tanaka, J. Chem. Phys. **145**, 211801 (2016).

[56] J. T. Berryman, M. Anwar, S. Dorosz, and T. Schilling, J. Chem. Phys. **145**, 211901 (2016).

[57] U. Gasser, E. R. Weeks, A. Schofield, P. N. Pusey, and D. A. Weitz, Science **292**, 258 (2001).

[58] U. Gasser, A. Schofield, and D. A. Weitz, J. Phys.: Condens. Matter **15**, 375 (2003).

[59] N. Jakse and A. Pasturel, Phys. Rev. Lett. **91**, 195501 (2003).

[60] D. R. Nelson and F. Spaepen, *Solid State Physics*, Vol. 42 (Elsevier, Amsterdam, 1989), pp. 1-90.

[61] H. Tanaka, Eur. Phys. J. E **35**, 113 (2012).

[62] J. Russo, F. Romano, and H. Tanaka, Phys. Rev. X **8**, 021040 (2018).

[63] A. Pasturel and N. Jakse, Npj Comput Mater **3**, 33 (2017).

[64] Y. Shibuta, S. Sakane, E. Miyoshi, S. Okita, T. Takaki, and M. Ohno, Nat. Commun. **8**, 10 (2017).

[65] G. Jug, A. Loidl, and H. Tanaka, Europhys. Lett. **133**, 56002 (2021).

[66] A. Kolmogoroff, Giornale dell'Istituto Italiano degli Attuari 4, 83 (1933).

[67] S. Becker, E. Devijver, R. Molinier, and N. Jakse, Sci. Rep. **12**, 3195 (2022).