Unsupervised local structure learning in elemental zirconium using Topological Data Analysis

Anonymous Author(s)
Affiliation
Address
email

Abstract

We discuss in this extended abstract a new application in materials science where topological data analysis (TDA) has proven to be relevant. TDA is used to analyze local atomic structures during homogeneous nucleation phenomena of pure zirconium in the undercooled liquid namely below its melting temperature. Persistent diagrams (PDs) are considered as descriptors to represent configurations of the physical system generated by large-scale molecular dynamics (MD) simulations up to one million atoms, and are evaluated by comparison to classical physical descriptors. This work is a first step towards a deeper understanding of the nucleation behavior of the system, where topological machine learning is a challenging but promising perspective.

1 Context

Physical motivation As crystal nucleation of matter from the liquid state is essentially initiated by heterogeneous nucleation from impurities, surfaces, or near grain boundaries, understanding the homogeneous crystal nucleation, driven by complex phenomena difficult to capture experimentally, remains a challenging issue. In [3], the glass formation (GF) has been studied as well as the onset of crystallization features of pure zirconium as a function of the degree of undercooling by means of large-scale MD simulations up to one million atoms. The GF results from a geometrical frustration and a strong competition between the crystal and icosahedral structural orderings. Setting up a topological machine learning approach to understand how this structural competition impacts the crystallization kinetics is the main aim of this work.

Data Homogeneous nucleation is observed along an isotherm at $T = 1250$ K at constant ambient pressure between the melting and the glass transition temperature ($T = 890$ K), corresponding to the lowest nucleation time [3] (see Fig. 2 in Supplementary Material). Configurations are recorded and their inherent structures are produced by an energy minimization using a conjugate gradient. From those data, the goal is to characterize, in an unsupervised way, the set of local structures, namely a central atom with its first neighbours up to 4.4 Å (corresponding to the first minimum of the pair correlation function and thus defining the coordination sphere, see Fig. 3 in Supplementary Material).

Descriptors To study these local structures we need descriptors which are invariant by translation and rotation. Usually in material science, the following classical descriptors are used: the Bond Angle Analysis (BAA) [1], the Bond-Orientational Order Analysis (BOOA) [13, 10] and the Common Neighbour Analysis [9] among others [2]. They characterize the radial and/or angular distribution of bonded pairs of atoms in the coordination sphere. In this work, we use for the first time TDA, and more precisely persistent homology, to study the topology of these local structures. The local environments of atoms are seen as a points cloud, and we draw PDs from them. PDs are usually

Submitted to the Topological Data Analysis and Beyond Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

compared using bottleneck or Wasserstein distances, which are stable under small perturbations of the point clouds [7]. However, the induced metric spaces are not isometrically embeddable (even with a bi-lipschitz map) into a finite dimensional Hilbert space [5]. We propose to use the signature introduced in [6], classically used to study 3D shapes, in order to get a stable representation of our PDs into a Hilbert space. For each pair of points (x, y) in a PD D, we consider

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

where $d_\Delta(\cdot)$ denotes the ℓ^∞ distance to the diagonal. The signature of D corresponds then to the vector with all those values sorted in decreasing order. Remark that the topological signature obtained here and the usual physical descriptors are highly correlated (see Fig. 5 in Supplementary Material).

2 Method and results

The topological signature is computed for 20 000 local structures, subsampled from the data generated in [3]. PDs have been denoised by removing points too close to the diagonal, using a threshold learned by bootstrap. Then, the unsupervised learning is done using a Gaussian Mixture Model (GMM), with full covariance matrices, fitted by an Expectation-Maximization algorithm (EM) on the topological descriptors. The number of clusters is selected using the Integrated Completed Likelihood criterion [4] (ICL)[1]. From our data, we get a signature vector with at most 17 coefficients for H_0 and at most 18 coefficients for H_1. This vector is illustrated by drawing several PDs in Fig. 4 in Supplementary Material. The GMM fits a model with 39 clusters (see Fig. 6 to see the ICL criterion in Supplementary Material).

