Competing nucleation pathways in nanocrystal formation

Carlos Salazar, Akshay Krishna Ammothum Kandy, Jean Furstoss, Jean Furstoss, Quentin Gromoff, Jacek Goniakowski, Julien Lam

Competing nucleation pathways in nanocrystal formation

Carlos R. Salazar,[1,2] Akshay Krishna Ammothum Kandy,[2] Jean Furstoss,[1] Quentin Gromoff,[2] Jacek Goniakowski,[3] and Julien Lam[1,2,*]

[1] *Univ. Lille, CNRS, INRA, ENSCL, UMR 8207, UMET, Unité Matériaux et Transformations, F 59000 Lille, France*
[2] *Centre d'élaboration des Matériaux et d'Etudes Structurales, CNRS (UPR 8011), 29 rue Jeanne Marvig, 31055 Toulouse Cedex 4, France*
[3] *CNRS, Sorbonne Université, Institut des NanoSciences de Paris, UMR 7588, 4 Place Jussieu, F-75005 Paris, France*

Despite numerous efforts from numerical approaches to complement experimental measurements, several fundamental challenges have still hindered one's ability to truly provide an atomistic picture of the nucleation process in nanocrystals. Among them, our study resolves three obstacles: (1) Machine-learning force fields including long-range interactions able to capture the finesse of the underlying atomic interactions, (2) Data-driven characterization of the local ordering in a complex structural landscape associated with several crystal polymorphs and (3) Comparing results from a large range of temperatures using both brute-force and rare-event sampling. Altogether, our simulation strategy has allowed us to study zinc oxide crystallization from nano-droplet melt. Remarkably, our results show that different nucleation pathways compete depending on the investigated degree of supercooling.

I. INTRODUCTION

Polymorphisms occur when the same material can be found in different structural forms. In protein crystals, the competition between each of the possible structures has dramatic consequences causing amyloid diseases [1–3] and toxicity of pharmaceutical compounds [4, 5]. Meanwhile, for technological applications associated with material science, each crystal phase has distinct physical and chemical properties, necessitating the stabilization of a specific polymorphic form. As the triggering mechanisms for the emergence of order, crystal nucleation should have been key for controlling polymorphic selection. However, its study remains extremely challenging because disparate lengths and time scales are simultaneously involved [6, 7]. On the one hand, in terms of size, an extended mother phase along with a small critical cluster made of few tens of atoms must be jointly studied. On the other hand, crystal nucleation combines both a long time scale for stochastic fluctuations to trigger the critical event and a short time scale for crystal growth.

This already complex picture is exacerbated for nanoscale systems. Indeed, the preponderance of surface effects expands the structural landscape of possible polymorphic structures. In addition, a competition between two different natures of nucleation can be found: homogeneous in the core and heterogeneous in the peripheral. While numerous experimental works have provided insights into this challenging nucleation process [8–11], numerical simulations should have been the ideal tool as it provides a dynamic picture at the single-particle level. However, simulating crystal nucleation in nanoparticles requires facing two major challenges. On the one hand, ab initio molecular dynamics simulations are too much computationally demanding to perform the necessary large-scale simulations while classical interaction potentials can not always precisely model both bulk and surface effects. On the other hand, nucleation involves overcoming a free energy barrier and is thus an intrinsically rare-event combining a long induction time (set by the nucleation rate) with a short transition period (set by the growth rate).

Our study focuses on the polymorphic competition in zinc oxide nanoparticles which exhibits promising electrochemical [12–14] and antibacterial [15–18] activities. For all these applications, a key feature of ZnO is its structural complexity associated with the downsizing to the nanoscale [19–23]. Indeed, while the Wurtzite structure is the most stable in bulk, it was found that a competition between different polymorphs exists with body-centered tetragonal structure being more preponderant at sufficiently small sizes. However, body-centered tetragonal is yet to be observed experimentally in a nanoparticle. This crystal phase was first theoretically discovered by **(author?)** [24], and it has been experimentally observed only on surface reconstructions [25] and nanosheets [26]. From the theoretical perspective, studies of the formation process of ZnO nanoparticles have mostly employed classical force-fields including ReaxFF and Buckingham [27–29]. More recently, more precise modeling was also achieved using machine-learning interaction potentials [30, 31] (MLIP) yet never focusing on crystal nucleation at the nanoscale.

In this work, we first constructed a machine-learning interaction potential including long-range interactions using the recently developed Physical LassoLars Interaction Potential (PLIP) methodology and demonstrated its higher accuracy when compared to simpler short-range MLIP. Then, we performed both brute-force molecular dynamics along with seeded simulations to unravel

* julien.lam@cnrs.fr

the competition between Wurtzite (WRZ) which is the most stable crystal in bulk, and body-centered tetragonal phase (BCT) which is a concurring phase only stable at sufficiently small nanoparticle sizes. Finally, to analyze the obtained results, we developed a data-driven clustering method based on a Gaussian-mixture model enabling for characterizing the local structure at the atomistic level. Altogether, by complementing brute-force simulations with the seeding approach, we managed to demonstrate the presence of two different nucleation pathways depending on the investigated temperatures: Multi-step process involving a metastable crystal phase and Classical nucleation picture respectively at high and moderate degrees of supercooling.

II. RESULTS

A. Validation of the machine-learning interaction potential

Many variations of MLIPs have demonstrated excellent performance in capturing short-range interactions by utilizing descriptors of local atomic environments during their model training. Nevertheless, employing these short-ranged local environment descriptors in MLIPs can pose difficulties when trying to simulate systems that entail substantial long-range interactions [32, 33]. For this study of ZnO, we develop a PLIP+Q model that combines the PLIP approach for short-range interactions [31, 34–36] along with a scaled point charge model to incorporate a long-range description of the interactions. Please see Methods A for further details on the PLIP and PLIP+Q models. We note that the superiority of the short-range PLIP against other classical force fields for ZnO was already determined in our previous work [31].

To begin, the error in lattice parameters with respect to density functional theory (DFT) calculations for all the ZnO polymorphs in the database are shown in Fig. 1.a. In general, both PLIP+Q and PLIP method performs well, as the errors remain below 1 %. Nevertheless, PLIP+Q seems to slightly improve the lattice parameter compared to PLIP, except for the sodalite (SOD) polymorph.

Further, we compute the phonon density of states (DOS) for WRZ, zincblend (ZBL), and BCT polymorphs using a supercell approach where atomic positions are slightly perturbed to measure the reaction forces. The following super cells were used: WRZ ($5 \times 5 \times 3$), ZBL($3 \times 3 \times 3$), and BCT ($3 \times 3 \times 5$). The phonon calculations are carried out using the PHONOPY package [37, 38]. Fig. 1b, shows the comparison between PLIP, PLIP+Q, and DFT phonon density of states. Qualitatively, the DOS from both PLIP and PLIP+Q methods show a good match with the DFT reference. In particular, the low-frequency acoustic DOS is almost an exact match between PLIP and PLIP+Q. However, for high-frequency optical modes, PLIP+Q shows a better agreement than PLIP, for peaks between 16-17 Hz for BCT, 15-16 Hz for WRZ, and 12 Hz for ZBL.

After studying the properties of ZnO crystals, we measure the accuracy of the model for disordered bulk structures. Ab initio molecular dynamics (AIMD) as well as MD with the obtained MLIP are carried out averaging through three different initial structures in the NVT ensemble at 1500 K for 4 ps and with a time-step of 1 fs. The partial radial distribution functions (RDF) obtained from both PLIP and PLIP+Q show a good agreement with AIMD results [See Fig. 1c]. It seems that a good description of short-range interactions can already provide structural information for disordered structures and that the addition of long-range interactions does not alter the agreement.

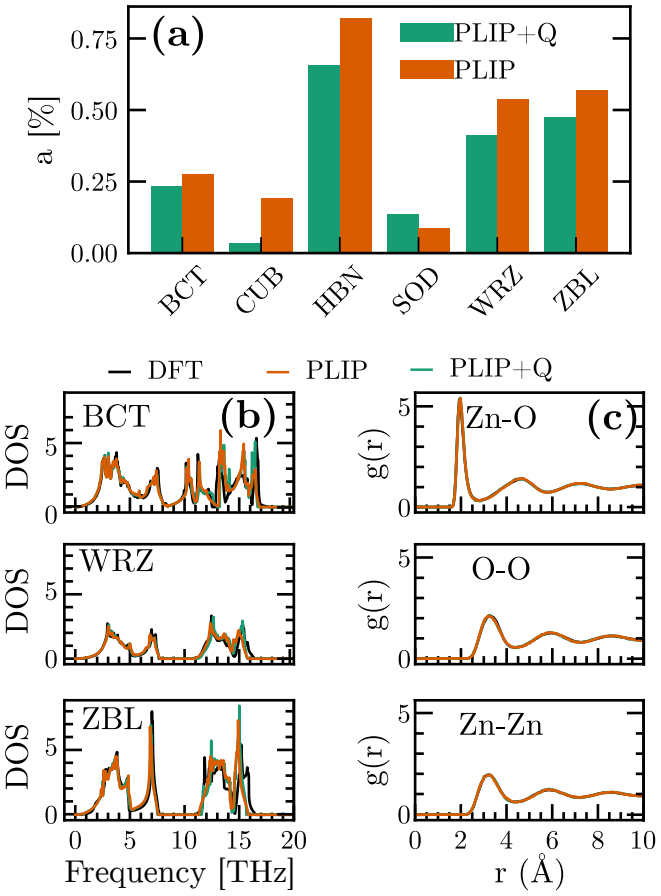

FIG. 1. Properties of bulk ZnO modeled with PLIP and PLIP+Q: (a) Lattice parameter, (b) Phonon density of states, and (c) Liquid radial distribution functions obtained at 1500 K. Please see Table VI A 3 for the nomenclature of each crystal phases.

Thus far, the performance of both PLIP and PLIP+Q has been almost indistinguishable. However, in the context of nucleation of nanoparticles, the relative solid-vacuum surface energies of low-index surfaces dictate the obtained morphology and should therefore be well reproduced too. In the specific case of zinc oxide, one must

put an emphasis on studying non-polar as well as low-index polar and polar-reconstructed surfaces. On the left part of Fig. 2.a, we show results for non-polar surfaces where one can see that the behavior of PLIP and PLIP+Q closely aligns. Then, for polar-reconstructed surfaces, although PLIP is able to perform quite well, PLIP+Q still provides slightly better agreement with the DFT results. However, a noticeable distinction emerges for polar surfaces, where PLIP+Q largely outperforms PLIP. Specifically, PLIP exhibits an error larger than 50 percent for those non-reconstructed polar surfaces, while the error is less than 10 percent for the PLIP+Q. More importantly, PLIP not only exhibits large percentage errors, it also incorrectly predicts that the two studied polar surfaces are the most stable ones [See Fig. 2b]. In contrast, PLIP+Q is able to retrieve the correct stability ordering when compared to DFT calculations, as demonstrated in Fig. 2.b.

