



OPEN

Unsupervised topological learning approach of crystal nucleation

Sébastien Becker^{1,2}, Emilie Devijver², Rémi Molinier³ & Noël Jakse¹✉

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 as a result of the strong bonding 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 Tomography³ 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 metals².

To identify translational and orientational orderings during homogeneous nucleation in MD simulations, an unsupervised learning approach⁶ 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 glasses⁹, ice¹⁰ and complex molecular liquids¹¹. 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, (already used with success to analyse MD simulations¹³) and its estimation by an Expectation Maximization (EM) algorithm¹⁴. The number of clusters (The word ‘cluster’ is used for groups detected by the machine learning method throughout the text.) is selected by Integrated Completed Likelihood (ICL¹⁵), a refinement for clustering of Bayesian Integrated Likelihood (BIC¹⁶). 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.

Results and discussion

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 simulations¹⁷ comprising one and ten million atoms were performed with a similar procedure used in our preceding work on pure Zr¹⁸ 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

¹Université Grenoble Alpes, CNRS, Grenoble INP, SIMaP, 38000 Grenoble, France. ²Université Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France. ³Université Grenoble Alpes, CNRS, IF, 38000 Grenoble, France. ✉email: noel.jakse@grenoble-inp.fr

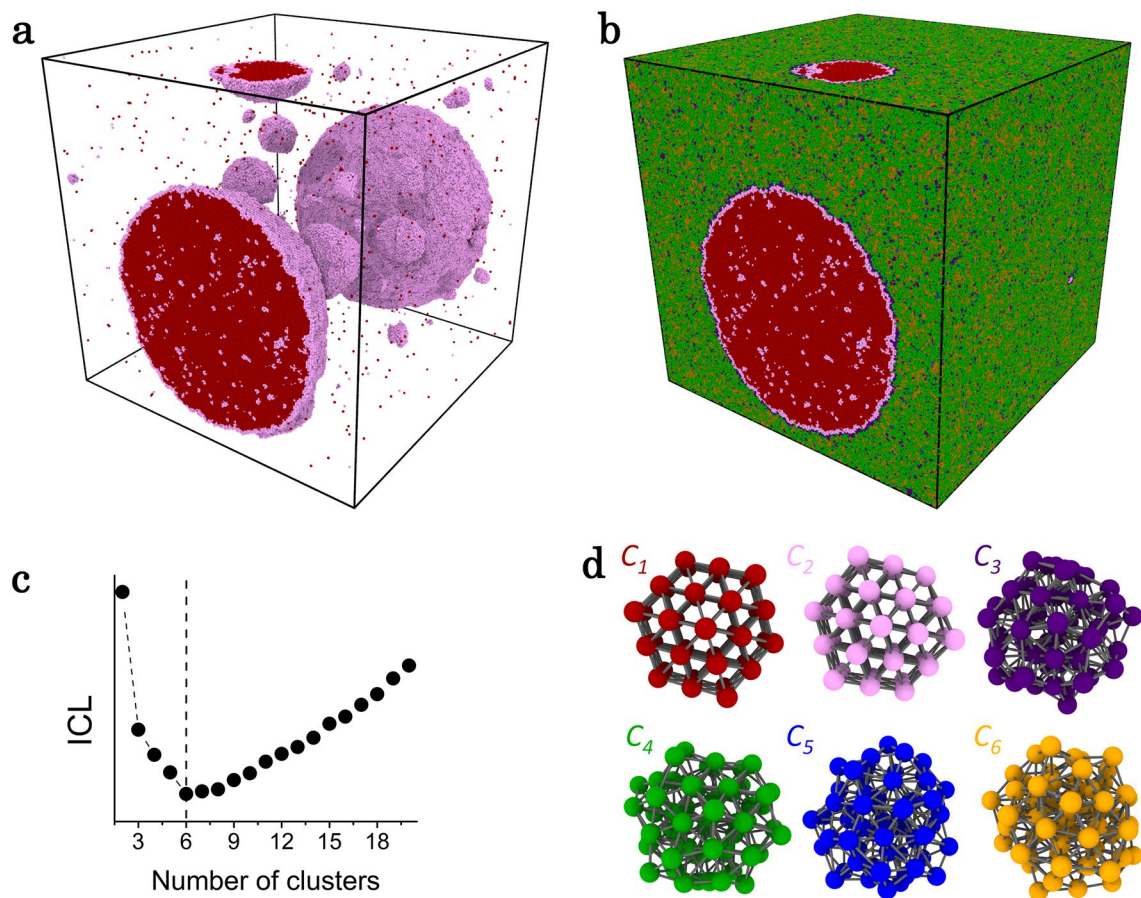


Figure 1. Unsupervised learning of homogeneous nucleation. Snapshot of a 10-million atom MD simulation of Ta during nucleation along the $T = 1900$ K isotherm (**a,b**). Independent local atomic structures within a cut-off-radius of 6.8 \AA form a train set represented in the descriptor space by 215 PH components up to dimension 2. (**c**) Evolution of the ICL criterion as a function of the 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.

sizes and a substantial amount of liquid, it is considered as representative of the phenomenon. From its inherent structure¹⁹, a training set of 5 000 non overlapping local spherical structures within a cutoff radius of 6.8 \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 Fig. 1d was automatically inferred to be representative of the system based on the minimum ICL criterion Fig. 1c. The snapshot of the simulation box in Fig. 1a displays only atoms of type C_1 and C_2 , as they 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. 1b. 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 method²⁰. 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 Fig. 2c,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 Fig. 2a,b, revealing evidently the nuclei and their evolution in time, recalling that solely the topological vector is describing the local structure. The nuclei morphologies show

