TimeSOAP: Tracking high-dimensional fluctuations in complex molecular systems via time variations of SOAP spectra

Accepted to J. Chem. Phys. 10.1063/5.0147025

*Time*SOAP: Tracking high-dimensional fluctuations in complex molecular systems *via* time-variations of SOAP spectra

Cristina Caruso,[1] Annalisa Cardellini,[2] Martina Crippa,[1] Daniele Rapetti,[1] and Giovanni M. Pavan[1,2]
[1)]*Department of Applied Science and Technology, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy*
[2)]*Department of Innovative Technologies, University of Applied Sciences and Arts of Southern Switzerland, Polo Universitario Lugano, Campus Est, Via la Santa 1, 6962 Lugano-Viganello, Switzerland*

(*Electronic mail: corresponding author: Giovanni M. Pavan (giovanni.pavan@polito.it))

(Dated: 6 May 2023)

Many molecular systems and physical phenomena are controlled by local fluctuations and microscopic dynamical rearrangements of the constitutive interacting units that are often difficult to detect. This is the case, for example, of phase transitions, phase equilibria, nucleation events, and defect propagation, to mention a few. A detailed comprehension of local atomic environments and of their dynamic rearrangements is essential to understand such phenomena, and also to draw structure-property relationships useful to unveil how to control complex molecular systems. Considerable progresses in the development of advanced structural descriptors (*e.g.*, Smooth Overlap of Atomic Position (SOAP), etc.) have certainly enhanced the representation of atomic-scale simulations data. However, despite such efforts, local dynamic environment rearrangements remain still difficult to elucidate. Here, exploiting the structurally rich description of atomic environments of SOAP and building on the concept of time-dependent local variations, we developed a SOAP-based descriptor, *Time*SOAP (τSOAP), which essentially tracks the time variations in the local SOAP environments surrounding each molecule (*i.e.*, each SOAP center) along ensemble trajectories. We demonstrate how analysis of the time-series τSOAP data and of their time-derivatives allows to detect dynamics domains and track instantaneous changes of local atomic arrangements (*i.e.*, local fluctuations) in a variety of molecular systems. The approach is simple and general, and we expect will help to shed light on a variety of complex dynamical phenomena.

I. INTRODUCTION

Structure-property relationships, at the heart of modern materials science, are hard to be elucidated in complex molecular systems. Multi-scale and many-body interactions among all the atoms make it challenging, yet inspiring, to reconstruct the macroscopic behavior of such systems from the underlying atomic structure.[1,2] Ranging from materials with an intrinsically dynamic character such as soft supramolecular architectures, to common crystal lattices, a thorough knowledge of atomic arrangements, including their structural and dynamic evolution, is required to increasingly unlock tangled material responses and features.[3–7] In crystalline solids, for instance, material's plasticity, viscosity, and microstructure's evolution are dictated by the energy and kinetics of defects[8], or structural imperfections.[9,10] Furthermore, atomically disordered domains such as surfaces, grain boundaries, and heterogeneous interfaces have been widely recognized to be linked to transport, mechanical, electronic, and optical properties.[11–14] Shedding light on the vital connection between the complex atomic arrangements and the material dynamic properties would clearly pave the way for novel design rules and optimization of molecular systems for tailored behaviors.[15,16] However, although desirable, such material design objective meets a number of practical and theoretical challenges thus standing for the most ambitious goal in material science.

In recent years, the advances in data availability and computational power have enabled the development of valuable tools for gaining a deeper understanding into chemical-physical phenomena occurring in materials[17,18] In particular, molecular dynamics (MD) simulations have been playing an increasingly significant role in the exploration of materials, providing a large source of potential information.[19–27] The use of MD simulations to elucidate structure-property relationships substantially embeds two-steps level protocol: (i) the translation of MD trajectories into a numerical representation of atomic neighborhood environments, resulting in high-detailed and high-dimensional data, known as *fingerprints* or *descriptors*; (ii) the extrapolation of meaningful information from the large volumes of generated data sets. Regarding the latter step, Machine Learning (ML) algorithms have often revealed promising advantages to handle the large and complex set of data, thereby achieving increased interest.[28–33] However, a low-dimensional representation facilitating the navigation and identification of hidden patterns and features would be desirable.

Within this framework, methods for adequately characterizing complex atomic arrangements from MD simulations have received a remarkable expansion. Over the last decades, many *descriptors* have been proposed relying on the use of order parameters or mathematical quantities.[1] Low-dimensional descriptors based on the use of order parameters often allow to gain very accurate information, though being dependent from *a priori* knowledge about systems' features. However, methods operating on structural environments (*i.e.*, order parameters) such as the coordination analysis, bond order analysis,[34] bond angle analysis (BAA),[35] common neighbor analysis, (CNA)[36] adaptive CNA (A-CNA),[37] and Voronoi analysis, generally struggle to identify different local coordination environments when the geometric symmetry is lost or exhibits a short-range nature (*e.g.*, in crystalline systems close to the melting temperature). On the other hand, coupling more mathematically-sophisticated de-

scriptors to ML approaches enables effective characterization of systems by exploiting the rich and high-dimensional data sets provided by MD simulations,[38–40] also being less dependent from *a priori* knowledge. Nevertheless, advanced mathematical descriptors such as the Behler-Parrinello symmetry functions (BP),[41] Chebyshev polynomial representations (CPR),[42] adaptive generalizable neighborhood informed features (AGNI),[43,44] smooth overlap of atomic positions, (SOAP)[45] and atomic cluster expansion (ACE)[46] generally operate on atomic environments, that still represent local properties and weakly capture global and dynamics pictures.

Among more mathematically related descriptors, SOAP turned out to be very efficient in the characterization of a wide plethora of systems[47–50] including soft disordered and complex assemblies.[51–54] Despite being strongly connected to the structural features of local environments, the SOAP fingerprint coupled with unsupervised clustering approaches and statistical analyses has been recently used also to reconstruct the dynamics of complex systems such as, *e.g.*, metal surfaces,[55] metal nanoparticles,[56] soft supramolecular polymers,[51,57] self-assembled micelles[52] and complex hierarchical superlattices, to cite a few.[54,58] Since SOAP descriptors are typically high-dimensional, both linear and non-linear dimensionality reduction (DR) approaches are often employed for facilitating both analyses and data visualization.[59–62] However, DR represents the fundamental roadblock because it inherently leads to a loss of information, resulting in a challenging characterization of systems where ordered and disordered domains coexist in dynamic exchange and equilibrium. In addition, beyond some valuable techniques,[63,64] the time evolution of structural changes, including rare fluctuations, still remains weakly explored by simply classifying datasets with unsupervised and sophisticated ML tools.

Time-dependent descriptors offer a different approach. For example, a recently developed descriptor - Local Environments and Neighbors Shuffling (LENS)[65] - monitors how much the microscopic surrounding of each molecular unit changes over time in terms of neighbor individuals/identities along an MD trajectory. LENS allows to identify dynamic domains and detect local fluctuations in a variety of systems tracking events of addition/subtraction of neighbors within a certain cutoff over time. However, LENS does not contain structural information on, *e.g.*, the relative position or arrangements of the neighbors inside the cutoff sphere. In this way, it does not capture, *e.g.*, local structural rearrangement, adjustment, or rattling. A time-dependent descriptor capable of retaining rich structural information and of efficiently monitoring structural changes over time would be desirable.

Building on such a concept, here we report a time-dependent descriptor, *Time*SOAP (τSOAP), which essentially exploits the structurally rich description of molecular/atomic environments guaranteed by the SOAP vectors and measures to what extent the SOAP power spectra of each unit within a complex molecular system change over time. An ML-based analysis of the time-series τSOAP data allows us to robustly and efficiently detect, *e.g.*, structural transitions, phase transitions, and the coexistence of phases in a variety of systems with rich and diverse intrinsic dynamics. Noteworthy, the time derivative of τSOAP also provides sharp signals identifying local fluctuations, highlighting local and rare events that may be overlooked with other approaches. The paper is organized as follows. In Section II (Methods), we present our τSOAP and τSOAP-based descriptors and the coupled ML-based workflow. In Section III, we discuss the results obtained by performing our τSOAP analysis on various systems characterized by solid/liquid coexisting phases, solid-like and fluid-like behaviors, respectively. Our tests indicate that τSOAP analyses are flexible and robust, and can shed light on complex molecular/atomic systems with non-trivial multilayered dynamics providing insights that are difficult to attain with other approaches.

II. METHODS

A. SOAP as a descriptor of atomic environments

Recently, data-driven approaches capturing the structural complexity of materials from equilibrium MD trajectories have been proposed. A generic MD trajectory is represented by an ordered list of N atomic coordinates $\mathbf{R}$(t) in the 3D space at each simulation time step, where N is the number of particles in the system. In order to characterize complex atomic arrangements, descriptors of atomic neighborhood environments have been widely employed. By associating a *feature* vector to each $\mathbf{R}$(t), the descriptors enable to pass from the 3D *coordinate* space to an S-dimensional *feature* space. Importantly, these representations are required (1) to be permutationally, translationally and rotationally invariant, in order for physically equivalent configurations to be recognised as such, and (2) to smoothly vary with small changes in atomic positions. Among many developed descriptors, we adopt the Smooth Overlap of Atomic Position (SOAP) to examine our sample of materials ranging from crystalline to soft and liquid states. SOAP is a state-of-art, high-dimensional representation of atomic environments and it has recently provided valuable insights on both properties and structural features[49,57,66,67].
The SOAP descriptor centers Gaussian density distributions on each atom. For a given atom, a smooth representation of the neighbor density is generated from the sum of Gaussians centered on each surrounding atom, namely:

$$\rho^i(\mathbf{r}) = \sum_j \exp\left[\frac{-|\mathbf{r}-\mathbf{r}_{ij}|^2}{2\sigma^2}\right] f_{rcut}(|\mathbf{r}-\mathbf{r}_{ij}|), \qquad (1)$$

where to each neighbor center j, located at a distance $\mathbf{r}=\mathbf{r}_{ij}$ from the i-th center, a Gaussian function is associated. σ is the distribution width of each Gaussian. The environment related to each center i incorporates information up to a fixed cutoff, *rcut*, where the function f_{rcut} smoothly goes to 0. Then, by expanding the Eq. (1) in the basis of orthonormal radial functions $R_n(r)$ and spherical harmonics $Y_{l,m}(\hat{\mathbf{r}})$, the corresponding SOAP power-spectrum is calculated. For the i-th center, it

can be expressed as:

$$\gamma^{i}_{nn'l} \propto \frac{1}{\sqrt{2l+1}} \sum_{m=-l}^{+l} (c^{i}_{nlm})^{*} c^{i}_{n'lm}, \tag{2}$$

with c^{i}_{nlm} representing the expansion coefficients of the neighbor density associated to the i-th center. The parameters n and n' range from 1 to $nmax$, while l index runs from 1 to $lmax$. From the values of $nmax$ and $lmax$ it is possible to derive the dimension S of the full SOAP feature vector, which can be written as:

$$\mathbf{p}_i = \{\gamma^{i}_{nn'l}\}, \tag{3}$$

representing the SOAP descriptor associated to the i-th center, which includes all the contributions from the Eq. (2). Here, we used in-house code, SOAPify,[68] to compute the SOAP vectors, with $nmax$, $lmax$=8, and different $rcut$ values depending on the characteristics of the considered system (see supplementary material, Table S1). From the 3D coordinate vector corresponding to each MD simulation time, we calculate the SOAP vector $\mathbf{p}_i$ for a selected set $\{i\}$ of centers (referred to as SOAP centers). In summary, we obtain a dataset containing S-dimensional SOAP vectors describing the structural arrangements related to the $\{i\}$ selected sites at each sampled configuration. Since these SOAP vectors encode the information about the atomic environments surrounding each center, SOAP is referred to as a "local" descriptor.

