All-around local structure classification with supervised learning: The example of crystal phases and dislocations in complex oxides

Jean Furstoss, Carlos Salazar, Philippe Carrez, Pierre Hirel, Julien Lam

All-around local structure classification with supervised learning: The example of crystal phases and dislocations in complex oxides

Jean Furstoss, Carlos Salazar, Philippe Carrez, Pierre Hirel, Julien Lam

All-around local structure classification with supervised learning: The example of crystal phases and dislocations in complex oxides

Jean Furstoss[a,b,*], Carlos R. Salazar[a], Philippe Carrez[a], Pierre Hirel[a], Julien Lam[a]

[a] *Univ. Lille, CNRS, INRA, ENSCL, UMR 8207, UMET, Unité Matériaux et Transformations, F 59000 Lille, France*
[b] *Université de Poitiers, ISAE-ENSMA, CNRS, PPRIME, Poitiers, France*

Abstract

To accurately identify local structures in atomic-scale simulations of complex materials is crucial for the study of numerous physical phenomena including dynamic plasticity, crystal nucleation and glass formation. In this work, we propose a data-driven method to characterize local atomic environments, and assign them to crystal phases or lattice defects. After constructing a reference database, our approach uses descriptors based on Steinhardt's parameters and a Gaussian mixture model to identify the most probable environment. This approach is validated against several test cases : polymorph identification in alumina, and dislocation and grain boundary analysis in the olivine structure.

1. Introduction

The last two decades have seen a spectacular increase of computing power [1], and the rise of efficient, massively parallelized simulation softwares [2]. Both have fueled significant advances in the field of materials modeling where large-scale molecular dynamics (MD) simulations can now include up to hundreds of millions of atoms [3]. This has called for the development of increasingly efficient tools for analyzing simulations, such as central symmetry parameter [4], common neighbor analysis [5, 6], topological cluster classification [7], polyhedral template matching [8], which have been widely applied to face-centered cubic (fcc), body centered cubic (bcc), hexagonal close-packed (hcp), and diamond lattices. All these methods are based on assigning the local bonding network with a specific topological fingerprint.

Meanwhile, more agnostic approaches consists first in computing numerical descriptors able to characterize the local ordering (eg. Behler-Parinello symmetry functions [9], Steinhardt's bond orientational order parameters [10, 11], persistent homological descriptors [12] and effective entropy [13, 14]) and then in determining the structure of unknown atoms based on similarity in descriptor space. Herein, different ways to define and measure similarity have been developed including Euclidian distance [15], kernel metrics [16, 17, 18] and Gaussian mixture model [12]. We note that this set of methods include both supervised learning where a data-set of known structures is used for comparison and unsupervised learning where a clustering method enables for autonomously determining the local structure.

Beyond the characterization of crystal structures, being able to identify extended defects like dislocations represents a further challenging task. For that purpose, dislocation extraction algorithm [19, 20] has been a pivotal tool. Based on common neighbor analysis and a catalog of known slip system in metals-like fcc, bcc or hcp structures, such tool have proved its robustness and efficiency to identify and characterize dislocations microstructure.

A major drawback of all these methods is that they are specifically tailored for specific lattice types and a database of known defects. As a result, adding new types of structures or defects requires significant adjustments of the method. In particular, these tools are poorly suited for studies of defects in complex ceramics. In cubic binary oxides such as MgO or NiO with rock-salt lattice, one alternative to identify and visualize defects consists in applying one (or several) of the above tools to a sub-lattice, e.g. the oxygen sub-lattice. This bypass solution comes at the cost of a lower sampling and hence a decreased accuracy. In more complex materials however, such as compounds with olivine structure, no existing tool is able to identify the lattice itself nor its defects.

In this paper we introduce a new methodology for identifying local ordering, including extended defects, in complex crystalline materials. The method, that we named Steinhardt Gaussian Mixture Analysis (SGMA), consists in a supervised learning approach relying on Steinhardt's bond orientational order parameters in combination with a classical Gaussian mixture model. The accuracy and transferability of SGMA are demonstrated on two compounds of complex symmetry. On the one hand, alumina Al_2O_3 has many important technological applications both in its bulk (corundum α phase) and nanoscale forms [21, 22, 23]. It can exist under various crystalline phases that can not be discriminated with classical meth-

*Corresponding authors: jean.furstoss@univ-poitiers; julien.lam@cnrs.fr

ods of structure identification. In particular, we focus on the melting of α and γ nanoparticles and on the interfacial crystal growth. On the other hand, the olivine lattice can accommodate a wide variety of chemical compositions, notably as the most abundant solid phase of Earth's upper mantle $(Mg, Fe)_2SiO_4$ [24], as well as $LiMPO_4$ materials (M=metal) used as cathodes in Li-battery applications [25]. As such there is a wide interest in the physical and chemical properties of defects in this lattice, but a lack of reliable characterization methods. In the following we demonstrate the ability of SGMA to identify extended defects in olivine systems, with the example of forsterite Mg_2SiO_4.

2. Methodology

Our method is based on a supervised learning approach thus requiring an initial training step with carefully curated model systems. In this work, SGMA is applied to two different types of materials, alumina Al_2O_3 and forsterite which leads to the construction of two databases. All molecular dynamics (MD) and statics (MS) simulations are carried on with LAMMPS [2], and visualization is performed with OVITO [26]. As detailed below, the selected systems used for training are extracted along a linear heating ramps performed within classical thermostat and barostat of Nose-Hoover type as implemented into LAMMPS. It is worth noticing that by purposely sampling the same structures during the heating ramp, we induce a diversity in the structural landscape which is expected to ensure a higher transferability of SGMA especially towards different regimes of temperature and pressure.

2.1. Database for alumina Al_2O_3

Alumina Al_2O_3 is chosen as an archetype complex binary oxide with polymorphic stability crossover when reducing its size from bulk to nanoscale [27]. Our database incorporates four different types of structures, represented in Fig. 1: (a) α crystal or sapphire, which is hexagonal and the stable phase in ambient conditions; (b) metastable fcc-packing lattice γ, which is stabilized at nanoscale due to its lower surface energy [27]; (c) snapshots from liquid state; (d) snapshots of a liquid free surface. All calculations are performed using the interatomic potential developed by Streitz and Mintmire [28], which relies on a variable charge electrostatic potential associated with an embedded-atom method potential. We chose this potential for its ability to successfully model Al_2O_3 at nanoscale [29]. In practice, the liquid state is obtained by first heating sapphire at 10000 K during 500 ps to ensure a complete melting, and then thermalizing at 2000 K during 100 ps. Frozen liquid configurations are obtained by quenching several snapshots from MD through energy minimization. Finally, four of these initial configurations are submitted to a heating ramp, performed in isothermal-isobaric (NPT) ensemble, up to 2500 K during 1 ns. For the two crystal phases, we

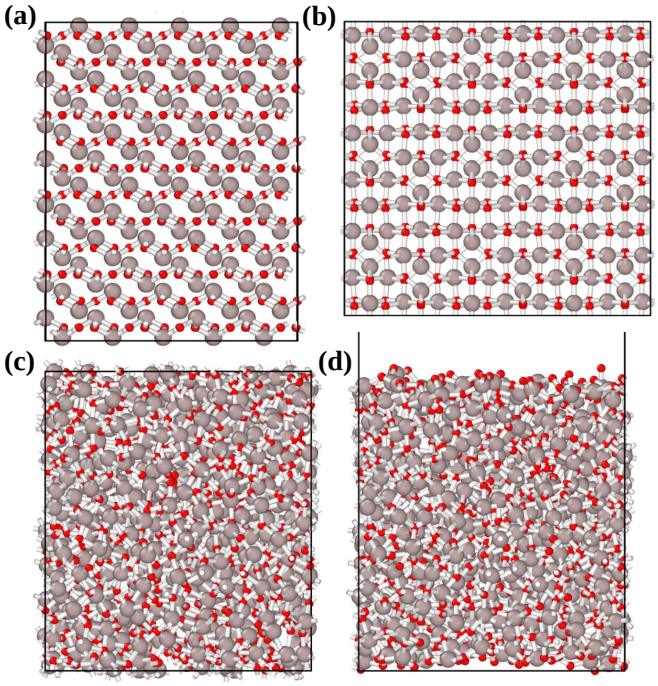

Figure 1: Illustration of the different structures considered in the aluminium oxyde database, (a) α phase, (b) γ phase, (c) Liquid regime and (d) Liquid free surface with a vacuum buffer of 25 Å.

proceed with the same heating ramp. The atomic configurations incorporated in the database are randomly selected during this last stage. The number of different reference points in the database was increased until achieving an accurate identification of defects in the different test cases (see Appendix A). At the end for Al_2O_3 system, our database contains 21 α and γ crystalline configurations, 21 liquid states and 21 free surfaces.