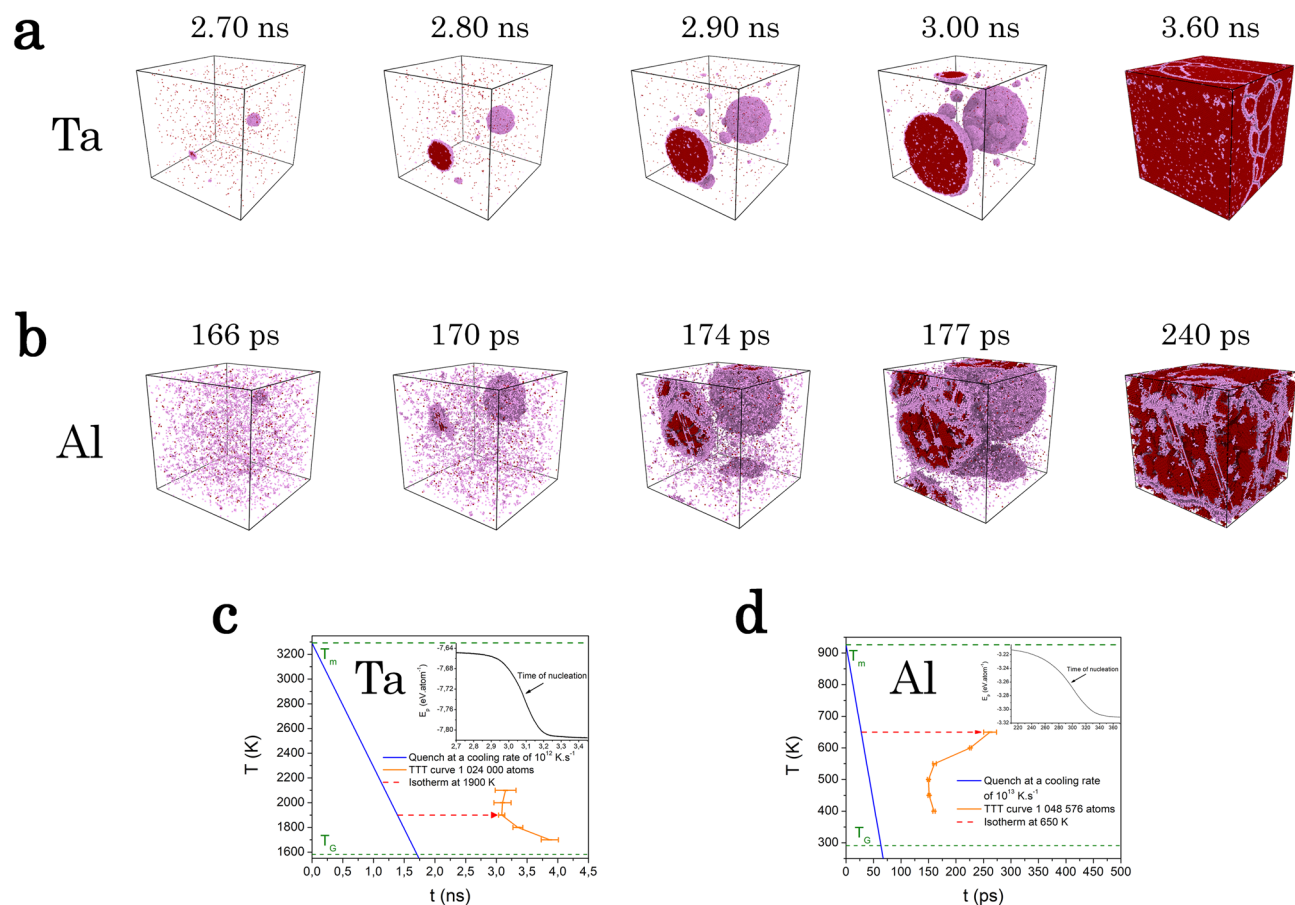


Figure 2. Homogeneous nucleation in Ta and Al undercooled liquids. Snapshots of the molecular dynamics simulations for Ta (**a**) and Al (**b**) with respectively 10 and 1 million atoms, during isothermal nucleation at different times at temperatures close to the nose of the Time-Temperature Transformation (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,d**). 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).

globular shapes that are rather spherical, characteristic of high ΔT , 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 simulation box for Ta allows to follow the nucleation process for a longer time, sufficient to observe more secondary nucleation events²¹. Direct estimation of the critical size is still unreachable by experiment, as nuclei can be detected only at larger size³. This is also scarcely studied by MD simulation as it is not easy to define their boundary from the surrounding liquid^{22,23}, especially in the case of non-spherical or ramified shape²⁴. 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 size of the nuclei that persist 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. 1d, 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 Cu²⁵ and fcc Zn²⁶ in similar high ΔT regime. For Al and Mg the simulations were performed at lower ΔT yielding obviously larger critical nuclei which are consistent with the Lennard–Jones case^{5,22} and also with Al but somewhat lower with respect to recent MD simulations²⁷.

