



**HAL**  
open science

# Unsupervised topological learning approach of crystal nucleation

Sébastien Becker, Emilie Devijver, Rémi Molinier, Noël Jakse

► **To cite this version:**

Sébastien Becker, Emilie Devijver, Rémi Molinier, Noël Jakse. Unsupervised topological learning approach of crystal nucleation. *Scientific Reports*, 2022, 10.1038/s41598-022-06963-5 . hal-03806284

**HAL Id: hal-03806284**

**<https://hal.science/hal-03806284v1>**

Submitted on 7 Oct 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Unsupervised topological learning approach of crystal nucleation

Sébastien Becker,<sup>1,2</sup> Emilie Devijver,<sup>2</sup> Rémi Molinier,<sup>3</sup> and Noël Jakse<sup>1</sup>

<sup>1</sup>*Université Grenoble Alpes, CNRS, Grenoble INP, SIMaP*

*F-38000 Grenoble, France*

<sup>2</sup>*Université Grenoble Alpes, CNRS, Grenoble INP, LIG*

*F-38000 Grenoble, France*

<sup>3</sup>*Université Grenoble Alpes, CNRS, IF*

*F-38000 Grenoble, France*

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 unravelled. Crystal nucleation, the early stages where the liquid-to-solid transition occurs upon undercooling, initiates at the atomic level on nanometre 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 monatomic metals, it shows that both translational and orientational ordering always come into play simultaneously when homogeneous nucleation starts in regions with low five-fold symmetry. It also reveals the specificity of the nucleation pathways depending on the element considered, with features beyond the hypothesis of Classical Nucleation Theory.

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, 2]. 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 [3] 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 [4, 5]. To reach statistically meaningful events, large scale simulations are required. This still remains challenging as only few studies are providing now million-to billion-atom simulations for monatomic metals [2].

To identify translational and orientational orderings during homogeneous nucleation in MD simulations, an unsupervised learning approach [6] based on topological data analysis (TDA) signatures was developed through persistent homology (PH) [7, 8]. PH is an intrinsically

flexible, yet highly informative, tool which detects meaningful topological features deduced from atomic configurations. It was successfully applied very recently to characterise structural environments in metallic glasses [9], ice [10] and complex molecular liquids [11]. 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. For the latter a model-based method is used, namely Gaussian Mixture Models (GMM) [12, Chapter 14] (already used with success to analyse MD simulations [13]) and its estimation by an Expectation Maximization (EM) algorithm [14]. The number of clusters [15] is selected by Integrated Criterion Likelihood (ICL, [16]), a refinement for clustering of Bayesian Integrated Likelihood (BIC, [17]). 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 the crystal nucleation.

With this unsupervised approach, the homogeneous nucleation process was studied in three monatomic metals chosen for the variety of their underlying crystalline phase, namely body-centered cubic (bcc) for Ta, face centred-cubic (fcc) for Al, and hexagonal-closed packed (hcp) for Mg. Large-scale molecular dynamics simulations [18] comprising one and ten million atoms were performed with a similar procedure used in our preceding work on pure Zr [4] and described in more details in Methods Section. Figure 1 depicts the methodology applied here to Ta. A rapid quenching at constant pressure brings the liquid from  $T = 3300$  K down to  $T = 1900$  K close to the time-temperature-transformation (TTT) nose. Crystal nucleation is observed along an isothermal process during which a configuration of the simulation is chosen for the clustering. 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 [20], a training set of 5 000 non overlapping local spherical structures within a cutoff radius of  $6.8 \text{ \AA}$  was sampled for the unsupervised learning (see Supplementary Information), 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 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 atoms of type  $C_1$  and  $C_2$ , as they



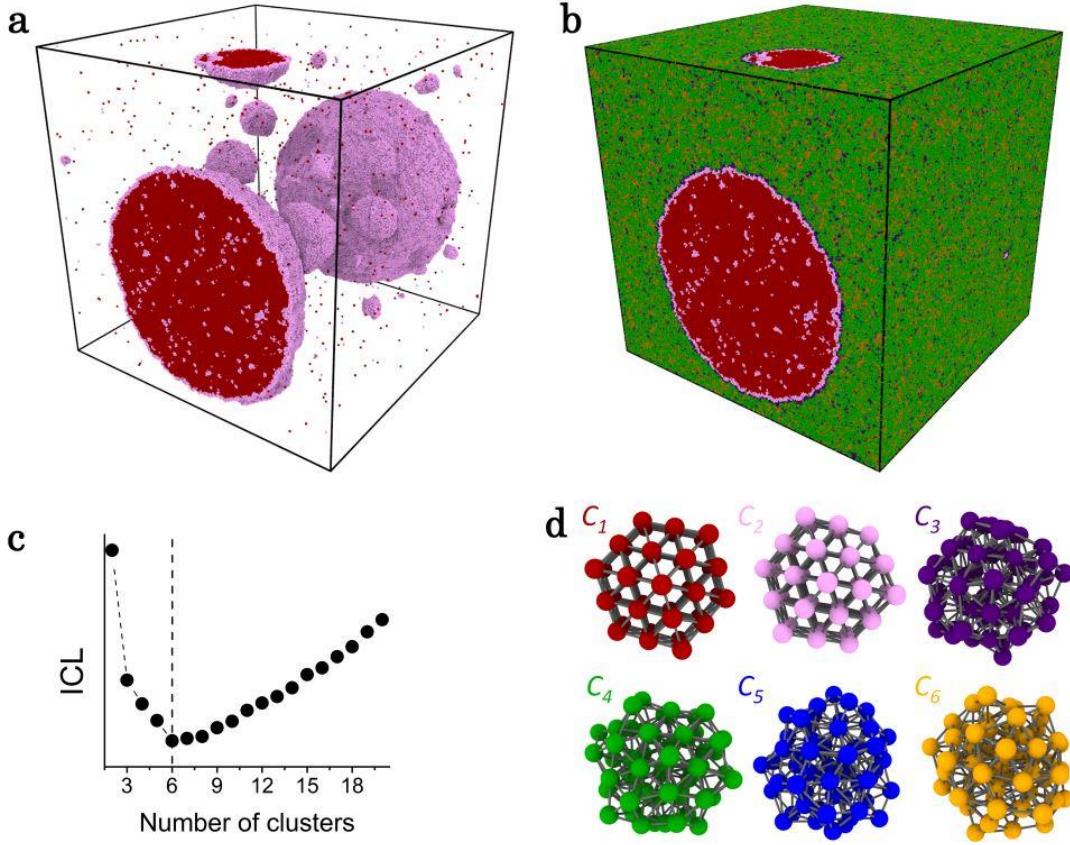


FIG. 1. **Unsupervised learning of homogeneous nucleation.** Snapshot of a ten-million atom MD simulation of Ta during nucleation along the  $T = 1900$  K isotherm (a and b). Independent local atomic structures within a cut-off-radius of  $6.8 \text{ \AA}$  form a train set represented in the descriptor 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 autonomously 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 (see text), while in (b) atoms of all clusters are displayed showing that those in cluster  $C_3$  are located mainly at the border of the nuclei and  $C_4$ ,  $C_5$  and  $C_6$  correspond to the surrounding liquid with various topological characteristics.

show clearly a crystalline order, refraining at this stage from characterising it. 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 to one of the six clusters (the one with the highest probability). Such a clustering training is performed independently for each metal 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.