2.2. Database for forsterite Mg_2SiO_4

As mentioned in the introduction, forsterite Mg_2SiO_4 is chosen in this study to highlight the efficiency of SGMA method to identify extended defects. For that purpose, we include in the database the following configurations, as presented in Fig. 2: (a) perfect crystal of forsterite with various applied elastic strains; (b) a high-angle $60°//[100](011)$ tilt grain boundary (GB) taken from [30]; (d,e) infinite straight screw dislocations of Burgers vectors (d) [100] and (e) [001]. It is worth noticing that the [100] and [001] dislocations are the only possible intracrystalline dislocations in Mg_2SiO_4 [31, 32] making our database complete regarding slip systems in forsterite. Calculations are carried out using the rigid-ion potential developed by Pedone and co-workers [33], which includes long-range Coulomb interactions, and where a Morse function mimics short-range interactions. This interatomic potential accurately reproduces the physical properties of forsterite crystal [34], and was used recently to model tilt and twist GB in forsterite [30]. Sampling of the atomic systems is achieved by selecting atomic configurations at random during linear heating

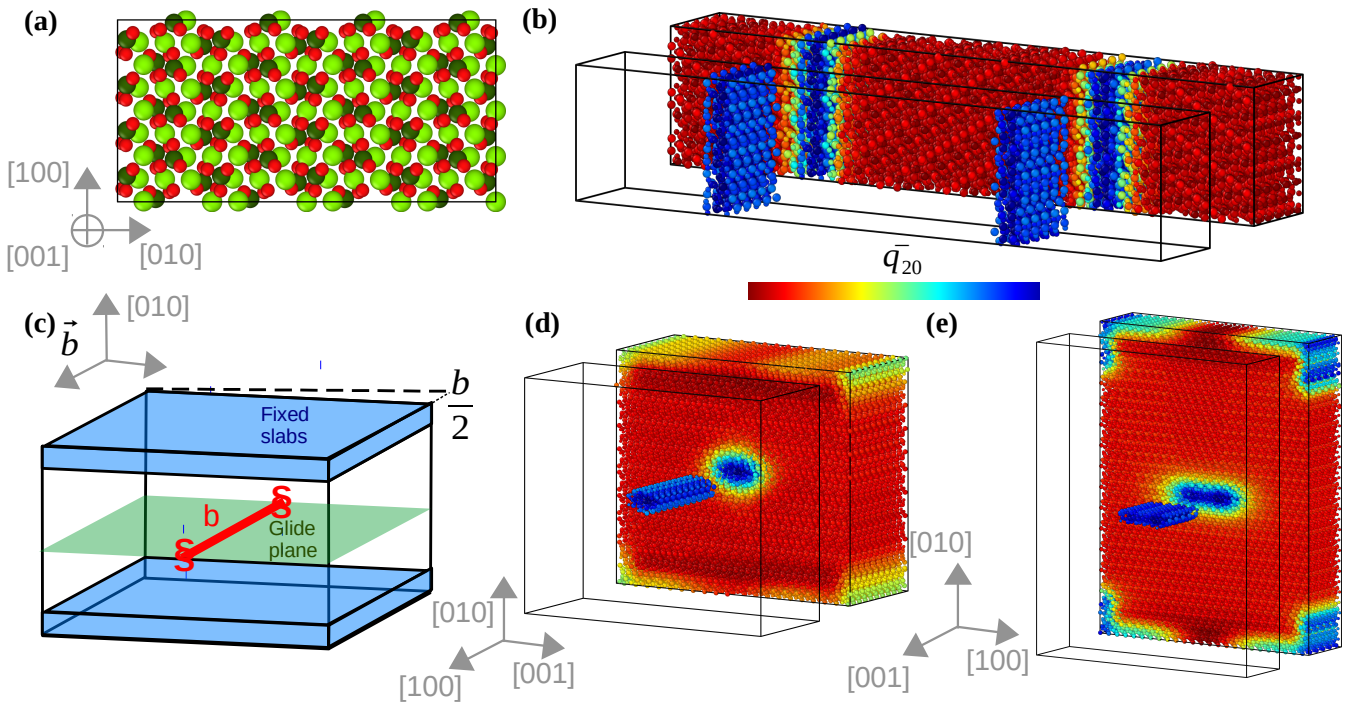

Figure 2: Illustration of the different structures considered in the forsterite database, (a) Defect-free crystal phase, (b) Grain boundary system, (c-e) [100] and [001] dislocations along with the schematic representation.

up to 800 K for 10 ps. When searching the number of point in the database needed for an accurate identification of the different test cases, we found that the Mg_2SiO_4 system requires much larger database than the Al_2O_3 one (see Appendix A). For the perfect crystal, 98 configurations are included into the database, corresponding to various elastic deformations, up to 10% for simple shear ϵ_{ij}, and up to 15% in uniaxial shear ϵ_{ii}. For the GB, 251 configurations are extracted from the heating MD simulation. Finally, 501 screw dislocations configurations are extracted. A so-called periodic cluster approach is used, where the system is periodic along the dislocation line, and top and bottom layers are maintained fixed (see e.g. [35, 36]). A single screw dislocation of Burgers vector $\mathbf{b} = [100]$ or [001] is introduced in the middle of the cell, as depicted in Fig. 2c. To substantially increase the sampling of atomic arrangements along the dislocation lines, the line length is extended to $8b$.

2.3. Descriptors and Gaussian mixture model

Fig. 3 outlines the steps followed by SGMA. Once the database is constructed, each structure must be encoded into fingerprints respecting physical symmetries of rotation, translation and permutation invariances. In SGMA we choose to work with Steinhard's parameters. In comparison with other types of structural descriptors, like Behler Parinello descriptor [9] or the Smooth Overlap of Atomic Positions [37] developed especially for the design of machine-learning interaction potentials, Steinhard's parameters are physically more sensible and offer a good compromise between completeness and computational cost. Moreover, they have been widely employed for the identification of crystal and liquid structures [38, 11, 10, 15]. In this work, we employ the averaged version of these parameters, as originally introduced by Lechner and Dellago [10]:

$$\bar{q}_l^i = \sqrt{\frac{4\pi}{2l+1}\sum_{m=-l}^{l}\left|\frac{1}{|N^i(r_c)|}\sum_{k=0}^{N^i(r_c)} q_{lm}^k\right|^2} \tag{1}$$

where the sum over k counts ion i itself and its neighbors inside the cutoff radius r_c (represented by $N^i(r_c)$), and q_{lm}^i represent the non-averaged Steinhardt's parameters defined as follows:

$$q_{lm}^i = \frac{1}{|N^i(r_c)|}\sum_{j=1}^{N^i(r_c)} Y_{lm}(\theta_{ij}, \phi_{ij}) \tag{2}$$

where Y_{lm} are the spherical harmonics, and θ_{ij} and ϕ_{ij} are the colatitude and azimuthal angles between ions i and j, respectively. The cutoff radii are chosen with respect to the radial distribution functions of the different crystals. Their values allow to enclose the main peaks of these functions. Then, the cutoff radius is set to $r_c = 5$ Å in Al_2O_3 and $r_c = 8$ Åin Mg_2SiO_4. For the values of the spherical harmonics degrees l, they are chosen after increasing value of l and until the different structures do not strongly overlap in the descriptor space. This step is