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

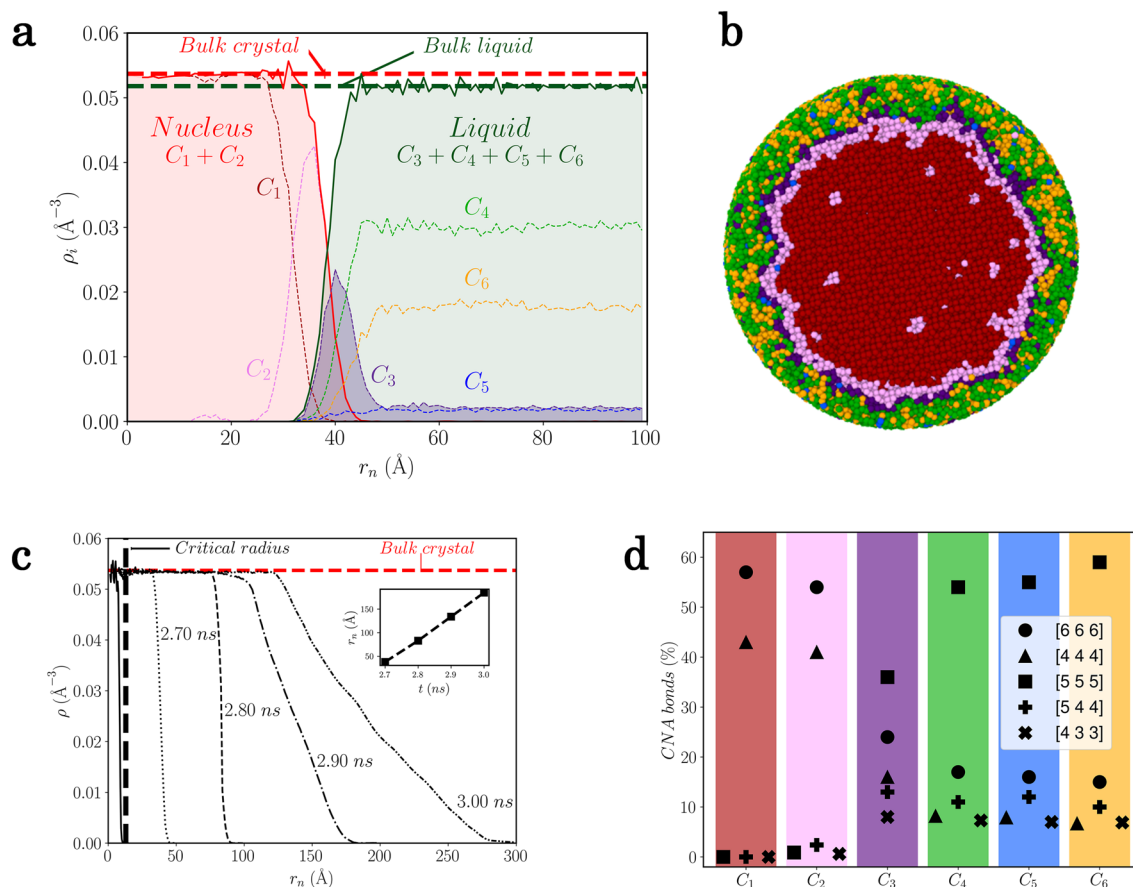


Figure 3. Translational and bond-orientational order parameters for Ta. (a) Radial density profile of the largest nucleus during the growth at 2.7 ns along the $T = 1900$ K isotherm. The red and blue dashed horizontal lines correspond respectively to the average bulk crystalline density and average bulk 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. 1d). (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 Analysis²⁸ for each cluster of the model.

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 $\Delta r = 1$ \AA . Considering the nucleation process of Ta as an illustration, Fig. 3a depicts the density profiles $\rho_i(r)$ for all 6 clusters for the largest nucleus shown in Fig. 2a and its surrounding liquid at time 2.7 ns. The corresponding slice of the nucleus through its centre is drawn in Fig. 3b. 