Since our analysis is performed on the structure obtained during crystallization, we focus our attention on body-centered cubic (BCC) structures which are described in CNA by the signature [444]: 6 and [666]: 8. Among the 39 clusters, 6 of them have a pure BCC structure from the CNA signature (see Fig. 8 in Supplementary Material), thus would have been merged when only considering CNA; but have different signatures in TDA has illustrated in Fig. 9 in Supplementary Material. It illustrates the strength of TDA, which is able to detect perfect BCCs (where coefficients should be zero, as all the points in H_0 and H_1 should be equal), as well as their distorted counterparts due to temperature effects in the simulation where coefficients are nonzero. From Fig. 9, we remark that the cluster 16 seems to be the most perfect, with the smallest coefficients.

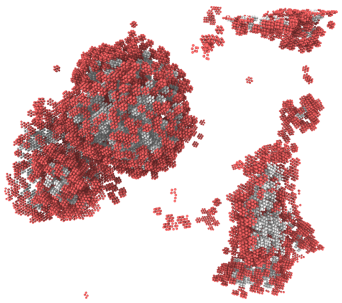

Figure 1: Representation of the system obtained with the OVITO software [14]. We plot only the pure BCC clusters for clarity. The perfect cluster is in the center of the nuclei, whereas higher values in TDA correspond to particles around the nuclei. Other particles (not BCC) are outside of the nuclei.

By analyzing the physical system, we notice that several nuclei are forming, showing the crystallization of the system. We represent particles with a complete BCC local structure in Fig. 1, where particles in white are from cluster 16, and in red from other pure BCC clusters. Particles from cluster 16 corresponds to the center of the nuclei, and particles from clusters with higher TDA signatures (clusters 1, 4, 14, 23, 37) surround them. In particular, this shows that TDA detects a gradual distortion of BCC local structures which are physically relevant.

Clusters with BCCs but not pure are also interpretable from the CNA signature. We detect that they are composed of structures close to BCC, with missing bonds or neighbours. They are also mainly outside the main nuclei in the 3D representation. See Fig. 10 and Fig. 11 in Supplementary Material.

3 Conclusion and future work

This application shows a first attempt to understand the local structures which arise in elemental zirconium during solidification, using TDA as a relevant descriptor. The TDA is found to represent thin important patterns for local structure, even more than classical material science features, which are discriminated using a GMM. This first step show good advantages of the descriptor, and we plan to take benefit of it also for a study of the global structure.

[1] In practice, we use Python with `Gudhi` [11] and `scikit-learn` [12] libraries.

References

[1] G. J. Ackland and A. P. Jones, "Applications of local crystal structure measures in experiment and simulation," Phys. Rev. B, vol. 73, no. 5, p. 054104, 2006, doi: 10.1103/PhysRevB.73.054104.

[2] A. P. Bartók, R. Kondor, G. Csányi. "On representing chemical environments", Phys. Rev. B, vol. 87, no. 18, 2013.

[3] S. Becker, E. Devijver, R. Molinier, and N. Jakse, "Glass-forming ability of elemental zirconium", Phys. Rev. B, vol. 102, 104205, 2020, doi: 10.1103/PhysRevB.102.104205

[4] C. Biernacki, G. Celeux and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 7, pp. 719-725, 2000, doi: 10.1109/34.865189.

[5] M. Carrière, U. Bauer. "On the Metric Distortion of Embedding Persistence Diagrams into Separable Hilbert Spaces", 35th International Symposium on Computational Geometry (SoCG 2019), vol. 129, pp. 21:1–21:15, 2019, doi:10.4230/LIPIcs.SoCG.2019.21.

[6] M. Carrière, S.Y. Oudot and M. Ovsjanikov, "Stable Topological Signatures for Points on 3D Shapes", Computer Graphics Forum, 34: 1-12, 2015, doi:10.1111/cgf.12692

[7] D. Cohen-Steiner, H. Edelsbrunner, J. Harer. "Stability of Persistence Diagrams", Discrete & Computational Geometry, vol. 37, no. 1, pp. 103-120, 2007, doi:10.1007/s00454-006-1276-5.