In order to evaluate how similar are two environments centered in two sites, a similarity measure has been defined by means of a linear kernel of their neighbor density representations:

$$K^{SOAP}(i,j) = (\mathbf{q}_i \cdot \mathbf{q}_j). \tag{4}$$

Since $\mathbf{q} = \frac{\mathbf{p}}{|\mathbf{p}|}$, that is, the unit-normalized SOAP vector, $K^{SOAP}(i,j)$ goes from 0 for no overlapping to 1 for completely superimposed vectors. Furthermore, from Eq. (4), a metric referred to as "SOAP distance" between two environments can be defined:

$$d^{SOAP}(i,j) = \sqrt{2 - 2 \cdot K^{SOAP}(i,j)} \propto \sqrt{2 - 2\mathbf{p}_i\mathbf{p}_j}. \tag{5}$$

Importantly, $\mathbf{p}_i$ and $\mathbf{p}_j$ describe the local environments related to two *different* SOAP *centers*. Besides the SOAP kernel, this distance representation provides a bounded similarity measure between two local environments, indicating how their local densities match in the S-dimensional feature space.

B. Tracking dynamical SOAP variations with *Time*SOAP

The output dataset containing the S-dimensional SOAP vectors is typically high-dimensional, and although rich in information on the atomic/molecular arrangements, it requires a crucial pre-processing to both facilitate the interpretation of the results and effectively identify relevant molecular patterns. For this reason, after estimating $\mathbf{p}_i$ (Eq. (3)) for the whole set of SOAP centers $\{i\}$ at each sampled configuration of the MD trajectory, a SOAP-based pattern recognition procedure typically relies on two successive key phases: (1) use of a dimensionality reduction (DR) of SOAP spectra by means of, for instance, Principal Component Analysis[69,70] (PCA); (2) employment of unsupervised clustering techniques for the identification of molecular motifs. Despite providing some information on a wide range of molecular systems, this approach presents some key shortcomings: (i) since the time information is not emphasized, insights on consequential transition events as well as the temporal persistence of the individual molecular configurations are not retained, thus hindering a detailed comprehension of the rate of change of every individual molecular configuration; (ii) on such low-dimensional SOAP-based data set, some poorly-populated configurations may remain undetected by (*e.g.*, density-based) unsupervised clustering approaches; (iii) low-dimensional embedding of atom-density representations can fail in faithfully preserving valuable information, such that a high number of principal components would be desirable.[71] This makes detecting local fluctuations and rare events typically awkward with such approaches.

In this work, we propose an alternative procedure allowing to retain the time information from the high-dimensional SOAP vectors. Building upon the SOAP distance $d^{SOAP}(i,j)$ introduced above, we present a new SOAP-based fingerprint, named "*Time*SOAP (τSOAP)", which quantifies the local environment variation, over time, of each individual SOAP center i. Indicating by λ_i the variable form of τSOAP, its instantaneous value is defined as:

$$\lambda_i^{t+\Delta t} = \frac{\sqrt{2 - 2 \cdot K^{SOAP}(i^t, i^{t+\Delta t})}}{\Delta t} \propto \frac{\sqrt{2 - 2\mathbf{p}_i^t \mathbf{p}_i^{t+\Delta t}}}{\Delta t}. \tag{6}$$

Differently than the Eq. (5), here both $\mathbf{p}_i^t$ and $\mathbf{p}_i^{t+\Delta t}$ describe the local environments related to the *same unit* (*i.e.*, the i-th SOAP center) but at *different simulation times*, t and $t+\Delta t$, respectively. Thus, $\lambda_i^{t+\Delta t}$ measures how similar the i-th SOAP vector calculated at time t is to that calculated at the next sampled timestep ($t+\Delta t$). We analyze consecutive frames, namely adjacent points, where Δt represents the MD sampling timestep (different for the various systems, see Molecular Dynamics Simulations for more details). As a result, τSOAP evaluates how the i-th local environment changes, in terms of SOAP descriptor, at every consecutive time interval Δt. We thus obtain $\lambda_i(t)$, namely a τSOAP signal over time for each individual in the selected set $\{i\}$, thereby allowing to track the evolution of each SOAP constituent unit center along the trajectory.

We can take a further step by estimating $\dot{\lambda}_i$, namely the first time-derivative of τSOAP signal. Using the NumPy[72] Python package, we have:

$$\dot{\lambda}_i^{t+\Delta t} = \frac{\lambda_i^{t+\Delta t} - \lambda_i^t}{\Delta t}. \tag{7}$$

By computing it along the MD trajectory, we get $\dot{\lambda}_i(t)$. What $\dot{\lambda}_i$ represents is the *rate* of local environment changes

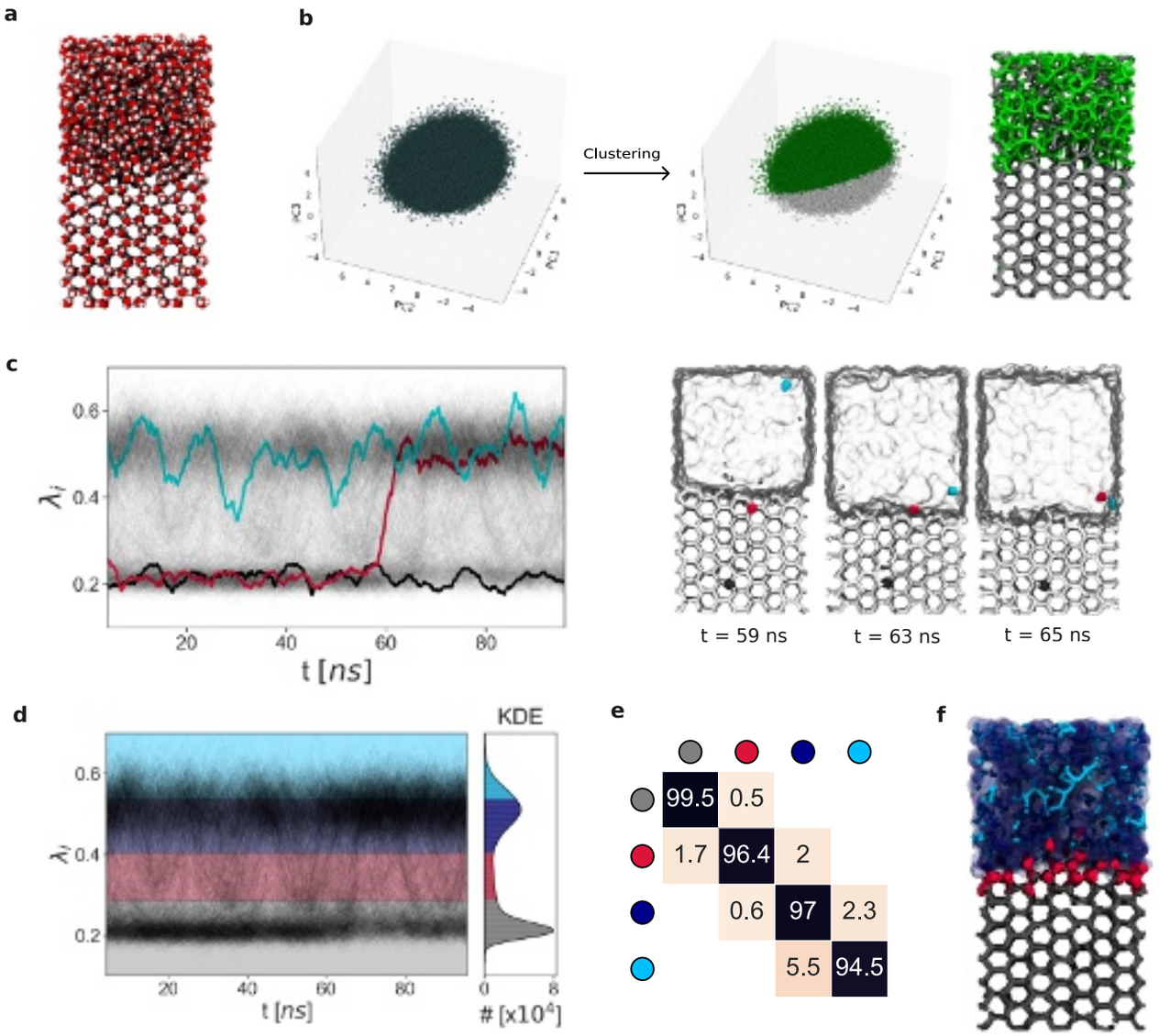


FIG. 1. Automatic detection of molecular motifs in ice/liquid water coexistence. (a) MD snapshot of ice/liquid water simulation box made of 2048 TIP4P/ICE water molecules at $T = 268$ K. Color code: red for oxygen and white for hydrogen atoms. (b) Example of a typical SOAP-based pattern recognition procedure. Left: PCA projection of the SOAP-based data-set estimated from the ice/liquid water system in (a). Right: clustering analysis - on the same data-set - carried out with KMeans. The two main detected clusters, colored in green and gray, are also visualized on the MD water snapshot (taken at 44 ns), showing the ice and water domains in gray and green, respectively. (c) On the left, time-series of τSOAP signals, $\lambda_i(t)$, shown for all the oxygen atoms in (a). The colored $\lambda_i(t)$ profiles are related to three explicative oxygen atoms, i.e. (i) black, (ii) cyan, and (iii) crimson, displayed on the right with the respective color code. The reported MD snapshots are around $t \sim 60$ ns. (d) τ SOAP-based analysis. $\lambda_i(t)$ profiles and their KDE are carried out for all the oxygen atoms of all water molecules in the system (a). The final $k = 4$ detected macro-clusters are shown as colored in gray, crimson, blue, and cyan. (e) Interconnection probability matrix of the final $k = 4$ identified macro-clusters. (f) MD snapshot (taken at 44 ns) showing the four main clusters identified by τ SOAP-based analysis (same color code of (d)): ice (in gray), solid/liquid interface (in crimson), liquid water (in blue), a distinct domain in the liquid phase (in cyan).

over time for the i-th SOAP center. This allows to highlight the relevant dynamic phenomena occurring along the trajectory, notably discriminating between local environments characterized by a constant variation and those exhibiting an increasing/decreasing variation.

To increase the signal-to-noise ratio (S/N), both $\lambda_i(t)$ and $\dot{\lambda}_i(t)$ are pre-processed by employing the Savitzky-Golay[73] filter from the SciPy[74] python package, thus obtaining smoothed signals. A common polynomial order parameter of p = 2 is used for each signal $\lambda_i(t)$, while different time-

windows are chosen depending on the analyzed system, in order to reach a compromise between an acceptable S/N value and a sufficiently smaller window compared to the length of signal (see supplementary material, Fig. S1 for details). Chosen the time-window for $\lambda_i(t)$, to adequately smooth its first-derivative $\dot{\lambda}_i(t)$ we keep the same time-window and use two applications of the filter (following a general rule: for the n-th derivative, use at least n+1 applications of the filter).

C. Dynamics domains detection

After increasing S/N, an ML-based analysis is performed on $\lambda_i(t)$ data to detect relevant dynamics domains in each system. As a clustering method, we opted to use the KMeans algorithm from the Scipy python package,[75,76] since it was demonstrated robust and capable of providing a good tradeoff between clustering quality and computational cost[65]. Nonetheless, it is worth noting that the analysis approach is versatile, and other clustering methods could be used. KMeans requires to pre-determine the number K of clusters to be created in the process. Here, with the aim of guaranteeing a wide variety of micro-clusters dynamics regardless of the analysed system, we start anyway from $K = 10$ clusters. On the basis of the exchange probability matrix and the dendrogram associated to clusters interconnection, then we hierarchically merged the K clusters *a posteriori*. The exchange probability matrix contains, indeed, the percentage probability of a unit i belonging to a given cluster to persist in that cluster or to jump to another cluster in the sampling timestep Δt; from this, by means of an "average" linkage method, we built the associated dendrogram connecting the dynamics domains which have a high probability of exchanging elements. Ultimately, to establish the cut-off point of the dendrogram, we used the Elbow Curve Method as an indicative guideline for selecting the optimal number of clusters k (see supplementary material, Fig. S2). However, for completeness, all the steps leading from the starting $K = 10$ clusters to the final k clusters are reported in supplementary material, Fig. S3 and Fig. S4.

