Unsupervised topological learning approach of crystal nucleation in pure Tantalum

Sébastien Becker
sebastien.becker@univ-grenoble-alpes.fr

Emilie Devijver
emilie.devijver@univ-grenoble-alpes.fr

Rémi Molinier
remi.molinier@univ-grenoble-alpes.fr

Noël Jakse
noel.jakse@univ-grenoble-alpes.fr
Université Grenoble Alpes, CNRS, Grenoble INP, SIMaP, LIG et IF
F-38000 Grenoble, France

Abstract

Nucleation phenomena commonly observed in our every day life are of fundamental, technological and societal importance in many areas, but some of their most intimate mechanisms remain however to be unraveled. Crystal nucleation, the early stages where the liquid-to-solid transition occurs upon undercooling, initiates at the atomic level on nanometer length and sub-picoseconds time scales and involves complex multidimensional mechanisms with local symmetry breaking that can hardly be observed experimentally in the very details. To reveal their structural features in simulations without a priori, an unsupervised learning approach founded on topological descriptors loaned from persistent homology concepts is proposed. Applied here to a monatomic metal, namely Tantalum (Ta), it shows that both translational and orientational ordering always come into play simultaneously when homogeneous nucleation starts in regions with low five-fold symmetry.

Understanding homogeneous crystal nucleation under deep undercooling conditions remains a formidable issue, as crystallization is essentially heterogeneous in nature and initiated from impurities, surfaces, or near grain boundaries that often hinder its occurrence [1]. Unreachable until very recently, experimental observations of early stages of nuclei was achieved by a *tour de force* using time tracking of three-dimensional (3D) Atomic Electron Tomography [2] of metallic nanoparticles. Those complex phenomena remain to date out-of-reach experimentally for bulk systems, thus hindering our theoretical understanding. This line of research still belongs mostly to the domain of atomic-level simulations and more particularly to molecular dynamics (MD) with generic interaction models [3, 4]. To reach statistically meaningful events, large scale simulations are required[1].

To identify translational and orientational orderings during homogeneous nucleation in MD simulations, an unsupervised learning approach based on topological data analysis (TDA) signatures, and more precisely persistent homology (PH) [5, 6] was developed. PH is an intrinsically flexible, yet highly informative, tool which detects meaningful topological features deduced from atomic

[1] This still remains challenging as only few studies are providing now million-atom simulations for monatomic metals [1].

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

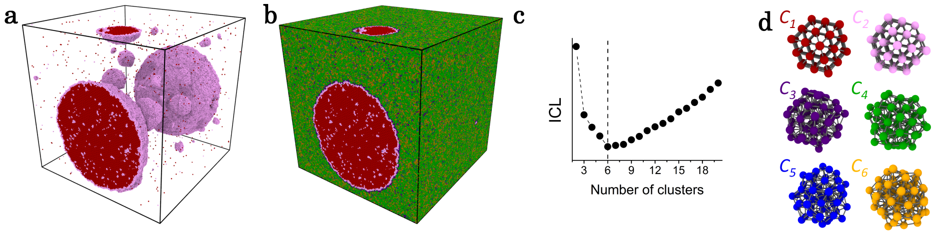

Figure 1: **Unsupervised learning of homogeneous nucleation.** Snapshot of a ten-million atom MD simulation of Ta during nucleation along the $T = 1900$ K isotherm (a and b). Independent local atomic structures form a train set represented in the descriptor space by 215 PH components up to the second order. (c) Evolution of the ICL criterion as a function of number of clusters is used to get the optimal number of clusters shown in (d). In (a) the snapshot is represented only with atoms in cluster C_1 and cluster C_2 revealing all nuclei, while in (b) atoms of all clusters are displayed.

configurations. It was successfully applied very recently to characterize structural environments in metallic glasses [7], ice [8] and complex molecular liquids [9]. Always used as a structural analysis in these studies, the originality here is to use PH as a translational and rotational invariant descriptor to encode the local structures required for the clustering method. More precisely, a persistence diagram is drawn from each local structure and then encoded into a topological vector as in [6]. Each coordinate of the topological vector is associated to a pair of points (x, y) in a persistence diagram D for a fixed level of homology, except the infinite point, and is calculated by

