Machine learning of microscopic structure-dynamics relationships in complex molecular systems

Machine learning of microscopic structure-dynamics relationships in complex molecular systems

IOP Publishing Mach. Learn.: Sci. Technol. 4 (2023) 045044 https://doi.org/10.1088/2632-2153/ad0fa5

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
13 September 2023

REVISED
6 November 2023

ACCEPTED FOR PUBLICATION
20 November 2023

PUBLISHED
6 December 2023

PAPER

Machine learning of microscopic structure-dynamics relationships in complex molecular systems

Martina Crippa[1], **Annalisa Cardellini**[2], **Matteo Cioni**[1], **Gábor Csányi**[3] 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
[3] Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: giovanni.pavan@polito.it

Keywords: molecular motifs, complex molecular systems, structure-dynamics relationships, microscopic analysis, high-dimensional descriptors, smooth overlap of atomic positions, complex dynamics

Supplementary material for this article is available online

Abstract

In many complex molecular systems, the macroscopic ensemble's properties are controlled by microscopic dynamic events (or fluctuations) that are often difficult to detect via pattern-recognition approaches. Discovering the relationships between local structural environments and the dynamical events originating from them would allow unveiling microscopic-level structure-dynamics relationships fundamental to understand the macroscopic behavior of complex systems. Here we show that, by coupling advanced structural (e.g. Smooth Overlap of Atomic Positions, SOAP) with local dynamical descriptors (e.g. Local Environment and Neighbor Shuffling, LENS) in a unique dataset, it is possible to improve both individual SOAP- and LENS-based analyses, obtaining a more complete characterization of the system under study. As representative examples, we use various molecular systems with diverse internal structural dynamics. On the one hand, we demonstrate how the combination of structural and dynamical descriptors facilitates decoupling relevant dynamical fluctuations from noise, overcoming the intrinsic limits of the individual analyses. Furthermore, machine learning approaches also allow extracting from such combined structural/dynamical dataset useful microscopic-level relationships, relating key local dynamical events (e.g. LENS fluctuations) occurring in the systems to the local structural (SOAP) environments they originate from. Given its abstract nature, we believe that such an approach will be useful in revealing hidden microscopic structure-dynamics relationships fundamental to rationalize the behavior of a variety of complex systems, not necessarily limited to the atomistic and molecular scales.

1. Introduction

The macroscopic behavior of complex systems is often influenced by fluctuations that, while being fundamental for comprehending the systems' dynamics, are challenging to detect and control. This also holds true at the molecular scale, where phenomena such as nucleation, defect propagation, and phase transitions are intricately linked to these fluctuations. The integration of advanced molecular descriptors with machine learning (ML) has been playing a key role in analyzing molecular trajectories, contributing to a better understanding of diverse nanoscale systems, ranging from atomistic to supramolecular levels [1–11]. Standard human-based descriptors, tailored for building detailed analyses and investigating specific systems like, i.e. ice-water interfaces [12] or metal clusters [13, 14], have increasingly left more and more space to abstract descriptors, [15–21] often combined with supervised and unsupervised ML methods [1–10]. These ML-based techniques offer valuable insights into the structural and dynamical properties of the systems. [5–10] While human-based approaches provide an accurate comprehension of intrigued physical–chemical

mechanisms, they heavily rely on in-depth prior knowledge of the system, limiting their transferability. On the contrary, the use of abstract descriptors allows more general representations and outlines a broader picture of the system behavior, eventually managing a large amount of high-dimensional data which are often difficult to rationalize. Widely recognized approaches based on dimensionality reduction principles (e.g. linear principal component analysis (PCA), kernel-PCA [22], t-distributed stochastic neighbor embedding [23]), are frequently employed to extract information from such descriptors related dataset, then classifying the reduced dataset with diverse clustering methods (e.g. kmeans [24], Gaussian mixture models (GMM) [25], DBSCAN [26], HDBSCAN [27]) to facilitate its interpretation. However, when relying on structure-based descriptors, these approaches have limitations: while they effectively detect dominant structural environments in the system, they may fail to capture local time-dependent events that are sparsely observed within the trajectory. These transitions, although statistically insignificant, have revealed a crucial role in the overall behavior of the system [28, 29]. The absence of an adaptive resolution that allows to catch non-dominant events presents two challenges: firstly, it leads to a loss of information by failing to detect fluctuations within the system, and secondly, these fluctuations may be inaccurately classified within the dominant clusters, thereby contaminating them.