Figure 2 shows typical homogeneous nucleation events in undercooled Ta and Al during an isothermal process close to the nose of the TTT, which can be done by standard MD simulations without the need of an accelerated methods such as the Forward-flux sampling method [21]. The liquids above the melting point  $T_M$  were first quenched down at ambient pressure to the glass transition sufficiently rapidly to avoid nucleation (see Table S1 in Supplementary Information). 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(c) and (d). An isotherm slightly above the TTT nose is chosen for the analysis, *i.e.*  $T = 1900$  K for Ta and  $T = 650$  K for Al. From chosen configurations during the nucleation and growth process, the clustering is obtained from application of the corresponding trained model as described above. For all metals considered here, strongly growing fraction of mainly two clusters, concomitant to the sharp drop of the energy, is observed. For Ta and Al, only local structures belonging to these clusters are drawn in Figures 2(a) and (b), revealing evidently the nuclei and their evolution in time, recalling that solely the topological vector are describing the local structure. The nuclei morphologies show globular shapes that are rather spherical, characteristic of high  $\Delta T$ , although obviously not strictly as revealed more quantitatively from a convex hull analysis. Interestingly, atoms from one of the two clusters (coloured in red) are mainly located inside the nuclei while atoms from the second one (coloured in pink) steadily remain essentially at the border upon growing. They stay finally at grain boundaries after full solidification of the simulation boxes. Its appearance inside the nuclei reveals also the presence of defaults, as it will be examined below.

The simulations of homogeneous nucleation shown in Fig. 2 were performed with 10 and 1 million atoms for Ta and Al, respectively. In both cases, the vast majority of the embryos dissolve back to the liquid while those attaining the critical size are rare and grow. The larger

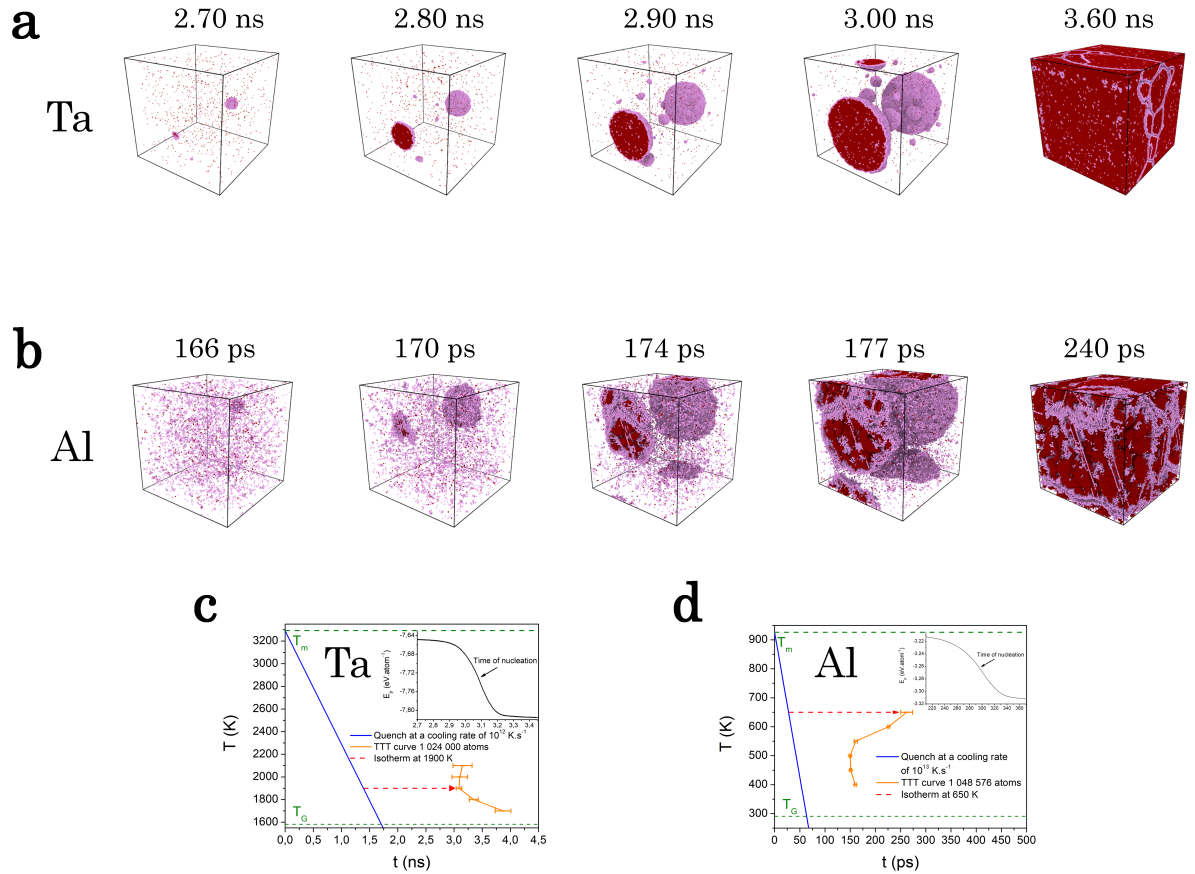


FIG. 2. **Homogeneous nucleation in Ta and Al undercooled liquids.** Snapshots of the molecular dynamics simulations for Ta (a) and Al (b) with respectively 10 and 1 million atoms, during isothermal nucleation at different times for temperatures close to the nose of the Time-Temperature Transformation (TTT) for Ta (c) and Al (d). From stored configurations during fast cooling (blue curves), nucleation events along several isotherms were observed by monitoring the sharp drop of the internal energy (insets in (c) and (d)). The average nucleation times  $\tau_N$  (symbols) were determined from 5 independent simulations for each temperature giving the TTT curves in the vicinity of the nose (orange lines).

simulation box for Ta allows to follow the nucleation process for a longer time, sufficient to observe more secondary nucleation events [22]. Direct estimation of the critical size is still unreachable by experiment, as nuclei can be detected only at larger size [3]. This is

also scarcely studied by MD simulation as it is not easy to define their boundary from the surrounding liquid [23, 24], especially in the case of non-spherical or ramified shape [25]. Here, the size distribution of nuclei was obtained by counting the number of atoms in overlapping structures identified as red and pink clusters within the cut-off radius. An estimation of the critical size was inferred from the nuclei's size that persists between the first and second configurations shown in Fig. 2, at least without losing atoms they contained initially. As it can be seen for Ta on Fig. 1(d), the local structures of the two clusters forming the nuclei are unambiguously crystalline (with only a slight distortion for structures from cluster  $C_2$ ) giving a clear definition of them. This is repeated in the subsequent consecutive pairs of configurations to refine statistics, and the results for all metals are gathered in Table S1 in Supplementary Information. For Ta, embryos with size less than 120 atoms always dissolve back to the liquid while the few nuclei found with size larger than 150 atoms always grew. Similar values of the critical radius were determined very recently for bcc Fe and fcc Cu [26] and fcc Zn [27] in similar high  $\Delta T$  regime. For Al and Mg the simulations were performed at lower  $\Delta T$  yielding obviously larger critical nuclei which are consistent with the Lennard-Jones case [5, 23] and also with Al but somewhat lower with respect to recent MD simulations [28].