On the other hand, the domain recognition on $\dot{\lambda}_i(t)$ data has been performed *via* a different approach. Obtained the $\dot{\lambda}_i$ distribution and the associated Kernel Density Estimate (KDE) for each system, we divide the KDE in deciles and consider the first and the tenth deciles to detect units significantly falling far from the mean local environment variation rate. The first decile and the tenth decile capture units highly decreasing and increasing, respectively, their local environment changes. This provides a clear distinction between domains moving toward more dynamic and those moving toward less dynamic configurations.

D. Molecular dynamics (MD) simulations

We test our τSOAP analysis on MD trajectories obtained for different systems with non-trivial and different dynamics: *i.e.*, a water-ice interface system in correspondence of the transition temperature, a gold nanoparticle at 200 K, a copper surface at 700 K, and DPPC lipid bilayer where liquid and gel phases coexist at 293 K of temperature.

The atomistic ice/liquid water phase coexistence at the solid/liquid transition temperature is obtained by employing the direct coexistence technique[77,78] using the GROMACS software[79]. In order to model both the ice and the liquid water phase, the TIP4P/ICE water model[80] is used. The direct coexistence technique is based on the idea to put in contact two or more phases (in this case, the phase of ice I_h and the liquid water phase) in the same simulation box and at constant pressure. Since the energy is constant at $T = 268$ K, while the system melts at $T = 269$ K[81], we set the temperature at $T = 268$ K and keep it constant by means of the v-rescale thermostat with a relaxation time of $t = 0.2$ ps.
To get the initial configuration of ice I_h the *Genice* tool proposed by Matsumoto *et al.*[82] is used, which generates a hydrogen-disordered lattice with zero net polarization satisfying the Bernal-Fowler rules. The solid lattice is equilibrated by performing a 10 ns-long anisotropic *NPT* simulation at ambient pressure (1 atm). The c-rescale barostat[83] is used with a time constant of $t = 20$ ps and compressibility of $9.1 * 10^{-6}$ bar^{-1}. On the other hand, the liquid phase is obtained from the same initial configuration of ice I_h, but performing a *NVT* simulation at $T = 400$ K in order to quickly melt the ice slab. Then, a 10 ns long simulation at $T = 268$ K is performed to equilibrate the liquid phase, using the c-rescale barostat in semi-isotropic conditions and compressibility of $4.5 * 10^{-5}$ bar^{-1}. Since the initial ice slab is composed of 1024 water molecules, both the solid and liquid phases have the same number of molecules and box dimensions. The two phases are put in contact and, then, the system is equilibrated for $t = 10$ ns employing the c-rescale pressure coupling at ambient pressure with the water compressibility ($4.5 * 10^{-5}$ bar^{-1}). The production *NPT* is carried out in semi-isotropic conditions, applying the pressure only in the direction perpendicular to the ice/water interface, thus reproducing the strictly correct ensemble for the liquid-solid equilibrium simulation by the direct coexistence technique.[84] Finally, a 100 ns-long production run is performed, with a sampling time interval of $\Delta t = 0.1$ ns.

The second case study analyzed in this work is an icosahedral Gold nanoparticle (Au-NP) composed of 309 atoms. The parameterization of the model is performed according to the Gupta potential[85]. The Au-NP system is simulated for $t = 2$ μs at $T = 200$ K sampling every $\Delta t = 0.1$ ns using the LAMMPS software[86]. The details are described in reference[56].

The third system, the atomistic model of Copper FCC surface Cu(210), is composed of 2304 Cu atoms and simulated at $T = 700$ K. A Neural Network potential built by means of the DeepMD platform[87] is employed to perform Deep-potential MD simulations of the Cu(210) surface with the LAMMPS software[86], as reported in reference[55]. The MD trajectory is conducted for 150 ns, using a sampling time interval of $\Delta t = 0.3$ ns.

Finally, the last case study is a DPPC lipid bilayer com-

posed of 1152 lipids simulated at $T = 293$ K. As detailed in reference[53], DPPC lipids are simulated and parametrized in explicit water by using the Martini2.2[88] Coarse-Grained (CG) force field. The CG-MD simulation is performed for $t = 1$ μs and sampled every 0.1 ns with the GROMACS software[79]. However, in our analysis, we use the last 500 ns of MD trajectory.

III. RESULTS AND DISCUSSION

Herein, we use the descriptor τSOAP to elucidate the dynamics of atomic/molecular structural environments which are often key determinants in global materials performances. In order to show the whole picture of dynamic information that can be extracted from τSOAP signals, we analyze MD trajectories of different systems exhibiting various structures and non-trivial behaviors, thus indicating the transferability of such approach to a wide range of materials. In particular, we first focus on ice/liquid water coexistence at the solid/liquid transition temperature, where structural and dynamic properties continuously alternate from solid-like to liquid-like character[89]. We also carry out our analysis on systems revealing solid-like dynamics, such as metal nanoparticles and surfaces well below the melting point. Ultimately, a fluid-like soft system is included by testing our approach on a lipid bilayer below the gel-to-liquid transition temperature.

A. Into the Dynamics of Ice/Liquid Water Phase Coexistence *via* τSOAP Signal

We start testing τSOAP on a system where crystalline ice and liquid water coexist at the solid/liquid transition temperature, while exhibiting a dynamic equilibrium between solid-like and liquid-like regime. We analyze a simulation box, in periodic boundary conditions, having 1024 hexagonal ice (*Ih*) molecules in contact with 1024 liquid water molecules (see Fig. 1a) at $T = 268$ K. We consider 1001 consecutive frames sampled every $\Delta t = 0.1$ ns along 100 ns of an MD trajectory. As a first step, we compute the SOAP vectors for the oxygen atoms of each water molecule (2048 TIP4P/ICE water molecules) along all frames of the trajectory (see Methods section for details). Before illustrating τSOAP analysis, we start by briefly discussing the results obtained via, *e.g.*, a SOAP + PCA pattern recognition procedure - widely used for studying molecular systems - on such ice/liquid water system, here presented as a first case study. Fig. 1b shows the results of this analysis which detects, from the SOAP-based data set, two main clusters, corresponding to the ice and liquid water domain (in green and gray). It is worth noting that DR (*via* PCA) to a 3-dimensional subspace allows already to capture $> 90\%$ of the cumulative variance of the SOAP-based data set in this case (see Fig. S5a). A systematic analysis on the effect of increasing the dimensionality provided essentially the same results, demonstrating how two main SOAP domains are detected (ice and liquid water) independently on the number of PCs used and of identified clusters (see also Fig. S5).

After computing SOAP vectors, τSOAP signals are estimated by capturing the variations of local SOAP environments in $\Delta t = 0.1$ ns (see Eq. 6). Fig. 1c reports, on the left, the resulting $\lambda_i(t)$ time-profiles related to each of the 2048 oxygen atoms, while, on the right, it shows the ice/liquid water MD snapshots at $t = 59$ ns, $t = 63$ ns, and $t = 65$ ns. Notably, three distinct $\lambda_i(t)$ profiles are highlighted in Fig. 1c, left: (i) the black signal oscillating around $\lambda_i = 0.2$; (ii) the cyan signal laying on the highest λ_i region; and (iii) the crimson signal which rapidly passes, at $t \sim 60$ ns, from low to high λ_i values. The oxygen atoms related to the latter three $\lambda_i(t)$ profiles are instead depicted on the MD snapshots in Fig. 1c, right, with the respective color code, *i.e.*, black, cyan, and crimson. The visualization of these selected atoms clearly shows that the black and cyan oxygen units belong to the ice and liquid water phase, respectively, regardless of the displayed time steps. On the other hand, the identified crimson oxygen represents an atom involved in the ice/liquid water transition occurring at $t \sim 60$ ns. By lightening the behavior of such atoms, we attempt to emphasize the potential meaning provided by τSOAP descriptor on the single unit dynamics: following the time variation of atomic structural environments, τSOAP allows both to distinguish atoms belonging to different phase states and to capture those one undergoing phase transitions.

In order to systematically detect the complete scenario of distinct dynamics behaviors in our water system, an ML-based analysis is carried out on all τSOAP signals. The results of the clustering investigation, performed *via* the KMeans algorithm, are shown in Fig. 1d. The final four identified clusters (gray, crimson blue, and cyan) are displayed both on the time-series of the $\lambda_i(t)$ data (Fig. 1d: left) and on the $\lambda_i(t)$ data distribution reported with the correlated KDE (Fig. 1d: right). As already pointed out, such four different dynamics domains identify those water molecules undergoing specific transitions: *i.e.*, instantaneously changing their local structural environments. In particular, the gray domain is dominated by oxygen units that are characterized by low λ_i values along the complete trajectory, *i.e.*, by a weak variability of their local atomic environments. On the other hand, oxygen atoms showcasing high changes of their structural atomic distributions, and hence high values of λ_i, belong to the blue cluster or cyan domain. Oxygen units that, instead, tend to reveal medium values of λ_i - because of their transition from one λ_i regime to the other one - are classified into a distinct crimson cluster. It is interesting to note how, differently from the SOAP + PCA pattern recognition procedure shown in Fig. 1b, an analysis of the time-series τSOAP data reveals this third dynamically different environment - *i. e.*, the ice/liquid water interface, which gets lost in SOAP + PCA-based analyses due to its reduced (negligible) statistical weight (see supplementary material, Fig. S5 for SOAP + PCA-based analyses with increased number of clusters). Ultimately, the cyan domain is detected as a different cluster of units having higher local environment changes. The graphical representation of such clusters is shown in Fig. 1f through an MD snapshot (see supplementary material, Movie S1). Not surprisingly, the gray cluster, characterized by the lowest τSOAP signal, corresponds to the ice phase; the blue domain, characterized by 0.4