In recent studies, we have developed dynamic descriptors that are very efficient in capturing the local dynamic environments of atoms in complex molecular systems from structural information/identity-based information. [28, 29] By monitoring these descriptors over time along the trajectory, we can effectively capture dynamic behaviors, including local and sparse events within the system. In particular, we introduced a dynamical descriptor, Local Environments and Neighbors Shuffling (LENS) [28], which considers the interacting particles as distinct individuals monitoring how much the list of neighbor particles (of each particle i) changes over time, for example at each sampled Δt. As follows, LENS provides information on the reshuffling of the local neighbor's environments surrounding each unit i in the system along the trajectory. However, at the same time, a descriptor like LENS retains very limited structural information: if, e.g. the neighbor units rearrange locally, while remaining within the neighborhood in Δt, LENS would not detect any signal (such events are vice-versa well captured by structural descriptors such as tSOAP [29]). Thus, while LENS can detect the local dynamics of the system, it does not allow to determine, e.g. the specific structural environment from which dynamic events originate.

Here we demonstrate how, combining structural (SOAP [15]) and dynamical (LENS [28]) descriptors, it is possible to obtain an improved characterization of the system. We compose a dataset where the SOAP spectra (n components each) are augmented with the LENS descriptor (an additional dimension), leading to significant technical and scientific advantages. Firstly, (i) it enables the separation of sparsely observed, but relevant, dynamic events/environments (e.g. fluctuations) from the noise in the SOAP dataset. As a result, (ii) the interpretation of SOAP and LENS (combined) not only provides a more accurate complete characterization, but the two descriptors improve each other: the addition of LENS yields an enhanced SOAP structural classification. Furthermore, (iii) this allows identifying unique microscopic structure-dynamics relationships, showing e.g. which local SOAP structural environments generate a certain type of dynamical event along the sampled molecular dynamics (MD) trajectory. In this work we demonstrate the efficiency and abstraction of this approach on diverse molecular systems, employed herein as case studies.

2. Results

As a first case study, we focus on a copper **Cu(211)** face-centered cubic (FCC) metal surface recently demonstrated to possess non-trivial internal atomic dynamics. Metals are known to display interesting dynamic behavior even well below the melting temperature [30, 31]. For example, when simulated at $T =$ 600 K, the **Cu(211)** FCC slab of figure 1 exhibits a surface with structurally diverse environments, as made evident by a simple coordination analysis, and a non-trivial internal atomic dynamics (figure 1(a), right: dynamical atomic rearrangements). Unveiling the underlying mechanism behind such dynamics is essential to understand the properties of these metal systems [32–34]. Moreover, the comprehension of structural-dependent features plays a fundamental role in practical applications such as heterogeneous catalysis, mechanical properties, etc [35–38]. SOAP-based and LENS-based ML analyses have been recently employed to analyze MD simulation trajectories of metals below the melting temperature (including, e.g. copper surfaces as that of figure 1) [9, 28, 29]. Although a structural-descriptor-based analysis, such as that one using SOAP combined with dimensionality reduction and density-based clustering, captures the most prevalent and dominant conformation domains within the system, a pure LENS analysis based on the reshuffling of the neighborhood over time, catches the dynamical features of the system (see figure 1(b)).

Here, we investigate a **Cu(211)** FCC copper slab using a preexisting MD trajectory composed of $N =$ 2400 atoms simulated via a DeepMD-based potential [39] for 150 ns (see Cioni *et al* [9] for details). To examine both the structural and dynamical properties of the **Cu(211)** system, we firstly adopted a similar