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 centre 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  $\Delta r = 1 \text{ \AA}$ . Considering the nucleation process of Ta as an illustration, 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 centre 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 centre of the nucleus, while atoms of  $C_2$  being mainly located at its border, as can be easily confirmed visually. 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

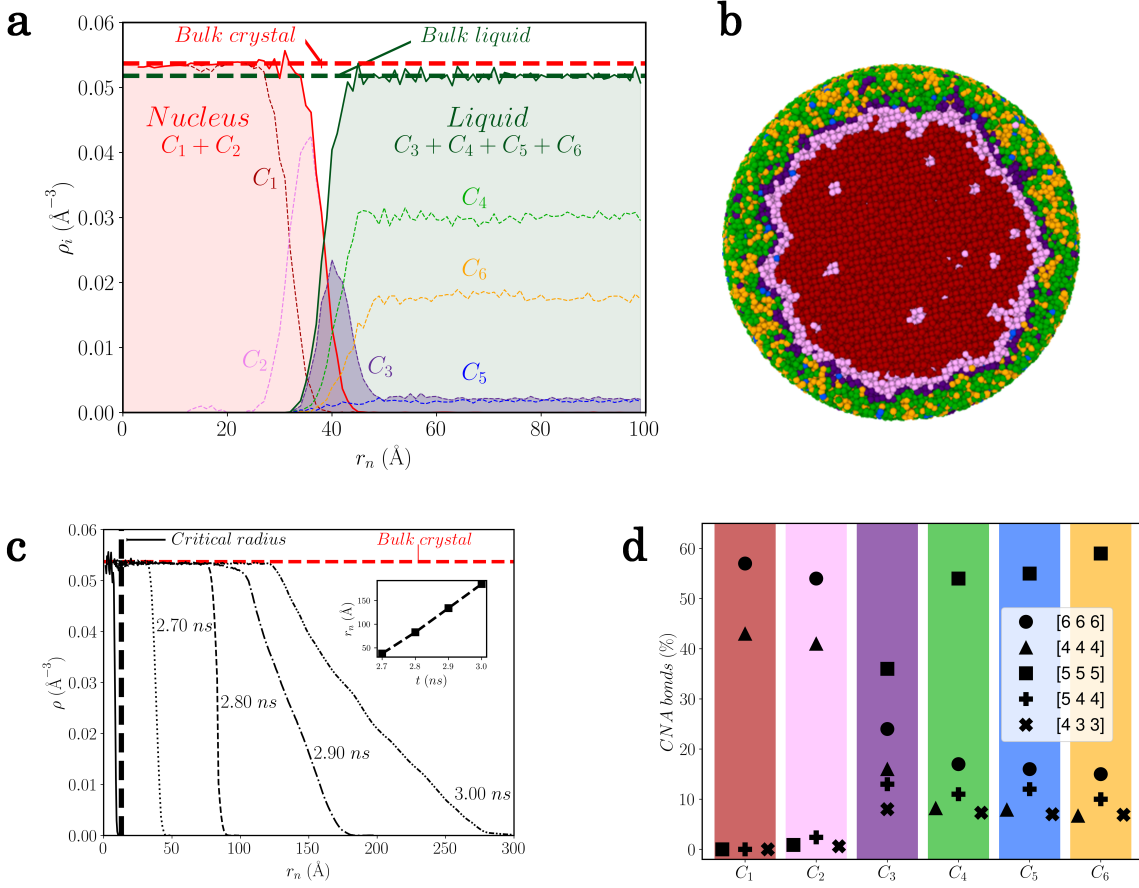


FIG. 3. **Translational and bond-orientational order parameters for Ta.** (a) Radial density profile of the largest nucleus during the growth at 2.7 ns along the  $T = 1900$  K isotherm. The red and blue dashed horizontal lines correspond respectively to the average bulk crystalline density and average bulk undercooled liquid without nucleation events, both being simulated at  $T = 1900$  K at ambient pressure (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 at times corresponding to Fig. 2 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 [29] for each cluster of the model.

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 evolution with time is shown in the inset, displaying a linear behaviour in agreement with CNT [2]. 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 local structure of type  $C_1$  or  $C_2$  corresponding to the minimal size of about 65 atomic structures identified by the TDA-GMM given the chosen cutoff radius (see Supplementary Information). This feature appears to be general as similar results are found for Al and Mg as shown in the Supplementary Information.

The BOO of each cluster is identified through the Common Neighbour Analysis (CNA) [29], chosen as a well-known and robust tool. The CNA signature [30] 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. 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 [31] and was interpreted as compatible with the A15 crystalline phase. This analysis in terms of CNA highlights and confirms that the TDA-GMM unsupervised learning approach is a powerful method to capture the structural picture in its finest details.

The peculiar spatial distribution of structure of type  $C_3$  shown in Fig. 3(a) deserves further attention. Firstly, its location at the boundaries of the nuclei is consistent with the mixed bcc and FFS orderings. This is however seen as an effect of the TDA-GMM procedure that picks up structures covering a part of the nucleus and of the liquid in the configuration used for the training. More interestingly, its presence in the whole simulation box indicates that

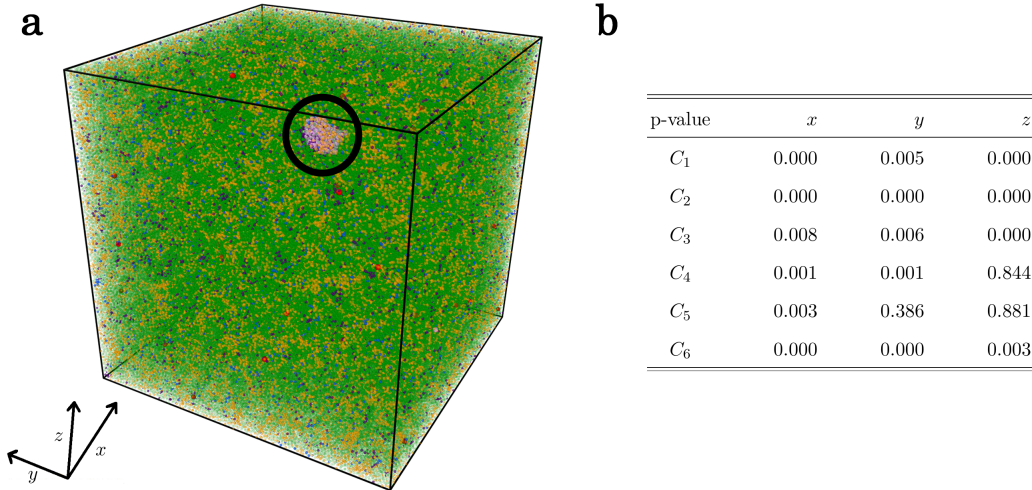


FIG. 4. **Early nucleation stage for Ta.** (a) Snapshot of the simulation with 1 million atoms along isotherm  $T = 1900K$  showing the onset of nucleation when the first nucleus starts to grow (highlighted by the black circle). Atoms have been coloured according to the cluster they belong to (see Fig. 1(d)). (b) p-values computed on the projection of atomic positions on each direction of the box from a Kolmogorov-Smirnov test against the uniform distribution.

in the undercooled liquid, some regions with higher bcc ordering might develop apart from the vicinity of the growing nuclei. Fig. 4(a) shows a snapshot of the simulation at the onset of nucleation when the first nucleus starts to grow, all atoms being coloured according to the cluster they belong to. Fig. 4(b) depicts a table of the concurrent p-values obtained, for each cluster, on the projection of atomic positions on the 3 directions of space, from a Kolmogorov-Smirnov test [32] against the uniform distribution. For a level 0.01, the test is always rejected in at least one direction, which proves that the distribution of the clusters in the box is not uniform, i.e. their heterogeneity. Focusing on atoms of type  $C_4$  (green) and  $C_6$  (yellow), which represent more than 90 % of the atoms at this stage (58% for  $C_4$  and 34% for  $C_6$ ), it clearly shows that the undercooled liquid embodies structural heterogeneities with varying degree of FFS. Moreover, higher bcc ordering characterized by structures of type  $C_3$  appears in localized regions of lower FFS (green) from which, in most of the cases, embryos formed by structures from clusters  $C_1$  and  $C_2$  emerge. The same conclusion of structural heterogeneity is obtained for Mg and Al, with particularly low p-values for Mg



(see Supplementary Information).

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 centred on the hard sphere and associated colloidal systems [33, 34]. For Ta, 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 at  $T = 1900$  K, a feature that also holds for the other metals investigated here. This provides evidence for 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 [35] for metallic systems. This view is consistent with the fact that, unlike hard spheres, metallic systems with strong bonding are more energy driven rather than entropy driven systems.