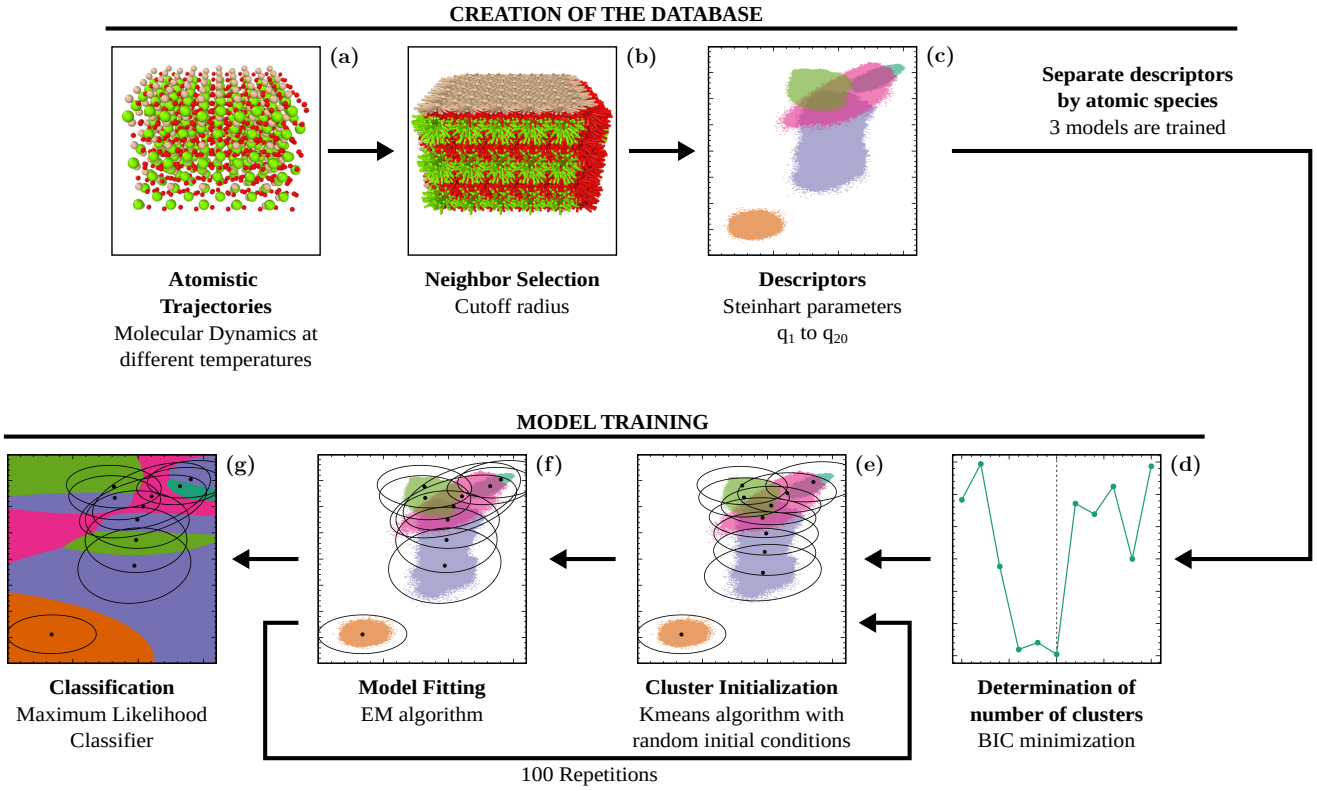

Figure 3: Schematic of the steps involved in SGMA. A database is created from (a) atomistic trajectories obtained using Molecular Dynamics. (b) The neighbors of each atom are found using a cutoff radius of 8 Å. (c) The Steinhardt parameters are then computed for all atoms in each snapshot of the trajectory. Multiple Gaussian Mixture Models are then trained on the points in the database corresponding to each atom species. (d) For each model, the number of Gaussian clusters to be used is determined by minimizing the BIC. (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 and on new test structures.

helped by graphical tools we developed in order to visualize database points (see example in Appendix C), which are provided in the release of the code. Here, the averaged Steinhardt parameters are computed with l up to 12 and 20 respectively for Al_2O_3 and Mg_2SiO_4 systems respectively, meaning that 12 or 20 numerical values per atom are used for training. It is worth mentioning that for the studied systems, using only two Steinhardt's parameters, such as the couple q_4/q_6 that is widely employed in simpler crystals, is not sufficiently informative and discriminant (see Appendix B). As such, a more extensive sampling is necessary for robust identification, due to the complexity of structural landscapes. Furthermore, while we can simply use the whole set of atoms for crystalline and liquid structures, a prior selection of atoms stored in the database for dislocations and grain boundary structures is performed based on the value of $\bar{q}_{20}$ (shown in Fig.2b, d and e). Finally, the database is divided per atomic species for which a separate classifier is trained leading to 2 and 3 distinct databases for Al_2O_3 and Mg_2SiO_4, respectively. This latter point enhances the quality of the structural analysis by allowing each sub-lattices to have different local atomic environments associated to a given structure.

These descriptors $\vec{q}_l^{\,i}$ are then transferred into a Gaussian Mixture Model (GMM), which is a machine learning clustering method, as implemented in the Python library SCIKIT-LEARN [39]. The unknown parameters of the GMMs are iteratively estimated using the Expectation-Maximization algorithm [40] with full covariance matrices and 100 k-means initializations. We note that applying the GMM does not necessary require that each structure in the database follows a unique Gaussian distribution. As such, the ideal number of Gaussian distributions must be estimated for each GMM. For that purpose, multiple models are trained with different numbers of Gaussian clusters, and the number of clusters that minimize the Bayes Information Criterion (BIC) [41] is finally chosen.

2.4. Softly-labeled Gaussian Mixture Model and classification

The NVT simulations for all structures present in the databases that we perform to obtain statistical diversity are conducted at temperature and pressure conditions where the structure is known to be stable, and the label on the data points is then conserved. During the simulations there may be atoms in a different environment to the rest,

but assuming that this happens rarely it will not be of statistical importance when performing the Gaussian fitting. In the case of the dislocations or GB, since they are very localized structures, a preliminary selection is made using the descriptors as tools to spatially locate the core of the defects (see section 2.2).

After computing the descriptors we then obtain clouds of data points with a certain label. On this database we perform the Gaussian fitting, which is not label aware. However, since we assigned labels to the original data point clouds, we can compare the obtained Gaussian clusters and assign them the label of the cloud they are covering. In this way we are also able to assess if our clustering was successful and all the database structures are represented.

Once the GMM are trained and labeled, classification is performed using the Maximum Likelihood Classifier (MLC) where the probability of an object x_i to belong to class ω_k is computed as:

$$p(\omega_k|\mathbf{x}_i) = \frac{\alpha_k \mathcal{N}(\mathbf{x}_i|\mathbf{m}_k, \mathbf{C}_k)}{\sum_{j=1}^{K} \alpha_j \mathcal{N}(\mathbf{x}_i|\mathbf{m}_j, \mathbf{C}_j)} \tag{3}$$

with α_k designating the mixture proportions, and $\mathbf{m}_k$ and $\mathbf{C}_k$ being 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 to belong to one of the structures contained in the database. Then probabilities are compared, and an atom is considered as belonging to the structure with the highest probability. As MLC generally provides near-100% values, we consider the identification unreliable if it falls below 95%, and exclude it during visualization. This misidentification may have different causes. First, it might reflect an important deviation from the local atomic environments present in the database. Second, it could be the results of the overlapping of the different structures in the descriptor space leading for local atomic environment in the overlap region to a reduced MLC. Finally, it could reflect that the local atomic environment associated to a low MLC does not represent any of the structures present in the database. In this case, that we do not believe to occur in our systems, the given local atomic environment can be isolated, further studied, and added in the database before fitting again the GMM to include the new structure in the SGMA. In the following section, the percentage of ions associated with MLC lower than 95% are provided for each test cases.

In closing, previous structure identifications also employed Gaussian Mixture Models [42, 43, 44]. In comparison, SGMA brings three innovative features. First, our atomic fingerprints are based on Steinhardt's parameters which are physical parameters, while the other works chose to employ machine-learning oriented structural descriptions (namely neural-network encoder and persistent homology). This enables for a higher control on the stability of the method and a lower computational cost. Second, we use a softly-labeled database instead of unsupervised learning, which allows to pinpoint the classified structures. And third, we rely on separate GMMs for each atom type, which proves to be crucial for complex chemistry systems.

A release of the code also containing extended documentation for retrieving the results presented in the article as well as using the method in other complex materials, is freely available at https://github.com/JeanFurstoss/AtomHIC/rele This C++ oriented object program uses OpenMP parallelization to efficiently compute Steinhardt's parameters as shown in Appendix D.

3. Results and discussion

3.1. Melting of alumina nanoparticles

We begin with the characterization of alumina nanoparticles. In particular, two different nanoparticles made of 5000 atoms are constructed by spherically cutting from a bulk of α and γ crystal phases. Molecular dynamics simulations are carried out in the NVT ensemble during 1 ns while increasing temperature from 2000 K to 3000 K.

After MD simulations, SGMA is used to identify local environments, and associate each atom with a phase from the database, i.e. crystalline α or γ phases, liquid state, or free surface. Fig. 4(a,b) displays snapshots where atoms are colored according to their identified environment. One can observe that the method correctly characterizes both crystal phases even at high temperature with large thermal noise. As expected, melting occurs from free surfaces, as evidenced by violet nuclei, and progressively propagates through the whole particle. Final snapshots (t = 650 ps) correspond to fully liquid particles, and atoms near free surfaces are still identified as such.