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 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.

Figure 3c 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 CNT². 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)²⁸, chosen as a well-known and robust tool. The CNA signature³⁰ given in Fig. 3d 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

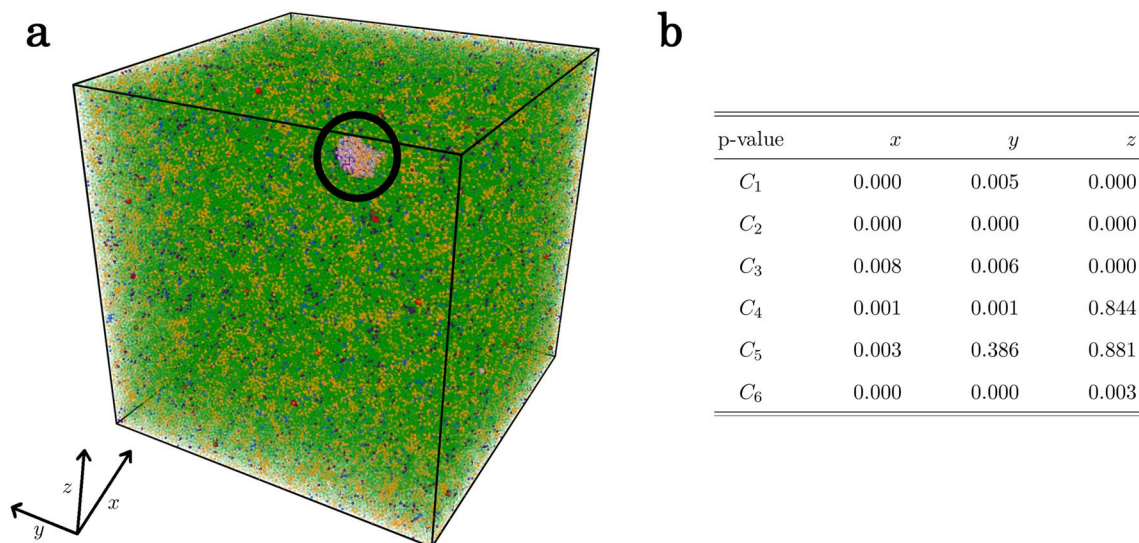


Figure 4. Early nucleation stage for Ta. **(a)** Snapshot of the simulation with 1 million atoms along isotherm $T = 1900\text{K}$ showing the onset of nucleation when the first nucleus starts to grow (highlighted by the black circle). Atoms have been coloured according to the cluster they belong to (see Fig. 1d). **(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.