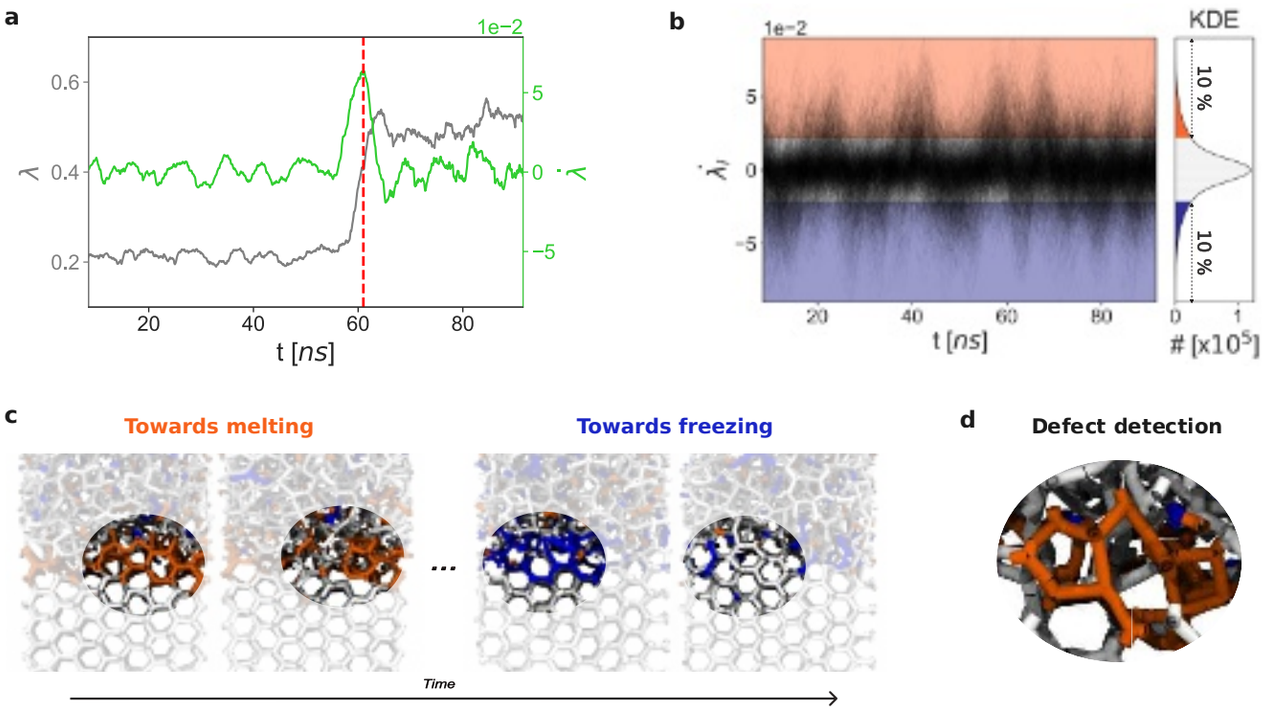


FIG. 2. First time-derivative of τSOAP signal for ice/liquid water system. (a) τSOAP ($\lambda(t)$) and its first time-derivative ($\dot{\lambda}(t)$) profiles associated with the same oxygen atom are shown in gray and green, respectively. (b) $\dot{\lambda}_i(t)$ signals and their KDE estimated for all the oxygen atoms in Fig. 1a. Clustering color code: (i) blue for environments corresponding to the first decile; (ii) orange for those corresponding to the tenth decile; (iii) white for $\dot{\lambda}_i$ values in all the other deciles. (c) MD snapshots displaying blue, orange, and white domains. Left: local detail of the orange cluster evolving toward melting (first and second snapshots). Right: local detail of blue cluster associated with a small disordered region evolving toward freezing (third and fourth snapshots). (d) Orange local environments identify ice molecules moving out of hexagonal ice configurations.

$< \lambda_i \leq 0.54$, is mainly correlated to the liquid phase; and the crimson one, including oxygen atoms with $0.2 < \lambda_i < 0.4$, is instead located at the solid/liquid interface. Finally, the cyan cluster ($\lambda_i > 0.54$), although sited in the same region of the liquid phase (blue cluster), is identified as presenting a different dynamic behavior. In the considered ice/liquid water system, the exchange probabilities among the final four clusters are displayed in the matrix in Fig. 1e: although the oxygen atoms exhibit a probability higher than $\sim 94\%$ to persist in a given cluster in the sampling timestep Δt (probabilities on the matrix diagonal), no negligible transient events occur between red-blue and cyan-blue clusters, demonstrating that a percentage of oxygen population is involved into instantaneous transitions among dynamics domains (out of diagonal probabilities).

After detecting the main dynamics clusters based on τSOAP signals, $\lambda_i(t)$, we carry out a further domain recognition analysis based on $\dot{\lambda}_i(t)$, that is, the instantaneous *rate* of local environment variations $\lambda_i(t)$. The key information which can be gathered from the time-derivative of $\lambda_i(t)$ is pointed out in Fig. 2a, where an explicative example is reported. Here, both $\lambda_i(t)$ and $\dot{\lambda}_i(t)$ time-profiles are associated with the same oxygen atom *i*: in gray, $\lambda_i(t)$ shows the atom undergoing the phase transition at $t \sim 60$ ns when the time-signal significantly and rapidly passes from the low to the high λ_i value region; in green, the first time-derivative of the gray profile exhibits a peak in correspondence of the phase transition, while fluctuating around zero in both the initial and the final stages of the trajectory. Clearly, $\dot{\lambda}_i(t)$ tracks a notable signal leading up to a substantial dynamic change in the system. The first time-derivative, indeed, offers a neat discrimination between small oscillations of $\lambda_i(t)$ - which are intrinsic to the constituent units, independently from the proper dynamics domain - and large fluctuations driving significant changes in the atomic structure. Notably, $\dot{\lambda}_i(t)$ also provides a detailed comprehension of the *directionality* of the local environment variations, *i.e.*, on the evolution of the material structures. While the presence of a peak, *i.e*, of a large fluctuation in $\dot{\lambda}_i(t)$ profile, suggests that a relevant event is occurring in that time interval, the sign of such fluctuation points out the evolution of a structural environment: a positive sign indicates that the atom is undergoing a local re-configuration toward a more dynamic domain; a negative sign means that a local environment re-configuration toward a more static domain is occurring.

Fig. 2b shows, on the left, the time-profiles of $\dot{\lambda}_i(t)$ related to each of the 2048 oxygen atoms, while, on the right, the KDE of the $\dot{\lambda}_i$ data distribution. We color in blue and orange

the domains corresponding to the first and the tenth decile, respectively, while we merge all the other deciles in a single white cluster. It is worth noting that the KDE distribution has a peak approximately in correspondence of $\dot{\lambda}_i = 0$, indicating that the local environment variations -$\lambda_i(t)$- are, on average, constant. On the other hand, atoms that significantly increase or decrease, frame by frame, their local environment changes are captured by positive (in the orange region) or negative peaks (in the blue region), respectively. In Fig. 2c, we visualize these three different domains (blue, orange and white) on some snapshots along the MD trajectory, thereby showing that the positive and negative peaks allow characterizing melting and freezing phenomena occurring within small solid-like and liquid-like regions. In the first snapshot of Fig. 2c, we represent a small portion of oxygen solid-like atoms (in orange) located at the ice/liquid water interface and exhibiting positive τSOAP (λ_i) variations (*i.e.*, undergoing rearrangements toward more dynamic configurations). Accordingly, in the second snapshot those rings appear as broken, thus proving a melting-type process. In the third snapshot of Fig. 2c, we report, instead, example of oxygen units presenting negative τSOAP (λ_i) variations (blue cluster), thus evolving toward more static configurations at the solid-liquid interface. Indeed, as shown in the fourth snapshot, an ordered ring structure forms, thus reproducing a typical freezing phenomenon. Ultimately, Fig. 2d shows a further detail potentially revealed by our analysis. In particular, water molecules exhibiting a high positive rate of change of their local SOAP environment (high $\dot{\lambda}_i$) turned out to be also associated with ice molecules that, at the interface with liquid water, undergo transitions out of the typical hexagonal packing: *i.e.*, forming interface ice defects (Fig. 2d)[90]. In summary, besides capturing the local atomic re-arrangements and characterizing their evolution, $\dot{\lambda}_i(t)$ seems to be also promising for defect detection purposes.

The previous results suggest how τSOAP descriptor and its first time-derivative are possible strategies to unveil some microscopic phenomena occurring at the ice/water interface in a dynamic equilibrium. In particular, by reliably detecting local fluctuations along with rearrangements and their evolution, the time-variations of structural atomic environments show considerable potential for tracking crystallization or melting processes from MD trajectories. In order to outline the main features of τSOAP and the differences respect to other analysis approaches often used to study the dynamics, we also compared τSOAP with a time-lagged Independent Component Analysis (tICA),[91,92] a DR approach used to process high-dimensional input data by retaining valuable temporal information (see supplementary material, Fig. S6). Concerning the study case of ice-liquid water transition, we projected the high-dimensional SOAP space on its highest-autocorrelation linear tICA subspace. The results in Figure S6 demonstrate how tICA essentially finds two main environments, corresponding to the ice and water domains. However, similar to a classical SOAP+PCA analysis (see Fig. 1b), such SOAP+tICA DR does not recognize the ice/water interface as a separate environment, nor does it capture the local individual transitions as done by τSOAP (Figs. 1, 2). This shows how such standard pattern-recognition approaches (*e.g.*, PCA or tICA coupled with clustering analyses) can effectively detect dynamic domains with dominant statistical weight, while sparse and local fluctuations get typically lost due to their negligible statistical occurrence. In this sense, τSOAP has the advantage to preserve any changes of local structural environments, from the slowest to the fastest visited along the studied trajectories, and thereby avoiding a specific screening of structural variations.

B. Application to discrete solid-like dynamics

As completely different test cases, we test our approach on systems revealing solid-like dynamics. We discuss the results of our analysis applied on MD trajectories of (i) a 309-atoms icosahedral Gold nanoparticle, denoted as Au-NP, at 200 K (Fig. 3a), and (ii) a Copper Cu(210) FCC surface at 700 K of temperature (Fig. 4a).

Regarding the case (i), we analyse 20000 consecutive frames of a 2-μs long MD trajectory sampled every $\Delta t = 0.1$ ns at $T = 200$ K. It is well known that metal nanoparticles may exhibit a not-trivial dynamics at room and at even sufficiently lower temperatures. Although the reduced atomic motion and, accordingly, the more stabilized ideal icosahedron architecture, some local fluctuations and atomic rearrangements can be observed in a Au-NP even at $T = 200$ K. τSOAP signals in Fig.3b, indeed, present a sudden increase after ~ 0.1 μs, demonstrating that some atoms are experiencing intense instantaneous local environment variations. Our cluster analysis on $\lambda_i(t)$ recognises five main dynamics domains whose transition probabilities are reported in Fig. 3c. This transition matrix proves a negligible attitude of the gold atoms to transfer from/toward diverse dynamics domains, while preferring to remain in their own cluster with probabilities higher than 98.4 %. The MD snapshots in Fig. 3d show that these clusters well identify distinct dynamics behaviors and structural domains (see also supplementary material, Movie S2). First, the cluster analysis is able to accurately distinguish the inner core of the Au-NP (in gray), namely a more static region characterised - not surprisingly - by low $\lambda_i(t)$ values along the whole simulation, from an interface region (in crimson) between the inner core and the outermost layer. Second, such clustering approach sharply separates the surface of the Au-NP in two coexisting regions (pink and blue) related to different characteristic $\lambda_i(t)$. While the pink face turned out to be more static, the blue domain reliably detects the portion of the surface where a fracture formation may occur, breaking down the symmetry (Fig. 3d, second MD snapshots on the right). Interestingly, τSOAP also identifies some local events such as the formation of concave "rosettes" (a vertex, having five neighbors in an ideal icosahedron, penetrates inside the NP surface, thus passing to six neighbors). In Fig. 3d (third snapshot on the right), two rosettes can be observed as belonging to a more dynamic cluster - highly varying their local environments - (in blue), while the associated vertices are identified as more stable (in crimson).