All these features allow us to propose a nucleation pathway for the metals considered here. For Ta, our findings show a single step process with an onset of homogeneous nucleation taking place in low FFS domains of the heterogeneous liquid, where emerging bcc embryos have simultaneously the density of the bulk solid. After reaching the critical size, the nuclei grows in a rather globular shape with a bcc structure with a small amount of defects and a diffuse interface with decreasing bcc ordering. During the growth the surrounding liquid keeps the bulk liquid density. A similar one step nucleation pathway also holds for Al in which embryos emerge from the low FFS regions directly with the fcc bond ordering. The growing nuclei have here a more patchy morphology and a significant amount of fcc stacking faults. For Mg, a two steps process is identified as can be seen in the Supplementary Information: an onset of nucleation showing embryos having mainly a bcc ordering followed by growth of nuclei with a mixed fcc/hcp structure and some bcc ordering at the surface of the nuclei. In this case, the scenario is more akin to the Lennard-Jones case [5, 23] following the Landau Theory in which the bcc precursor is favoured in the early stages of crystal nucleation [36] as well as the Ostwald step rule [37] for which the primary crystal phase nucleating from the liquid is not necessarily the thermodynamic stable one.

The present unsupervised learning approach was shown to be a powerful tool to unravel the atomic scale mechanisms of crystal nucleation in monatomic metals. It allowed us to



reveal general aspects in the homogeneous nucleation process as well as specificities depending on the metallic element under consideration. Our results are in line with the emerging idea that heterogeneities which exist in the undercooled liquid [34] play the foremost role in the onset of nucleation. For all metals, nucleation have been found to start in low FFS regions, which is consistent with Frank’s argument [38], 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 [2]. However, our analysis reveals also some aspects beyond the CNT, such as nuclei having a diffuse interface with the surrounding liquid and metals possessing their own nucleation pathways, involving *e.g.* for Mg a two step mechanism [37]. The complexity and richness found here for metals and in other systems [23, 33, 34] underline the future challenges in stepping forward in our theoretical understanding beyond the CNT. This promising methodology more generally opens the door to a deeper and autonomous investigation of atomic level mechanisms in materials science.

## METHODS

**Simulation method.** Molecular dynamics simulations were performed with the LAMMPS code [18] in a fully periodic situation. Verlet’s algorithm in the velocity form for the numerical integration of the phase space trajectory was used with a time step of 2 fs for Ta with a number of atoms  $N = 10^7$  ( $10^6$  for the training) and 1 fs for Al and Mg with  $N = 10^6$ . Interatomic interaction were taken in the Embedded Atom Model form and chosen for their ability to reproduce the liquid and solid properties as described in the Supplementary Information. Control of the thermodynamic conditions was done with the Nosé-Hoover thermostat and barostat [39] was used to maintain the ambient pressure whatever the temperature. The time-temperature transformation curves were first built for each metal following the procedure established recently [4]. Along an isotherm located slightly above the TTT nose, 6 configurations of interest were selected for the purpose of monitoring the crystal nucleation process. Before analysing the configurations, minimization of the energy

by means of a conjugate gradient algorithm has been performed to bring the system in a local minimum of the potential energy surface to suppress the thermal noise [20].

**Persistent homological descriptors’ space (TDA).** The unsupervised learning in the MD configurations is performed in terms of the local atomic environment of each atom (called the local structure) within a cut-off radius defined as the second minimum of the pair-correlation function  $g(r)$  in the liquid, as described in Fig. S1 of Supplementary Information. The use of two atomic neighbour shells to represent the local environment was shown to optimize the local structural information of descriptors at the expense of a loss of the spatial resolution [7]. In Persistent Homology [7, 8], components of homological dimensions  $H_0$ ,  $H_1$  and  $H_2$  are then used in the form of a topological vector of dimension  $n_{PH}$  to represent each local structure. Its components are calculated from the Persistent Diagrams (PD) representing the birth and death characteristics of each topological component, as shown in Fig. S2 of the Supplementary Information. More precisely, for each pair of points  $(x, y)$  in a PD,  $D$ , the values of the topological vector components are calculated, except for the infinite point, for a fixed level of homology [8] by

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

where  $d_\Delta(\cdot)$  denotes the  $\ell^\infty$  distance to the diagonal. The number of  $H_0$  is fixed by the number of neighbour atoms and the number of components of  $H_1$  and  $H_2$  is inferred from a subsampling approach as described in [41] to remove the noise.

**Clustering using a Gaussian mixture Model (GMM).** In order to build a training set for the learning, a sampling of 5 000 to 7 000 structures, that covers the entire simulation box by means of their central particles at least separated by two times a cut-off radius are extracted from a million atoms configuration chosen during the nucleation. From the build topological descriptors’ space as described above, a mixture of  $M$  Gaussian distributions  $(\phi(\cdot; \boldsymbol{\mu}_m, \Sigma_m))_{1 \leq m \leq M}$  of weights  $(\alpha_m)_{1 \leq m \leq M}$  as

$$\sum_{m=1}^M \alpha_m \phi(\cdot; \boldsymbol{\mu}_m, \Sigma_m), \quad (2)$$

where  $\boldsymbol{\mu}_m$  is the position of the mean and  $\Sigma_m$  the covariance matrix of the  $m$ th Gaussian distribution. The number of Gaussian components is set using the ICL criterion [16] and

full covariance matrices with 3 000 K-means initializations are used to construct a model for applications on configurations along the nucleation process.

## ACKNOWLEDGMENTS

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 Architected 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 CNRS 3584 (TherMatHT) are also acknowledged.

- 
- [1] Kelton, K. F. & Greer, A. L. *Nucleation in Condensed Matter: Applications in Materials and Biology* (Pergamon, 2010).
  - [2] Sosso, G. C. et al. Crystal Nucleation in Liquids: Open Questions and Future Challenges in Molecular Dynamics Simulations. *Chem. Rev.* 116, 7078–7116 (2016).
  - [3] Zhou, J. *et al.* Observing crystal nucleation in four dimensions using atomic electron tomography. *Nature* 570, 500–503 (2019).
  - [4] Auer, S. & Frenkel, D. Prediction of absolute crystal-nucleation rate in hard-sphere colloids. *Nature* 409, 1020–1023 (2001).
  - [5] 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).
  - [6] Ceriotti, M. Unsupervised machine learning in atomistic simulations, between predictions and understanding. *J. Chem. Phys.* 150, 150901 (2019).
  - [7] Motta, F. C. *Topological Data Analysis: Developments and Applications in Adv. Nonlinear Geosci.*, 369–391 (Tsonis A. A. ed., Springer International Publishing AG 2018).

- [8] Carrière, M., Oudot, S. Y. & Ovsjanikov, M. Stable topological signatures for points on 3D shapes. *Eurographics Symp. Geom. Process.* 34, 1–12 (2015).
- [9] 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).
- [10] Hong, S. & Kim, D. Medium-range order in amorphous ices revealed by persistent homology. *J. Phys. Condens. Matter* **31**, (2019).
- [11] 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).
- [12] Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc. (2001).
- [13] Boattini, E. et al. Autonomously revealing hidden local structures in supercooled liquids. *Nat. Commun.* **11**, 1–9 (2020).
- [14] 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).
- [15] The word 'cluster' is used for groups detected by the machine learning method throughout the text.
- [16] 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.
- [17] Schwarz, G., Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464 (1978).
- [18] S. J. Plimpton, *J. Comp. Phys.* **117**, 1 (1995); <http://www.lammps.sandia.gov>.
- [19] Becker, S., Devijver, E., Molinier, R. & Jakse, N. Glass-forming ability of elemental zirconium. *Phys. Rev. B* 102, 104205 (2020).
- [20] Stillinger, F. H. & T. A. Weber, T. A. Hidden structure in liquids. *Phys. Rev. A* **25**, 978 (1982).
- [21] R. J. Allen, C. Valeriani, & P. Rein ten Wolde, *J. Phys.: Condens. Matter.* **21**, 463102 (2009).
- [22] 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).