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. As shown in the Supplementary Information file, the use of Steinhardt parameters²⁹ leads to a similar result. Such a BOO of the four clusters associated to the liquid agrees well with *ab initio* molecular dynamics simulations³¹ 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. 3a deserves further attention. Firstly, its location at the boundaries of the nuclei could suggest the existence of a new type of ordering like in nucleation of ice³². However, here it is consistent with the mixed bcc and FFS orderings. This is then 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. As a matter of fact, increasing the number of clusters in the GMM reveal several clusters with various crystalline and FFS ordering as shown in Supplementary Information file using a Principal Component Analysis (PCA)¹². More interestingly, its presence in the whole simulation box indicates that in the undercooled liquid, some regions with higher bcc ordering might develop apart from the vicinity of the growing nuclei. Figure 4a 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. Figure 4b 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 test³³ 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^{2,34} and was debated during the last decade with a controversy essentially centred on the hard sphere and associated colloidal systems^{35–37}. 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 $T = 1900\text{K}$, 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^{35,38} for pure metallic systems. For more complex metallic alloys like Al–Ni³⁴, it was shown, using Steinhardt parameters, that nucleation is initiated by orientational ordering, followed by density ordering. However, our results are consistent with the fact that, at least for pure metallic systems with strong bonding, unlike hard spheres, are more energy driven rather than entropy driven systems.

All these features allow us to propose a nucleation pathway for the metals considered here, guided by their electronic structure characteristics and underlying stable crystalline structure. 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 grow in a rather globular shape with a bcc structure, a small amount of defects, and a diffuse interface with decreasing bcc ordering. During the growth the surrounding liquid keeps the bulk liquid density. This single-step nucleation might be explained by the strong stability of the bcc structure of Ta coming from its half filled 5d-band structure with one of the highest cohesive energy⁴⁷. 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. The difference in the nucleation pathways might be surprising as the fcc and ideal hcp structures displayed by Al and Mg have identical first and second neighbors atoms^{47,48}. In describing these sp-valent metals within the second-order perturbation theory, it was shown^{49,50} that their structural energy difference come from more distant neighbors and is small, and even smaller for Mg. Nevertheless, the relative stability with respect to the bcc structure is ~ 3 times larger for Al than for Mg^{47,48}. This might explain why the onset of nucleation is a one-step nucleation process due to the higher fcc stability, while it is a two-step process for Mg, starting with in the first step with embryos having locally preferred high symmetry bcc structure, due to the much smaller structural difference with respect to the hcp. The presence of defective hcp structures in fcc Al nuclei and fcc local structure in hcp Mg nuclei may result from the very small hcp-fcc energy difference. For Mg, the scenario is more akin to the Lennard–Jones case^{5,22} following the Landau Theory in which the bcc precursor is favoured in the early stages of crystal nucleation³⁹ as well as the Ostwald step rule⁴⁰ for which the primary crystal phase nucleating from the liquid is not necessarily the thermodynamic stable one.

Conclusion

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 liquid³⁷ 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 argument⁴¹ and also observed for Al–Ni alloys³⁴. For pure metals, translational and orientational ordering taking place simultaneously in emerging embryos. Given the fact that metals considered here have different crystalline ordering, this feature is induced by the bonding character of interactions. 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 theory². 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 mechanism⁴⁰. This may trigger further theoretical developments, and for instance in the diffuse interface approach². The complexity and richness found here for metals and in other systems^{22,36,37} 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 code¹⁷ 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 barostat⁴² 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 recently¹⁸. The procedure to compute the TTT curves follows the one from our previous work on Zr¹⁸. Along an isotherm located slightly above the TTT nose, 6 configurations of interest were selected for the purpose of monitoring the crystal nucleation process. The unsupervised method is applied to the inherent structures¹⁹ of the chosen configurations. The latter are obtained from a minimization of the energy by means of a conjugate gradient algorithm to bring the system in a local minimum of the potential energy surface. In this manner, the thermal noise is removed and allows us to reveal the genuine structural features. This is an important aspect for the unsupervised method used here, either for the construction of the topological descriptor or for the clustering. It avoids identifying several clusters with similar structural characteristics in the same local minimum of the PEL. With the system size considered here neither partial crystallisation nor change in the nucleation process were observed by applying the minimization.