Prompted by these results, we conduct an assessment of the performance of both PLIP and PLIP+Q on nanostructures. To begin, we use ZnO clusters obtained by **(author?)** [23]. We explore 3 different families of $(ZnO)_N$ structures, encompassing cuts of bulk crystalline structures namely BCT, WRZ, and SOD. The nanoparticles are optimized further with our DFT setting. The error in energy is measured for each system by comparing to the DFT optimized reference and MLIP optimized structure and is shown in Fig. 2.c. Although the training set is composed of bulk and surface configurations only, it is evident from Fig. 2.c that both MLIPs are transferable to these nanometric structures. Since these nanoparticles do not exhibit any polar surfaces, we additionally test both MLIPs using WRZ and ZBL nanostructures purposely constructed to expose polar surface terminations. In the case of ZBL, we design octahedral nanoparticles with polar (111) facets, featuring truncated corners to ensure their overall stoichiometry. The WRZ nanoparticles are made by top-down cuts of the bulk polymorph, as illustrated by **(author?)** [23]. From Fig.2.d, which displays the corresponding single point energy errors, it can be seen that while PLIP was able to correctly model nanoparticles with non-polar surfaces, it leads to much higher error than PLIP+Q for both types of nanoparticles exhibiting polar surfaces. The discrepancy is mainly driven by the substantial underestimation of surface energy of polar terminations in the local PLIP approach, and may lead to a spurious abundance of polar nanoparticles in MD simulations.

Altogether, the main drawback of PLIP is that polar surfaces are not only incorrectly reproduced in energy, but they are also considered the most stable ones. Such an issue has dramatic consequences when dealing with nanostructures. However, although very simple in its conceptual formalism, PLIP+Q is already able to rectify this error and model correctly both polar and non-polar surfaces as well as their subsequent nanostructures. Consequently, moving forward, the results presented in the remainder of the article will exclusively focus on calculations obtained using the PLIP+Q potential. We note that the computational efficacy was reduced by roughly 20 % upon adding the long-range interactions.

III. BRUTE-FORCE SIMULATIONS

To study the crystallization of ZnO nanoparticles we perform brute-force simulations for liquid nano-droplets of 500, 1000, 1500, and 2000 atoms at the same degree of supercooling ie. $T/T_{melt} = 0.625$. The simulations are carried out using different random instance for the initial atomic velocities so that different nucleation pathways could be explored. Please see Methods B for further details on the brute-force simulation protocol. Then, in order to analyze the structural composition of the system during nucleation and growth, we develop a supervised machine-learning method based on the combination of Gaussian mixture model for classification and Steinhardt bond order parameters for structural description. Further details on the procedure are to be found in Methods C. The composition of the nanoparticles as a function of time is shown for the system of 2000 atoms in Fig. 3 as well as snapshots at different key points in time. In all the obtained simulations, one can observe an induction time of approximately 300 ps after which a first nucleus sufficiently large is formed. Surprisingly, while WRZ is the most stable crystal structure both in bulk and in this size regime, the nuclei consist primarily [See Fig. 3(c,d)], and in some cases completely [See Fig. 3(a,b,e)], of atoms in the BCT structure. Additionally, in three cases [See Fig. 3(a,c,e)] where only one nuclei emerges at the early stages, WRZ competes and becomes the most preponderant crystal phase at the later stages. Meanwhile, in two others cases [See Fig. 3(b,d)] where more than one nuclei are formed, the system presents a slower growth rate and during the entire simulation is mostly composed of atoms in the BCT structure. These results clearly show that there is a competition in the formation of the BCT and WRZ crystal phases, where BCT is more predominant in the early stages of crystallization while WRZ forms later and becomes the main structure.

These observations are further supported by additional brute-force simulations of 500, 1000, and 1500 atoms [See Supplementary Figures (1-3)]. In all systems sizes, BCT forms first and WRZ appears later in some of the simulations. Such a two-step nucleation process, observed here in the crystal nucleation from a liquid nano-droplet, is consistent with our previous results obtained in bulk and with short-range PLIP [31]. It is also reminiscent of seminal findings in much more simple systems interacting with Lennard-Jones and hard-spheres force fields where, while the face-centered cubic is the most thermodynamically stable, it is the body-centered cubic that is often observed at the early stages of nucleation [39–42].

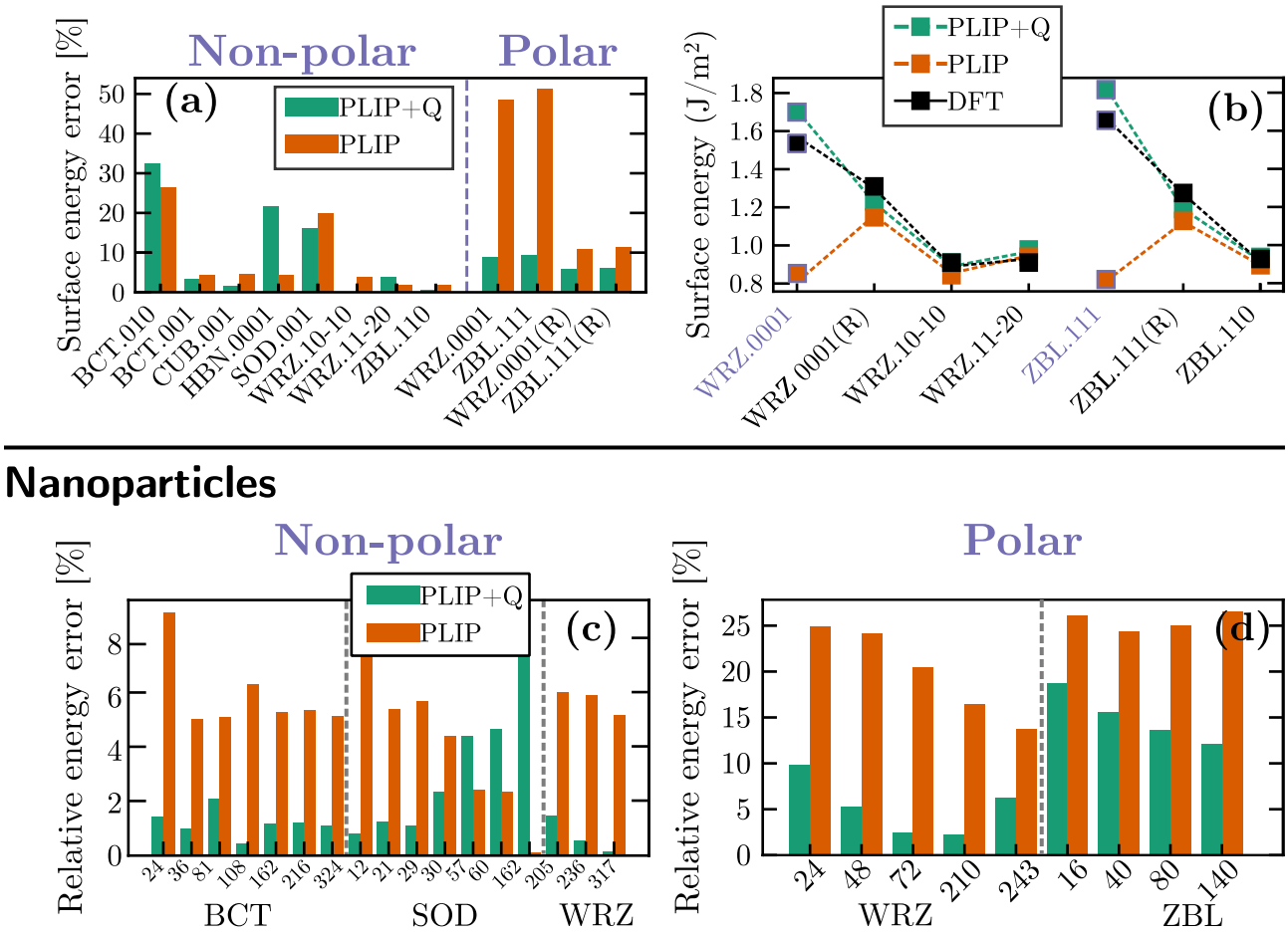

FIG. 2. (a,b) Solid-vacuum surface energy measured with PLIP and PLIP+Q for different ZnO polymorphs. (a) Surface energy error with non-polar and polar surfaces respectively on the left and on the right. (b) Value of the surface energy when focusing on WRZ and ZBL polymorphs. In violet, the polar surfaces are highlighted. "(R)" designates polar-reconstructed surfaces. (c,d) Nanoparticle energy when comparing PLIP and PLIP+Q. (c) Optimized nanoparticles without any polar surfaces as obtained by **(author?)** [23] and (d) Non-optimized nanoparticles created to exhibit polar surfaces.

IV. SEEDING SIMULATIONS

In the brute-force simulations, simulations must be performed at very low temperatures ie. in a deeply supercooled regime in order thus reducing both the free energy barrier and the associated induction time. In order to further investigate this competition and to verify its presence at moderate degrees of supercooling, one has to employ rare-event sampling techniques and we therefore perform seeding simulations [44]. In particular, a crystal seed is manually inserted in the fluid and its critical temperature is characterized as the temperature for which the seed neither shrinks nor grows [45]. By inserting a crystal seed, the free energy barrier of nucleation is artificially overcome, allowing for the study of more realistic conditions closer to the melting temperature. The seeding technique has already been applied to the study of crystal nucleation in many different systems ranging from condensed to soft matters [45–51]. But, to the best of our knowledge, the seeding technique has never been applied to nanocrystal simulations. Herein, we used the seeding technique to find the critical temperature of WRZ and BCT crystalline clusters which will allow us to address the competition between these two crystalline structures in nanoparticles [Please see Methods B for more details on the seeding simulation protocol].