[8] D. Faken and H. Jónsson, "Systematic analysis of local atomic structure combined with 3D computer graphics," Computational Materials Science, vol. 2, no. 2, pp. 279-286, 1994, doi: 10.1016/0927-0256(94)90109-0.

[9] J. D. Honeycutt and H. C. Andersen, "Molecular dynamics study of melting and freezing of small Lennard-Jones clusters," J. Phys. Chem., vol. 91, no. 19, pp. 4950-4963, 1987, doi: 10.1021/j100303a014.

[10] W. Lechner and C. Dellago, "Accurate determination of crystal structures based on averaged local bond order parameters," The Journal of Chemical Physics, vol. 129, no. 11, p. 114707, 2008, doi: 10.1063/1.2977970.

[11] C. Maria, J.-D. Boissonnat, M. Glisse, M. Yvinec. "The Gudhi Library: Simplicial Complexes and Persistent Homology", Mathematical Software, ICMS, 167-174, 2014.

[12] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python", JMLR 12, pp. 2825-2830, 2011.

[13] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, "Bond-orientational order in liquids and glasses," Phys. Rev. B, vol. 28, no. 2, pp. 784-805, 1983, doi: 10.1103/PhysRevB.28.784.

[14] A. Stukowski, “Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool,” Modelling Simul. Mater. Sci. Eng., vol. 18, no. 1, p. 015012, Jan. 2010, doi: 10.1088/0965-0393/18/1/015012.

A Data

A.1 Physical context

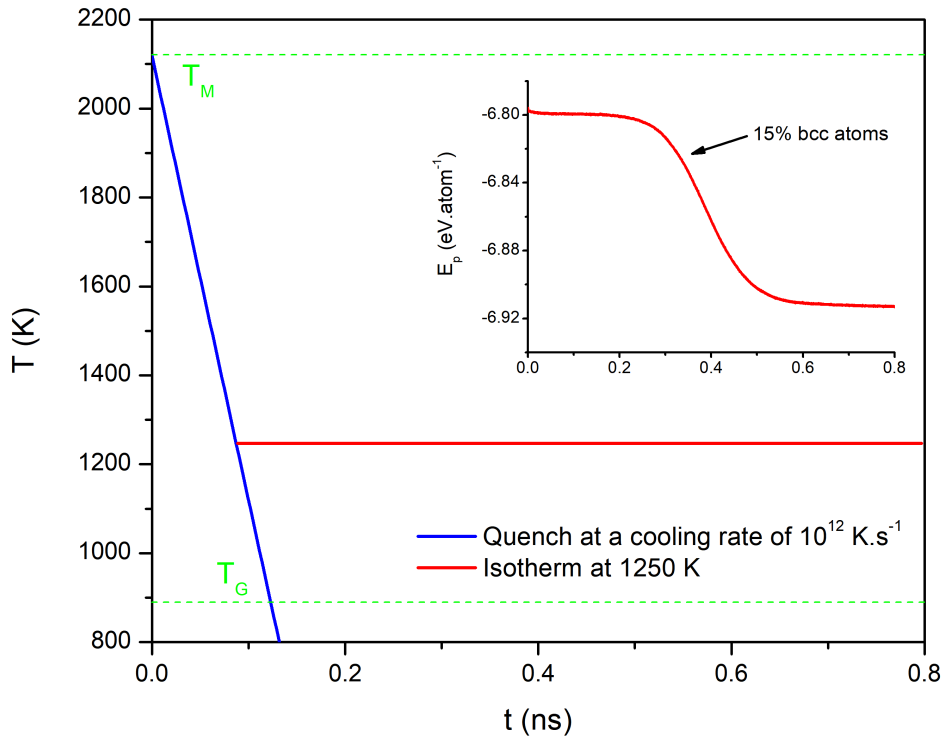

Figure 2: Time-temperature schematic representation of the quench and isotherm modelised by molecular dynamics simulations. The inset represents the evolution of the total potentiel energy along the isotherm at 1250 K.