Definition of the local structures. Figure 5 shows schematically 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 Å was set to capture topological informations with the help of the Python package `gudhi`⁴³ and `ripser.py`⁴⁴ up to the second neighbour shell. The local atomic structures were extracted with Python package `pyscal`⁴⁵.

In the context of classical descriptors like the averaged bond-orientational order analysis it was shown²⁹ 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

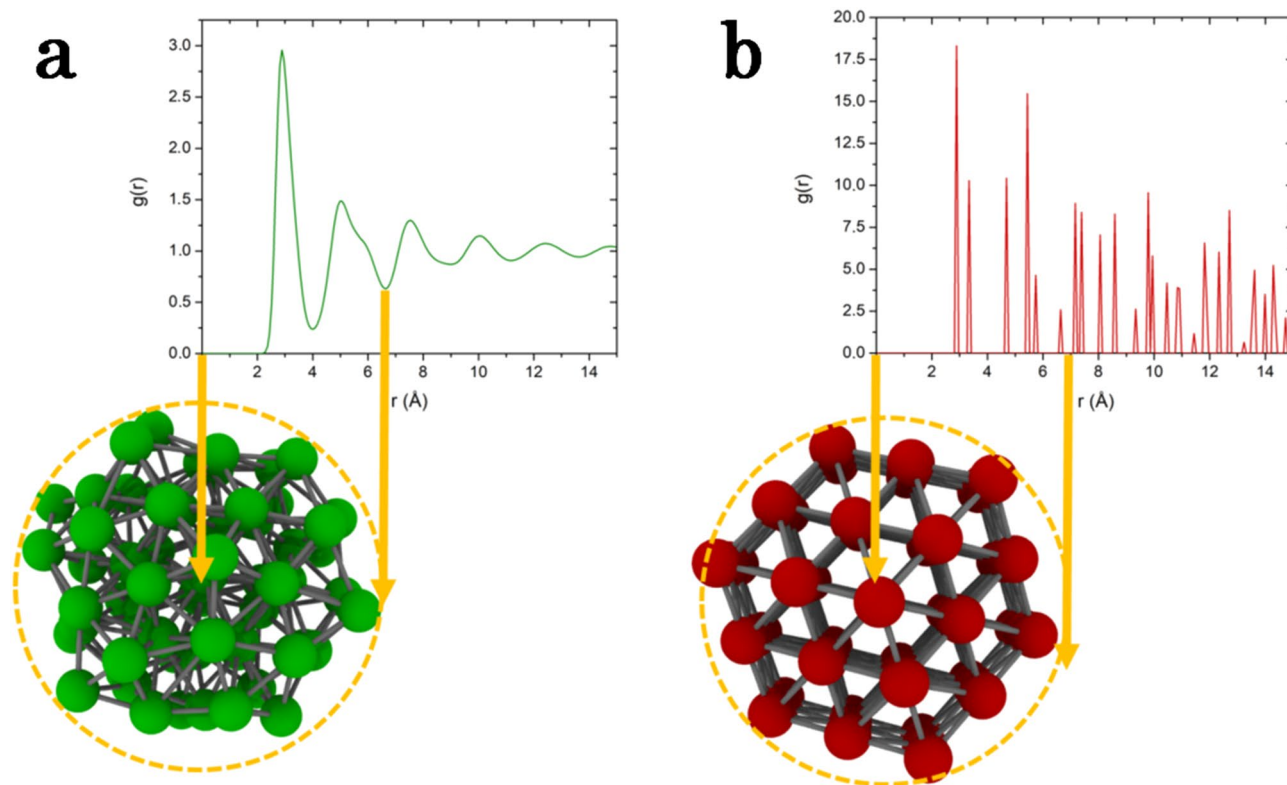


Figure 5. Cut-off radius for the clustering based on the $g(r)$ function with the second minimum leading to structures with two neighbors shells with the one associated in Ta to the preponderant liquid in green (a) and the one to bcc ordering in red (b).

which give rise to more H_0 and H_1 components as well as H_2 components which appear only when considering more than just one neighbour shell. It should be pointed out 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 findings²⁹.

Persistent homological descriptors' space (TDA). The unsupervised learning in the MD configurations is performed in terms of the local atomic environment (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 resolution²⁹. 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. 6. 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 homology⁸ by

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

where $d_\Delta(\cdot)$ denotes the ℓ^∞ 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 in⁴⁶ to remove the noise.