Furthermore, the estimation of $\dot{\lambda}_i$ (Fig. 3e) provides interesting details on the dynamic evolution of the system. A quite

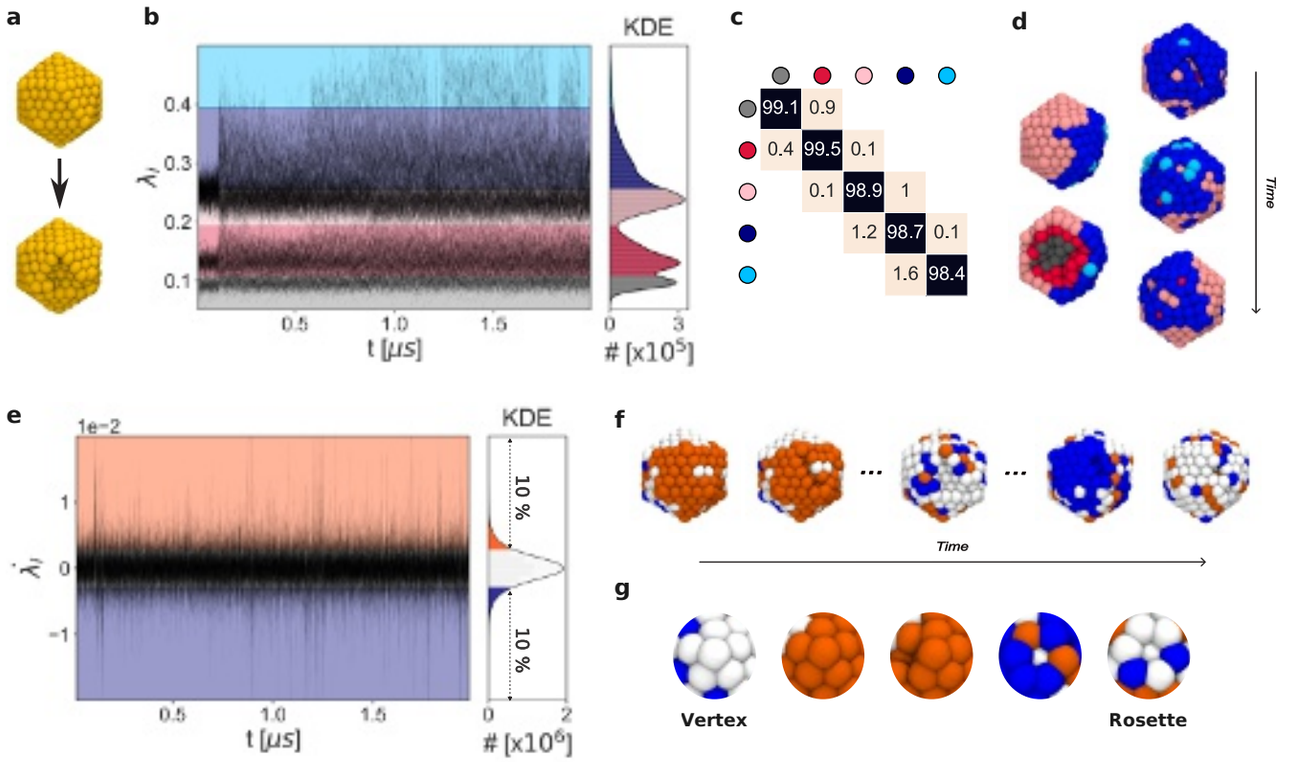


FIG. 3. τ SOAP-based analysis on 309-atoms icosahedral Gold nanoparticle (Au-NP). (a) MD snapshots of ideal Au-NP (top), and equilibrated one at $T = 200$ K (bottom). (b) $\lambda_i(t)$ profiles and the related KDE for the Au atoms in the system (a). The final $k = 5$ macro-clusters identified by KMeans are shown in gray, crimson, pink, blue, and cyan. (c) Exchange probability matrix of the final $k = 5$ detected macro-clusters. (d) MD snapshots with the five main clusters identified in (b): inner core in gray, interface region between the inner core and the outermost layer in crimson, more static surface face in pink, more dynamic surface face in blue, atoms undergoing the highest local environments changes in cyan. (e) Domain detection based on $\dot{\lambda}(t)$ profiles and their KDE: blue domain is associated with the first decile, the orange domain is linked to the tenth decile, and the white domain includes $\dot{\lambda}(t)$ in all the other deciles. (f) MD snapshots displaying the emergence of blue, orange, and white domains along the MD trajectory. On the left, the predominance of the orange cluster before (first snapshot) and during the symmetry breakdown (second snapshot). The central snapshot exhibits a prevalence of white domain, together with a balance between orange e blue ones. On the right, a prevalence of the blue domain can be observed (fourth snapshot) before the formation of a more static configuration (white cluster: fifth snapshot). (g) Blue, orange, and white domains associated with the rearrangement, over time, of a local configuration from "vertex" to "rosette".

large percentage of Au atoms is characterized by a constant variations of their surrounding environment ($\lambda_i(t)$ = const, and $\dot{\lambda}_i \sim 0$). Rare and sharp fluctuations are anyhow remarkable. To qualitatively illustrate some of these $\dot{\lambda}$ peaks, five MD snapshots presenting different predominant domains are shown Fig. 3f. In the first and second snapshots, the domain characterized by positive $\dot{\lambda}_i$ (in orange) prevails, suggesting that the atoms belonging to that cluster are collectively involved in a significant increase of the instantaneous local environment variations. Indeed, this predicts the symmetry breaking shown in the second snapshot. However, the prevalence of $\dot{\lambda}_i \sim 0$ represented by the white domain, along with a balance between positive (orange) and negative (blue) peaks, establishes a dynamic equilibrium leading to no relevant events along several trajectory frames (one example is presented in the third snapshot). In the last two snapshots, instead, a significant collective decrease of the instantaneous local variations (negative $\dot{\lambda}_i$) emerges (prevalence of blue domain in the fourth snapshot), thus predicting the evolution of the associated atoms toward more static environments (shown in the final snapshot). Ultimately, the information on the directionality of local rearrangements is also highlighted in Fig. 3g: while positive $\dot{\lambda}_i$ values (in orange) mark a vertex evolving toward a less stable configuration where a missing atom appears, negative $\dot{\lambda}_i$ (in blue) predict the rearrangement of the structure toward a stable rosette-like configuration.

For case (ii), we use 502 consecutive frames of 150 ns long MD simulation of a Cu(210) surface composed of 2304 Cu atoms (Fig. 4a) sampled every $\Delta t = 0.3$ ns at $T = 700$ K. Although metals tend to be traditionally considered as hard matter, it is now well known that their constituent surface atoms may exhibit a non-trivial dynamics, undergoing rearrangements well below the melting temperature.[55,93] Our clustering procedure applied on τSOAP profiles identifies three main do-

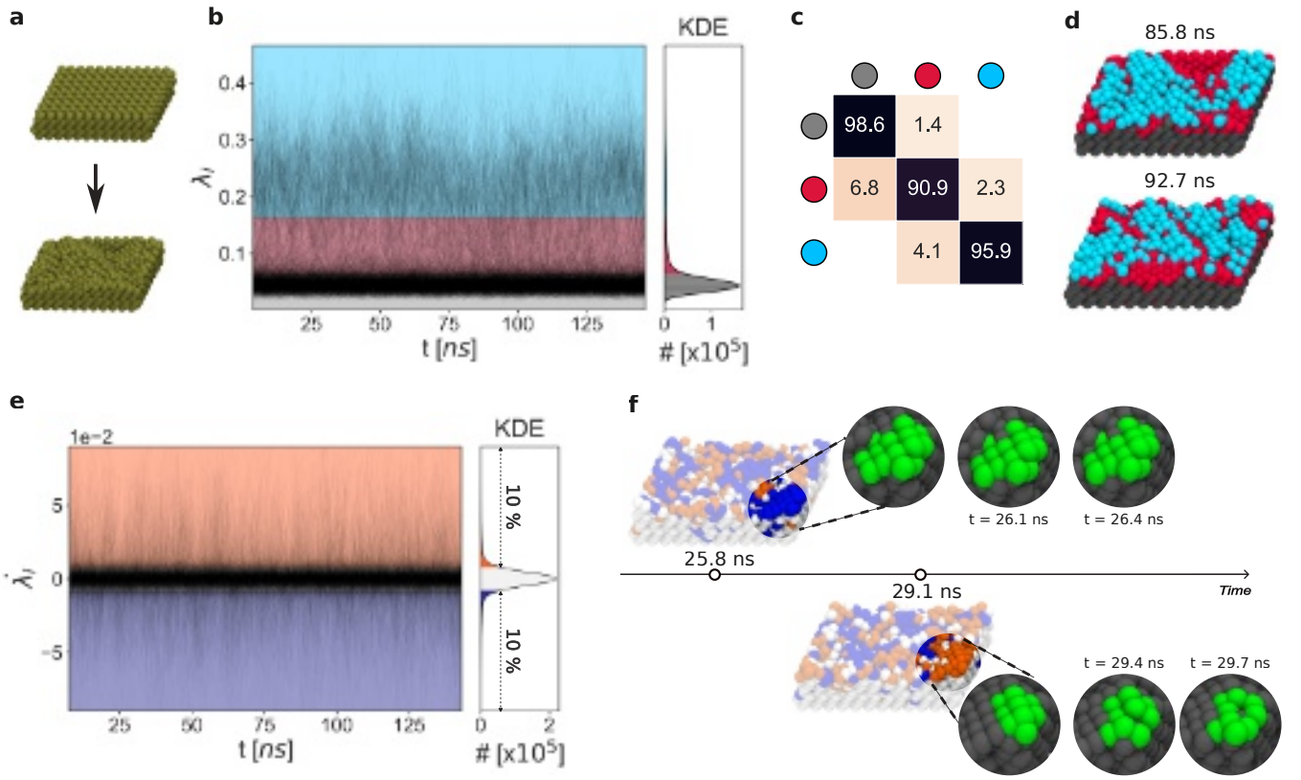


FIG. 4. τSOAP based analysis on a Copper FCC surface, Cu(210), composed of 2304 atoms. (a) MD snapshots of ideal Cu(210) surface (top), and equilibrated one at $T = 700$ K (bottom). (b) $\lambda_i(t)$ signals and the related KDE for Cu atoms of system in (a). The final $k = 3$ macro-clusters identified by KMeans are shown in gray, crimson, and cyan. (c) Exchange probability matrix of the final $k = 3$ detected macro-clusters. (d) MD snapshots showing the three main clusters identified in b): crystalline bulk in gray, sub-surface region in crimson, more dynamic surface atoms in cyan. (e) Domain detection based on $\dot{\lambda}(t)$ profiles and the related KDE: the blue domain is associated to the first decile, the orange cluster is linked to the tenth decile, the white domain is related to all the other deciles. (f) MD snapshots of blue, orange and white domains (in transparency) related to two different frames. Top: at $t = 25.8$ ns, the circled portion of the surface exhibits a prevalence of blue domain, thus predicting stable reconfigurations in the two successive frames (green atoms). Bottom: at $t = 29.1$ ns, the same circled portion exhibits a predominance of orange cluster, thus predicting dyanmic reconfigurations in the two successive frames (green atoms).

mains related to Cu atoms exhibiting very competing behaviors (Fig. 4b): one dense and more static cluster in gray along with two less populated but more dynamic domains in red and cyan. The exchange probability matrix in Fig. 4c points out that the transient events among diverse domains mainly engage Cu atoms belonging to the red and cyan clusters. Fig. 4d graphically represents the identified clusters at two explicative time steps, $t = 85.8$ ns and $t = 92.7$ ns: not surprisingly, the gray domain corresponds to the crystalline bulk of the Cu(210) surface, reasonably detected by our analysis as the most static with small local environment variations (low $\lambda_i(t)$ values); on the other hand, the surface atoms are identified as more dynamic clusters, thereby including all $\lambda_i(t) > 0.07$. However, two sub-surface regions are recognized by KMeans: in crimson, a domain characterized by $0.07 < \lambda_i(t) \leq 0.16$, and in cyan, a cluster with the highest local environment variations. The two MD snapshots in Fig. 4d show the significant correspondence between that more static surface region (crimson) and more stable surface atomic arrangements with increased coordination (see also supplementary material, Movie S3).