- [23] Ten Wolde, P. R., Ruiz-Montero, M. J. & Frenkel, D. Numerical calculation of the rate of crystal nucleation in a Lennard-Jones system at moderate undercooling. *J. Chem. Phys.* 104, 9932–9947 (1996).
- [24] Báez, L. A. & Clancy, P. The kinetics of crystal growth and dissolution from the melt in Lennard-Jones systems. *J. Chem. Phys.* 102, 8138–8148 (1995).
- [25] Toxvaerd, S. The role of local bond-order at crystallization in a simple supercooled liquid. *Eur. Phys. J. B* 93, 1–8 (2020).
- [26] Louzguine-Luzgin, D. V. & Bazlov, A. I. Crystallization of fcc and bcc liquid metals studied by molecular dynamics simulation. *Metals (Basel)*. 10, 1–11 (2020).
- [27] Zhou, L. L. et al. Crystallization characteristics in supercooled liquid zinc during isothermal relaxation: A molecular dynamics simulation study. *Sci. Rep.* 6, 31653 (2016).
- [28] Mahata, A., Zaeem, M. A. & Baskes, M. I. Understanding homogeneous nucleation in solidification of aluminum by molecular dynamics simulations. *Model. Simul. Mater. Sci. Eng.* **26**, (2018).
- [29] 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).
- [30] 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).
- [31] Jakse, N., Le Bacq, O. & Pasturel, A. Prediction of the local structure of liquid and supercooled tantalum. *Phys. Rev. B* **70**, 174203 (2004).
- [32] Kolmogoroff, A. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell’Istituto Italiano degli Attuari* **4**, 83-91 (1933).
- [33] 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).
- [34] 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).
- [35] Russo, J. & Tanaka, H. The microscopic pathway to crystallization in supercooled liquids. *Sci. Rep.* 2, 505 (2012).

- [36] Alexander, S. & McTague, J. P. Should all crystals be bcc? Landau theory of solidification and crystal nucleation. *Phys Rev Lett* **41**, 702–705 (1978).
- [37] Ostwald, W. The formation and changes of solids (Translated from German). *Z. Phys. Chem.* **22**, 289–330 (1897).
- [38] Frank, F.C. Proc. Supercooling of liquids *Roy. Soc. London* **A215**, 43 (1952).
- [39] Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids: Second Edition.* (Oxford University Press, 2017).
- [40] Lechner, W. & Dellago, C. Accurate determination of crystal structures based on averaged local bond order parameters. *The Journal of Chemical Physics* **129**, 114707 (2008).
- [41] Fasy, B. T. et al. Confidence sets for persistence diagrams. *Ann. Statist.* **42**, (2014).

## Supplementary Information File

Additional information are presented here to support data and figures of the main text.

### SUPPLEMENTARY INFORMATIONS ON THE METHODOLOGY

#### Molecular dynamics simulations

Table SI presents some characteristic properties related to each system described by potentials that are used (Ta [1], Al [2] and Mg [3]) in the present classical molecular dynamics simulations. Namely: the melting temperature  $T_m$ ; the glass transition temperature  $T_g$  which were extracted from the quenching of the liquid at ambient pressure with a cooling rate  $Q$  up to the amorphous state; the isotherm  $T_{\text{iso}}$  along which the nucleation process was studied; the ratios  $T_{rg} = T_g/T_m$  and  $\Delta T = (T_m - T)/T_m$ ; the estimated critical size of the nuclei  $n_c$ ; and the critical cooling rate  $Q_c$ , up to which the crystallization can be avoided, inferred from the nose of the TTT curves.

The procedure to compute the TTT curves follows the one from our previous work on Zr [4].

	$T_m$ (K)	$T_g$ (K)	$T_{\text{iso}}$ (K)	$T_{rg}$	$\Delta T$	$n_c$	$Q$ (K/s)	$Q_c$ (K/s)
Ta	3290 <sup>a</sup>	1582	1900	0.48	0.42	140-150	$10^{12}$	$4.2 \times 10^{11}$
Al	926 <sup>b</sup>	291	650	0.31	0.30	400-800	$10^{13}$	$1.8 \times 10^{12}$
Mg	918 <sup>c</sup>	303	600	0.33	0.35	310-350	$10^{12}$	$3.2 \times 10^{11}$

TABLE SI. Characteristic features of the classical molecular dynamics potentials as described in the text. Melting temperatures  $T_m$  are taken from Refs. <sup>a</sup>[1]; <sup>b</sup>[2]; <sup>c</sup>[3]

### Definition of the local structures

Figure S5 shows the pair-correlation function  $g(r)$  of the undercooled liquid and crystalline states of Ta at  $T = 1900$  K with the respective mean structure assigned to the clusters  $C_4$  (preponderant liquid) and  $C_1$  (preponderant bcc). A cut-off radius of  $6.8 \text{ \AA}$  was set to capture topological informations with the help of the Python package `gudhi` [5] and `riprser.py` [6] up to the second neighbour shell.

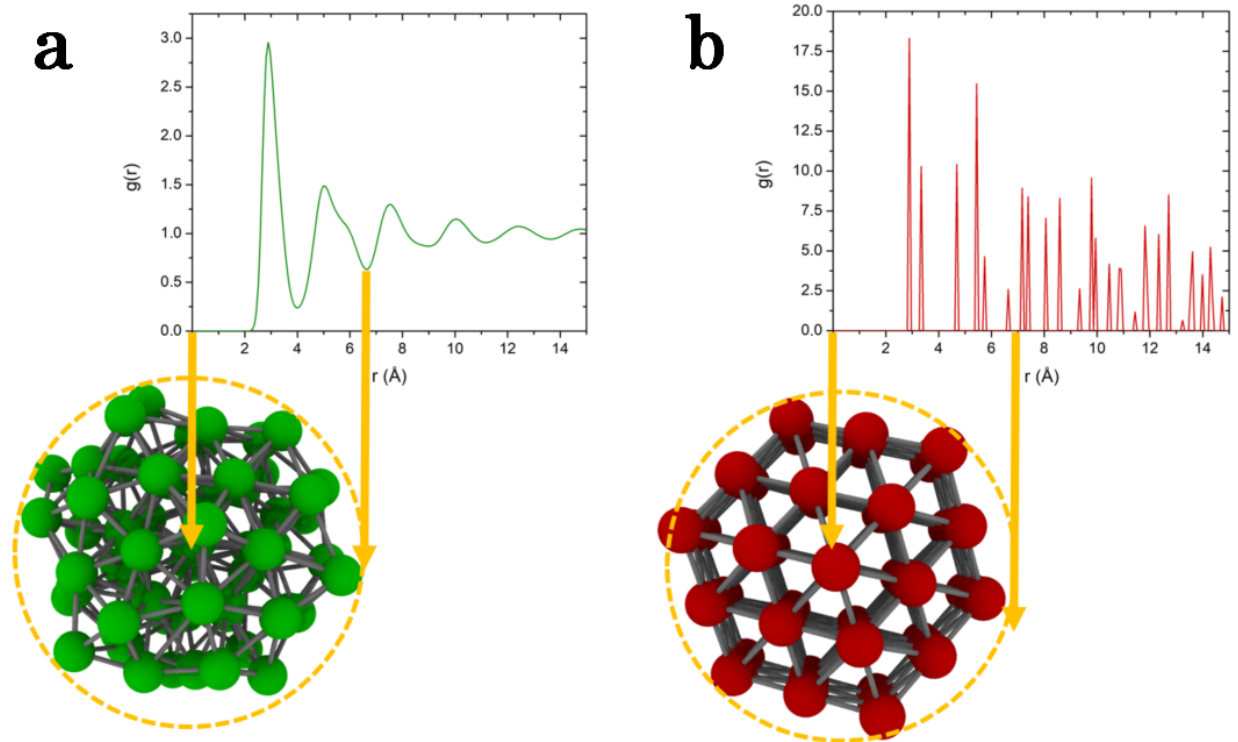


FIG. S5. Cut-off radius for the clustering based on the  $g(r)$  function with the second minimum leading to structures with two neighbors shells with the one associated in Ta to the preponderant liquid in green (a) and the one to bcc ordering in red (b).