Clustering using a Gaussian mixture model (GMM). In order to build a training set for the learning, a sampling of 5000–7000 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 built topological descriptors' space as described above, a mixture of M Gaussian distributions $(\phi(\cdot; \mu_m, \Sigma_m))_{1 \leq m \leq M}$ of weights $(\alpha_m)_{1 \leq m \leq M}$ is written as

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

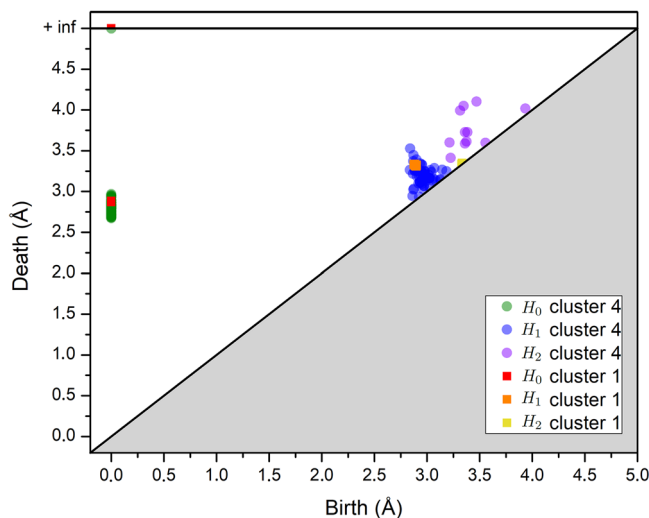


Figure 6. Persistence diagrams of homological dimensions H_0 , H_1 and H_2 for the mean structures of C_1 and C_4 which represent respectively the bcc ordering and the preponderant liquid structure.

where μ_m is the position of the mean and Σ_m the covariance matrix of the m th Gaussian distribution. The number of Gaussian components is set using the ICL criterion¹⁵ and full covariance matrices with 3 000 K-means initializations are used to construct a model for applications on configurations along the nucleation process. Doubling manually the number of clusters to be identified, even if it is not optimal in the sense of the ICL, leads to the same evolution of the CNA of the clusters as in Fig. 3d with additional clusters having intermediate orderings displaying a “more continuous” evolution as shown in the Supplementary Information file.

Received: 15 November 2021; Accepted: 27 January 2022

Published online: 24 February 2022

References

- Kelton, K. F. & Greer, A. L. *Nucleation in Condensed Matter: Applications in Materials and Biology* (Pergamon, 2010).
- Sosso, G. C. *et al.* Crystal nucleation in liquids: Open questions and future challenges in molecular dynamics simulations. *Chem. Rev.* **116**, 7078–7116 (2016).
- Zhou, J. *et al.* Observing crystal nucleation in four dimensions using atomic electron tomography. *Nature* **570**, 500–503 (2019).
- Auer, S. & Frenkel, D. Prediction of absolute crystal-nucleation rate in hard-sphere colloids. *Nature* **409**, 1020–1023 (2001).
- 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).
- Cerioti, M. Unsupervised machine learning in atomistic simulations, between predictions and understanding. *J. Chem. Phys.* **150**, 150901 (2019).
- Motta, F. C. Topological data analysis: Developments and applications. In *Advances in Nonlinear Geosciences* (ed. Tsonis, A. A.) 369–391 (Springer, 2018).
- Carrière, M., Oudot, S. Y. & Ovsjanikov, M. Stable topological signatures for points on 3D shapes. *Eurograph. Symp. Geom. Process.* **34**, 1–12 (2015).
- 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).
- Hong, S. & Kim, D. Medium-range order in amorphous ices revealed by persistent homology. *J. Phys.: Condens. Matter* **31**, 25 (2019).
- Sasaki, K., Okajima, R. & Yamashita, T. Liquid structures characterized by a combination of the persistent homology analysis and molecular dynamics simulation. *AIP Conf. Proc.* **20**, 020015 (2018).
- Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, 2001).
- Boattini, E. *et al.* Autonomously revealing hidden local structures in supercooled liquids. *Nat. Commun.* **11**, 1–9 (2020).
- Dempster, A., Laird, N. & Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **39**(1), 1–38 (1977).
- Biernacki, C., Celeux, G. & Govaert, G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(7), 719–725. <https://doi.org/10.1109/34.865189> (2000).
- Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
- Plimpton, S. J. *J. Comp. Phys.* **117**, 1 (1995). <http://www.lammps.sandia.gov>
- Becker, S., Devijver, E., Molinier, R. & Jakse, N. Glass-forming ability of elemental zirconium. *Phys. Rev. B* **102**, 104205 (2020).
- Stillinger, F. H. & Weber, T. A. Hidden structure in liquids. *Phys. Rev. A* **25**, 978 (1982).
- Allen, R. J., Valeriani, C. & Rein ten Wolde, P. Forward flux sampling for rare event simulations. *J. Phys. Condens. Matter* **21**, 463102 (2009).
- 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).
- 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).
- 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).