Results of the growth/melting curves are shown in Fig. 4 for a 2000-atoms system and for initial crystal seeds of different sizes and crystal structures [Please see Supplementary Figures (4-6) for results obtained with different numbers of atoms inside the droplet]. For each growth/melting curves, standard deviations are showed in shaded colors and are obtained through 5 different simulations using the same state conditions (initial crystalline seed, temperature and total droplet size) yet with different initial velocities. First, we note that in all seeding simulations, the size of the biggest crystalline cluster N increases during the relaxation step. This is required

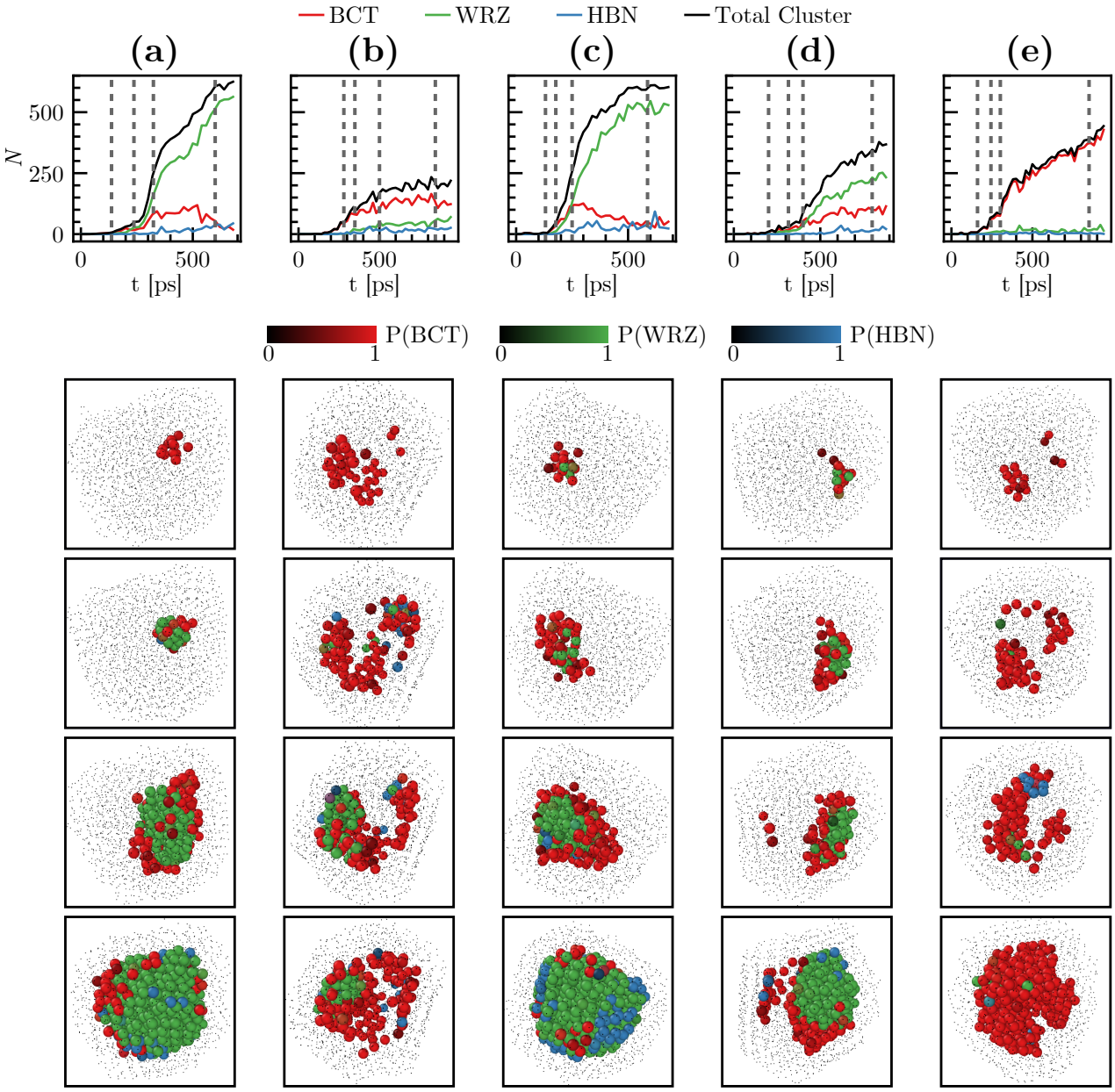

FIG. 3. Brute force nucleation simulations of ZnO nanoparticles made of 2000 atoms. Five simulations are shown using different random instances for the initial velocities in graphs (a-e) where the simulation time is shown on the x-axis and the number of atoms is shown on the y-axis. Snapshots of the clusters are shown below each graph corresponding to times indicated by the vertical dashed lines. Atoms are colored according to their probabilities to be in BCT structure (red), WRZ structure (green), and h-BN (HBN) structure (blue). Images are plotted using Ovito software [43].

to fine-tune the initial size N_0 in order to reach the desired studied size N_c. Then, despite this fine-tuning of the relaxation process, Fig. 4 shows that for a given set of conditions (i.e. N_c, T and crystal structure), there is still a large standard deviation around each plain line and subsequently a large discrepancy between simulations with different initial velocity conditions. This justifies the necessity to generate several growth/melting curves per set of conditions (5 in our case).

For each of the inserted seeds, one can deduce from the growth/melting curves a critical temperature T_{crit} located between the lowest temperature at which the cluster shrinks and the highest temperature at which it grows. Ideally, the critical temperature is such that the slope of N as a function of time is zero. In practice, a first temperature range is roughly selected and growth/melting simulations are performed at the upper and lower limits of the range, as well as in the middle.

Then, the cluster sizes are plotted and the temperature for the next simulation is chosen as the value in between the temperatures with opposite slopes of N. We iterate this method until reaching a temperature range of 16 K which defines our reported error on the measurement of T_{crit} and we use the middle between the two extreme values of the obtained range. The critical temperatures found for different crystalline cluster sizes are shown in Fig. 5 together with the critical temperatures found for seeds in differently sized droplets. Using the results for the 2000-atoms system and our estimate of the melting temperature, the system is located in a degree of supercooling T/T_{melt} between 0.8 and 0.9 which is way larger compared to the 0.625 studied with the brute-force simulations. The first observation is that the critical temperature increases with size meaning that lower temperatures are necessary to stabilize the smallest critical seeds. For all cluster sizes in the 2000-atoms and 1500-atoms systems, WRZ has a significantly higher critical temperature when compared to BCT. Accordingly, there is a range of temperatures where a crystal seed of a similar size will be more stable in the WRZ phase than in the BCT phase. In the case of the 1000-atoms and 500-atoms systems, we observe that the difference in critical temperature between BCT and WRZ crystal clusters becomes less significant. More specifically, it is observed that the critical temperatures for BCT and WRZ on the 500-atoms system converge, which is consistent with what has been observed in the brute-force simulations and energy minimization at 0 K [23]. In addition, the largest BCT crystal seed made of approximately 300 atoms in the 2000-atoms system shows an unstable behavior by sharply decreasing in size at the beginning of the growth/melting simulation [See Fig. 4.a]. As such, it can be conjectured that at these higher temperature regimes where the free energy barrier is the highest, the only possible nucleation pathway is the one starting with WRZ seeds. In closing, Fig. 5.b shows that a strong dependence of the critical temperature on the system size as for a similar critical cluster size there is a difference of almost 300 K in critical temperature between the 500-atoms and 2000-atoms system. This finding strongly suggest that the nanoscale reduction associated with finite-size effects and surface preponderance is already at play in the investigated size regime.

Ultimately, our seeding simulations reveal a completely different nucleation behavior compared to brute force results that focused in a more deeply supercooled regime. Indeed, we showed that BCT and WRZ are respectively favored in deeply supercooling conditions (brute-force simulations) and in moderate supercooling conditions (seeding simulations). Therefore, nucleation mechanisms are highly driven by the investigated degree of supercooling thus advocating for the necessity to combine brute force and rare-event sampling approaches. A similar observation was also made when studying nucleation from a dilute phase as in a gas or in the presence of a non-reactive solvent like NaCl in water. In both cases, depending on the saturation regime, one can indeed either directly nucleate crystalline clusters or start with a so-called high-density amorphous precursor [52–55].

V. GROWTH MECHANISMS

Along with determining the critical temperature as a function of the critical size, seeding simulations also allow for studying the subsequent growth mechanisms. Herein, we extended the previous growth curves during 84 ps after the relaxation procedure and focused on temperature regimes where the crystal seed grows.

At first, in the WRZ case, Fig. 6 shows that the cluster composition was consistent for all of the considered seed sizes ie. the cluster core is composed of atoms in the WRZ phase at the beginning and remains so until the end of the simulations. This is shown on the different graphs on which it can be noted that the number of atoms in the WRZ phase is always equal to or bigger than the number of atoms in the BCT phase. On the other hand, the outer layers of the cluster are often composed of atoms in the BCT and h-BN (HBN) phases. These atoms are located at the interface between the crystal and the liquid. The presence of few atoms predicted to be in the HBN phase at the surface of the crystalline cluster is probably a consequence of the large resemblance between HBN and WRZ structures. On the other hand, the scarce presence of atoms in the BCT phase is reminiscent of the BCT vs. WRZ competition. We note that these BCT atoms are mostly located at the surface of the crystalline cluster.

Then, the BCT case exhibits a much more complex picture. Indeed, from Fig. 7, it can be seen that the composition of the crystalline cluster evolves with time. Right after the relaxation procedure, the crystalline clusters are composed almost entirely of atoms in the BCT structure as it is the structure introduced in the seeding initial phase. This means that our relaxation procedure correctly stabilizes the interface between crystal and liquid without changing the crystal structure. The number of atoms in the BCT phase then increases, but given enough time it starts to decline, and the number of atoms in the WRZ phase increases. Finally, at the end of the simulation, the center of the crystalline cluster is composed mostly of atoms in the WRZ structure, while the outer layers are composed of atoms in HBN and BCT structures. This observation suggests that the preferred crystallized structure for ZnO in these conditions remains WRZ even when starting with a BCT seed.