The Cu(210) domain characterization based on $\dot{\lambda}_i$ confirms the effectiveness of this analysis in providing some key information on the time evolution of the material structure. Fig. 4e highlights, also in this case, that the average rate of the local environment variations is null, *i.e.*, most of the Cu atoms in Cu(210) show a steady-state behavior of $\lambda_i(t)$. In addition, the cluster representation in Fig. 4f suggests that the domain with $\dot{\lambda}_i \sim 0$ essentially corresponds to the ice crystalline bulk (in white). On the other hand, most of the surface atoms are highly dynamic, and consequentially a balance between domains with positive (orange) or negative (blue) $\dot{\lambda}_i$ is established over time. Consistent with the test cases discussed above, this dynamic balance indicates that any substantial reconfiguration toward more stable/dynamic arrangements is not occurring. Nevertheless, some cluster details are interesting to be noticed in Fig. 4f: the snapshot on the top, corresponding to $t = 25.8$ ns, exhibits a portion of the surface with a clear predominance of atoms evolving toward more static configurations (in blue); the zoom onto that portion clarifies

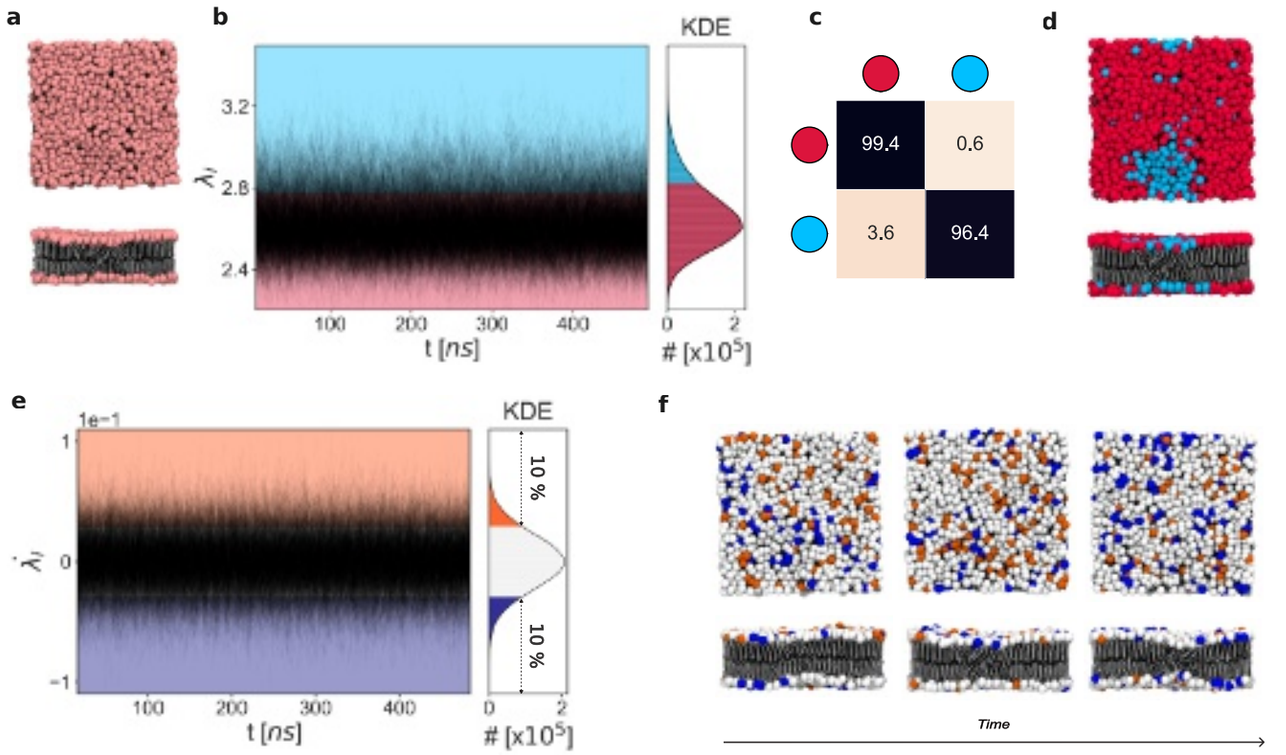


FIG. 5. τ SOAP-based analysis on a lipid bilayer composed of 1152 DPPC lipids at $T = 293$ K. (a) MD snapshots of DPPC lipid bilayer (top and lateral views). (b) $\lambda_i(t)$ signals and the related KDE for all the phosphate atoms of all lipid molecules in the DPPC bilayer (a), along the last 500 ns of the MD trajectory. KMeans clustering identifies $k = 2$ final macro-clusters shown with crimson and cyan. (c) Exchange probability matrix of the final $k = 2$ macro-clusters. (d) MD snapshot of the two detected domains (same color code as (b)): gel phase in crimson, and liquid phase in cyan. (e) Domain detection based on $\dot{\lambda}(t)$ profiles and the related KDE: the blue domain is associated to the first decile; the orange cluster is linked to the tenth decile; the white domain is related to all the other deciles. (f) MD snapshots of blue, orange, and white domains related to different frames along the trajectory.

its stability in the two successive sampled times ($t = 26.1$ and $t = 26.4$). On the contrary, when the same surface region is characterized, after some frames, by a prevalence of positive $\dot{\lambda}_i$ (in orange in the snapshot on the bottom at $t = 29.1$ ns), the associated atoms experience an evident rearrangement as highlighted by the green atoms onto the zoom. Again on this system, $\dot{\lambda}_i$ was revealed to be useful in predicting the evolution toward more static/more dynamic configurations.

C. Phases coexistence in soft dynamical systems

As a final test case, we apply our τSOAP-based dynamics domain recognition on a soft system characterized by a two-phase coexistence, *i.e.*, gel and liquid. Specifically, we analyze the last 500 ns of 1 μs-long CG-MD simulation of a DPPC lipid bilayer composed of 1152 self-assembled DPPC lipids (see Fig. 5a) at $T = 293$ K, thus considering the last 5001 consecutive frames ($\Delta t = 0.1$ ns). Although the gel-to-liquid transition temperature of a DPPC membrane is at ~ 315 K, here we investigate the dynamics of the lipid bilayer at a slightly lower temperature, thereby avoiding addressing the critical dynamics issues occurring at the transition temperature.

Our clustering analysis, displayed in Fig. 5b on $\lambda_i(t)$ profiles and on the related KDE, identifies two main dynamics domains: one colored in crimson including $\lambda_i < 2.8$; and the other one in cyan, containing the highest values of τSOAP fingerprints. Fig. 5c reports, instead, the interconversion matrix between the two clusters. Beyond a small probability (3.6 %) to transient from cyan to red cluster, the lipids manifest relatively high stability to preserve, along the complete trajectory, a specific local environment variation (λ_i), typical of each individual dynamics cluster. The graphical representation of the lipid bilayer in Fig. 5d suggests a close link between the dynamics domains and the phase states: the crimson cluster characterized by small λ_i is indeed associated to a more static - gel - phase; while the cyan domain, with higher local environment variations, is connected to a more dynamic - liquid - phase. Although the ability of SOAP to distinguish environments characterized by diverse structural features is known [53,57], this case demonstrates how τSOAP is able to clearly detect the nucleation and emergence of distinct dynamic do-

mains in intrinsically disordered systems, in a very agile and efficient way. This offers an additional proof of the versatility and robustness of this descriptor.

The further analysis on $\dot{\lambda}_i(t)$, reported in Fig. 5e, detects a predominance of $\dot{\lambda}_i \sim 0$ (white cluster) along with a balance between positive (orange) and negative (blue) peaks over time. We recall indeed that a null time-derivative of $\dot{\lambda}_i(t)$ represents the behavior of those units exhibiting a constant variation of their local environment, with some statistical oscillations classified in the orange and blue domain. Within such a resulting scenario, the proposed analysis predicts gel-liquid phases coexistence in a dynamic equilibrium, as shown in the MD snapshots in Fig. 5f. In other words, the lipids are not evolving toward a more static/dynamic configuration, whereas each remains in its proper dynamics domain.

IV. CONCLUSIONS

Investigating the dynamics of individual units in many atomic/molecular systems is essential to understand the behavior of complex molecular systems, their physical and chemical properties, collective transitions, as well as to design next-generation materials and molecular systems with desirable dynamical behaviors[94]. However, because of the complexity of local structural environments along with their dynamics in such systems, a general approach is still lacking. Although faithful representations of atomic neighborhood environments - such as the SOAP descriptor - are available and widely employed, here we want to draw attention to the time evolution of these structures, which is typically overlooked in molecular motif recognition procedures.

In this work, we propose an alternative perspective allowing us to track the dynamical changes in atomic structural environments of the interacting sub-units, thus enhancing the detection of dynamics domains and emerging phenomena. Building upon the SOAP descriptor, we implement τSOAP, a new fingerprint that quantifies the variations of local SOAP environments surrounding each constituent unit along its MD trajectory. τSOAP, indeed, retains the time information from the high-dimensional SOAP vectors, thereby aiming at emphasizing the importance of consequential events for reconstructing dynamics and detecting rare fluctuations. Coupled to an ML-based analysis, we demonstrate the potentiality of such an approach to identify domains with different structural and dynamical behaviors. Ranging from an ice/liquid water system where solid-like and fluid-like domains coexist in a dynamic equilibrium, to solid-like materials, and soft matter presenting gel and liquid coexisting phases, we prove that our analysis reliably addresses phase transitions, rare dynamic events, and coexisting phases. Moreover, by estimating the first time derivative of τSOAP signal, we gain further information on the direction of the local structural changes. Indeed, besides detecting local rearrangements, the first time-derivative of τSOAP enables the characterization of their evolution toward either highly or weakly dynamic environments. Finally, we can envisage that descriptors like τSOAP, and its first time derivative, may be also interesting in enhanced sampling methods, where they can offer degrees of freedom along which accelerating systems' variations/transitions.

Nonetheless, τSOAP-based investigation presents some limitations. Although τSOAP signal tracks the evolution of each constituent unit along the whole MD trajectory, thus providing time history data, the coupled ML-based approach relies on the instantaneous values of local environment changes, without performing a time-series clustering for identifying dynamics domains. Importantly, time-series clustering and classification based on the frequency/duration of local environment variations could have a striking advantage in discriminating fluctuations leading up to significant structural changes in the system. Notably, by including in our ML-based framework the first time derivative of τSOAP we start providing some further insights on predicting the evolution of local changes, and specifically how selected environments reconstruct or evolve in time. In summary, our approach turned out to be robust and versatile to capture fluctuating environments from SOAP spectra in a variety of systems by means of a completely agnostic and data-driven analysis.

SUPPLEMENTARY MATERIAL

The supplementary material contains details about: the length of MD simulation trajectories and SOAP vectors parameters; the Elbow Curve Method profiles for the identification of the optimal number of final clusters; KMeans clustering analysis on the τSOAP data starting from $K = 10$ clusters with their relative transition probabilities and the associated dendrograms; SOAP+PCA based analyses related to TIP4P/ICE ice/liquid water system; SOAP+tICA based analyses related to TIP4P/ICE ice/liquid water system.

ACKNOWLEDGMENTS

G.M.P. acknowledges the support received by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement no. 818776 - DYNAPOL) and by the Swiss National Science Foundation (SNSF Grant IZLIZ2_183336).

DATA AVAILABILITY STATEMENT

Complete details of all molecular models used for the simulations, and of the simulation parameters (input files, etc.), as well as the complete *Time*SOAP analysis code, are available at: `https://github.com/GMPavanLab/TimeSOAP` (this temporary folder will be replaced with a definitive Zenodo archive upon acceptance of the final version of this paper). Further details on the analyses are provided in the supplementary material.