The onset of melting is usually characterized by an abrupt change of the system's internal energy. Fig. 4(c,d) shows the evolution of the internal energy during the simulation (black curves) with a rapid increase of energy occurring around 600 ps. SGMA also provides a count of atoms in each of the identified structures. Fig. 4(c,d) shows the temporal evolution of the number of atoms N_X associated with each phase, and normalized to the total number of atoms N_{tot}. We observe that the number of atoms belonging to the surface (green curves) remains constant throughout the simulation. The number of atoms in a crystalline phase (red or blue curves), initially large, shows a rapid decrease around 400−600 ps, associated with a rapid increase of atoms belonging to the liquid state (violet curves). Thus a good correlation is found between the energy increase and the atom count, which demonstrates the validity of our approach.

3.2. Phase growth in alumina

Now we model the interface between two crystal phases α and γ, submitted to a temperature ramp from 2000 to 3000 K as before. In the whole simulation box, the number of atoms is equal to 5520. Again after MD simulations,

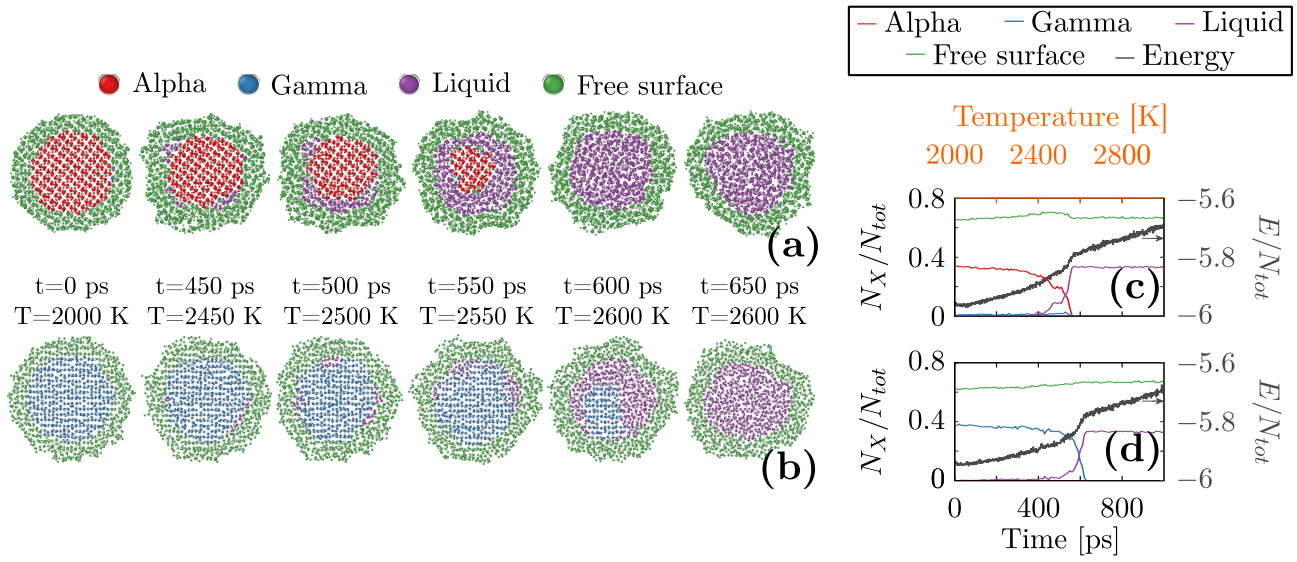

Figure 4: (a-b) Slices of Al_2O_3 nanoparticles during melting simulations. Atoms are coloured according to their environment identified by SGMA : α (red), γ (blue), liquid (violet), and free surface (green). (c-d) Temporal evolution of internal energy per atom (E/E_{tot}, black curve) and number of atom in each phase (N_X/N_{tot}) : α (red curve), γ (blue curve), liquid (violet curve), free surface (green curve).

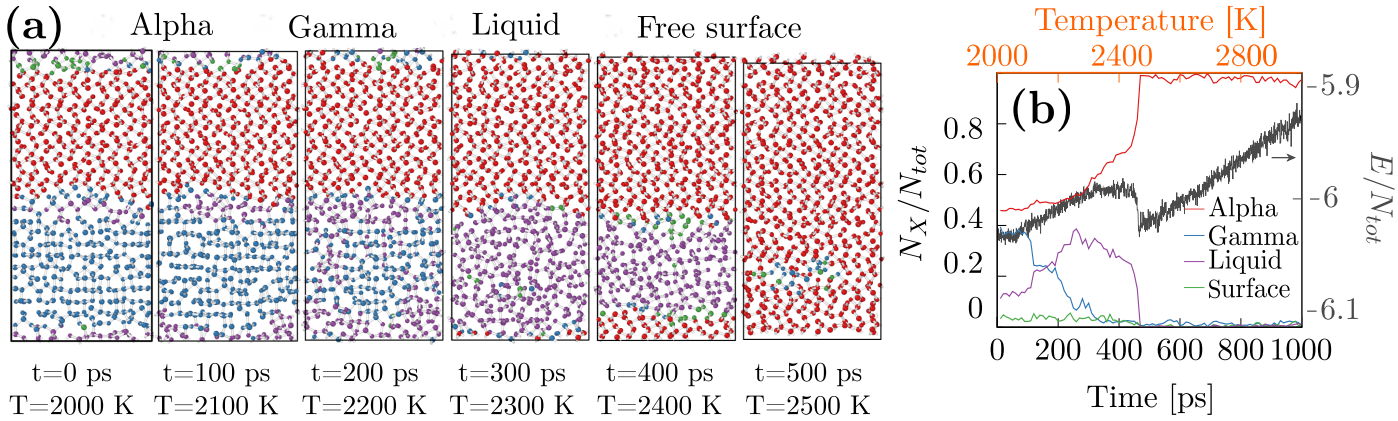

Figure 5: (a) Slice of the $\alpha//\gamma$ interface during melting. Coloring is obtained by the structural analysis methodology. (b) Corresponding temporal evolution of the different structure proportion and of the energy per number of atom.

SGMA is applied to identify local environments. Fig. 5(a) illustrates the obtained identification. Initially both α and γ phases are correctly identified. Throughout the simulation we observe melting of the γ phase, followed by crystallization into the α phase. At the end of the simulation ($t > 500$ ps) the system contains a single crystal of α alumina.

Fig. 5(b) shows the internal temporal evolution of the internal energy. We observe a drop in the energy around 450 ps which corresponds to the liquid to α phase transition. Again SGMA provides a count of atoms in each phases, represented in Fig. 5(b) (coloured curves). At $t = 0$ ps most atoms are identified as belonging to either the α or γ crystal, with only few atoms assigned to free surface. Between 150−300 ps the number of atoms in γ phase decreases sharply, while the number of atoms in liquid increases, consistent with what was observed before. Then around 450 ps, the number of atoms in liquid decreases sharply to the point of vanishing, while almost 100% of atoms are identified as belonging to the α phase. We note that some atoms at the interface are wrongly identified as belonging to a free surface (atoms in green in Fig. 5(a)). However this artefact does not hinder visualization, and involves only a very small number of atoms (green curve in Fig. 5(b)).

In this simulation, the observed melting of the γ phase and a subsequent recrystallization into the α phase is consistent with experimental results[45]. Yet, in experiments, the phase transformation occurs at a lower temperature regime. Such difference can be attributed to a flaw in the employed force field or to the employed heating ramp that is much faster than in experimental conditions. It remains that our simulation still serves the purpose of showing the capabilities of SGMA in complex multi-phase systems.

3.3. Dislocations in forsterite

We now proceed with the study of forsterite Mg_2SiO_4 with olivine lattice. As a first benchmark, we test our method on a bulk crystal of forsterite containing four dislocations. The dislocations are of pure screw character with straight lines, with Burgers vectors $\pm$[001] in a quadrupolar arrangement. After initial relaxation, the system is equilibrated at 800 K and 0 GPa during 50 ps in the NPT ensemble. Then a constant shear strain rate of 10^8 s^{-1} is applied for 1 ns to promote glide in (010) planes. Because such a configuration of dislocations is unstable under stress, dislocations of opposite Burgers vectors glide in opposite directions.