VI. DISCUSSION

The comparison between results from brute force and seeding simulations exhibits fundamental insights on the nucleation process. Indeed, at low temperatures associated with the deepest degree of supercooling, crystal nucleation seems to follow a multi-step process where a

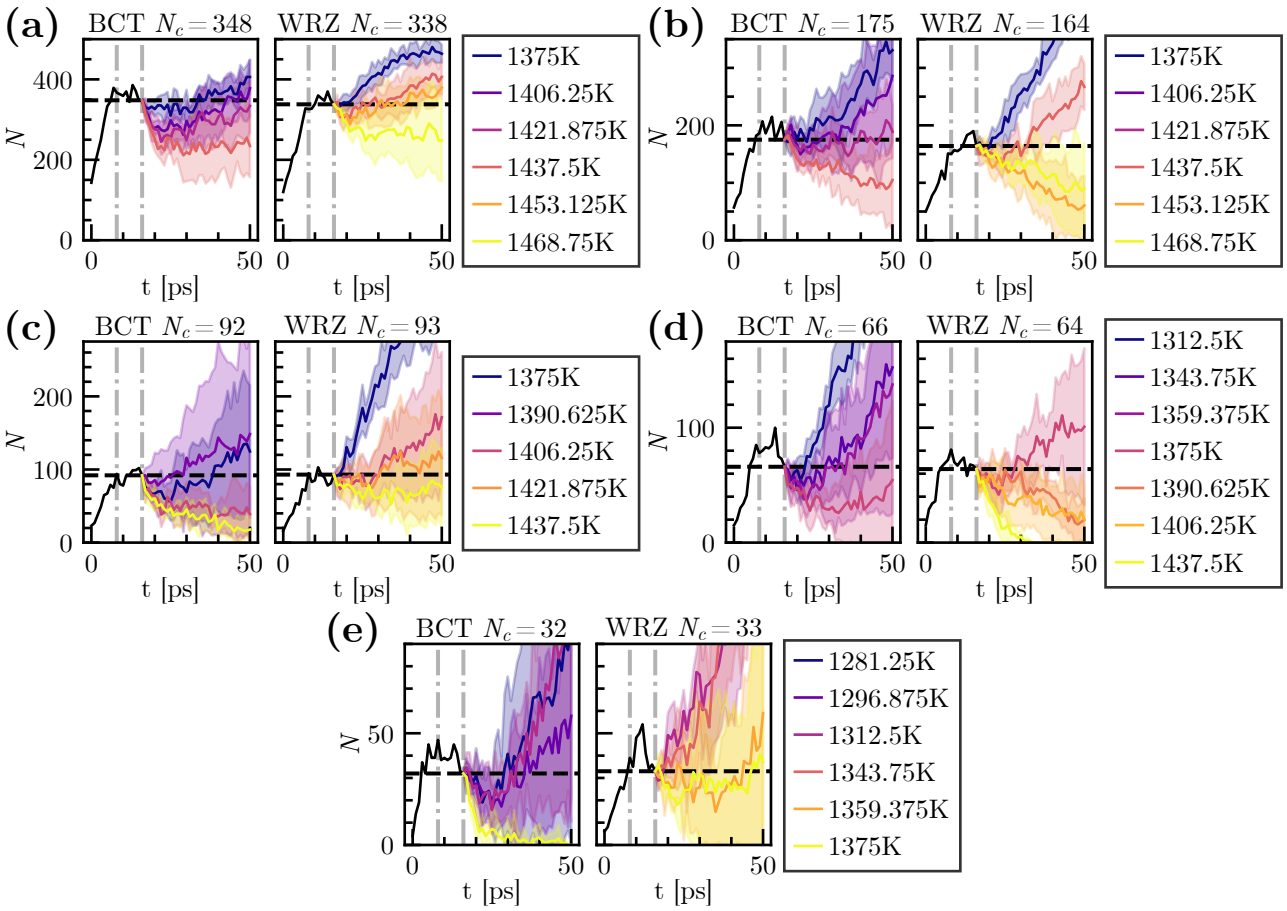

FIG. 4. (a-e) Growth/Melt simulations for BCT and WRZ clusters of different sizes are shown in subfigures with a droplet made of 2000 atoms. The relaxation steps are shown in each graph with a black line, after which the growth/melt simulations are shown for different temperatures. In each subfigure, two clusters of similar critical cluster size N_c but different crystal structures are compared. In particular, on the left (resp. on the right), the seed is made of BCT (resp. WRZ) crystalline atoms. Meanwhile, for sub-figures a to e, size of the crystalline seeds is respectively around 343, 170, 92, 65 and 32 atoms.

metastable phase (BCT) emerges first before being replaced by the most stable phase (WRZ). Meanwhile, at higher temperatures involving larger free energy barriers, the BCT seeds are in fact less stable than the WRZ ones. Indeed, we showed that their critical temperature is consistently lower than that of WRZ and that even when artificially inserted the BCT structure rapidly turns into WRZ during the subsequent growth. These two observations suggest that nucleation occurs in a single step for those moderate degrees of supercooling. To the best of our knowledge, such a change in nucleation pathway has so far only been found in crystal condensation from dilute systems like in the gas phase or in the presence of a surrounding liquid solvent. In the latter case, the observed metastable phase was made of a dense amorphous phase while here, it is an ordered phase made of different crystal polymorph.

Despite the breadth of our findings, interpreting them remains a very challenging task. On the one hand, the preponderance of WRZ at moderate degrees of supercooling, ie. in the presence of free energy barrier, is consistent with a classical nucleation picture where the nucleating crystal is also the most stable crystal. On the other hand, it is more surprising that an alternative nucleation pathway involving a different crystal structure emerges at deeper degrees of supercooling when the free energy barrier is almost vanishing. As a possible explanation, a classical picture based on the capillary approximation would evoke the possibility that the crystal/liquid interface is more favorable for BCT than for WRZ which only becomes important under these temperature conditions since the nucleus is small enough to enhance surface effects. As such, it would be appealing to characterize bulk properties including crystal/liquid interface, migration rate or elastic stresses. While appealing for qualitative understanding, characterizing the crystal/liquid interface is rendered almost impossible because in this supercooled regime the interface is not stable and the crystal will spontaneously grow. Additionally, herein, the nucleus is composed of relatively few atoms (less than 50 atoms) and therefore it can not be characterized with any sort of bulk properties. More generally, the complexity

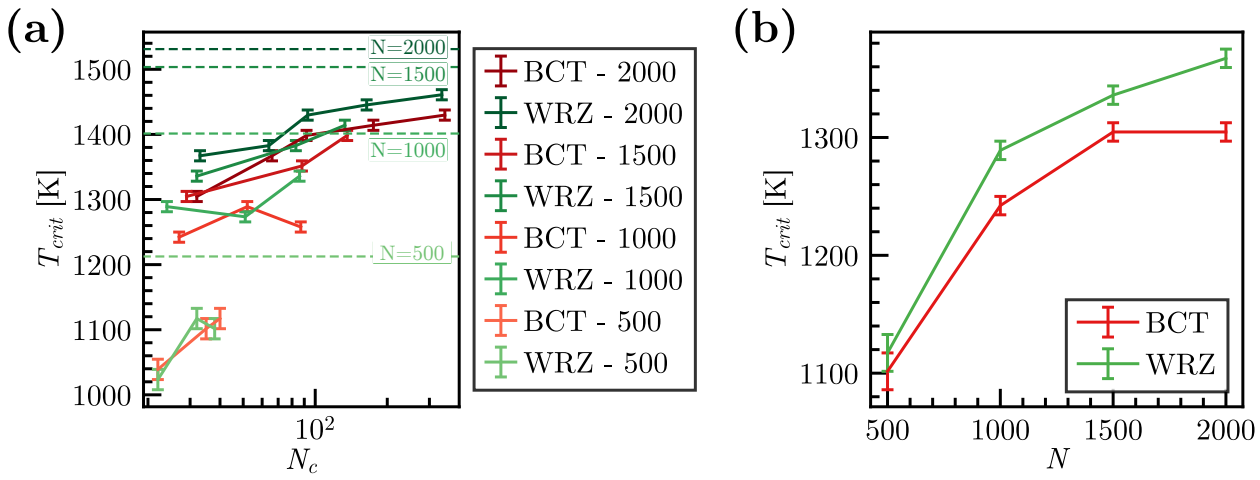

FIG. 5. (a) Critical temperature results for different cluster structures and droplet sizes as a function of critical cluster size. The critical temperature is shown on the y-axis while the critical cluster size is shown in logarithmic scale on the x-axis. The melting temperatures of WRZ nanoparticles are shown in dashed lines where the color represents the size of the nanoparticle. (b) Dependence of the critical temperature on the nano-droplet size using a similar initial crystal seed size of approximately 30 atoms.

herein is that even the concept of nucleus is no longer relevant in this regime of vanishing free energy barrier. As such, the current state of the art do not allow us to provide a quantitative explanation for our findings. However, it remains that from an empirical viewpoint, WRZ and BCT crystal symmetry can very well be compared structurally to face centered cubic (FCC) and body centered cubic (BCC). For the latter, the emergence of BCC when the free energy barrier vanishes has been observed in several occasions and can be explained using symmetry arguments and mean-field approaches[56].

From the technical viewpoint, our work proposes three solutions to allow for studying nucleation in complex nanoscale systems. Firstly, we developed an approach for modeling long-range interactions that exploits Lasso-Lars fitting for obtaining an effective static point charge Coulomb interaction. Our comparison with short-range PLIP shows that this approach enables us to better capture subtle charge effects related to polar surfaces which are crucial when dealing with nanoparticles where surface effects become preponderant. Secondly, we combined Gaussian Mixture Model with Steinhardt parameters to classify structures within a complex structural landscape made of 7 crystal polymorphs. We expect that the same methodology can be applied to characterize polymorphs in different materials as well as defects including grain boundaries or dislocations. Thirdly, we managed to explore nucleation at different degrees of supercooling by using brute-force simulations as well as the seeding technique. Altogether, our promising results advocate for transferring the proposed simulation strategy to the formation of different types of nanocrystals including quantum dots and nanoalloys.