AUTHORS' CONTRIBUTIONS

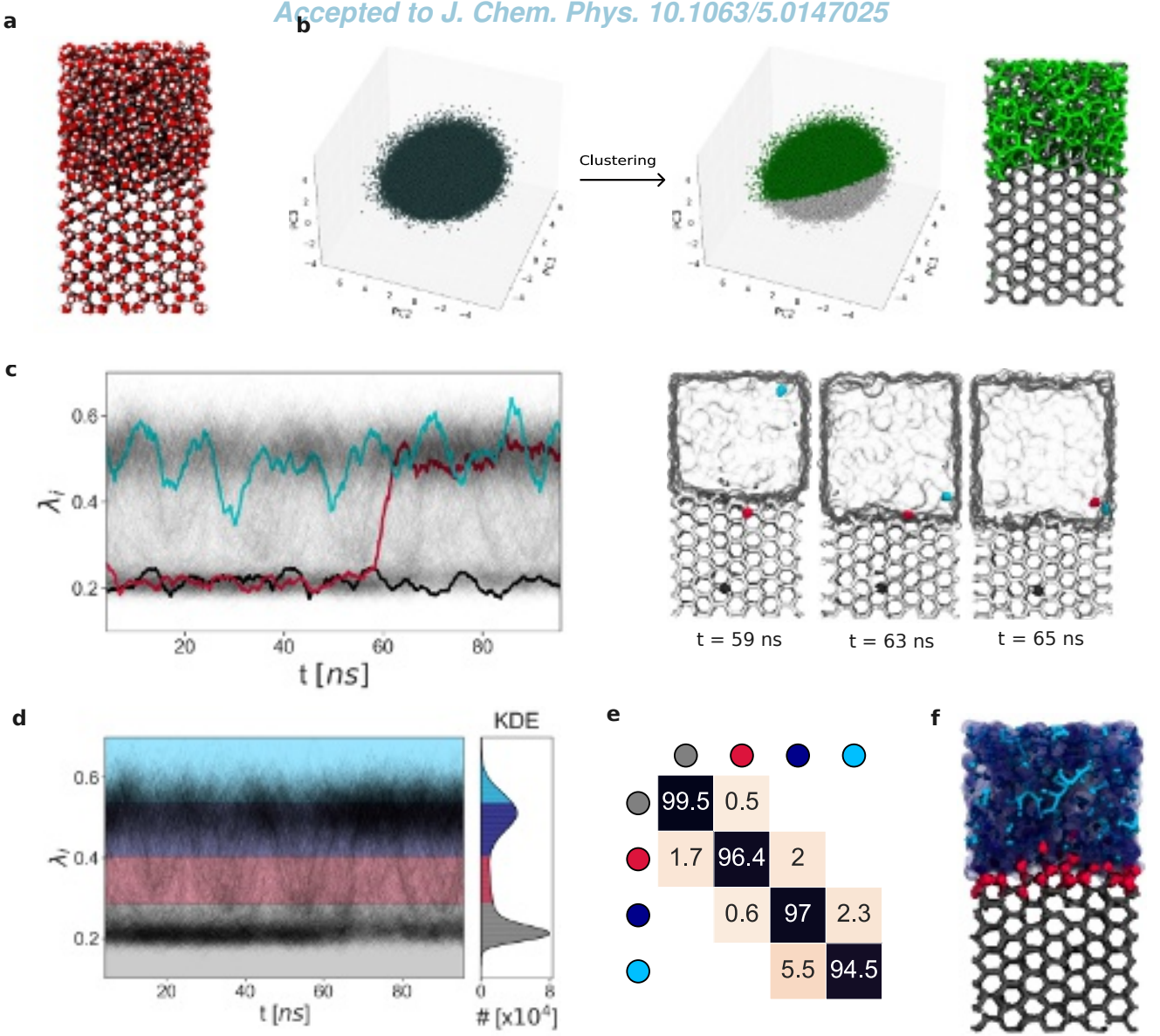
G.M.P. conceived this research and supervised the work. C.C. developed the descriptor and implemented the analysis. C.C., M.C. D.R., and A.C. performed the simulations and the analyses. All authors analyzed and discussed the results. C.C., A.C., and G.M.P. wrote the manuscript.

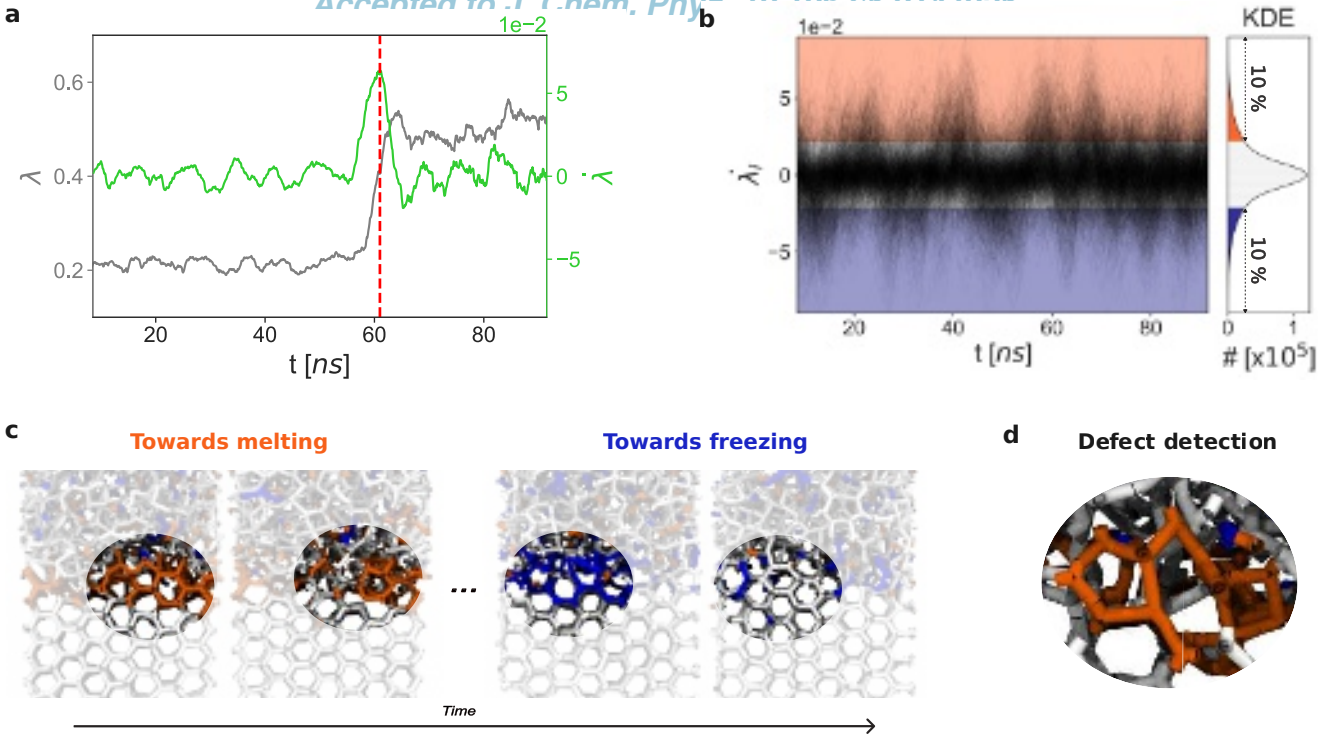
[1] J. Chapman, N. Goldman, and B. C. Wood, "Efficient and universal characterization of atomic structures through a topological graph order parameter," npj Computational Materials **8**, 1–12 (2022).
[2] S. K. Stavroglou, A. A. Pantelous, H. E. Stanley, and K. M. Zuev, "Unveiling causal interactions in complex systems," Proceedings of the National Academy of Sciences **117**, 7599–7605 (2020).
[3] D. Bochicchio, M. Salvalaglio, and G. M. Pavan, "Into the dynamics of a supramolecular polymer at submolecular resolution," Nature Communications **8**, 147 (2017).
[4] D. Bochicchio, S. Kwangmettatam, T. Kudernac, and G. M. Pavan, "How defects control the out-of-equilibrium dissipative evolution of a supramolecular tubule," ACS nano **13**, 4322–4334 (2019).
[5] L. Albertazzi, D. van der Zwaag, C. M. Leenders, R. Fitzner, R. W. van der Hofstad, and E. Meijer, "Probing exchange pathways in one-dimensional aggregates with super-resolution microscopy," Science **344**, 491–495 (2014).
[6] D. Wang, M. Hermes, S. Najmr, N. Tasios, A. Grau-Carbonell, Y. Liu, S. Bals, M. Dijkstra, C. B. Murray, and A. van Blaaderen, "Structural diversity in three-dimensional self-assembly of nanoplatelets by spherical confinement," Nature communications **13**, 6001 (2022).
[7] G. C. Sosso, T. F. Whale, M. A. Holden, P. Pedevilla, B. J. Murray, and A. Michaelides, "Unravelling the origins of ice nucleation on organic crystals," Chemical science **9**, 8077–8088 (2018).
[8] A. M. Goryaeva, C. Lapointe, C. Dai, J. Dérès, J.-B. Maillet, and M.-C. Marinica, "Reinforcing materials modelling by encoding the structures of defects in crystalline solids into distortion scores," Nature communications **11**, 4691 (2020).
[9] I. Lifshitz and A. Kosevich, "The dynamics of a crystal lattice with defects," Reports on Progress in Physics **29**, 217 (1966).
[10] M. T. Dove, *Introduction to lattice dynamics*, 4 (Cambridge university press, 1993).
[11] Z. H. Stachurski, "On structure and properties of amorphous materials," Materials **4**, 1564–1598 (2011).
[12] Q. Li, Y. Xu, S. Zheng, X. Guo, H. Xue, and H. Pang, "Recent progress in some amorphous materials for supercapacitors," Small **14**, 1800426 (2018).
[13] W.-X. Zhou, Y. Cheng, K.-Q. Chen, G. Xie, T. Wang, and G. Zhang, "Thermal conductivity of amorphous materials," Advanced Functional Materials **30**, 1903829 (2020).
[14] S. Yan, K. Abhilash, L. Tang, M. Yang, Y. Ma, Q. Xia, Q. Guo, and H. Xia, "Research advances of amorphous metal oxides in electrochemical energy storage and conversion," Small **15**, 1804371 (2019).
[15] C. W. Rosenbrock, E. R. Homer, G. Csányi, and G. L. Hart, "Discovering the building blocks of atomic systems using machine learning: application to grain boundaries," NPJ Computational Materials **3**, 29 (2017).
[16] A. Cardellini, F. Jiménez-Ángeles, P. Asinari, and M. Olvera de la Cruz, "A modeling-based design to engineering protein hydrogels with random copolymers," ACS nano **15**, 16139–16148 (2021).
[17] F. Tantakitti, J. Boekhoven, X. Wang, R. V. Kazantsev, T. Yu, J. Li, E. Zhuang, R. Zandi, J. H. Ortony, C. J. Newcomb, *et al.*, "Energy landscapes and functions of supramolecular systems," Nature materials **15**, 469–476 (2016).
[18] K. Carter-Fenk, K. U. Lao, K.-Y. Liu, and J. M. Herbert, "Accurate and efficient ab initio calculations for supramolecular complexes: Symmetry-adapted perturbation theory with many-body dispersion," The journal of physical chemistry letters **10**, 2706–2714 (2019).
[19] P. W. Frederix, I. Patmanidis, and S. J. Marrink, "Molecular simulations of self-assembling bio-inspired supramolecular systems and their connection to experiments," Chemical Society Reviews **47**, 3470–3489 (2018).
[20] O.-S. Lee, V. Cho, and G. C. Schatz, "Modeling the self-assembly of peptide amphiphiles into fibers using coarse-grained molecular dynamics," Nano letters **12**, 4907–4913 (2012).
[21] K. K. Bejagam and S. Balasubramanian, "Supramolecular polymerization: a coarse grained molecular dynamics study," The Journal of Physical Chemistry B **119**, 5738–5746 (2015).
[22] C. Perego, L. Pesce, R. Capelli, S. J. George, and G. M. Pavan, "Multiscale molecular modelling of atp-fueled supramolecular polymerisation and depolymerisation," ChemSystemsChem **3**, e2000038 (2021).
[23] D. Bochicchio and G. M. Pavan, "From cooperative self-assembly to water-soluble supramolecular polymers using coarse-grained simulations," ACS nano **11**, 1000–1011 (2017).
[24] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Physical review letters **98**, 146401 (2007).
[25] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Physical review letters **104**, 136403 (2010).
[26] M. Crippa, C. Perego, A. L. de Marco, and G. M. Pavan, "Molecular communications in complex systems of dynamic supramolecular polymers," Nature Communications **13**, 2162 (2022).
[27] P. C. Souza, R. Alessandri, J. Barnoud, S. Thallmair, I. Faustino, F. Grünewald, I. Patmanidis, H. Abdizadeh, B. M. Bruininks, T. A. Wassenaar, *et al.*, "Martini 3: a general purpose force field for coarse-grained molecular dynamics," Nature methods **18**, 382–388 (2021).
[28] C. Abrams and G. Bussi, "Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration," Entropy **16**, 163–199 (2013).
[29] A. L. Ferguson, "Machine learning and data science in soft materials engineering," Journal of Physics: Condensed Matter **30**, 043002 (2017).
[30] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature **559**, 547–555 (2018).
[31] N. E. Jackson, M. A. Webb, and J. J. de Pablo, "Recent advances in machine learning towards multiscale soft materials design," Current Opinion in Chemical Engineering **23**, 106–114 (2019).
[32] F. Häse, L. M. Roch, P. Friederich, and A. Aspuru-Guzik, "Designing and understanding light-harvesting devices with machine learning," Nature Communications **11**, 4587 (2020).
[33] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, and A. Laio, "Unsupervised learning methods for molecular simulation data," Chemical Reviews **121**, 9722–9758 (2021).
[34] P. J. Steinhardt and P. Chaudhari, "Point and line defects in glasses," Philosophical Magazine A **44**, 1375–1381 (1981).
[35] G. Ackland and A. Jones, "Applications of local crystal structure measures in experiment and simulation," Physical Review B **73**, 054104 (2006).
[36] J. D. Honeycutt and H. C. Andersen, "Molecular dynamics study of melting and freezing of small lennard-jones clusters," Journal of Physical Chemistry **91**, 4950–4963 (1987).
[37] A. Stukowski, "Structure identification methods for atomistic simulations of crystalline materials," Modelling and Simulation in Materials Science and Engineering **20**, 045021 (2012).
[38] B. Cheng, R.-R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter, *et al.*, "Mapping materials and molecules," Accounts of Chemical Research **53**, 1981–1991 (2020).
[39] M. Ceriotti, "Unsupervised machine learning in atomistic simulations, between predictions and understanding," The Journal of chemical physics **150**, 150901 (2019).
[40] J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. De Fabritiis, F. Noé, and C. Clementi, "Machine learning of coarse-grained molecular dynamics force fields," ACS central science **5**, 755–767 (2019).
[41] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," The Journal of chemical physics **134**, 074106 (2011).
[42] N. Artrith, A. Urban, and G. Ceder, "Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species," Physical Review B **96**, 014112 (2017).
[43] R. Batra, H. D. Tran, C. Kim, J. Chapman, L. Chen, A. Chandrasekaran, and R. Ramprasad, "General atomic neighborhood fingerprint for machine learning-based methods," The Journal of Physical Chemistry C **123**, 15859–15866 (2019).
[44] A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, and R. Ramprasad, "Solving the electronic structure problem with machine learning,"