$$m_D(x, y) = \min\{\|x - y\|_\infty, d_\Delta(x), d_\Delta(y)\}, \tag{1}$$

where $d_\Delta(\cdot)$ denotes the ℓ^∞ distance to the diagonal, and those coordinates are sorted by decreasing order. For the clustering, a model-based method is used, namely Gaussian Mixture Models (GMM) [10, Chapter 14] and its estimation by an Expectation Maximization (EM) algorithm [11]. The number of clusters is selected by Integrated Criterion Likelihood (ICL, [12]), a refinement for clustering of Bayesian Integrated Likelihood (BIC, [13]). The inferred model from the method called hereafter TDA-GMM, is used to identify and describe the structural and morphological properties of the nuclei as well as their liquid environment at various steps of crystal nucleation. With this unsupervised approach, the homogeneous nucleation process was studied in liquid Ta, a monatomic metal having an underlying body-centered cubic (bcc) crystalline phase. Large-scale molecular dynamics simulations comprising ten million atoms were performed. Figure 1 depicts the result of the methodology. A configuration of the simulation is chosen during crystal nucleation as described below. As it contains many nuclei with different sizes and a substantial amount of liquid, it is considered as representative of the phenomenon. From its inherent structure [14], a training set of 5 000 non overlapping local spherical structures (encoded trough their topological vectors) within a cutoff radius of 6.8 Å was sampled for the unsupervised learning, with the constraints of covering the entire simulation box uniformly and randomly. Among all possible sets upon applying the GMM, the one with 6 clusters (later on denoted by C_1 through C_6) shown in 1 (d) was automatically inferred to be representative of the system based on the minimum ICL criterion 1(c). The snapshot of the simulation box in Fig. 1(a) displays only local structure from clusters C_1 and C_2, as they show clearly a crystalline order. They reveal all nuclei as it will be seen below, along with their structure, size and morphology out of the simulation box displayed in Fig. 1(b). From this model, each atom of each configuration generated by the MD simulation can be assigned, when considered with its surrounding local structure, to one of the six clusters (the one with the highest probability). Such a clustering training is performed and shows that more than 99.99 % of the structures have a probability to belong to the most probable Gaussian component greater than 0.999, even for structures not in the initial training set. An analysis of the eigenvalues of the covariance matrices shows elliptical shapes, which proves the necessity of the GMM with general covariance matrices compared to simpler unsupervised algorithms (e.g., k-means would only fit hyperspheres).

Figure 2 shows typical homogeneous nucleation events in undercooled Ta during an isothermal process close to the nose of the time-temperature-transformation (TTT) curve[2]. The liquid above the melting point T_M (at $T = 3300$ K) was first quenched down at ambient pressure to the glass

[2] which can be done by standard MD simulations without the need of an accelerated method [15].

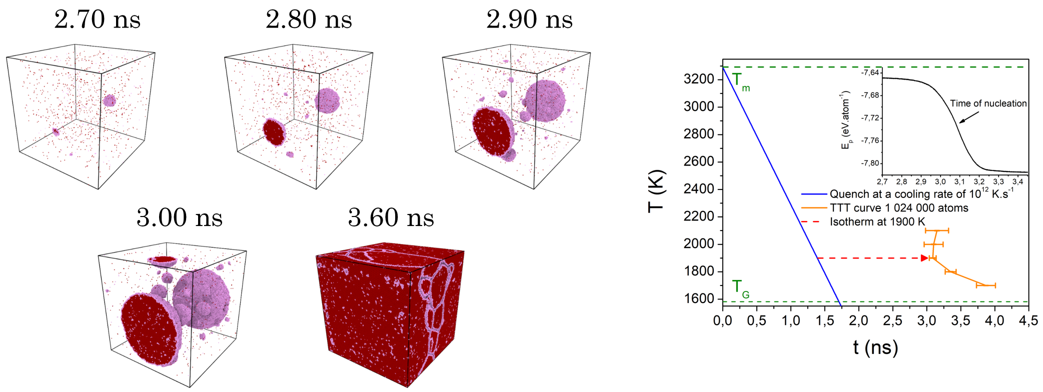

Figure 2: **Homogeneous nucleation in Ta undercooled liquids.** Snapshots of the MD simulations, during isothermal nucleation at different times for temperatures close to the nose of the Time-Temperature Transformation (TTT). From stored configurations during fast cooling (blue curves), nucleation events along several isotherms were observed by monitoring the sharp drop of the internal energy (inset). The average nucleation times τ_N (symbols) were determined from 5 independent simulations for each temperature giving the TTT curves in the vicinity of the nose (orange lines).