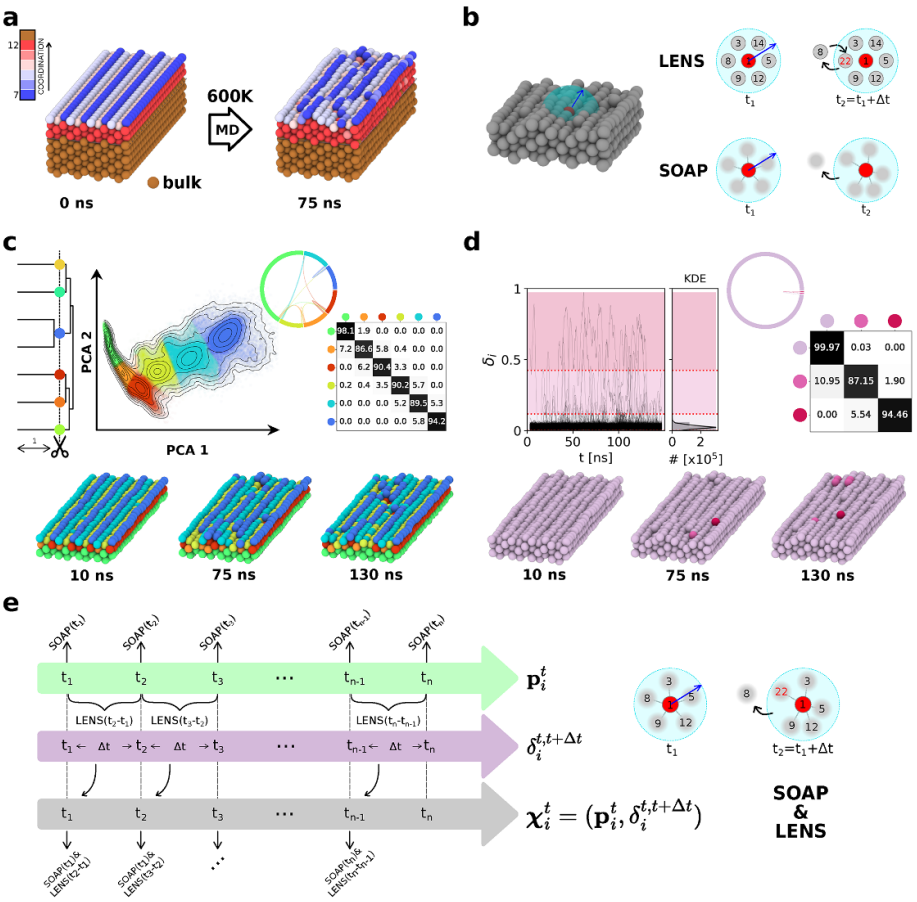

Figure 1. Flow of the analysis. (a) FCC 211 copper slab snapshots colored by atom coordination (excluding bulk) at $t = 0$ ns and after 75 ns of an MD trajectory at $T = 600$ K. (b) LENS and SOAP: given the local neighborhood (cyan sphere) of each atom (red atom) in the system, LENS tracks the *identity* of the neighbor atoms within it (no information on their geometrical organization is retained), while SOAP captures their structural arrangement (without tracking their identity: it is a permutationally-invariant description). (c) SOAP-based analysis of **Cu(211)** system. Left: Hierarchical Clustering based dendrogram (from an HDBSCAN* classification, see figure S1(a)) and dendrogram cutting, defining the merged macro-clusters. Middle: PCA of the SOAP dataset (first two principal components), colored based on the detected macro-clusters. Right: chord diagram (fluxes) and transition probability matrix for the dynamical transitions between the macro-clusters (SOAP environments). Bottom: surface MD snapshots where atoms are colored based on the classification: bulk atoms in green, sub-surface in orange and red, surface 'valleys' in yellow, faces in cyan, and edges in blue. (d) LENS analysis of **Cu(211)** at 600 K. Left: LENS time-series and classification [28]. Right: chord diagram (fluxes) and transition probability matrix. Bottom: MD snapshots with atoms colored based on the LENS clusters: more/less dynamic atoms in brighter/lighter colors. (e) Scheme of the SOAP&LENS combined data set: the SOAP power spectrum of each particle at every time step ($\mathbf{p}_i^t$) is combined with the LENS scalar value calculated at the subsequent time-interval ($\delta_i^{t+\Delta t}$), obtaining a new dataset $\boldsymbol{\chi}_i^t$.

bottom–up protocol as described in the study by Cioni *et al* [9]. This strategy includes, as a first step, a representation of the system via the SOAP descriptor. One SOAP spectrum is extracted for each of the 900 atoms (three top-most layers, although the SOAP spectra also consider the presence of the 1500 bottom-side atoms as neighbors, they are not included in the analysis because we are interested in the dynamics of the surface [9]) in 482 snapshots taken every $\Delta t = 0.3$ ns along 145 ns of MD simulation, for a total of 482×900 spectra. A PCA is then used to reduce the dimensionality of the SOAP spectra dataset, considering the first n-PCA components in order to retain at least 99.5% of the variance (see table S1 in supporting information for details). Unsupervised clustering algorithm (HDBSCAN* [27] or GMM [25]) can be finally adopted to rationalize the data and to identify the dominant atomic environments on the surface (colored clusters in the PCA of figure S1(a)). From the atoms' transition between the clusters over time, we compute a transition probability matrix. This reports the probability of an atom that is in a certain cluster at time t to remain in the same environment at time $t + \Delta t$ (i.e. after Δt: the temporal resolution of our analysis) or to undergo transition into a different cluster (see figure S1(a) for the micro-clusters transition matrix). From the transition probability matrix, we construct a Hierarchical Clustering based dendrogram merging the clusters with high dynamic interconnection (figure S1(a)). The dendrogram is cut in order to retrieve only meaningful clusters, colored accordingly in the PCA plot of figure 1(c), where only the first two PCA components are reported. The results demonstrate how SOAP can successfully distinguish diverse structural

environments within this system, including the bulk (green), subsurface (orange and red), surface valleys (yellow), faces and edges (cyan and blue), identified in different colors in figure 1(c). The dynamic interconnections between the various clusters (atomic environments) on the surface are also represented by the cord diagrams in figure 1(c) on the right: in these cord diagrams the dimensions of the arcs stand for the population of the various clusters, while the dimensions of the strings connecting them give visual information on how pronounced the atomic flow is in Δt, and thus on their dynamic interconnection. Moreover, we also obtained the transition probabilities matrix (% to undergo transition in $\Delta t = 0.3$ ns) between the Hierarchical Clustering -merged clusters (figure 1(c) right).