METHODS

A. Machine-learning interaction potentials with long-range physics

1. *Short range PLIP*

To begin, we start with a concise description of the chosen short-range MLIP, before delving into long-range electrostatic interactions. To start, a linear model is employed to estimate each atomic energy (E_i) which is represented by a weighted sum of descriptors

$$E^i_{\text{short}} = \sum_n \omega_n \chi^i_n \tag{1}$$

where the coefficients (ω_n) are the fitting parameters. The descriptors for the PLIP model explicitly follow a many-body order expansion:

$$[\chi^{2\text{B}}]^i_n = \sum_j f_n(r_{ij}) \times f_{\text{c}}(r_{ij}), \tag{2}$$

$$[\chi^{3\text{B}}]^i_{n,l} = \sum_j \sum_k f_n(r_{ij}) f_{\text{c}}(r_{ij}) f_n(r_{ik}) f_{\text{c}}(r_{ik}) cos^l(\theta_{ijk}), \tag{3}$$

$$[\chi^{\text{NB}}]^i_{n,m} = \left(\sum_j f_n(r_{ij}) \times f_{\text{c}}(r_{ij}) \times f_{\text{s}}(r_{ij}) \right)^m, \tag{4}$$

where r_{ij} is the distance between atoms i and j, θ_{ijk} is the angle centered around the atom i, f_n are a set of Gaussian functions with different widths and central positions, f_{s} is a polynomial function that allows for setting the N-body interactions to 0 at short range and l and m are two positive integers. The cutoff function is defined as:

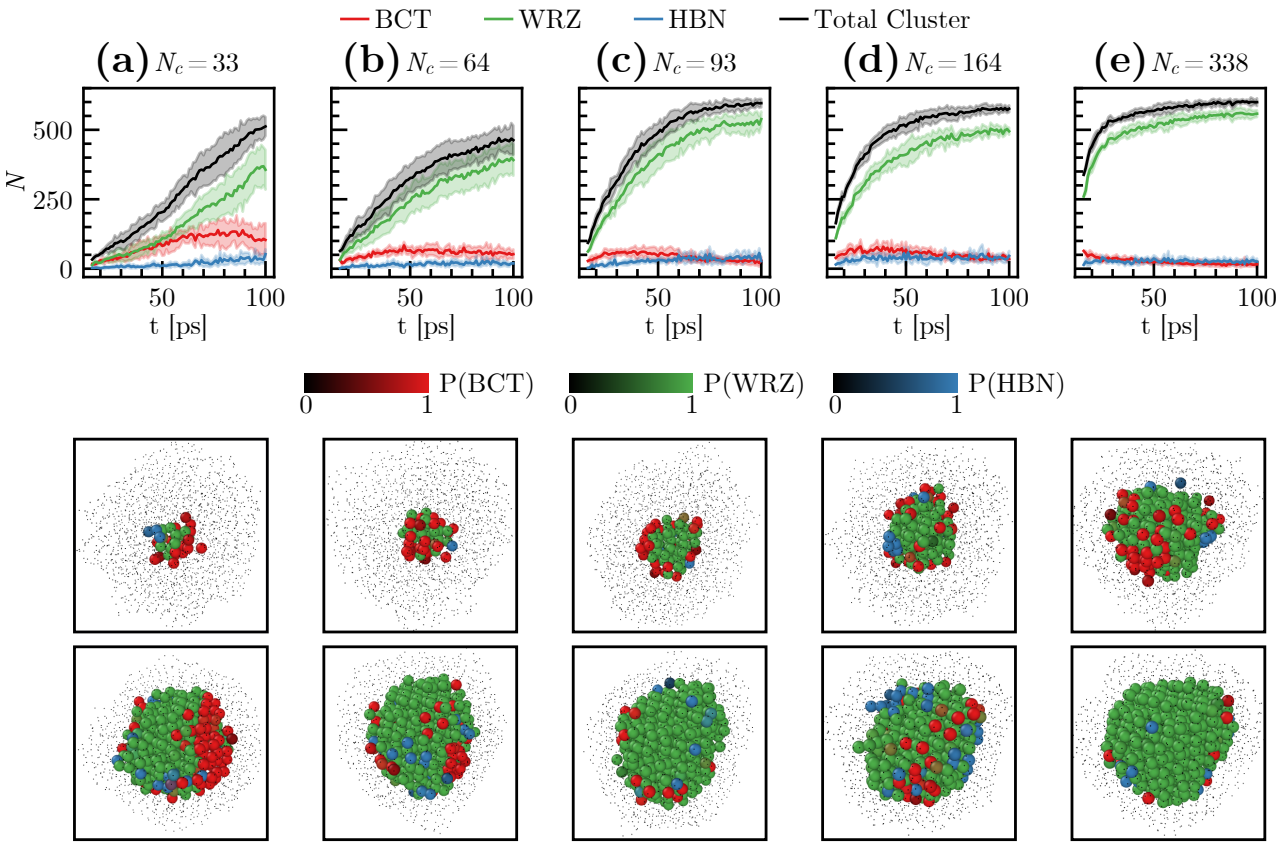

FIG. 6. Extended growth simulations following the insertion of a WRZ crystal seed at 1000 K for five different seed sizes in a 2000-atoms system. The composition of the crystalline clusters is shown in the graphs in each column. The first and last snapshots of each simulation are shown below the corresponding graphs. Atoms are colored according to their probabilities to be in BCT structure (red), WRZ structure (green), and HBN structure (blue).

$f_c(r_{ij}) = 0.5\,(1 + \cos(\pi(r_{ij}/r_{cut})))$. In the following, we use $r_{cut} = 6$ Å and impose that $l \leq 5$ and $m \leq 7$. For more details on the short-range PLIP model, one can refer to the following references [31, 34–36].

2. *Electrostatic PLIP+Q*

The field of extending MLIP models to include long-range interactions is actively advancing, and recent developments in this area have been summarized by **(author?)** [57]. In most of these approaches, a database of effective point charges is computed from electron structure calculations using different methods including Mulliken and Löwdin population analysis [58], Bader analysis [59], and the Density Derived Electrostatic and Chemical (DDEC) charge model [60]. Then, a machine-learning model is constructed in order to determine on-the-fly the charge values based on the local environment surrounding each atom. More recently, further progress was obtained by training the long-range machine-learning model on susceptibility instead of charge values which enables for capturing subtle charge transfer mechanisms [61–63]. While these approaches remain the most accurate to date, they might suffer from implementation difficulty and computational costs.

In our PLIP+Q model, we chose to use static point charges that are fixed with time and the local environment. In practice, we begin by setting the value of initial charges which can be for instance the oxidation number or deduced from electron density calculations. Then, we compute a fictive electrostatic contribution to the energies denoted E^i_{el} using the point-charge Coulomb model along with the Ewald summation method [64]. In order to determine how much this fictive electrostatic interaction indeed contributes to the overall interactions, we next define an overall charge scaling factor named γ leading to the following total energy:

$$E^{\mathrm{i}} = E^i_{\mathrm{short}} + \gamma E^{\mathrm{i}}_{\mathrm{el}} \tag{5}$$

As such, the scaling factor γ becomes an additional linear coefficient that can be fitted along with those associated with the short-range interactions. We finally use the LassoLars regression algorithm to determine simultaneously the linear coefficients required for the short range interactions and the value of γ thus enabling us to empiri-

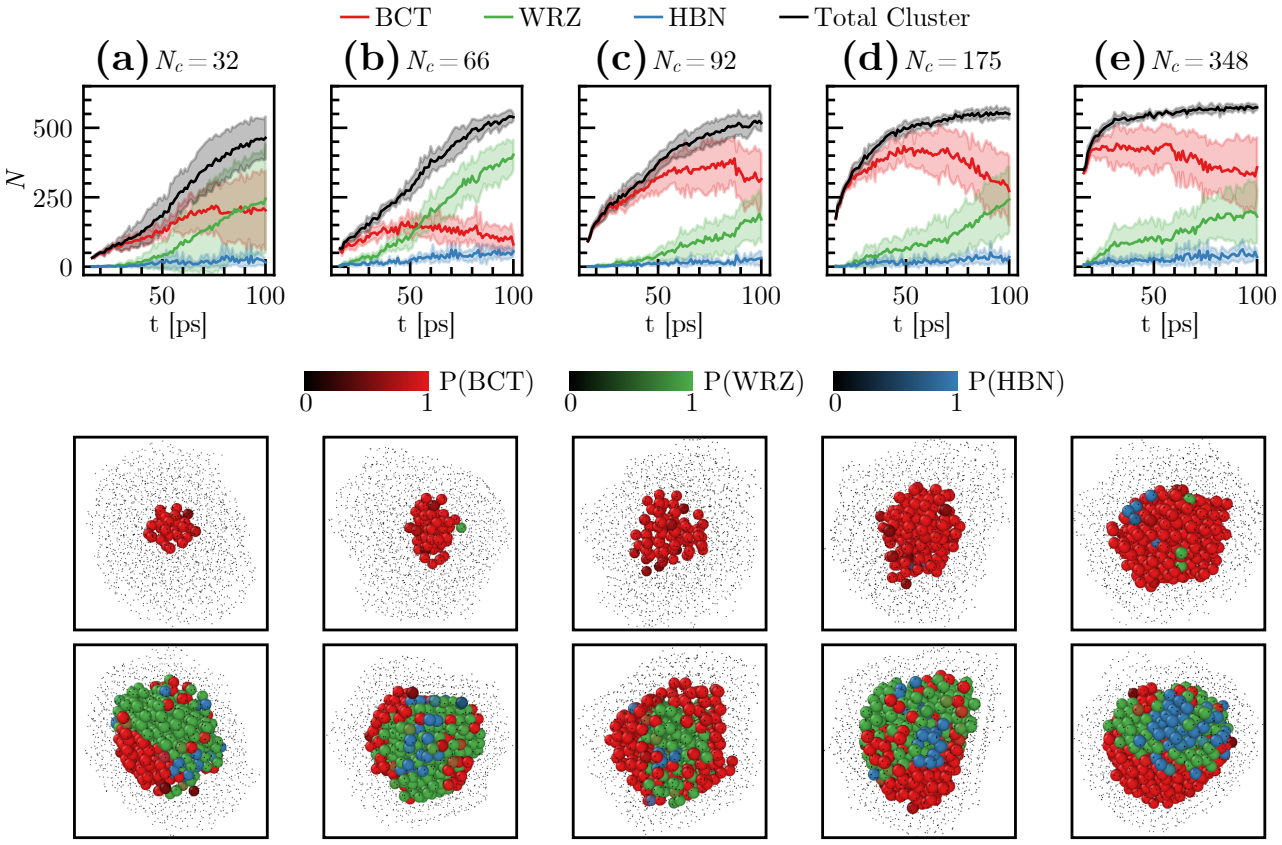

FIG. 7. Extended growth simulations following the insertion of a BCT crystal seeds at 1000 K for five different seed sizes in a 2000-atoms system. The composition of the crystalline clusters is shown in the graphs in each column. The first and last snapshots of each simulation are shown below the corresponding graphs. Atoms are colored according to their probabilities to be in BCT structure (red), WRZ structure (green), and HBN structure (blue).

cally obtain an effective value for the static point charge. Often in attempts at modeling long-range interactions using machine-learned charges, the short-range interactions is treated as a substantial difference between the long-range interactions and the total quantum accurate energy. Here, we chose instead to consider the long-range interactions simply as an additional descriptor without imposing its presence in the final model. As such, when using the LassoLars regression that selects only a subset of the most preponderant descriptors, the long-range interactions will be considered solely if necessary. In our present case, the initial charges are set as the oxidation number ie. at plus and minus 2 respectively for zinc and oxygen, and following the LassoLars fitting, the rescaled value of the charges becomes ± 0.64, to be compared with $\pm$ 1.16 obtained from the Bader decomposition of DFT charge densities. Although it is difficult to interpret this value because the short-range PLIP also contributes to the overall two-body interaction, it remains interesting to note that the LassoLars regression scheme is able to retrieve a value at a reasonable order of magnitudes.