transition sufficiently rapidly to avoid nucleation. From stored configurations during cooling, the TTT curves in the vicinity of the nose were built from observation of the nucleation along several isotherms as shown in Figs. 2. An isotherm slightly above the TTT nose is chosen for the analysis ($T = 1900$ K). From chosen configurations during the nucleation and growth process, the clustering is obtained from application of the trained model. Strongly growing fraction of mainly two clusters, concomitant to the sharp drop of the energy, is observed. Only local structures belonging to these clusters are drawn in Figures 2, revealing evidently the nuclei and their evolution in time, recalling that solely the topological vector is describing the local structure. The nuclei morphologies show globular shapes that are rather spherical, characteristic of high ΔT. Interestingly, atoms from one of the two clusters (in red) are mainly located inside the nuclei while atoms from the second one (in pink) steadily remain essentially at the border upon growing. They stay finally at grain boundaries after full solidification of the simulation boxes. The vast majority of the embryos[3] seen in Fig. 2 dissolves back to the liquid while those attaining the critical size are rare and grow. The large simulation box allows to follow the nucleation process for a longer time, sufficient to observe more secondary nucleation events [16].

The nucleation process is characterized at least by two order parameters, the translational order (TO) and the crystalline ordering called hereafter the bond orientational order (BOO). A dedicated representation of the TO is the number density. It is primarily applied to the embryos and the nuclei at different stage of the growth, through the radial partial atomic density profiles $\rho_i(r) = N_i(r)/\frac{4\pi}{3}[(r+\Delta r)^3 - r^3]$ as a function of distance r of the estimated center of the nucleus, $N_i(r)$ being the number of atoms belonging to cluster C_i in a spherical shell of radius r and thickness $\Delta r = 1$ Å. Fig. 3(a) depicts the density profiles $\rho_i(r)$ for all 6 clusters for the largest nucleus shown in Fig. 2(a) and its surrounding liquid at time 2.7 ns. The corresponding slice of the nucleus through its center is drawn in Fig. 3(b). Thus, the nucleus is defined by atoms belonging to clusters C_1 and C_2 as described above, atoms of C_1 forming the center of the nucleus, while atoms of C_2 being mainly located at its border. It should be noted that atoms of cluster C_3 are mainly located at the boundary of the nucleus, but they cannot be considered as being part of it, as they are also present in the entire box. From the total density profile of the nucleus $\rho_N(r) = \rho_1(r) + \rho_2(r)$, it can be seen clearly that the density of nucleus has already reached at this stage the one of the bulk crystal at the same temperature. Defining the remaining clusters (C_3 to C_6) as belonging to the liquid yields to a total density profile $\rho_L(r) = \sum_{i=3}^{6} \rho_i(r)$ showing that even in the vicinity of the nucleus the liquid is negligibly influenced by its presence, keeping the density of the bulk undercooled liquid. Fig. 3(c) shows the evolution of the density profile $\rho_N(r)$ at different times of the growing process. The average radius r_N of the nucleus is taken as the value of r at half-maximum of $\rho_N(r)$ and its

[3] Nuclei smaller than the critical size of 65 atoms.

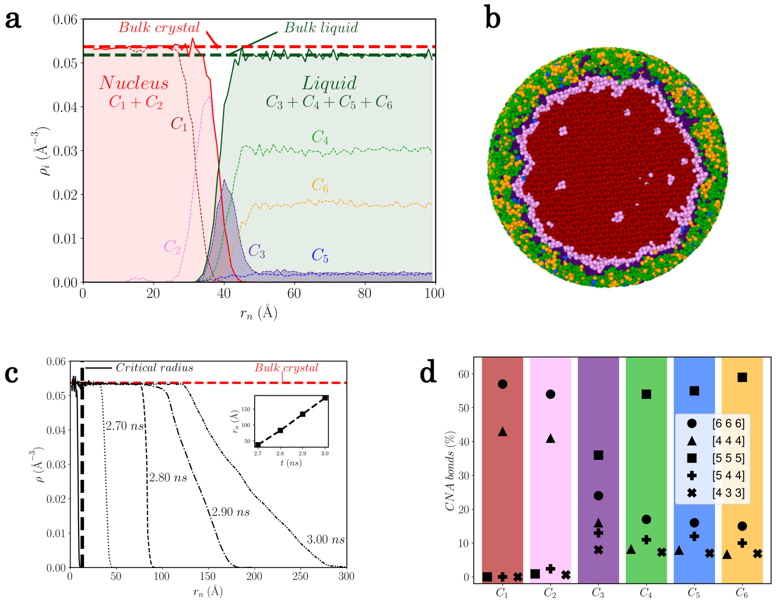

Figure 3: **Translational and bond-orientational order parameters.** (a) Radial density profile of the largest nucleus during the growth at 2.7 ns along the $T = 1900$ K isotherm. The red and blue dashed horizontal lines correspond respectively to the average bulk crystalline density and average bulk undercooled liquid without nucleation events (b) Corresponding slice of the nucleus through its centre and the surrounding liquid where atoms have been coloured according to the cluster they belong to (see Fig. 1(d)). (c) Total radial density profile of the largest nucleus during growth before solidification. Inset: time evolution of the radius of the nucleus. (d) Bond-orientational order in terms of bonded pairs of the Common-Neighbor Analysis [17] for each cluster of the model.