npj Computational Materials **5**, 1–7 (2019).
[45]A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Physical Review B **87**, 184115 (2013).
[46]R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials," Physical Review B **99**, 014104 (2019).
[47]B. Monserrat, J. G. Brandenburg, E. A. Engel, and B. Cheng, "Liquid water contains the building blocks of diverse ice phases," Nature communications **11**, 5757 (2020).
[48]A. Offei-Danso, A. Hassanali, and A. Rodriguez, "High-dimensional fluctuations in liquid water: Combining chemical intuition with unsupervised learning," Journal of Chemical Theory and Computation **18**, 3136–3150 (2022).
[49]S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," Physical Chemistry Chemical Physics **18**, 13754–13769 (2016).
[50]A. Reinhardt, C. J. Pickard, and B. Cheng, "Predicting the phase diagram of titanium dioxide with random search and pattern recognition," Physical Chemistry Chemical Physics **22**, 12697–12705 (2020).
[51]P. Gasparotto, D. Bochicchio, M. Ceriotti, and G. M. Pavan, "Identifying and tracking defects in dynamic supramolecular polymers," The Journal of Physical Chemistry B **124**, 589–599 (2019).
[52]A. Cardellini, M. Crippa, C. Lionello, S. P. Afrose, D. Das, and G. M. Pavan, "Unsupervised data-driven reconstruction of molecular motifs in simple to complex dynamic micelles," J. Phys. Chem. B **127**, 2595–2608 (2023).
[53]R. Capelli, A. Gardin, C. Empereur-Mot, G. Doni, and G. M. Pavan, "A data-driven dimensionality reduction approach to compare and classify lipid force fields," The Journal of Physical Chemistry B **125**, 7785–7796 (2021).
[54]C. Lionello, C. Perego, A. Gardin, R. Klajn, and G. M. Pavan, "Supramolecular semiconductivity through emerging ionic gates in ion–nanoparticle superlattices," ACS Nano **17**, 275–287 (2023).
[55]M. Cioni, D. Polino, D. Rapetti, L. Pesce, M. Delle Piane, and G. M. Pavan, "Innate dynamics and identity crisis of a metal surface unveiled by machine learning of atomic environments," J. Chem. Phys **158**, 124701 (2023).
[56]D. Rapetti, M. Delle Piane, M. Cioni, D. Polino, R. Ferrando, and G. M. Pavan, "Machine learning of atomic dynamics and statistical surface identities in gold nanoparticles. ChemRxiv [Preprint]," (2022), `https://chemrxiv.org/engage/chemrxiv/article-details/63642e6aac45c7a2a9a45332`.
[57]A. Gardin, C. Perego, G. Doni, and G. M. Pavan, "Classifying soft self-assembled materials via unsupervised machine learning of defects," Communications Chemistry **5**, 82 (2022).
[58]T. Bian, A. Gardin, J. Gemen, L. Houben, C. Perego, B. Lee, N. Elad, Z. Chu, G. M. Pavan, and R. Klajn, "Electrostatic co-assembly of nanoparticles with oppositely charged small molecules into static and dynamic superstructures," Nature chemistry **13**, 940–949 (2021).
[59]E. Facco, M. d'Errico, A. Rodriguez, and A. Laio, "Estimating the intrinsic dimension of datasets by a minimal neighborhood information," Sci. Rep. **7**, 1–8 (2017).
[60]J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," science **290**, 2319–2323 (2000).
[61]R. R. Coifman and S. Lafon, "Diffusion maps," Applied and computational harmonic analysis **21**, 5–30 (2006).
[62]B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural computation **10**, 1299–1319 (1998).
[63]C. R. Schwantes and V. S. Pande, "Modeling molecular kinetics with tica and the kernel trick," Journal of chemical theory and computation **11**, 600–608 (2015).
[64]S.-T. Tsai, Z. Smith, and P. Tiwary, "Sgoop-d: Estimating kinetic distances and reaction coordinate dimensionality for rare event systems from biased/unbiased simulations," Journal of Chemical Theory and Computation **17**, 6757–6765 (2021).
[65]M. Crippa, A. Cardellini, C. Caruso, and G. M. Pavan, "Detecting dynamic domains and local fluctuations in complex molecular systems via time-lapse neighbors shufflingt. arXiv [Preprint]," (2022), `https://arxiv.org/abs/2212.12694`.
[66]B. A. Helfrecht, P. Gasparotto, F. Giberti, and M. Ceriotti, "Atomic motif recognition in (bio) polymers: Benchmarks from the protein data bank," Frontiers in molecular biosciences **6**, 24 (2019).
[67]M. J. Willatt, F. Musil, and M. Ceriotti, "Atom-density representations for machine learning," The Journal of chemical physics **150**, 154110 (2019).
[68]"Soapify," `https://github.com/GMPavanLab/SOAPify`.
[69]K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," The London, Edinburgh, and Dublin philosophical magazine and journal of science **2**, 559–572 (1901).
[70]H. Hotelling, "Analysis of a complex of statistical variables into principal components." Journal of educational psychology **24**, 417 (1933).
[71]C. Zeni, A. Anelli, A. Glielmo, and K. Rossi, "Exploring the robust extrapolation of high-dimensional machine learning potentials," Physical Review B **105**, 165141 (2022).
[72]C. Harris, K. Millman, S. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, and S. Berg, "Smith 474 nj," Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del R'io JF, Wiebe M, Peterson P, G'erard-475 Marchant P, et al. Array programming with NumPy. Nature **585**, 357–362 (2020).
[73]W. H. Press and S. A. Teukolsky, "Savitzky-golay smoothing filters," Computers in Physics **4**, 669–672 (1990).
[74]P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," Nature Methods **17**, 261–272 (2020).
[75]S. Lloyd, "Least squares quantization in pcm," IEEE Trans. Inf. Theory **28**, 129–137 (1982).
[76]J. MacQueen, "Classification and analysis of multivariate observations," *5th Berkeley Symp. Math. Statist. Probability*, , 281–297 (1967).
[77]A. Ladd and L. Woodcock, "Triple-point coexistence properties of the lennard-jones system," Chemical Physics Letters **51**, 155–159 (1977).
[78]A. Ladd and L. Woodcock, "Interfacial and co-existence properties of the lennard-jones system at the triple point," Molecular Physics **36**, 611–619 (1978).
[79]B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl, "Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation," Journal of chemical theory and computation **4**, 435–447 (2008).
[80]J. Abascal, E. Sanz, R. García Fernández, and C. Vega, "A potential model for the study of ices and amorphous water: Tip4p/ice," The Journal of chemical physics **122**, 234511 (2005).
[81]R. García Fernández, J. L. Abascal, and C. Vega, "The melting point of ice i h for common water models calculated from direct coexistence of the solid-liquid interface," The Journal of chemical physics **124**, 144506 (2006).
[82]M. Matsumoto, T. Yagasaki, and H. Tanaka, "Genice: hydrogen-disordered ice generator," (2018).
[83]M. Bernetti and G. Bussi, "Pressure control using stochastic cell rescaling," The Journal of Chemical Physics **153**, 114107 (2020).
[84]D. Frenkel, "Simulations: The dark side," The European Physical Journal Plus **128**, 10 (2013).
[85]R. P. Gupta, "Lattice relaxation at a metal surface," Physical Review B **23**, 6265 (1981).
[86]A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, *et al.*, "Lammps-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales," Computer Physics Communications **271**, 108171 (2022).
[87]H. Wang, L. Zhang, J. Han, and E. Weinan, "Deepmd-kit: A deep learning package for many-body potential energy representation and molecular dynamics," Computer Physics Communications **228**, 178–184 (2018).
[88]S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. De Vries, "The martini force field: coarse grained model for biomolecular simulations," The journal of physical chemistry B **111**, 7812–7824 (2007).
[89]T. Bryk and A. Haymet, "Ice 1h/water interface of the spc/e model: Molecular dynamics simulations of the equilibrium basal and prism interfaces," The Journal of chemical physics **117**, 10258–10268 (2002).

[90]C. Moritz, P. L. Geissler, and C. Dellago, "The microscopic mechanism of bulk melting of ice," The Journal of chemical physics **155**, 124501 (2021).
[91]L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," Physical review letters **72**, 3634 (1994).
[92]G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for markov model construction," The Journal of chemical physics **139**, 015102 (2013).
[93]T. D. Daff, I. Saadoune, I. Lisiecki, and N. H. de Leeuw, "Computer simulations of the effect of atomic structure and coordination on the stabilities and melting behaviour of copper surfaces and nano-particles," Surface science **603**, 445–454 (2009).
[94]T. Xie, A. France-Lanord, Y. Wang, Y. Shao-Horn, and J. C. Grossman, "Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials," Nature communications **10**, 2667 (2019).
[95]R. Capelli, F. Muniz-Miranda, and G. M. Pavan, "Ephemeral ice-like local environments in classical rigid models of liquid water," The Journal of Chemical Physics **156**, 214503 (2022).

Accepted to J. Chem. Phys. 10.1063/5.0147025





Accepted to J. Chem. Phys. 10.1063/5.0147025