In the context of classical descriptors like the averaged bond-orientational order analysis it was shown [7] that information from the second neighbour shell increase the accuracy in the discrimination of local structures, but at the expense of a loss in the spatial resolution. This is also observed in the topological descriptors set up here with give rise to more  $H_0$  and  $H_1$  components as well  $H_2$  components which appear only when considering more than just one neighbour shell. It should be pointed our that when increasing further the cut-off radius up to the third neighbour shell and beyond, the benefit gained in topological information is counterbalanced by a too large spatial extension leading to a loss of resolution in the Gaussian Mixture Model (GMM) clustering. This compromise between the accuracy and the spatial resolution bring us the optimal choice of the second neighbour shell to define local structures consistently with earlier findings [7].



The local atomic structures were extracted with Python package `pyscal` [8]. For comparison, the persistent homological information are depicted on the persistence diagram shown in Figure S6 for the two mean local structures assigned to  $C_4$  and  $C_1$ . The differences can be seen here between a disordered liquid structure and a perfect periodic lattice where all the pairs (birth, death) are concurrent for each homological dimensions.

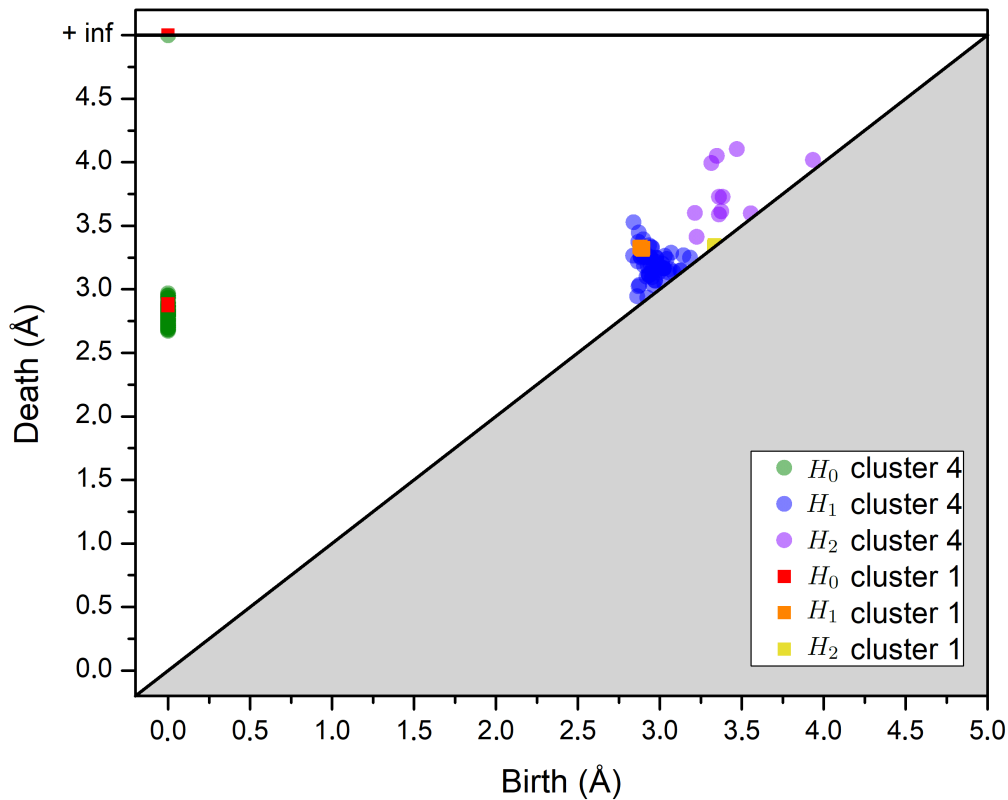


FIG. S6. Persistence diagrams for the mean structures of  $C_1$  and  $C_4$  which represent respectively the bcc ordering and the preponderant liquid structure.

## Clustering using a Gaussian mixture Model (GMM)

Figures S7 and S8 shows the clustering with the TDA-GMM method applied respectively to Al and Mg configuration during nucleation. The resulting Local atomic structures assigned to each cluster are shown. The number of clusters is determined using the ICL criterion. The clustering is performed with Python package `scikit-learn` [9].

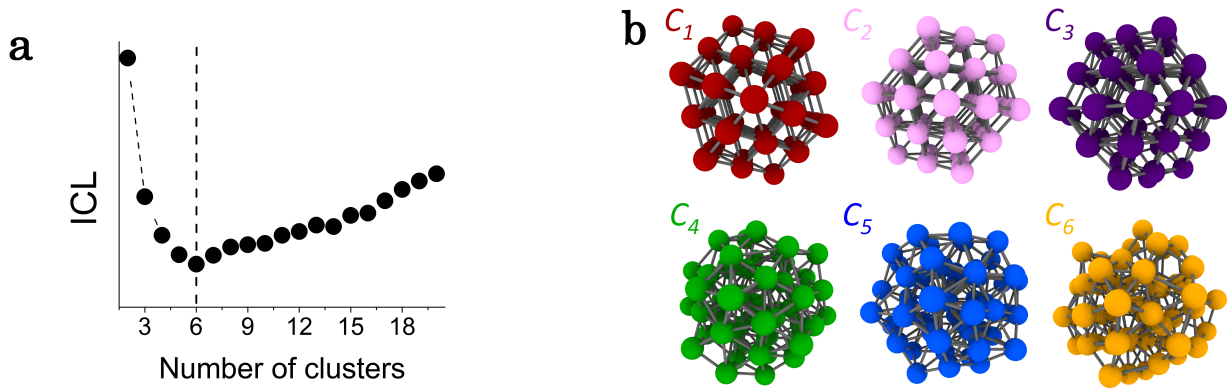


FIG. S7. TDA-GMM clustering of Al. (a) Evolution of the Integrated Completed Likelihood (ICL) criterion as a function of number of clusters. (b) Independent local atomic structures within a cut-off-radius of  $6.3 \text{ \AA}$  form a train set represented in the descriptor space by 173 PH components up to the second order.

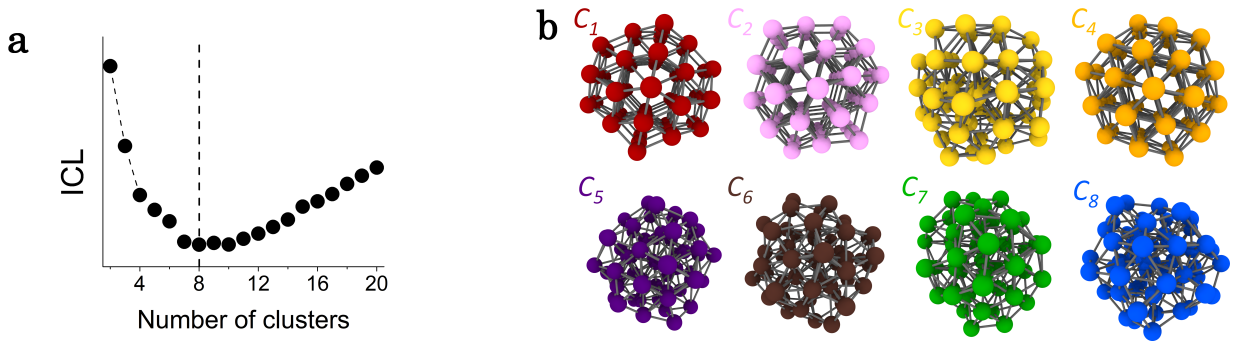


FIG. S8. TDA-GMM clustering of Mg. (a) Evolution of the Integrated Completed Likelihood (ICL) criterion as a function of number of clusters. (b) Independent local atomic structures within a cut-off-radius of  $6.9 \text{ \AA}$  form a train set represented in the descriptor space by 199 PH components up to the second order.