24. Toxvaerd, S. The role of local bond-order at crystallization in a simple supercooled liquid. *Eur. Phys. J. B* **93**, 1–8 (2020).
25. 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).
26. Zhou, L. L. *et al.* Crystallization characteristics in supercooled liquid zinc during isothermal relaxation: A molecular dynamics simulation study. *Sci. Rep.* **6**, 31653 (2016).
27. 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**, 20 (2018).
28. Faken, D. & Jónsson, H. Systematic analysis of local atomic structure combined with 3D computer graphics. *Comput. Mat. Sci.* **2**, 279–286 (1994).
29. Lechner, W. & Dellago, C. Accurate determination of crystal structures based on averaged local bond order parameters. *J. Chem. Phys.* **129**, 114707 (2008).
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. Russo, J., Romano, F. & Tanaka, H. New metastable form of ice and its role in the homogeneous crystallization of water. *Nat. Mater.* **13**, 733–739 (2014).
33. Kolmogoroff, A. Sulla determinazione empirica di una legge di distribuzione. *Giornale Isti. Itali. Degli Attuari* **4**, 83–91 (1933).
34. Hu, Y.-C. & Tanaka, H. Physical origin of glass formation from multicomponent systems. *Sci. Adv.* **6**, 20 (2020).
35. Kawasaki, T. & Tanaka, H. Formation of a crystal nucleus from liquid. *Proc. Natl. Acad. Sci. USA* **107**, 14036–14041 (2010).
36. 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).
37. 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).
38. Russo, J. & Tanaka, H. The microscopic pathway to crystallization in supercooled liquids. *Sci. Rep.* **2**, 505 (2012).
39. Alexander, S. & McTague, J. P. Should all crystals be bcc? Landau theory of solidification and crystal nucleation. *Phys. Rev. Lett.* **41**, 702–705 (1978).
40. Ostwald, W. The formation and changes of solids (Translated from German). *Z. Phys. Chem.* **22**, 289–330 (1897).
41. Frank, F. C. Supercooling of liquids. *Proc. R. Soc. Lond. A* **215**, 43 (1952).
42. Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids* 2nd edn. (Oxford University Press, 2017).
43. 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).
44. Tralie, C., Saul, N. & Bar-On, R. Ripser.py: A lean persistent homology library for Python. *JOSS* **3**, 925 (2018).
45. Menon, S., Leines, G. & Rogal, J. pyscal: A python module for structural analysis of atomic environments. *JOSS* **4**, 1824 (2019).
46. Fasy, B. T. *et al.* Confidence sets for persistence diagrams. *Ann. Stat.* **42**, 20 (2014).
47. Hafner, J. *From Hamiltonians to Phase Diagrams* (Springer, 1987).
48. McMahan, A. K. & Moriarty, J. A. Structural phase stability in third-period simple metal. *Phys. Rev. B* **27**, 3235 (1983).
49. Moriarty, J. A. & McMahan, A. K. High-pressure structural phase transitions in Na, Mg, and Al. *Phys. Rev. Lett.* **48**, 809 (1982).
50. Pettifor, D. G. & Ward, M. A. An analytic pair potential for simple metals. *Solid State Commun.* **49**, 291 (1984).

Acknowledgements

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 Multi-functional 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 networks in high-temperature thermodynamics GDR CNRS 3584 (TherMatHT) and in artificial intelligence in materials science GDR CNRS 2123 (IAMAT) are also acknowledged.

Author contributions

S.B., E.D., R.M. and N.J. conceived the methodology, S.B. perform the calculations, N.J., wrote the manuscript, and all authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-06963-5>.

Correspondence and requests for materials should be addressed to N.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022