To study the solidification process of a material, a rapid quench from the liquid to the glassy state must be performed. Then, along this quench, several metastable undercooled configurations of the simulation box, at specific times and temperatures, are extracted in order to do isotherms and let them evolve, through nucleation, to the solid stable state. For this study, we have extracted one configuration at the onset of crystallization, where 15 % of solid bcc atoms have been formed. The Fig. 2 summarize this steps, with T_M the thermodynamic melting temperature, T_G the glass transition temperature and in inset the evolution of the potential energy of the isotherm at 1250 K through the time of nucleation.

For an atom set at the origin, the radial distribution function $r \mapsto g(r)$ gives the distribution of the other atoms in a sphere around it. A typical appearance of this function is shown for the liquid state of our material in Fig. 3. Graphically, one can observes a succession of peaks, each of which represents successive layers of neighbours around the central atom. In the case of the local structure, the first minimum of this function is set as a cut-off radius delimiting the end of the first neighbours shell.

A.2 Descriptors

We plot 3 persistence diagrams in Fig. 4 that reflect the diversity in our data. We consider an observation in cluster 16, the perfect one discussed in the main paper, an observation in cluster 14 which is a pure BCC cluster with higher values in the TDA signature, and an observation in cluster 6

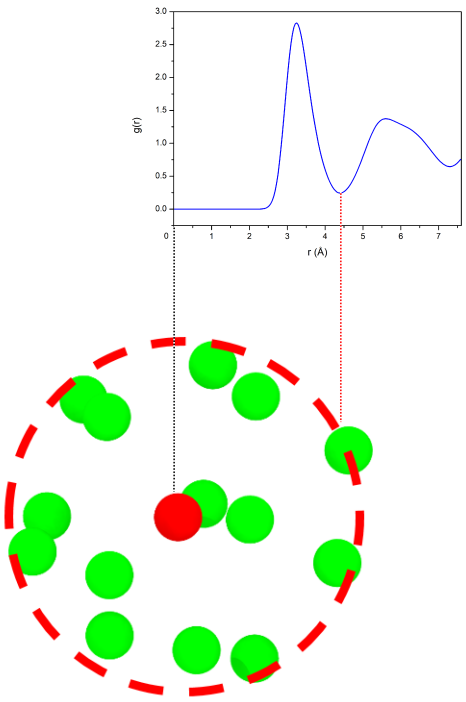

Figure 3: First neighbours shell of a central particle (in red) with a cut-off radius determined by the radial distribution function $r \mapsto g(r)$ of undercooled Zirconium at 1250 K.

which is a cluster with BCC but also other local structures. More precisely, we select the observation which is the closest to the center of each cluster.

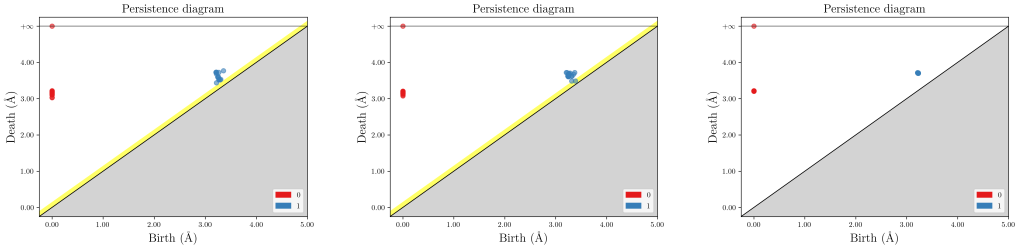

Figure 4: Persistence diagram on our data for observations in the clusters 6 (left), 14 (middle) and 16 (right). The yellow band is representing the noise detected by bootstrap.

To compare TDA descriptor with physical descriptors, we plot the empirical correlation matrix between BAA, CNA, TDA and BOAA in Fig. 5. We remark that TDA is highly correlated with all the other descriptors.