Snapshots of the MD simulation are processed with SGMA. To validate results, we also compute the disregistry functions ϕ [46] in the (010) planes containing the dislocations, as well as their derivatives $d\phi/dx$ along the glide direction. As Mg and O sub-lattices have respectively 2 and 3 different crystallographic sites in perfect crystal, we compute the disregistry function using only the Si sublattice. Fig. 6 shows the resulting SGMA results, where atoms in perfect crystal are hidden for sake of clarity, along with the derivative of disregistries. In the initial state, the four dislocations are well identified by SGMA, at positions that coincide with those obtained from disregistry functions. The extensions of the cores also seem consistent, with a width of approximately 10 Åinferred from the disregistries. As shear is increased and dislocations glide, SGMA successfully identifies and tracks the [001] dislocations, with positions that match those obtained from disregistries but with higher spatial extension in the glide plane and in the direction of motion. This discrepancy with the dislocation core spreading from disregistry could be explained by the fact that the disregistries are computed considering Si sub-lattice only while the core of the moving dislocations might be more extended in the Mg and O sub-lattices. Eventually as dislocations of opposite Burgers vectors cross periodic boundaries and meet, they annihilate and leave only perfect crystal. Again SGMA correctly recognizes that there are no more dislocation in the system, and assigns all atoms to perfect crystal environment, even though it is highly strained ($\epsilon_{yz} = 0.18$, see Fig. 6d). It is worth noticing that despite the elastic field of dislocations that causes long-range distortions of the lattice [47], the method still recognizes defect-free crystal.

3.4. Dislocation loops in forsterite

Our second test related to forsterite Mg_2SiO_4 deals with dislocation loops with Burgers vector [100]. A glissile circular loop is introduced in a periodic crystal of forsterite in both the (010) and the (001) planes. Such a glissile loop is unstable and would collapse due to line tension and strong attraction between opposite loop segments. To circumvent that, shear stress σ_{xz}=2 GPa is applied. Molecular dynamics is then performed at 800 K for 50 ps.

Fig. 7(b,c) illustrates the identification of these loops following SGMA. Again disregistry is computed in two planes passing through the loops for comparison. In both cases, SGMA is able to identify atoms in the vicinity of dislocations, at positions that match those obtained from disregistries. Remarkably, edge and mixed components are well captured, even though the database included only dislocations of screw character. This highlights the robustness of the method to assign the correct Burgers vector to a dislocation, no matter its character, and even when not included in the database. Thanks to SGMA, visualization clearly shows that the loops in two different planes adopt different curvatures. In (010) the loop seems circular, while in (001) it is almost square shaped. This is due to higher lattice friction in (001), which favors straight dislocation lines with perfect edge or screw characters [36, 48].

In the case of the loop in (001), some atoms are wrongly identified as belonging to a dislocation with [001] Burgers vector (blue atoms in Fig. 7b). Moreover, a relatively high number of ions are associated with MLC lower than 95% (i.e. 19.2%). This is not unexpected given that the system is highly distorted. This may also be caused by a system size effect, dislocation segments being very close to the boundaries and therefore to periodic images of the loop. Again this artifact is not crippling as it affects only a few atoms, and therefore does not prevent efficient visualization of defects.

3.5. Twist grain boundary in forsterite

Grain boundaries are another important type of defects present in materials. Twist GB are a particular type of GB where two grains share equivalent surfaces and are twisted around the axis normal to their interface. When the twist angle is low (typically below 15°), such GB relaxes into a network of intersecting screw dislocations [49]. Hence they offer a good testing ground for our identification method.

For that purpose, we use atomic configurations of a (010) twist GB in forsterite with disorientation angle of 4°, that we published earlier [50]. After thermalization at 800 K and 0 GPa during 50 ps in the NPT ensemble, we apply simple shear with a rate equal to 10^9 s^{-1} in a direction maximizing resolved stress in the (010)[001] slip system. Fig. 8.a displays the results of our SGMA analysis, where atoms belonging to perfect crystal are hidden. The method clearly identifies two sets of dislocations, those with [100] Burgers vector (in red), and those with [001] Burgers vector (in blue). As shear is increased, SGMA correctly tracks the movement of the [001] dislocations along the GB plane. Despite the elevated temperature, large applied deformation, and the complex dislocation junctions, the method successfully discriminates between perfect crystal and dislocation environments, with close to none mislabeling.

3.6. Polycrystalline forsterite

More generally, grain boundaries in polycrystalline materials correspond to a wide distribution of disorientations between adjacent grains. Here we constructed a 2D polycrystal consisting of tilt GB disoriented around the [100]

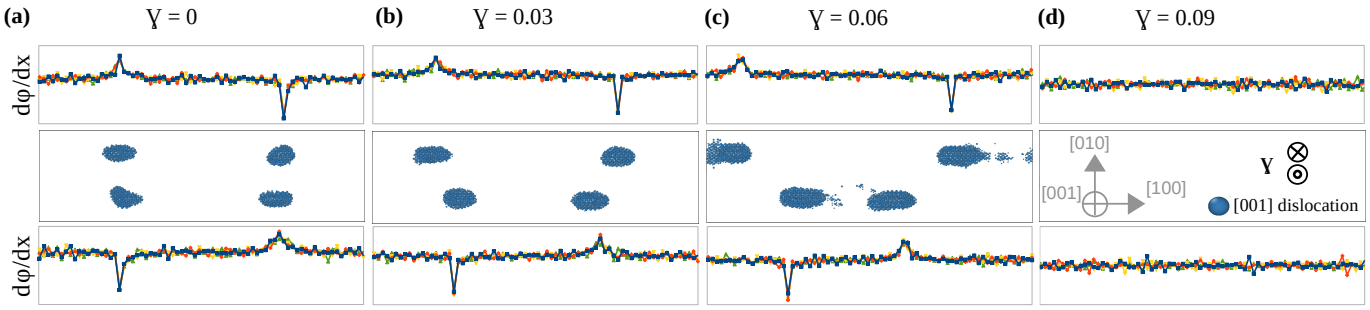

Figure 6: [001] screw dislocation quadrupole (a) sheared at a rate equal to 10^8 s^{-1} at 3 (b), 8 (c) and 9 % (d) of finite shear strain, the curves represents the spatial derivative of disregistry functions computed for the upper and lower dislocation dipoles considering only the silicon sub-lattice. The different colors in the disregistry curves represent the function along the dislocation lines. Atoms identified as the defect-free forsterite are hidden for visual clarity. In this systems about 6.4% of ions are associated with a MLC lower than 95%.

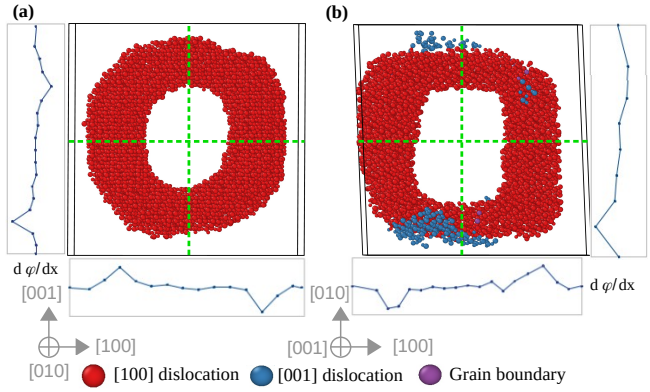

Figure 7: [100] dislocation loop respectively in (010) (a) and in (001) (b) planes after 50 ps thermalization, the blue curves represent the derivative of disregistry functions along the green lines. Atoms identified as the defect-free forsterite are hidden for visual clarity. About 6.2% and 19.2% of ions are associated with a MLC lower than 95% in (a) and (b), respectively.

axis using the atomsk software [51]. In order to ensure the electroneutrality of the system, we remove ions having very closed neighbors (typically at the grain boundaries) until retrieving the stoichiometry of Mg_2SiO_4. The obtained polycrystal is then relaxed at 0K and analyzed using the SGMA, as presented in Fig. 9. While the majority of the ions in the vicinity of GB are well identified by the method, a non-negligible part of GB ions are identified as belonging to [100] dislocation structure. This mislabeling can originate from the structural proximity of [100] tilt grain boundaries with [100] screw dislocation. Nevertheless, only one complexion of Mg_2SiO_4 tilt GB was used for the construction of the database and the SGMA successfully identified GB structurally different from this complexion. In order to improve the identification, other type of GB might be considered in the database to account for the structural diversity of GB in such type of materials.

4. Conclusion and perspectives

We present an innovative method named Steinhard Gaussian Mixture Analysis (SGMA), based on the combination of physically driven bond-order parameters for structural fingerprinting and Gaussian mixture model for classification. We demonstrated the ability of the method to characterize local environments in binary and ternary oxides. In alumina the different polymorphic phases are accurately identified, even during high-temperature molecular dynamics simulations. In forsterite, the method can discriminate between dislocations with different Burgers vectors and track them during high-temperature deformation. Planar defects such as general grain boundaries can also be monitored. A key feature of the method is that atoms can still be correctly assigned to defective regions, even if the defect was not explicitly included in the database, if the temperature is high, or if the system is highly distorted, which makes it very robust.