evolution with time is shown in the inset, displaying a linear behaviour in agreement with CNT [1]. Whatever the size of the nuclei, the density of the inner part is close to the bulk crystal. More importantly, this is all the more true for all the embryos below the critical size up to a single atomic structure corresponding to the minimal size of about 65 atoms belonging to cluster C_1 or C_2 as identified from their local structure. The BOO of each cluster is identified through the Common Neighbor Analysis (CNA) [17], chosen as a well-known and robust tool. The CNA signature [18] given in Fig. 3(d) reveals that structures from clusters C_1 and C_2 possess respectively a perfect and slightly distorted bcc crystalline ordering confirming the above analysis of nucleation and growth in terms of topological descriptors. Local structures from clusters C_4, C_5 and C_6 display various high degrees of five-fold symmetry (FFS) characteristic of the liquid state together with a small but non negligible degree of bcc ordering, while structures from cluster C_3 retains both FFS and bcc order in similar proportions. Such a BOO of the four clusters associated to the liquid agrees well with *ab initio* molecular dynamics simulations [19] and was interpreted as compatible with the A15 crystalline phase. This analysis highlights and confirms that the TDA-GMM unsupervised learning approach is a powerful method to capture the structural picture in its finest details.

The question whether the onset of nucleation is initiated primarily by translational or by orientational ordering is still open, and was debated during the last decade with a controversy essentially centered on the hard sphere and associated colloidal systems [20, 22]. The small emerging embryos at the onset of nucleation, corresponding to one structure of 55 to 70 atoms belonging to C_1 or C_2 with bcc crystalline BOO, show bond lengths of their bcc lattice close to the density of the bulk crystal. This provides evidence given the size of embryos that can be detected here: translational and bond-orientational orders appear simultaneously and rule out the scenario in which homogeneous nucleation is driven by BOO first [23] for metallic systems[4].

[4] which is consistent with the fact that, unlike hard spheres, metallic systems with strong bonding are more energy driven rather than entropy driven systems.

The present unsupervised learning approach was shown to be a powerful tool to unravel the atomic scale mechanisms of crystal nucleation in Ta. Other unsupervised methods can retrieve the dissociation between solid and liquid-like structure. For example, a simple Principal Component Analysis discriminates those two states on the first axis, as well as the famous t-SNE [21] that represents the points such that liquid related particles are closer . However, there is no clear frontier between them (whereas our clusters are well defined, as given by the a posteriori probabilities), and there is for example no distinction between cluster 3 and 4, although the interpretation is clear. Our results are in line with the emerging idea that heterogeneities which exist in the undercooled liquid [22] play the foremost role in the onset of nucleation. Nucleation have been indeed found to start in low FFS regions, which is consistent with Frank's argument [24], with translational and orientational ordering taking place simultaneously in emerging embryos. Moreover, embryos as well as nuclei during the growth possess the bulk crystal density driven by the metallic bond length while the surrounding liquid keeps the bulk liquid density in accordance with the classical nucleation theory [1]. However, our analysis reveals also some aspects beyond the CNT, such as nuclei having a diffuse interface with the surrounding liquid. This promising methodology more generally opens the door to a deeper and autonomous investigation of atomic level mechanisms in materials science. The nucleation analysis on multicomponent systems is, for example, especially relevant to enhance materials design. Also, it would be interesting to extend the method to learn the time evolution, e.g. through recent generalization of the persistent homology to time series [25].

Acknowledgments

We first acknowledge the two reviewers from their helpful remarks, that allow us to improve this paper. We acknowledge the CINES and IDRIS under Project No. INP2227/72914, as well as CIMENT/GRICAD for computational resources. This work was performed within the framework of the Centre of Excellence of Multifunctional Architectured Materials "CEMAM"ANR-10-LABX-44-01 funded by the "Investments for the Future" Program. This work has been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003). Fruitful discussions within the French collaborative network in high-temperature thermodynamics GDR CNRS3584 (TherMatHT) are also acknowledged.