## CRYSTAL NUCLEATION

### Identification and extraction of the structures during nucleation

Tables SII, SIII and SIV show respectively the evolution of proportion of the structures assigned to each clusters previously identified by the TDA-GMM method in Ta, Al and Mg. As can be seen in each case, clusters  $C_1$  and  $C_2$  follow a fast growth of their proportion along the nucleation process until they are the majority in the final bulk. Referring to Figure 1 in the main text and the previous Figures S7 and S8, these two clusters are indeed represented by local atomic structures on a pure and distorted crystalline structures. In the case of Mg, one can notice that clusters  $C_3$  and  $C_5$  are also growing along  $C_1$  and  $C_2$ . This is explained by the fact that part of the structures assigned to these two clusters shares bonds with the growing nuclei and carry partial crystalline structure owing to the use a cut-off corresponding to the second neighbour shell.

Time (ns)	2.70	2.80	2.90	2.96 <sup>(M)</sup>	3.00	3.60 <sup>(S)</sup>
$C_1$ (%)	0.09	1.03	5.18	13.44	15.84	83.52
$C_2$ (%)	0.06	0.33	1.25	4.81	3.27	12.25
$C_3$ (%)	4.13	4.10	4.12	5.61	4.18	2.92
$C_4$ (%)	58.11	57.10	53.72	45.56	45.88	0.85
$C_5$ (%)	3.39	3.38	3.25	3.05	2.84	0.20
$C_6$ (%)	34.23	34.06	32.48	27.53	28.00	0.27

TABLE II. Proportion of each cluster for Ta at different times during the nucleation process. Superscripts (M) and (S) correspond respectively to the configuration used to train the TDA-GMM model and the solidified configuration.

Time (ps)	166	170	174	175 <sup>(M)</sup>	177	240 <sup>(S)</sup>
$C_1$ (%)	0.05	0.54	3.42	4.24	10.07	36.55
$C_2$ (%)	0.73	2.29	7.14	9.10	14.34	22.43
$C_3$ (%)	19.78	19.42	19.74	19.7	20.29	23.57
$C_4$ (%)	33.22	32.46	29.32	28.85	23.76	9.95
$C_5$ (%)	44.38	43.53	38.94	36.92	30.60	7.48
$C_6$ (%)	1.84	1.76	1.43	1.19	0.95	0.02

TABLE III. Proportion of each cluster for Al at different times of the nucleation process. Superscripts (M) and (S) correspond respectively to the configuration used to train the TDA-GMM model and the solidified configuration.

Time (ps)	940	960	980	990 <sup>(M)</sup>	1000	1500 <sup>(S)</sup>
$C_1$ (%)	0.01	0.13	0.74	3.82	4.60	20.59
$C_2$ (%)	0.11	0.34	1.32	3.64	5.28	14.58
$C_3$ (%)	0.31	1.13	3.00	3.98	6.39	12.25
$C_4$ (%)	0.20	1.37	4.38	3.64	2.38	1.13
$C_5$ (%)	4.14	5.03	6.68	8.56	10.56	18.30
$C_6$ (%)	23.91	23.15	21.59	20.78	20.68	19.32
$C_7$ (%)	36.64	35.30	32.12	28.56	25.30	5.16
$C_8$ (%)	34.68	33.56	30.17	27.01	24.81	8.68

TABLE SIV. Proportion of each cluster for Mg at different times of the nucleation process. Superscripts (M) and (S) correspond respectively to the configuration used to train the TDA-GMM model and the solidified configuration.

Figure S9 shows the evolution in Mg of the central particles assigned to  $C_1$  and  $C_2$  through the nucleation process along with the TTT curve.

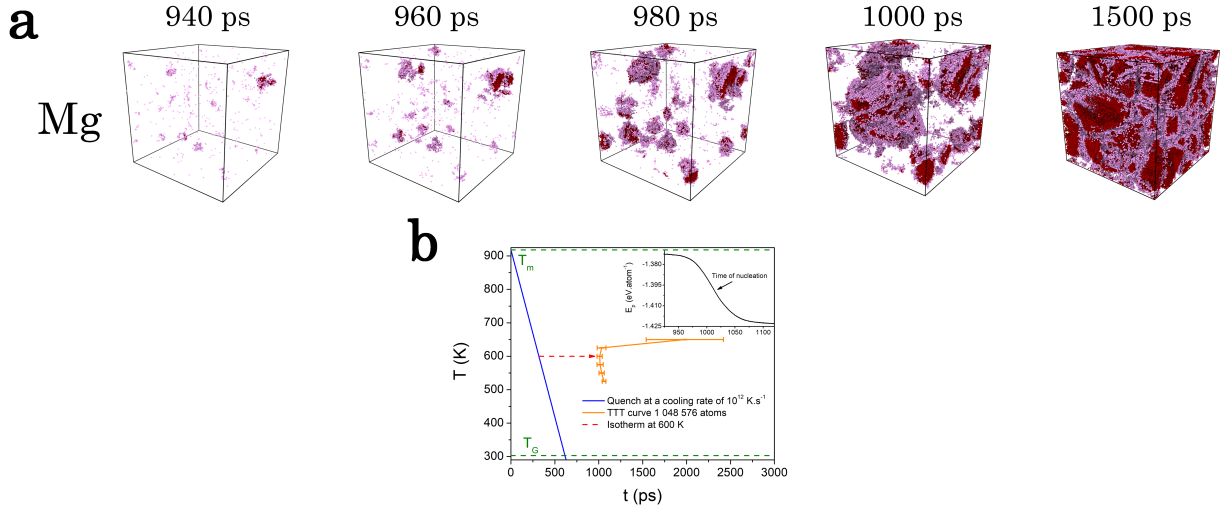


FIG. S9. Homogeneous nucleation events in undercooled Mg during an isothermal process (a) at the nose of the TTT curve (b).

## Translational and orientational orderings

Following the procedure described in the main text, a general behaviour for the translational and orientational orderings of Al and Mg is depicted on the Figures S10 and S11. All the nuclei are driven by a concurrent emergence of this two symmetries which correspond respectively to the density of the crystal bulk and the geometrical bonds related to the crystalline local structure.