This novel methodology paves the way for the study of extended defects in crystals with complex chemical composition or low-symmetry lattice. We hope it will be useful to the community for analyzing and visualizing simulations of deformation or diffusion in complex materials, with an efficiency comparable to long-time existing tools developed for metals or semiconductors. Meanwhile, because the same tool also allows for probing phase transformations even in complex systems, we anticipate that it will open up new avenues for the investigation of mechanisms such as crystal nucleation.[52, 53] along with amorphization[54, 55]. In particular, innovative collective variables employed to force the crossing of free energy barriers in rare-event sampling could be developed based on SGMA[56, 57, 58].

Acknowledgements

J.F., P.H. and P.C. acknowledge funding from the French government through the Programme Investissement d'Avenir (I-SITE ULNE / ANR-16-IDEX-0004 ULNE) and from

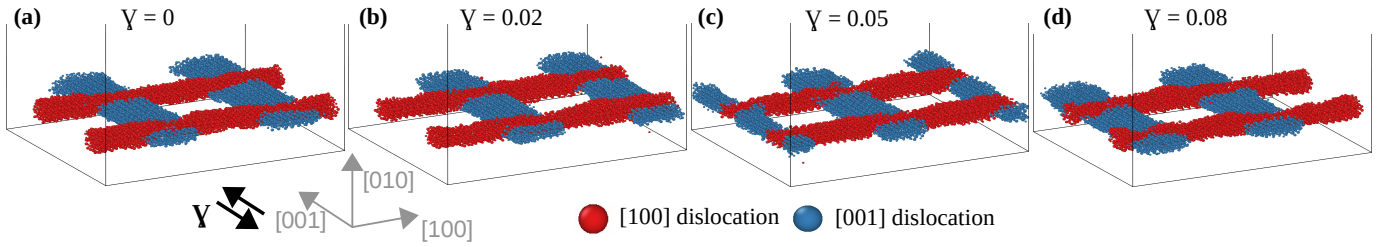

Figure 8: 4° low-angle twist (010) grain boundary (a) sheared at a shear rate equal to 10^9 s^{-1} at 2 (b), 5 (c) and 8 % (d) of finite shear strain. In this figure, atoms identified as the defect-free forsterite are hidden for visual clarity. In this systems about 5.6% of ions are associated with a MLC lower than 95%.

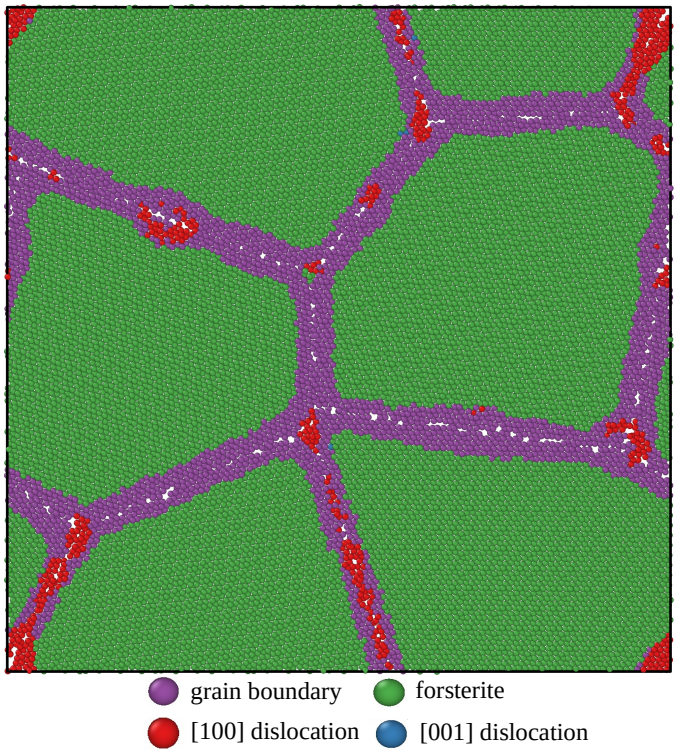

Figure 9: Structural analysis of a Mg_2SiO_4 polycrystal composed of 4 grains randomly disoriented around the [100] axis and relaxed at 0K. In this system about 3.7% of ions are associated with a MLC lower than 95%.

the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 787198 – TimeMan under the project name NuMoGO. C.R.S. and J.L. acknowledge funding from the French government through the Programme Investissement d'Avenir and the French National Research Agency (ANR) (NanoX / ANR-17-EURE-0009 and ANR-21-CE09-0006). Computational resources have been provided by the DSI at Université de Lille, CALMIP, Jean-Zay and TGCC.

Data Availability Statement

The method is implemented in the homemade code AtomHIC that is freely available at: https://github.com/JeanFurstoss/AtomHIC/releases

Appendix A. Number of point in database

As no generic method exists to determine the number of datapoint needed for a complete representation of the different structure in the descriptor space, we adopt here an incremental approach. By progressively increasing the number of point in the database, we compare the results obtained by the SGMA method on the different test cases presented section 3. For the Al_2O_3 system, a relatively small number of datapoint was need to achieve an accurate identification of the different structures. In comparison, the Mg_2SiO_4 system requires much more points in the database to identify correctly the different local atomic environments. As an example, Fig.A.10 represents the results of the SGMA method for the [100] dislocation loop in (001) plane (see section 3.4) test case of the Mg_2SiO_4 system and for different database with increasing number of point. It is clearly shown that an accurate identification of all part of the dislocation loop requires a database with a high number of points.

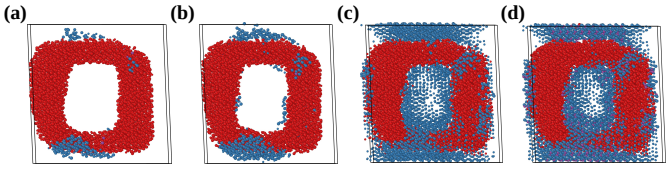

Figure A.10: [100] dislocation loop in (001) plane (see section 3.4) identified with the SGMA method with different size of the database for GMM training. For the largest database (used for all cases presented in the main article) with 1.4 millions of ions (a), with 700,000 (b), 350,000 (c) and 100,000 ions (d).

Appendix B. The limitation of the classical Q_4/Q_6 structural analysis

A classical analysis in ordered or disordered materials lies with the separation of different structures or environ-

ments in the Q_4/Q_6 subspace of Steinhardt parameters [59]. Relying on the computation of only 2 Steinhart parameters, such a method may appear computationally efficient. However, The Q_4 and Q_6 values do not necessarily allow to discriminate the various structures, specially in complex materials such as those investigated here (as shown Fig.B.11). Therefore, it justifies the current development of the SGMA method.

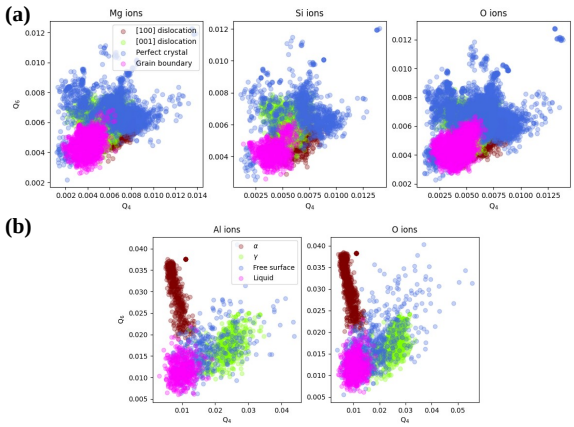

Figure B.11: Q_4 and Q_6 values for Mg_2SiO_4 (a) computed with l = 20 and $r_c = 8$ Åfor Mg (left), Si (center) and O (right) ions and for Al_2SiO_4 (b) computed with l = 12 and $r_c = 5$ Åfor Al (left) and O (ions), for the different structures considered in this work.