3. *Training database*

The employed training database is exactly the same as in our two previous applications of PLIP to ZnO [31, 34] In brief, the reference DFT calculations are performed using the VASP software, employing the PW91 exchange-correlation functional and the projector augmented wave method. The standard zinc and soft oxygen pseudopotentials are employed, with an energy cutoff of 270 eV [65–67]. The total database is made of both ordered and disordered structures starting from bulk and surface structures.

In particular, six different ZnO polymorphs namely WRZ, zinc blend (ZBL), body-centered tetragonal (BCT), sodalite (SOD), h-BN (HBN), and cubane (CUB) crystallographic structures are included. We employ relatively large supercells in the range of 16-19 Å forming parallelepipeds that contain between 320 to 480 atoms. The Brillouin zone sampling is done at a single Γ point. We conduct atomic coordinate relaxation until forces reduce below 0.01 eV/Å while maintaining stress tensor components below 0.01 eV/Å. To study nucleation in nanoparticles, it is also essential to include low-coordinated struc-

tures in the database. In this regard, for each of the six ZnO polymorphs, low-index nonpolar surfaces are also considered. In all our computations, we utilize slabs composed of 6 to 12 atomic layers, separated by approximately 15 Å of vacuum, and we conduct a thorough relaxation of atomic coordinates of all ions.

From all of these equilibrium structures, we perform molecular dynamics (MD) simulations using either a classical potential or a previously obtained MLIP to sample additional structures where forces are then determined afterward with single-point DFT calculations.

Shortname	**Fullname**
BCT	body-centered tetragonal
CUB	cubane
HBN	h-BN
SOD	sodalite
WRZ	wurtzite
ZBL	zinc blend

TABLE I. Nomenclature defining the acronyms used to describe each of the considered crystal structures.

B. Brute-force and seeding simulations

Brute-force simulations For the brute-force simulations, we used the NVT ensemble in the LAMMPS software [68] with a Nose-Hoover thermostat and a timestep of 1 fs. In order to work at a consistent degree of supercooling, we measure the melting temperature for a WRZ nanoparticle of each nanoparticle size using a linear temperature ramp [2000 K.ns^{-1}] and fitting the crystal size with a hyperbolic tangent function. The melting temperatures are 1213 K, 1401 K, 1503 K, and 1531 K for sizes of 500, 1000, 1500, and 2000 atoms, respectively. We note that our MD simulations as well as experimental measurements seems to indicate that the ZnO melting is congruent[69]. Despite imposing the same degree of supercooling, the nucleation induction time slightly differs for each nanoparticle and we used different simulation duration, between 1 and 10 ns, as necessary for the different droplet sizes. By varying initial atomic velocities we obtained multiple simulation trajectories for the same system and temperature. By doing this we are able to assess the average behavior of this type of system at the chosen temperature.

Seeding simulations After initializing the system with a crystalline seed of a chosen size N_0 inserted inside a liquid nano-droplet, the relaxation protocol enabling for reaching the desired temperature T consists of three different steps. Firstly, the crystalline cluster is kept static while MD is performed on the liquid nano-droplet in the NVT ensemble at the relaxation temperature T_0, which was chosen always lower than the expected critical temperature and changed depending on the size of the droplet and the crystalline cluster. This step is performed for a duration that is adjusted according to the system size in order to avoid complete crystallization during this step. Because only the liquid droplet can move, the size of the crystal seed can increase through interfacial growth. Furthermore, the gap created during the insertion of the seed between the crystal and the liquid is filled during this first step. Secondly, atoms in the crystalline cluster are also allowed to move in the NVT ensemble at an increasing temperature. In this second step, of the same duration as the first one, two separate Nose-Hoover thermostats are used: (1) for the liquid droplet at T_0 and (2) for the crystalline seed going from 100 K to T_0. During this step, we attempt to keep the size of the cluster constant. The purpose of this step is to slowly increase the temperature of the crystalline cluster until T_0 is achieved for the whole system. Thirdly, the temperature of the system needs to go from T_0 to T. For that purpose, we can not directly set the thermostat temperature at T as it would put the system out of equilibrium. In addition, because nucleation is a stochastic process, it was crucial to be able to study the same system (ie. size of crystalline seed + temperature) with different initial velocities. For these reasons, the Langevin thermostat [70] is used at an increasing temperature from T_0 to T and with 5 different random values for the thermostat. At the end of these three steps, the apparent size of the crystalline seed that is used for the analysis is always different from N_0 and is denoted N_c. After the Langevin temperature ramp, NVT simulations using the traditional Nose-Hoover thermostat are carried out at a temperature of interest T. At the end, the obtained growth/melting curves are averaged over the different random values of the Langevin thermostat thus accounting for the stochasticity of the nucleation process. We used those 5 different simulations to also compute a standard deviation associated to each growth/melting curves. The duration of each step and the relaxation temperatures used are presented in Supplementary Table 1.

C. Gaussian-mixture model to characterize polymorphic crystal ordering

In order to analyze local ordering in the obtained simulations, it is crucial to use a numerical tool capable of distinguishing 7 different polymorphs of ZnO among which 6 are already employed in the DFT training database as well as the rock salt (RCK) structure that was considered only for the structure identification. We propose a supervised learning method that we call Steinhardt Gaussian Mixture Analysis (SGMA) [See Fig. 8 for a schematic picture].

In particular, we start by creating a database consisting of crystalline structures sampled around their equilibrium positions. For that purpose, NVT simulations are performed during 10 ps with an increasing temperature from 200 K to 1500 K controlled via a Nosé-Hoover thermostat. The duration and the upper temperature are chosen so that none of the seven considered crystals

melt. In addition, liquid structures are also sampled in our database using NVT simulation at 2500 K. 21 snapshots are extracted along those simulations for each of the crystal polymorphs and for the liquid.

For each of these snapshots, we then compute the averaged Voronoi weighted Steinhardt parameters [71] using the *Pyscal* [72] library in *Python*. We augmented the generic Steinhardt values with homo nuclear ones calculated after removing each hetero atom type. In both cases, we used Steinhardt parameters indexed from 2 to 8 thus making a list of 14 order parameters to characterize the local ordering of each atom in a given snapshot.

The training of the database is next performed using the Gaussian Mixture Model (GMM) as implemented in the *scikit-learn* [73] library in *Python*. The unknown parameters of the GMM were iteratively estimated using the Expectation-Maximization algorithm [74]. The GMM was trained using full covariance matrices and 100 k-means initializations. In our case, instead of letting the GMM determine the number of Gaussian components automatically, we chose to impose it equal to the number of structure types in our database ie. 8. In this way, we give priority to the physical meaning of our database. For classification, the Maximum Likelihood Classifier is utilized, in which the probability of an object x_i to belong to class ω_k is computed as: $p(\omega_k|\mathbf{x}_i) = \frac{\alpha_k \mathcal{N}(\mathbf{x}_i|\mathbf{m}_k,\mathbf{C}_k)}{\sum_{j=1}^{K} \alpha_j \mathcal{N}(\mathbf{x}_i|\mathbf{m}_j,\mathbf{C}_j)}$, where α_k are mixture proportions and $\mathbf{m}_k$ and $\mathbf{C}_k$ are the mean vector and the covariance matrix of each Gaussian component ω_k. The mixture proportions satisfy the conditions $0 \leq \alpha_k \leq 1$ and $\sum_{k=1}^{K} \alpha_k = 1$. These values can then be interpreted as the probability of an atom being in one of the 8 structure types in the database. In the Maximum Likelihood Classifier, the probabilities are usually compared and an object is said to belong to a category for which it has the highest probability. In this work, we chose a more severe classification rule and considered an atom to belong to a cluster only when its probability is higher than 50%.

Parameters for the different Gaussian clusters in the model are obtained following the training procedure. However, only the index of each cluster is known and no information is given as to what structure each cluster represents. To do this, the model is tested by predicting the Gaussian cluster to which the perfect crystals and the liquid belong. In this way, the labels of each Gaussian cluster are obtained. This is another way in which our method differs from previous uses of the GMM. We first train a model, and since the clusters in our database are approximately Gaussian, it is expected that after training a Gaussian cluster will be assigned to each structure type. This model can then be used to analyze systems different from the ones encountered in the database and obtain specific structural predictions in a supervised manner.

Altogether, our SGMA methodology allows us to predict crystal structures in a system by creating a database with the different known polymorphs. In this application, we rely on the assumption that each structure in our database can be approximated by a single Gaussian cluster, as opposed to other applications of the GMM where the number of Gaussian clusters is found automatically [75–77]. For future applications, the method can also be adapted to automatically find the number of required Gaussian clusters when a structure is better represented with more than one. Our method is also characterized by the significantly large number of descriptors that are computed and used for analysis. Compared to other methods, we do not make use of dimensionality reduction techniques to decrease the complexity of the data at the clustering step [31, 76–83]. If the computational cost demands it, it is possible to reduce the complexity of the model by carefully choosing the descriptors that distinguish the structures in the database the best.

To test the usage of SGMA in the seeding technique we analyzed the structure of seeded nanoparticles right after the initialization and before the relaxation step for BCT and WRZ crystal seeds, illustrated in Fig. 9. It was found that in both cases our model correctly identifies the crystalline cluster surrounded by the liquid. In the case of the BCT crystal seed, the number of atoms predicted to be in the BCT phase is lower than the number of inserted atoms. This is expected from our model since at the interface between crystal and liquid, the atomic environments differ significantly from the environment represented in the database. Even though the atoms at the interface still present some order in their structure, according to the Maximum Likelihood Classifier described previously, they are considered to be closer to the liquid structure. Similarly, in the case of the WRZ crystal seed, the number of atoms predicted to be in the crystalline phase is lower than the number of inserted atoms. This time, however, not all atoms in the crystalline cluster are predicted to be in the WRZ phase. At the surface of the crystalline cluster, some atoms are predicted to be in the HBN or BCT phases instead. This is due to the large resemblance between WRZ and HBN phases. With these tests, we have shown the physical meaning of the predictions performed using the GMM. Crystalline atoms were correctly identified when surrounded by a liquid, with the exception of atoms at the interface between liquid and WRZ phases that can be also labeled as HBN. These results confirm that the training parameters of the model were appropriately chosen, and the physical meaning of the database was retained.

VII. DATA AVAILABILITY

The authors declare that the data supporting the findings of this study are available within the article and its supplementary information files or from the corresponding authors on reasonable request.