The correlation matrix associated to the TDA shows some specificity of the data: all local neighbours contain at least 10 particles, so the first 10 components of H_0 are highly correlated. There are correlated with high values of H_1, because most of the local structures have few H_1 components (50% of the population have less than 5 components).

The correlation matrix associated to the CNA is sparse, because only few coefficients (the first ones) are present in most of the structures. It also explains the low correlation between the last coefficients of CNA with the other descriptors.

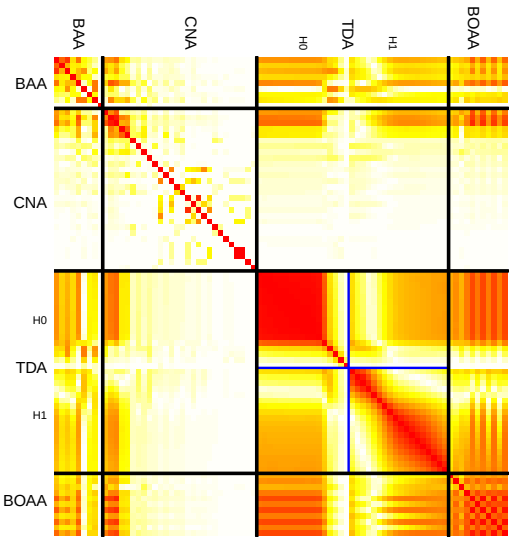

Figure 5: Empirical correlation matrix (absolute value) between several descriptor families: BAA, CNA, TDA and BOAA. Low values are white, high values are red. We distinguish between the descriptors by black lines, and between levels of homology with blue line (H_0 and H_1).

B Method

B.1 Number of clusters

We plot the ICL criterion in Fig. 6 to illustrate the model selection criterion. The criterion seems to be constant from a large enough dimension, but we smooth this curve (by a polynomial function) and compare several model selection criterion (BIC, ICL and the Slope Heuristics), which were focusing on the same value, 39 clusters.

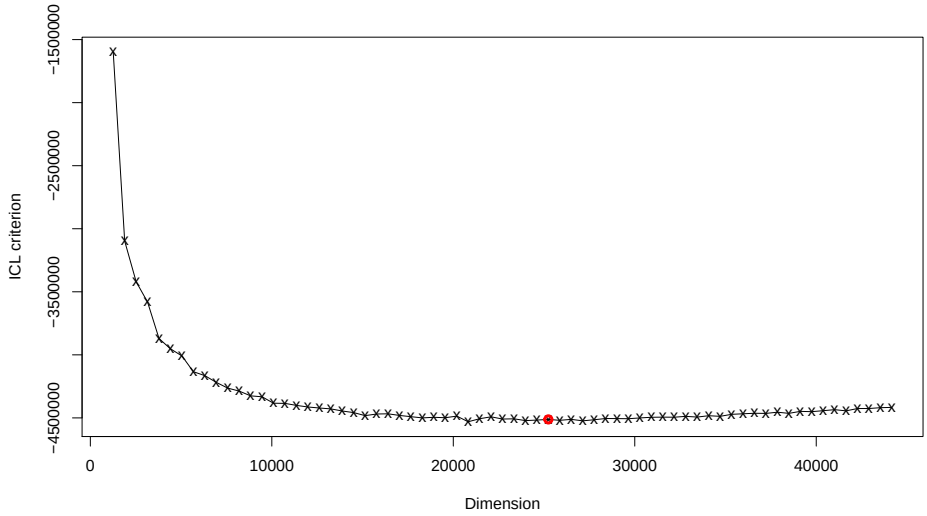

Figure 6: ICL criterion for each model in the collection with respect to the dimension of the model (number of parameters to be estimated). The red point corresponds to the selected model.