References

[1] Sosso, G. C. *et al.* Crystal Nucleation in Liquids: Open Questions and Future Challenges in Molecular Dynamics Simulations. Chem. Rev. **116**, 7078–7116 (2016).

[2] Zhou, J. *et al.* Observing crystal nucleation in four dimensions using atomic electron tomography. Nature **570**, 500–503 (2019).

[3] Auer, S. & Frenkel, D. Prediction of absolute crystal-nucleation rate in hard-sphere colloids. Nature **409**, 1020–1023 (2001).

[4] ten Wolde, P. R., Ruiz-Montero, M.J. & Frenkel D. Numerical evidence for b.c.c. or ordering at the surface of a critical f.c.c. nucleus. Phys Rev Lett **75**, 2714–2717 (1995).

[5] Motta, F. C. Topological Data Analysis: Developments and Applications in Adv. Nonlinear Geosci., 369–391 (Tsonis A. A. ed., Springer International Publishing AG 2018).

[6] Carrière, M., Oudot, S. Y. & Ovsjanikov, M. Stable topological signatures for points on 3D shapes. Eurographics Symp. Geom. Process. **34**, 1–12 (2015).

[7] Hirata, A., Wada, T., Obayashi, I. & Hiraoka, Y. Structural changes during glass formation extracted by computational homology with machine learning. Commun. Mater. **1**, 1–4 (2020).

[8] Hong, S. & Kim, D. Medium-range order in amorphous ices revealed by persistent homology. J. Phys. Condens. Matter **31**, (2019).

[9] Sasaki, K., Okajima, R. & Yamashita, T. Liquid structures characterized by a combination of the persistent homology analysis and molecular dynamics simulation. AIP Conf. Proc. 020015 (2018).

[10] Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc. (2001).

[11] Dempster, A., Laird, N.,& Rubin, D. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **39(1)**, 1-38 (1977).

[12] Biernacki, C., Celeux, G. & Govaert, G. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**, 7, 719-725 (2000).

[13] Schwarz, G., Estimating the dimension of a model. The Annals of Statistics **6**, 461-464 (1978).

[14] Stillinger, F. H. & T. A. Weber, T. A. Hidden structure in liquids. Phys. Rev. A **25**, 978 (1982).

[15] Allen, R. J. , Valeriani, C. & Ten Wolde, P.R. J. Phys.: Condens. Matter. **21**, 463102 (2009).

[16] Shibuta, Y. *et al.* Heterogeneity in homogeneous nucleation from billion-atom molecular dynamics simulation of solidification of pure metal. Nat. Commun. **8**, 1–8 (2017).

[17] Faken, D. & Jónsson H. Systematic analysis of local atomic structure combined with 3D computer graphics, Comput. Mat. Sci., Computational Materials Science **2**, 279-286 (1994).

[18] Jakse, N. & Pasturel, A. Local Order of Liquid and Undercooled Transition metal based systems: ab initio molecular dynamics study. Mod. Phys. Lett. B **20**, 655–674 (2006).

[19] Jakse, N., Le Bacq, O. & Pasturel, A. Prediction of the local structure of liquid and supercooled tantalum. Phys. Rev. B **70**, 174203 (2004).

[20] Berryman, J. T., Anwar, M., Dorosz, S. & Schilling, T. The early crystal nucleation process in hard spheres shows synchronised ordering and densification. J. Chem. Phys. **145**, 211901 (2016).

[21] van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. JMLR **9**(86):2579-2605 (2008).

[22] Russo, J. & Tanaka, H. Crystal nucleation as the ordering of multiple order parameters Crystal nucleation as the ordering of multiple order parameters. J. Chem. Phys. **145**, 211801 (2016).

[23] Russo, J. & Tanaka, H. The microscopic pathway to crystallization in supercooled liquids. Sci. Rep. **2**, 505 (2012).

[24] Frank, F.C. Proc. Supercooling of liquids Roy. Soc. London **A215**, 43 (1952).

[25] Ravishanker, N. & Chen, R. An introduction to persistent homology for time series. WIREs Computational Statistics **13**, 3, e1548 (2021).

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Data have been simulated by ourselves. We do not provide the code of the method, as it is the combination of standard tools that are available through standard packages (Gudhi and scikit-learn).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We have axed the presentation on the methodology, which was fast. If we would have taken into consideration the generation of the data, it would explode the running time, whereas it is by itself a research project of high interest.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We are the creators of the data.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]