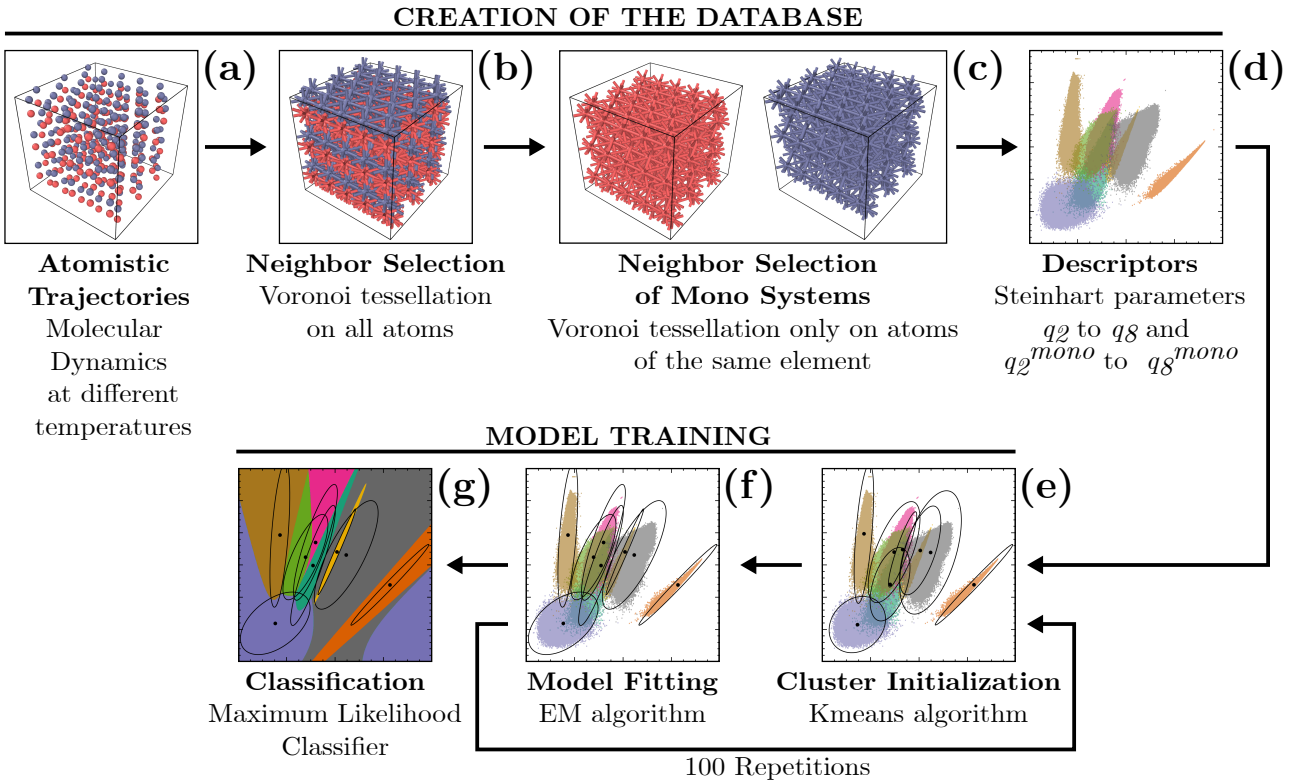

FIG. 8. Schematic of the steps in SGMA. A database is created from (a) atomistic trajectories obtained using Molecular Dynamics. (b) The neighbors of each atom are found using Voronoi tessellation, (c) as well as the neighbors of only the same element. (d) The Steinhardt parameters are then computed for all atoms in each snapshot of the trajectory. A Gaussian Mixture Model is then trained on the database. (e) The parameters of the Gaussian clusters are initialized using the Kmeans algorithm and are then optimized using the Expectation-Maximization algorithm (f). Steps (e) and (f) are performed 100 times and the parameters with the best results are kept. (g) Classification is then performed on the structures in the database or on new test structures.

VIII. CODE AVAILABILITY

The implementation of (1) PLIP+Q, (2) the structural analysis, (3) Seeding and brute forces simulations inputs and outputs will be shared with community upon request.

[1] Petkova, A. T. *et al.* Self-Propagating, Molecular-Level Polymorphism in Alzheimer's ß-Amyloid Fibrils. *Science* **307**, 262–265 (2005).

[2] Close, W. *et al.* Physical basis of amyloid fibril polymorphism. *Nat. Commun.* **9**, 1–7 (2018).

[3] Fändrich, M. *et al.* Amyloid fibril polymorphism: a challenge for molecular imaging and therapy. *J. Intern. Med.* **283**, 218–237 (2018).

[4] Zhang, L. L., Yang, S., Wei, W. & Zhang, X. J. Genetic polymorphisms affect efficacy and adverse drug reactions of DMARDs in rheumatoid arthritis. *Pharmacogenet. Genomics* **24**, 531 (2014).

[5] Morissette, S. L., Soukasene, S., Levinson, D., Cima, M. J. & Almarsson, Ö. Elucidation of crystal form diversity of the HIV protease inhibitor ritonavir by high-throughput crystallization. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 2180–2184 (2003).

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

[7] Finney, A. R. & Salvalaglio, M. Molecular simulation approaches to study crystal nucleation from solutions: Theoretical considerations and computational challenges. *WIREs Comput. Mol. Sci.* **14**, e1697 (2024).

[8] Ramamoorthy, R. K. *et al.* The role of pre-nucleation clusters in the crystallization of gold nanoparticles. *Nanoscale* **12**, 16173–16188 (2020).

[9] Schiener, A. *et al.* In situ investigation of two-step nucleation and growth of CdS nanoparticles from solution. *Nanoscale* **7**, 11328–11333 (2015).

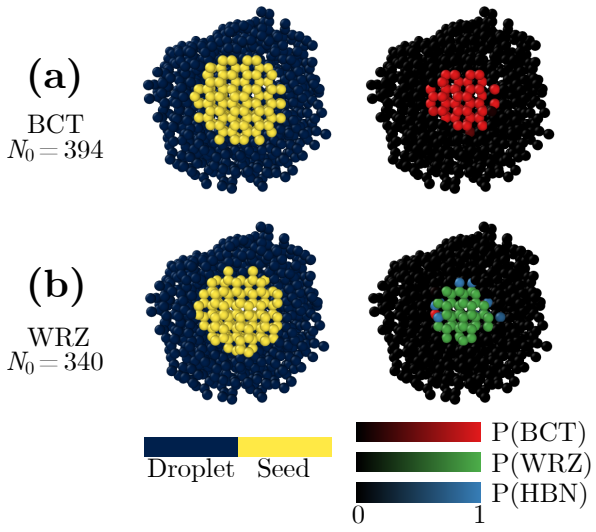

FIG. 9. Test of our structural analysis methodology on BCT (a) and WRZ (b) seeded nanoparticles. The left column illustrates the crystalline atoms that were inserted in the droplet system. The right column shows the predicted structure of each atom as computed with our methodology.

[10] Ibrahimkutty, S., Wagener, P., Menzel, A., Plech, A. & Barcikowski, S. Nanoparticle formation in a cavitation bubble after pulsed laser ablation in liquid studied with high time resolution small angle x-ray scattering. *Appl. Phys. Lett.* **101** (2012).

[11] Albrecht, W., Van Aert, S. & Bals, S. Three-Dimensional Nanoparticle Transformations Captured by an Electron Microscope. *Acc. Chem. Res.* **54**, 1189–1199 (2021).

[12] Zhou, X.-Q. *et al.* Zinc Oxide Nanoparticles: Synthesis, Characterization, Modification, and Applications in Food and Agriculture. *Processes* **11**, 1193 (2023).

[13] Nagajyothi, P. C. *et al.* Green route biosynthesis: Characterization and catalytic activity of ZnO nanoparticles. *Mater. Lett.* **108**, 160–163 (2013).

[14] Sun, Y. *et al.* The Applications of Morphology Controlled ZnO in Catalysis. *Catalysts* **6**, 188 (2016).

[15] Matinise, N., Fuku, X. G., Kaviyarasu, K., Mayedwa, N. & Maaza, M. ZnO nanoparticles via Moringa oleifera green synthesis: Physical properties & mechanism of formation. *Appl. Surf. Sci.* **406**, 339–347 (2017).

[16] Pushpalatha, C. *et al.* Zinc Oxide Nanoparticles: A Review on Its Applications in Dentistry. *Front. Bioeng. Biotechnol.* **10**, 917990 (2022).

[17] Islam, F. *et al.* Exploring the Journey of Zinc Oxide Nanoparticles (ZnO-NPs) toward Biomedical Applications. *Materials* **15** (2022).

[18] Gudkov, S. V. *et al.* A Mini Review of Antibacterial Properties of ZnO Nanoparticles. *Front. Phys.* **9**, 641481 (2021).

[19] Wang, L.-Y. *et al.* Size and Morphology Modulation in ZnO Nanostructures for Nonlinear Optical Applications: A Review. *ACS Appl. Nano Mater.* **6**, 9975–10014 (2023).

[20] Chen, M. & Dixon, D. A. Machine-Learning Approach for the Development of Structure–Energy Relationships of ZnO Nanoparticles. *J. Phys. Chem. C* **122**, 18621–18639 (2018).

[21] Zagorac, D. & Schön, J. C. Energy landscapes of pure and doped ZnO: from bulk crystals to nanostructures. In *Frontiers of Nanoscience*, vol. 21, 151–193 (Elsevier, Walthm, MA, USA, 2022).

[22] Leitner, J., Bartůněk, V., Sedmidubský, D. & Jankovský, O. Thermodynamic properties of nanostructured ZnO. *Appl. Mater. Today* **10**, 1–11 (2018).

[23] Viñes, F., Lamiel-Garcia, O., Illas, F. & Bromley, S. T. Size dependent structural and polymorphic transitions in ZnO: from nanocluster to bulk. *Nanoscale* **9**, 10067–10074 (2017).

[24] Wang, J. *et al.* Molecular dynamics and density functional studies of a body-centered-tetragonal polymorph of ZnO. *Phys. Rev. B* **76**, 172103 (2007).

[25] He, M.-R., Yu, R. & Zhu, J. Reversible Wurtzite-Tetragonal Reconstruction in ZnO(1010) Surfaces. *Angew. Chem. Int. Ed.* **51**, 7744–7747 (2012).

[26] Wang, F. *et al.* Nanometre-thick single-crystalline nanosheets grown at the water–air interface. *Nat. Commun.* **7**, 1–7 (2016).

[27] Gao, Y., Fan, Q., Wang, L., Sun, S. & Yu, X. Molecular dynamics simulation of oxidation growth of ZnO nanopillars. *Comput. Mater. Sci.* **219**, 112008 (2023).

[28] Baguer, N. *et al.* Study of the nucleation and growth of TiO2 and ZnO thin films by means of molecular dynamics simulations. *J. Cryst. Growth* **311**, 4034–4043 (2009).