B.2 Pure clusters of BCC

In Fig. 7, we provide boxplots for CNA and TDA descriptor for all the observations. It illustrates the diversity of the structures we are looking at, despite of the focus we are doing on BCC. It also illustrates the analysis we did for the empirical correlation matrices, with range of values for each descriptor.

In Fig. 8 and 9, we provide boxplots for the 6 clusters with only BCC structures. Although there is no difference on the CNA descriptor (there are labeled BCC), the TDA signatures are very different from a cluster to another. We particularly remark that the cluster 16 has very small values. We have illustrated in Fig. 1 that they are corresponding to the structure which are inside the nuclei and surrounded by structure from other BCC clusters.

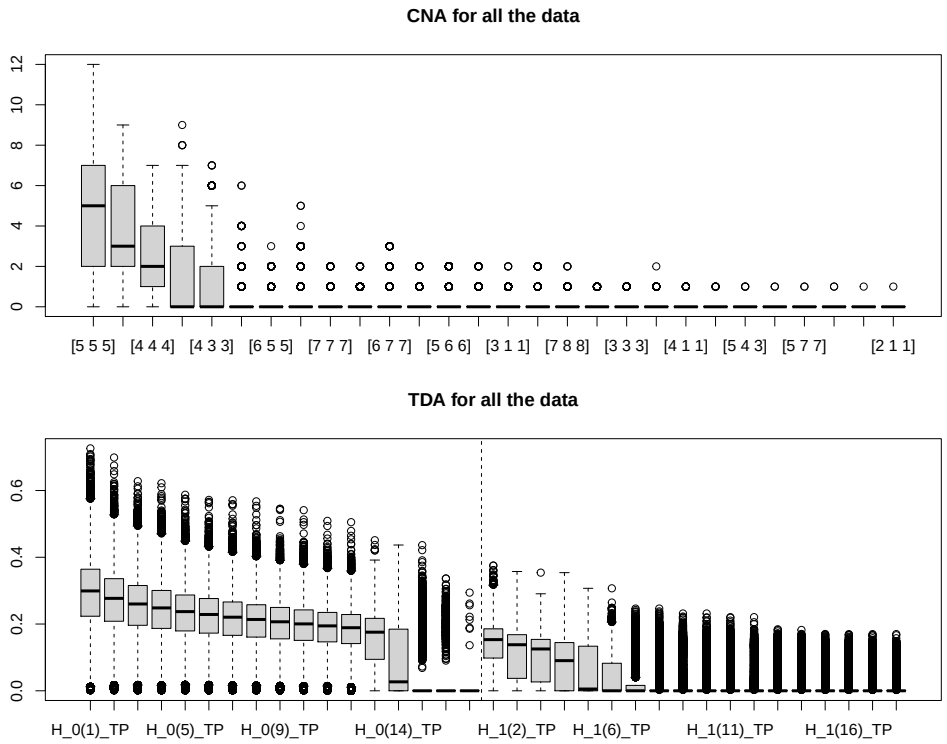

Figure 7: Boxplot of CNA and TDA descriptors for all the observations.