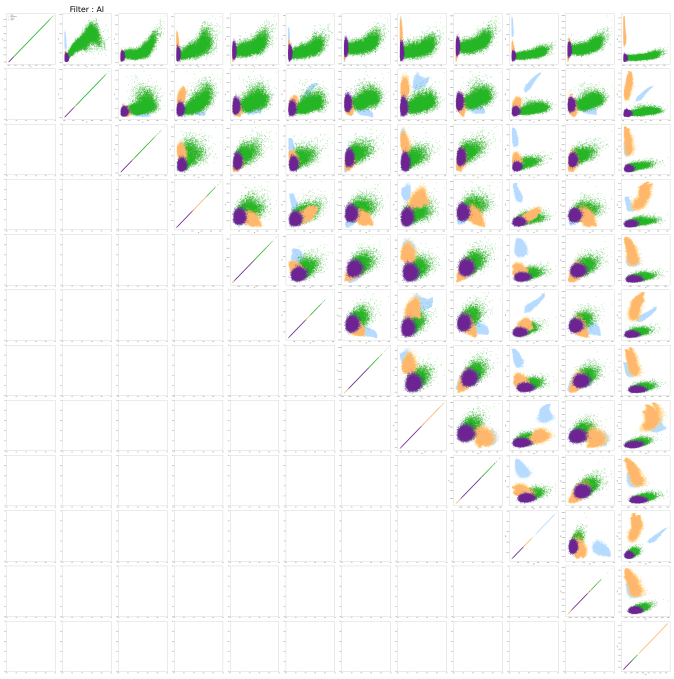

Figure C.12: Example of a matrix-type plot representing 2D slices in the descriptor space for Al ions in the Al_2O_3 systems. Colors refer to the different structures considered for this system. Such type of plot can be generated using a python script provided with the AtomHIC code.

Appendix C. Choice of spherical harmonics degree

The choice of the spherical harmonics degree l for the computation of the average Steinhardt's parameters is done for effectively separating the different structures in the descriptor space. This is helped by visualizing the data through a matrix-type plot representing 2D slices in the descriptor space. An example of such type of plot is given in Fig.C.12 for the Al ions in the Al_2O_3 system studied here. The python script used to generate such kind of plot is provided in the release of the AtomHIC code (see documentation in https://github.com/JeanFurstoss/AtomHIC/rele

Appendix D. Computational cost

The main computational cost of the SGMA lies with the computation of the Steinhardt parameters and the fitting of the GMM. In fact the classification, which is done through the computation of the Maximum Likelihood Classifier is numerically very efficient. Thus, as the GMM fitting is only performed once, the most time consuming part of the SGMA is the computation of the descriptors themselves. The AtomHIC C++ code (https://github.com/JeanFurstoss/AtomHIC/releases), developed and used for all steps of SGMA method includes optimized neighbor research and OpenMP parallelization. Figure D.13 shows typical computational time for the computation of the Steinhardt parameters as function of the number of atoms, the degree of the parameters and the cutoff radius. As shown by these results, the algorithm complexity is $\mathcal{O}(N, l^2, r_c^3)$ where N is the number of atom, l is the degree of the Steinhardt parameters and r_c is the cutoff radius. The parallelization efficiently reduces the computational cost by a factor almost equal to the number of threads used for the calculation.

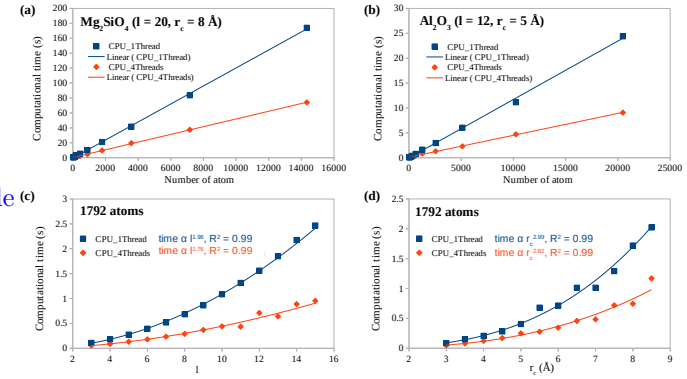

Figure D.13: Computational times as a function of the number of atom for Mg_2SiO_4 (a) and Al_2O_3 (b) systems, the degree l of the Steinhardt parameters (c) and the cutoff radius r_c (d).

References

[1] T. M. Conte, E. Track, E. DeBenedictis, Rebooting computing: New strategies for technology scaling, Computer 48 (12) (2015) 10–13.

[2] S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, Journal of computational physics 117 (1) (1995) 1–19.

[3] L. A. Zepeda-Ruiz, A. Stukowski, T. Oppelstrup, V. V. Bulatov, Probing the limits of metal plasticity with molecular dynamics simulations, Nature 550 (7677) (2017) 492–495.

[4] C. L. Kelchner, S. Plimpton, J. Hamilton, Dislocation nucleation and defect structure during surface indentation, Physical review B 58 (17) (1998) 11085.

[5] J. Dana. Honeycutt, H. C. Andersen, Molecular dynamics study of melting and freezing of small Lennard-Jones clusters, J. Phys. Chem. 91 (19) (1987) 4950–4963. doi:10.1021/j100303a014.

[6] D. Faken, H. Jónsson, Systematic analysis of local atomic structure combined with 3D computer graphics, Comput. Mater. Sci. 2 (2) (1994) 279–286. doi:10.1016/0927-0256(94)90109-0.

[7] A. Malins, S. R. Williams, J. Eggers, C. P. Royall, Identification of structure in condensed matter with the topological cluster classification, J. Chem. Phys. 139 (23) (Dec. 2013). doi:10.1063/1.4832897.

[8] P. M. Larsen, S. Schmidt, J. Schiøtz, Robust structural identification via polyhedral template matching, Model. Simul. Mater. Sci. Eng. 24 (5) (2016) 055007. doi:10.1088/0965-0393/24/5/055007.

[9] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, J. Chem. Phys. 134 (7) (Feb. 2011). doi:10.1063/1.3553717.

[10] W. Lechner, C. Dellago, Accurate determination of crystal structures based on averaged local bond order parameters, J. Chem. Phys. 129 (11) (Sep. 2008). doi:10.1063/1.2977970.

[11] P. J. Steinhardt, D. R. Nelson, M. Ronchetti, Bond-orientational order in liquids and glasses, Phys. Rev. B 28 (2) (1983) 784–805. doi:10.1103/PhysRevB.28.784.

[12] S. Becker, E. Devijver, R. Molinier, N. Jakse, Unsupervised topological learning for identification of atomic structures, Phys. Rev. E 105 (4) (2022) 045304. doi:10.1103/PhysRevE.105.045304.

[13] P. M. Piaggi, O. Valsson, M. Parrinello, Enhancing Entropy and Enthalpy Fluctuations to Drive Crystallization in Atomistic Simulations, Phys. Rev. Lett. 119 (1) (2017) 015701. doi:10.1103/PhysRevLett.119.015701.

[14] P. M. Piaggi, M. Parrinello, Entropy based fingerprint for local crystalline order, J. Chem. Phys. 147 (11) (Sep. 2017). doi:10.1063/1.4998408.

[15] J. Goniakowski, S. Menon, G. Laurens, J. Lam, Nonclassical Nucleation of Zinc Oxide from a Physically Motivated Machine-Learning Approach, J. Phys. Chem. C 126 (40) (2022) 17456–17469. doi:10.1021/acs.jpcc.2c06341.

[16] A. M. Goryaeva, C. Lapointe, C. Dai, J. Dérès, J.-B. Maillet, M.-C. Marinica, Reinforcing materials modelling by encoding the structures of defects in crystalline solids into distortion scores, Nat. Commun. 11 (4691) (2020) 1–14. doi:10.1038/s41467-020-18282-2.

[17] P. M. Piaggi, M. Parrinello, Calculation of phase diagrams in the multithermal-multibaric ensemble, J. Chem. Phys. 150 (24) (Jun. 2019). doi:10.1063/1.5102104.

[18] K. Rossi, G. D. Förster, C. Zeni, J. Lam, Modeling and characterization of the nucleation and growth of carbon nanostructures in physical synthesis, Carbon Trends 11 (2023) 100268. doi:10.1016/j.cartre.2023.100268.

[19] A. Stukowski, K. Albe, Extracting dislocations and non-dislocation crystal defects from atomistic simulation data, Model. Simul. Mater. Sci. Eng. 18 (8) (2010) 085001. doi:10.1088/0965-0393/18/8/085001.

[20] A. Stukowski, V. V. Bulatov, A. Arsenlis, Automated identification and indexing of dislocations in crystal interfaces, Model. Simul. Mater. Sci. Eng. 20 (8) (2012) 085007. doi:10.1088/0965-0393/20/8/085007.

[21] D. K. Koli, G. Agnihotri, R. Purohit, A Review on Properties, Behaviour and Processing Methods for Al- Nano Al2O3 Composites, Procedia Mater. Sci. 6 (2014) 567–589. doi:10.1016/j.mspro.2014.07.072.

[22] A. Z. Ziva, Y. K. Suryana, Y. S. Kurniadianti, A. B. D. Nandiyanto, T. Kurniawan, Recent Progress on the Production of Aluminum Oxide (Al2O3) Nanoparticles: A Review, 1. 1 (2) (2021) 54–77. doi:10.31603/mesi.5493.