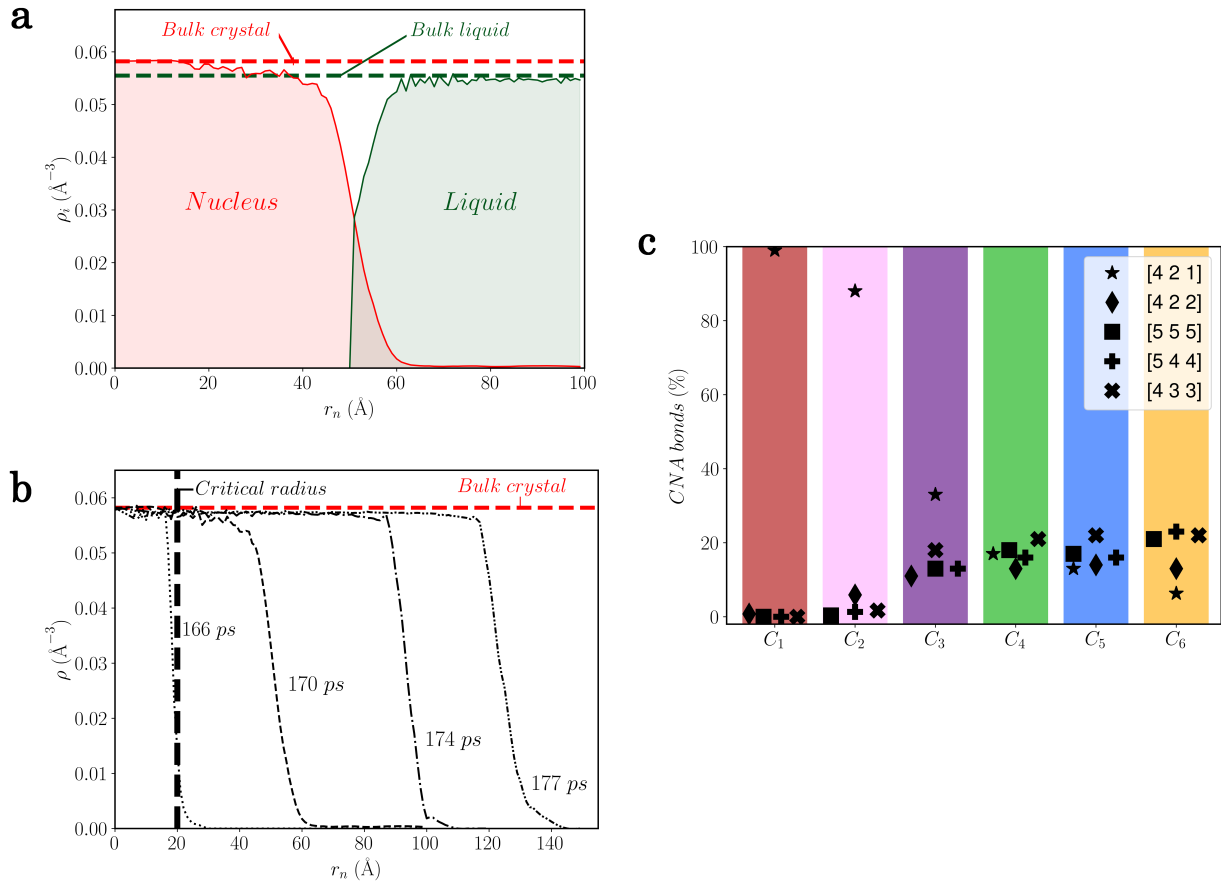


FIG. S10. Typical translational (a) and bond-orientational (c) order parameters for Al. An analysis of the density profile at various times of the biggest growing nuclei (b) shows that the translational order is concurrent with the orientational order at the onset of nucleation.

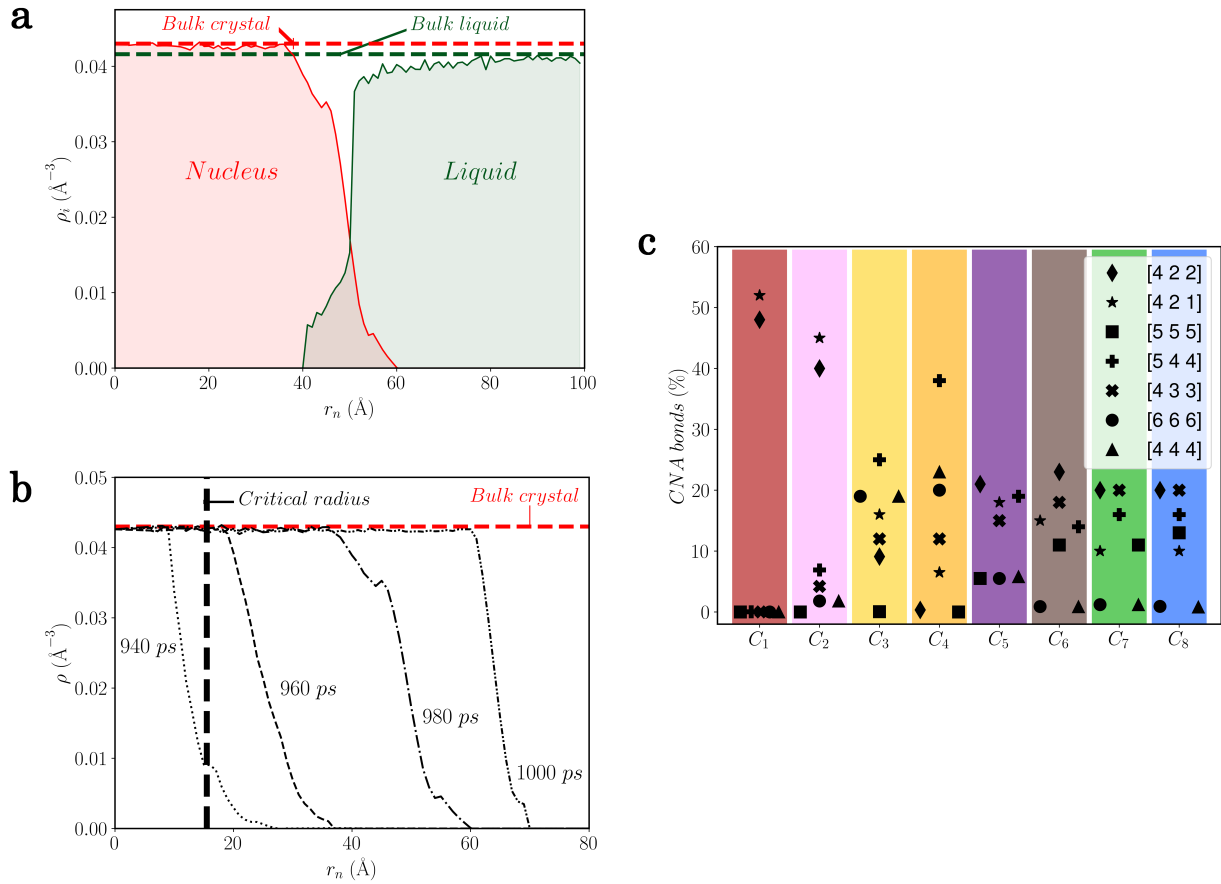


FIG. S11. Typical translational (a) and bond-orientational (c) order parameters for Mg. An analysis of the density profile for various times of the biggest growing nuclei (b) shows that the translational order is concurrent with the orientational order at the onset of nucleation.

- 
- [1] Zhong, L., Wang, J., Sheng, H., Zhang, Z. & Mao, S. X. Formation of monatomic metallic glasses through ultrafast liquid quenching. *Nature* 512, 177–180 (2014).
- [2] Mendelev, M. I., Kramer, M. J., Becker, C. A. & Asta, M. Analysis of semi-empirical interatomic potentials appropriate for simulation of crystalline and liquid Al and Cu. *Philosophical Magazine* 88, 1723–1750 (2008).

- [3] Wilson, S. R. & Mendeleev, M. I. A unified relation for the solid-liquid interface free energy of pure FCC, BCC, and HCP metals. *J. Chem. Phys.* 144, 144707 (2016).
- [4] Becker, S., Devijver, E., Molinier, R. & Jakse, N. Glass-forming ability of elemental zirconium. *Phys. Rev. B* 102, 104205 (2020).
- [5] Maria, C., Boissonnat, J.-D., Glisse, M. & Yvinec, M. The Gudhi Library: Simplicial Complexes and Persistent Homology. in *Mathematical Software – [Research Report] RR-8548, INRIA*, (2014).
- [6] Tralie, C., Saul, N. & Bar-On, R. Ripser.py: A Lean Persistent Homology Library for Python. *JOSS* 3, 925 (2018).
- [7] Lechner, W. & Dellago, C. Accurate determination of crystal structures based on averaged local bond order parameters. *The Journal of Chemical Physics* 129, 114707 (2008).
- [8] Menon, S., Leines, G. & Rogal, J. pysical: A python module for structural analysis of atomic environments. *JOSS* 4, 1824 (2019).
- [9] Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *JMLR* 12, 2825 (2011).
- [10] Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO—the Open Visualization Tool. *Modelling Simul. Mater. Sci. Eng.* 18, 015012 (2010).