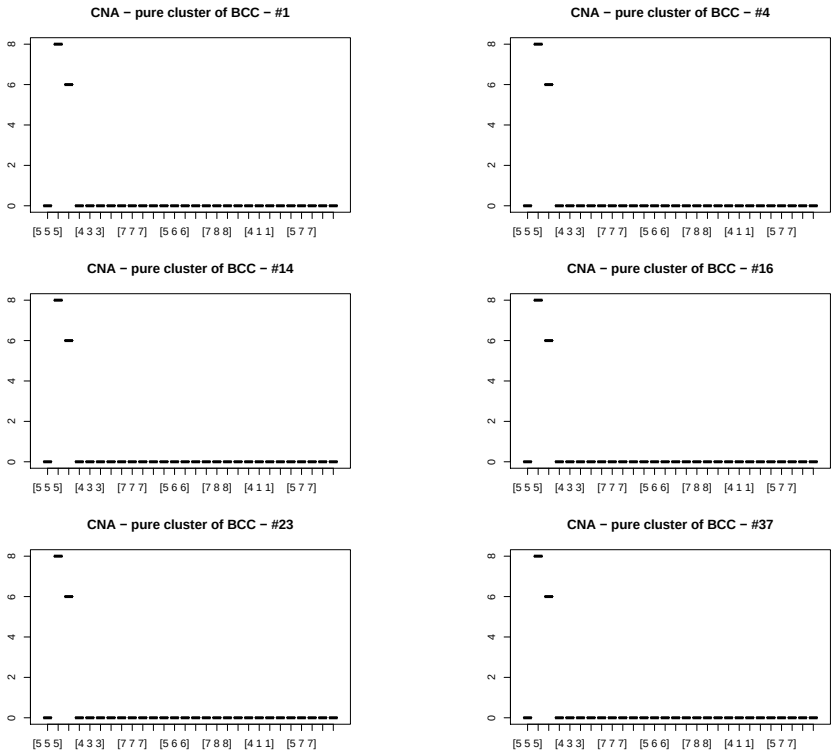

Figure 8: Boxplot for the CNA descriptor for the 6 pure clusters with BCCs.

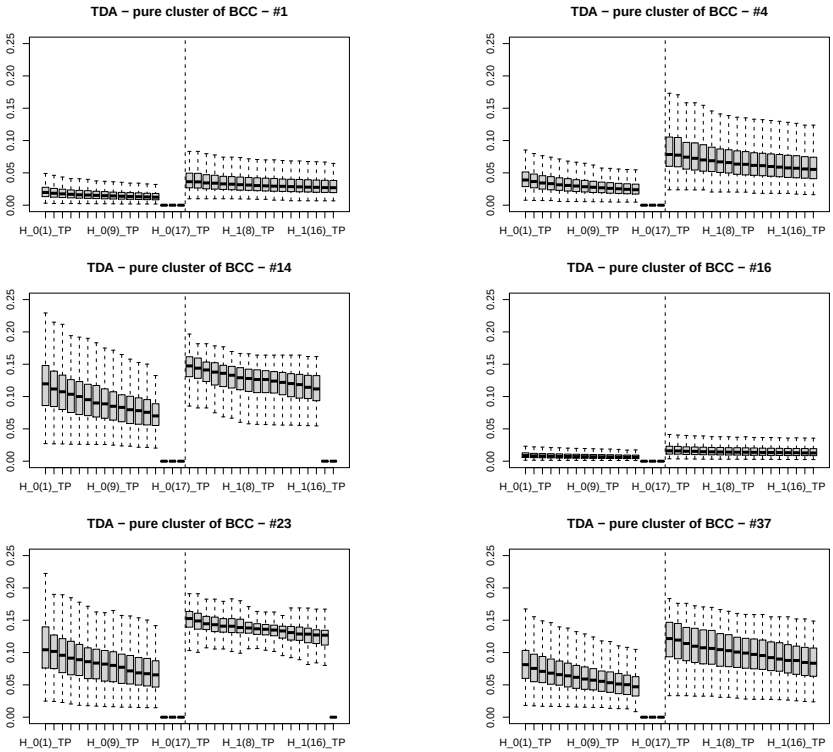

Figure 9: Boxplot for the TDA descriptor for the 6 pure clusters with BCCs.

B.3 Clusters of BCC (not pure)

In Fig. 10, we provide the CNA signature for observations in clusters that contain BCC structures (but which are not pure). We remark that those signatures are close to the BCC ones ([444]: 6 and [666]: 8), particularly for clusters 8 and 32. We discussed that in the main paper.

In Fig. 11, we provide the TDA signature for observations in the same clusters. We remark that the discrimination is mainly over H_1 coefficients (zero or nonzero).