[29] Barcaro, G., Monti, S., Sementa, L. & Carravetta, V. Modeling Nucleation and Growth of ZnO Nanoparticles in a Low Temperature Plasma by Reactive Dynamics. *J. Chem. Theory Comput.* (2019).

[30] Artrith, N., Morawietz, T. & Behler, J. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B* **83**, 153101 (2011).

[31] Goniakowski, J., Menon, S., Laurens, G. & Lam, J. Nonclassical Nucleation of Zinc Oxide from a Physically Motivated Machine-Learning Approach. *J. Phys. Chem. C* **126**, 17456–17469 (2022).

[32] Behler, J. & Csányi, G. Machine learning potentials for extended systems: a perspective. *Eur. Phys. J. B* **94**, 1–11 (2021).

[33] Behler, J. Four Generations of High-Dimensional Neural Network Potentials. *Chem. Rev.* **121**, 10037–10072 (2021).

[34] Kandy, A. K. A., Rossi, K., Raulin-Foissac, A., Laurens, G. & Lam, J. Comparing transferability in neural network approaches and linear models for machine-learning interaction potentials. *Phys. Rev. B* **107**, 174106 (2023).

[35] Benoit, M. *et al.* Measuring transferability issues in machine-learning force fields: the example of gold–iron interactions with linearized potentials. *Mach. Learn.: Sci. Technol.* **2**, 025003 (2020).

[36] Tallec, G., Laurens, G., Fresse-Colson, O. & Lam, J. Potentials based on linear models. In *Quantum Chemistry in the Age of Machine Learning*, 253–277 (Elsevier, Walthm, MA, USA, 2023).

[37] Togo, A. First-principles Phonon Calculations with Phonopy and Phono3py. *J. Phys. Soc. Jpn.* **92**, 012001 (2022).

[38] Togo, A., Chaput, L., Tadano, T. & Tanaka, I. Implementation strategies in phonopy and phono3py. *J. Phys.: Condens. Matter* **35**, 353001 (2023).

[39] Pusey, P. N. *et al.* Hard spheres: crystallization and glass formation. *Philos. Trans. Royal Soc. A* **367**, 4993–5011

(2009).
[40] Sanz, E. *et al.* Crystallization Mechanism of Hard Sphere Glasses. *Phys. Rev. Lett.* **106**, 215701 (2011).
[41] Trudu, F., Donadio, D. & Parrinello, M. Freezing of a Lennard-Jones Fluid: From Nucleation to Spinodal Regime. *Phys. Rev. Lett.* **97**, 105701 (2006).
[42] Desgranges, C. & Delhommelle, J. Controlling Polymorphism during the Crystallization of an Atomic Fluid. *Phys. Rev. Lett.* **98**, 235502 (2007).
[43] Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool. *Model. Simul. Mater. Sci. Eng.* **18**, 015012 (2009).
[44] Espinosa, J. R., Vega, C., Valeriani, C. & Sanz, E. Seeding approach to crystal nucleation. *J. Chem. Phys.* **144** (2016).
[45] Espinosa, J. R., Vega, C., Valeriani, C. & Sanz, E. The crystal-fluid interfacial free energy and nucleation rate of NaCl from different simulation methods. *J. Chem. Phys.* **142** (2015).
[46] Bai, X.-M. & Li, M. Calculation of solid-liquid interfacial free energy: A classical nucleation theory based approach. *J. Chem. Phys.* **124** (2006).
[47] Knott, B. C., Molinero, V., Doherty, M. F. & Peters, B. Homogeneous Nucleation of Methane Hydrates: Unrealistic under Realistic Conditions. *J. Am. Chem. Soc.* **134**, 19544–19547 (2012).
[48] Pereyra, R. G., Szleifer, I. & Carignano, M. A. Temperature dependence of ice critical nucleus size. *J. Chem. Phys.* **135** (2011).
[49] Sanz, E. *et al.* Homogeneous Ice Nucleation at Moderate Supercooling from Molecular Simulation. *J. Am. Chem. Soc.* **135**, 15008–15017 (2013).
[50] Espinosa, J. R., Sanz, E., Valeriani, C. & Vega, C. Homogeneous ice nucleation evaluated for several water models. *J. Chem. Phys.* **141** (2014).
[51] Zimmermann, N. E. R., Vorselaars, B., Quigley, D. & Peters, B. Nucleation of NaCl from Aqueous Solution: Critical Sizes, Ion-Attachment Kinetics, and Rates. *J. Am. Chem. Soc.* **137**, 13352–13361 (2015).
[52] Bulutoglu, P. S. *et al.* An investigation of the kinetics and thermodynamics of NaCl nucleation through composite clusters. *PNAS Nexus* **1**, pgac033 (2022).
[53] Addula, R. K. R. & Punnathanam, S. N. Molecular Theory of Nucleation from Dilute Phases: Formulation and Application to Lennard-Jones Vapor. *Phys. Rev. Lett.* **126**, 146001 (2021).
[54] Jiang, H., Debenedetti, P. G. & Panagiotopoulos, A. Z. Nucleation in aqueous NaCl solutions shifts from 1-step to 2-step mechanism on crossing the spinodal. *J. Chem. Phys.* **150** (2019).
[55] Iida, Y., Hiratsuka, T., Miyahara, M. T. & Watanabe, S. Mechanism of Nucleation Pathway Selection in Binary Lennard-Jones Solution: A Combined Study of Molecular Dynamics Simulation and Free Energy Analysis. *J. Phys. Chem. B* **127**, 3524–3533 (2023).
[56] Klein, W. & Leyvraz, F. Crystalline Nucleation in Deeply Quenched Liquids. *Phys. Rev. Lett.* **57**, 2845–2848 (1986).
[57] Anstine, D. M. & Isayev, O. Machine Learning Interatomic Potentials and Long-Range Physics. *J. Phys. Chem. A* **127**, 2417–2431 (2023).
[58] Mulliken, R. S. Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. *J. Chem. Phys.* **23**, 1833–1840 (1955).
[59] Bader, R. F. W. Atoms in molecules. *Acc. Chem. Res.* **18**, 9–15 (1985).
[60] Manz, T. A. & Limas, N. G. Introducing DDEC6 atomic population analysis: part 1. Charge partitioning theory and methodology. *RSC Adv.* **6**, 47771–47801 (2016).
[61] Ko, T. W., Finkler, J. A., Goedecker, S. & Behler, J. Accurate Fourth-Generation Machine Learning Potentials by Electrostatic Embedding. *J. Chem. Theory Comput.* **19**, 3567–3579 (2023).
[62] Ko, T. W., Finkler, J. A., Goedecker, S. & Behler, J. General-Purpose Machine Learning Potentials Capturing Nonlocal Charge Transfer. *Acc. Chem. Res.* **54**, 808–817 (2021).
[63] Ko, T. W., Finkler, J. A., Goedecker, S. & Behler, J. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* **12**, 1–11 (2021).
[64] Ewald, P. P. Die berechnung optischer und elektrostatischer gitterpotentiale. *Annalen der physik* **369**, 253–287 (1921).
[65] Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).
[66] Kresse, G. & Furthmuller, J. Efficient iterative schemes for ab initio total energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
[67] Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
[68] Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics* **117**, 1–19 (1995).
[69] Mukhanov, V. A. *et al.* Congruent melting and rapid single-crystal growth of ZnO at 4 GPa. *CrystEngComm* **15**, 6318–6322 (2013).
[70] Schneider, T. & Stoll, E. Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions. *Phys. Rev. B* **17**, 1302–1322 (1978).
[71] Mickel, W., Kapfer, S. C., Schröder-Turk, G. E. & Mecke, K. Shortcomings of the bond orientational order parameters for the analysis of disordered particulate matter. *J. Chem. Phys.* **138** (2013).
[72] Menon, S., Leines, G. D. & Rogal, J. pyscal: A python module for structural analysis of atomic environments. *Journal of Open Source Software* **4**, 1824 (2019). URL `https://doi.org/10.21105/joss.01824`.
[73] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
[74] Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**, 1–22 (1977).
[75] Becker, S., Devijver, E., Molinier, R. & Jakse, N. Unsupervised topological learning for identification of atomic structures. *Phys. Rev. E* **105**, 045304 (2022).
[76] Boattini, E. *et al.* Autonomously revealing hidden local structures in supercooled liquids. *Nat. Commun.* **11**, 1–9 (2020).
[77] Boattini, E., Dijkstra, M. & Filion, L. Unsupervised learning for local structure detection in colloidal systems. *J. Chem. Phys.* **151** (2019).
[78] Coslovich, D., Jack, R. L. & Paret, J. Dimensionality reduction of local structure in glassy binary mixtures. *J. Chem. Phys.* **157** (2022).

[79] Gasparotto, P., Meißner, R. H. & Ceriotti, M. Recognizing Local and Global Structural Motifs at the Atomic Scale. *J. Chem. Theory Comput.* **14**, 486–498 (2018).
[80] Pipolo, S. *et al.* Navigating at Will on the Water Phase Diagram. *Phys. Rev. Lett.* **119**, 245701 (2017).
[81] Reinhart, W. F. Unsupervised learning of atomic environments from simple features. *Comput. Mater. Sci.* **196**, 110511 (2021).
[82] Sarupria, S., Hall, S. W. & Rogal, J. Machine learning for molecular simulations of crystal nucleation and growth. *MRS Bull.* **47**, 949–957 (2022).
[83] Tamura, R. *et al.* Structural analysis based on unsupervised learning: Search for a characteristic low-dimensional space by local structures in atomistic simulations. *Phys. Rev. B* **105**, 075107 (2022).

IX. ACKNOWLEDGMENTS

This study was supported by the French National Research Agency (ANR) in the framework of its "Jeunes chercheuses et jeunes chercheurs" program, ANR-21-CE09-0006. This work was performed using HPC/AI resources from GENCI-[IDRIS/TGCC] (Grant 2021,2022-A0110913010) and using CALMIP (Grant 2021,2022-P21004).

X. AUTHOR CONTRIBUTIONS

C. R. S. performed and analysed the molecular dynamics simulations and developed the approach for structural analysis. A. K. A. K. established the PLIP+Q method and performed the comparison with PLIP. Both C. R. S. and A. K. A. K. contributed equally to the work. J. F. participated in the development of the approach for structural analysis. Q. M. participated in the comparison between PLIP and PLIP+Q. J. G. performed the DFT calculations that were required to establish the database along with the comparison with DFT. J. L. supervised the work and funded all parts of the project. All authors contributed to the writing and reviewing of the manuscript.

XI. COMPETING INTERESTS

The authors declare no competing interests.