[23] A. Prakash, S. Satsangi, S. Mittal, B. Nigam, P. K. Mahto, B. P. Swain, Investigation on Al2O3 Nanoparticles for Nanofluid Applications- A Review, IOP Conf. Ser.: Mater. Sci. Eng. 377 (1) (2018) 012175. doi:10.1088/1757-899X/377/1/012175.

[24] S. Demouchy, Defects in olivine, European Journal of Mineralogy 33 (3) (2021) 249–282.

[25] N. Tolganbek, Y. Yerkinbekova, S. Kalybekkyzy, Z. Bakenov, A. Mentbayeva, Current state of high voltage olivine structured limpo4 cathode materials for energy storage applications: A review, Journal of Alloys and Compounds 882 (2021) 160774. doi:https://doi.org/10.1016/j.jallcom.2021.160774.
URL https://www.sciencedirect.com/science/article/pii/S0925838821021836

[26] A. Stukowski, Visualization and analysis of atomistic simulation data with ovito–the open visualization tool, Modelling and Simulation in Materials Science and Engineering 18 (1) (2009) 015012. doi:10.1088/0965-0393/18/1/015012.
URL https://dx.doi.org/10.1088/0965-0393/18/1/015012

[27] J. M. McHale, A. Navrotsky, A. J. Perrotta, Effects of Increased Surface Area and Chemisorbed H2O on the Relative Stability of Nanocrystalline γ-Al2O3 and α-Al2O3, J. Phys. Chem. B 101 (4) (1997) 603–613. doi:10.1021/jp9627584.

[28] F. H. Streitz, J. W. Mintmire, Electrostatic potentials for metal-oxide surfaces and interfaces, Phys. Rev. B 50 (16) (1994) 11996–12003. doi:10.1103/PhysRevB.50.11996.

[29] G. Laurens, D. Amans, J. Lam, A.-R. Allouche, Comparison of aluminum oxide empirical potentials from cluster to nanoparticle, Phys. Rev. B 101 (4) (2020) 045427. doi:10.1103/PhysRevB.101.045427.

[30] J. Furstoss, P. Hirel, P. Carrez, P. Cordier, Complexions and stoichiometry of the 60.8°//[100](011) symmetrical tilt grain boundary in mg2sio4 forsterite: A combined empirical potential and first-principles study, American Mineralogist 107 (11) (2022) 2034–2043.

[31] D. Kohlstedt, C. Goetze, W. Durham, J. Vander Sande, New technique for decorating dislocations in olivine, Science 191 (4231) (1976) 1045–1046.

[32] A. Mussi, P. Cordier, S. Demouchy, Characterization of dislocation interactions in olivine using electron tomography, Philosophical Magazine 95 (4) (2015) 335–345.

[33] A. Pedone, G. Malavasi, M. C. Menziani, A. N. Cormack, U. Segre, A new self-consistent empirical interatomic potential model for oxides, silicates, and silica-based glasses, The Journal of Physical Chemistry B 110 (24) (2006) 11780–11795.

[34] P. Hirel, J. Furstoss, P. Carrez, A critical assessment of interatomic potentials for modelling lattice defects in forsterite mg 2 sio 4 from 0 to 12 gpa, Physics and Chemistry of Minerals 48 (12) (2021) 46.

[35] V. Bulatov, W. Cai, Computer Simulations of Dislocations, Osmm Series, OUP Oxford, 2006.
URL https://books.google.fr/books?id=AdfbPfihVJsC

[36] S. Mahendran, P. Carrez, S. Groh, P. Cordier, Dislocation modelling in mg2sio4 forsterite: an atomic-scale study based on the thb1 potential, Modelling and Simulation in Materials Science and Engineering 25 (5) (2017) 054002.

[37] A. P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, Phys. Rev. B 87 (18) (2013) 184115. doi:10.1103/PhysRevB.87.184115.

[38] S. Menon, G. D. Leines, J. Rogal, pyscal: A python module for structural analysis of atomic environments, Journal of Open Source Software 4 (43) (2019) 1824.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[40] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society: Series B (Methodological) 39 (1) (1977) 1–22. arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1977.tb01600.x, doi:https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.
URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x

[41] G. Schwarz, Estimating the dimension of a model, The Annals of Statistics 6 (2) (1978) 461–464.
URL http://www.jstor.org/stable/2958889

[42] S. Becker, E. Devijver, R. Molinier, N. Jakse, Unsupervised topological learning for identification of atomic structures, Phys. Rev. E 105 (4) (2022) 045304. doi:10.1103/PhysRevE.105.045304.

[43] E. Boattini, M. Dijkstra, L. Filion, Unsupervised learning for local structure detection in colloidal systems, J. Chem. Phys. 151 (15) (Oct. 2019). doi:10.1063/1.5118867.

[44] E. Boattini, S. Marín-Aguilar, S. Mitra, G. Foffi, F. Smallenburg, L. Filion, Autonomously revealing hidden local structures in supercooled liquids, Nat. Commun. 11 (5479) (2020) 1–9. doi:10.1038/s41467-020-19286-8.

[45] S. Lamouri, M. Hamidouche, N. Bouaouadja, H. Belhouchet, V. Garnier, G. Fantozzi, J. F. Trelkat, Control of the γ-alumina to α-alumina phase transformation for an optimized alumina densification, Bol. Soc. Esp. Cerám. Vidrio 56 (2) (2017) 47–54. doi:10.1016/j.bsecv.2016.10.001.

[46] R. Peierls, J. Yoccoz, Proc. phys. soc. (1940).

[47] J. Hirth, J. Lothe, Theory of Dislocations, Krieger Publishing Company, 1992.
URL https://books.google.fr/books?id=LFZGAAAAYAAJ

[48] S. Mahendran, P. Carrez, P. Cordier, On the glide of [100] dislocation and the origin of 'pencil glide' in mg2sio4 olivine, Philosophical Magazine 99 (22) (2019) 2751–2769.

[49] W. T. Read, W. Shockley, Dislocation models of crystal grain boundaries, Physical review 78 (3) (1950) 275.

[50] J. Furstoss, P. Hirel, P. Carrez, K. Gouriet, V. Meko-Fotso, P. Cordier, Structures and energies of twist grain boundaries in mg2sio4 forsterite, Computational Materials Science 233 (2024) 112768. doi:https://doi.org/10.1016/j.commatsci.2023.112768.
URL https://www.sciencedirect.com/science/article/pii/S0927025623007620

[51] P. Hirel, Atomsk: A tool for manipulating and converting atomic data files, Computer Physics Communications 197 (2015) 212–219.

[52] K. E. Blow, D. Quigley, G. C. Sosso, The seven deadly sins: When computing crystal nucleation rates, the devil is in the details, J. Chem. Phys. 155 (4) (Jul. 2021). doi:10.1063/5.0055248.

[53] G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen, A. Michaelides, Crystal Nucleation in Liquids: Open Questions and Future Challenges in Molecular Dynamics Simulations, Chem. Rev. 116 (12) (2016) 7078–7116. doi:10.1021/acs.chemrev.5b00744.

[54] E. Boattini, F. Smallenburg, L. Filion, Averaging Local Structure to Predict the Dynamic Propensity in Supercooled Liquids, Phys. Rev. Lett. 127 (8) (2021) 088007. doi:10.1103/PhysRevLett.127.088007.

[55] E. Boattini, S. Marín-Aguilar, S. Mitra, G. Foffi, F. Smallenburg, L. Filion, Autonomously revealing hidden local structures in supercooled liquids, Nat. Commun. 11 (5479) (2020) 1–9. doi:10.1038/s41467-020-19286-8.

[56] S. Bhakat, Collective variable discovery in the age of machine learning: reality, hype and everything in between, RSC Adv. 12 (38) (2022) 25010–25024. doi:10.1039/D2RA03660F.

[57] J. Lam, F. Pietrucci, Critical comparison of general-purpose collective variables for crystal nucleation, Phys. Rev. E 107 (1) (2023) L012601. doi:10.1103/PhysRevE.107.L012601.

[58] A. France-Lanord, H. Vroylandt, M. Salanne, B. Rotenberg, A. M. Saitta, F. Pietrucci, Data-driven path collective variables, arXiv (Dec. 2023). arXiv:2312.13868, doi:10.48550/arXiv.2312.13868.

[59] P. J. Steinhardt, D. R. Nelson, M. Ronchetti, Bond-orientational order in liquids and glasses, Physical Review B 28 (2) (1983) 784.