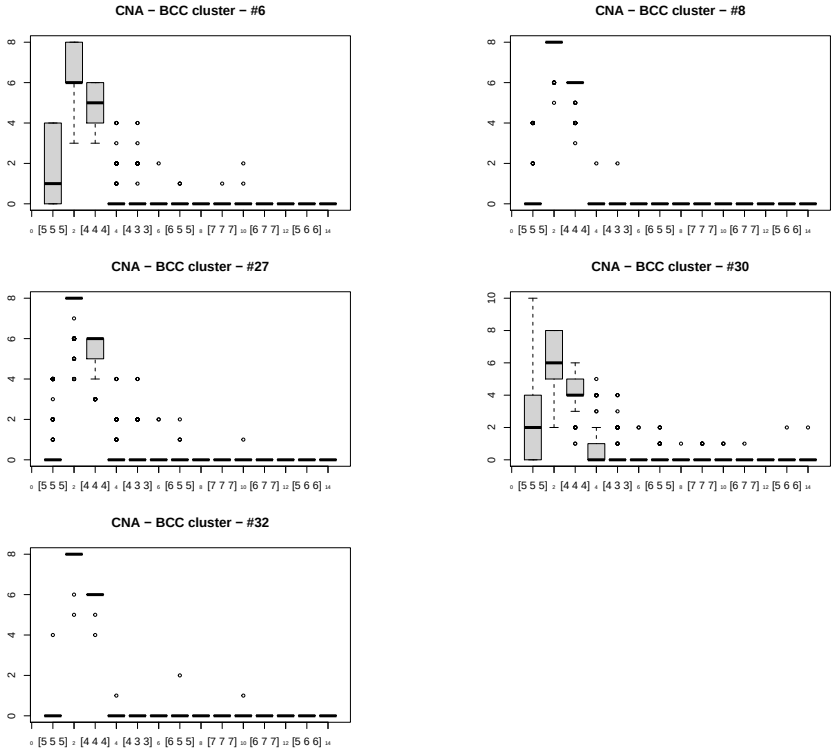

Figure 10: Boxplot for the CNA descriptor for the clusters with BCCs.

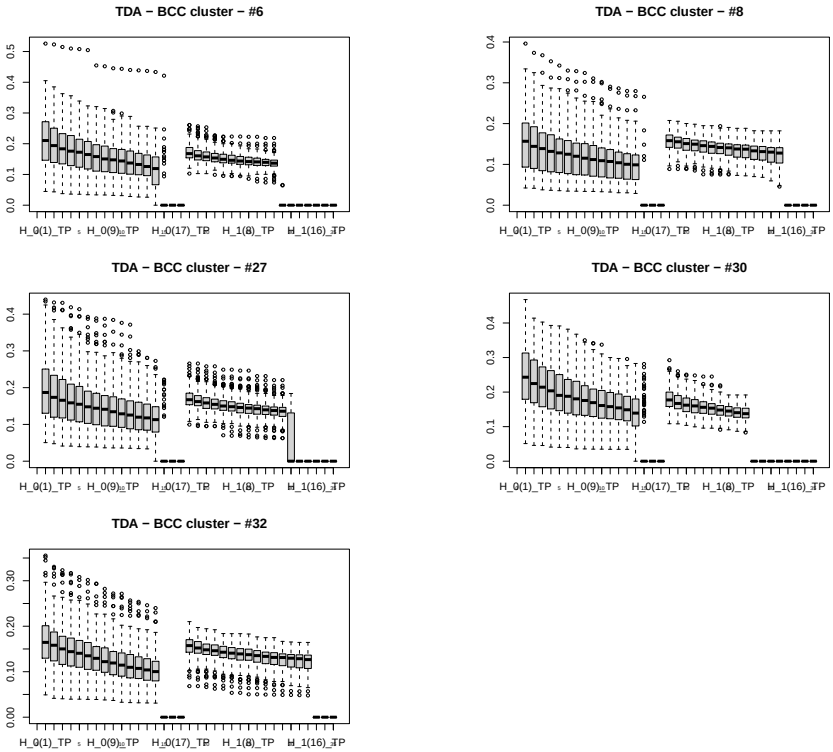

Figure 11: Boxplot for the TDA descriptor for the clusters with BCCs.