Separately, we also perform a LENS analysis on the same 482 snapshots extracted by the same MD trajectory. A LENS analysis of the system reveals intriguing surface events that are not captured (or highlighted) by the static SOAP-based analysis of structure as described above. Specifically, a few Cu atoms are seen to detach from the crystalline structure of the **Cu(211)** surface and to diffuse on it very fast. On the one hand, since these diffusing atoms are characterized by a high rate of reshuffling of their neighbors, they are clearly identified by LENS as a separate environment in the dataset (figure 1(d)). On the other hand, a comparison of figures 1(c) and 1(d) shows how, since these points are sparse and have negligible statistical weight in the dataset, these are overlooked in a pattern recognition approach such as that, e.g. of figure 1(c). In particular, in the SOAP analysis of figure 1(c), it is possible to note that the diffusing atoms (magenta in figure 1(d)), are merged to the SOAP cluster identifying the edges of **Cu(211)** surface. To address this limitation, here we developed a combined approach based on the basic assumption that a structural environment at a certain time might influence the dynamical events within the subsequent time interval. As shown schematically in figure 1(e), starting for example at time t_1, a SOAP spectrum $\mathbf{p}_i^{t_1}$ is computed for each particle i in the system. We also calculate its LENS value for the immediately subsequent time interval $\delta_i^{t_2-t_1}$. By including the LENS term as an extra-component into each SOAP power spectrum, we thus obtain a new vector $\boldsymbol{\chi}_i^{t_1} = (\mathbf{p}_i^{t_1}, \delta_i^{t_2-t_1})$ containing information on the structural properties in the neighbor environment surrounding atom i at time t_1 and its evolution in the subsequent time interval $t_2 - t_1$. The SOAP spectrum and LENS scalar component are opportunely normalized to have the same weight in the dataset (see Methods for details). Iterating this procedure for the whole trajectory, we thus obtain a new dataset (SOAP&LENS dataset) comprising $N = N_{\text{particle}} \times N_{\text{frames}}$ vectors, each one of dimension $n + 1$, where n is the SOAP spectrum dimension (structural information) and 1 the LENS (dynamical) component. Such updated dataset effectively contains information on the instantaneous environments surrounding each particle i and how they are prone to change over time at the resolution Δt (0.3 ns) of our analysis.

This method allows us to delineate a new concept for classification, as reported in figure 2. On the left side, figure 2(a) shows the PCA of the SOAP dataset projected onto the first two PCA components. On the right side, figure 2(a) shows the same projection for the new SOAP&LENS combined dataset (see Methods section for additional details). Notably, while the majority of the data has an almost identical distribution on the two PCAs, a distinct cloud of points appears as evidently separated from the rest in the combined dataset (top-right: highlighted by the red circle). Shown in figure 2(b), unsupervised HDBSCAN* clustering combined with Hierarchical Clustering based merging (in general, any other suitable clustering algorithm, e.g. GMM, DBSCAN, kmeans, would also work) reveals that such a separated domain on the SOAP&LENS PCA identifies a distinct, specific local environment. Note that the clustering parameters used for the analyses of figures 1(c) and 2(b) are exactly the same (see methods for details). This comparison shows how the classification of figure 1(c) (SOAP only) is enriched via the detection of a new LENS environment identified by the pink color (highlighted by the arrows in the transition matrix and chord diagram of figure 2(b)). As done for both the SOAP and LENS independent analyses, we can reconstruct the evolution of the detected environments by following the atomic environment belonging to all atoms at every time step (see the chord diagram and transition probability matrix in figure 2(b), right).

This analysis based on combining SOAP and LENS in a unique dataset offers distinct advantages over the purely SOAP-based approach. The decoupling of this additional pink LENS environment not only provides a more complete description of what happens in the **Cu(211)** surface at 600 K, but also improves the statistical precision in the classification of the SOAP environments. In fact, in differentiating the structural from the dynamical environments, the detection of the SOAP atomic environments in the SOAP&LENS dataset benefits from a reduced error. Notably, the PCA area identified by the red oval in figure 2(b), which corresponds in this analysis to a well-defined LENS atomic environment, merges into the SOAP atomic environments in the PCA of figure 1(c), creating errors and increased uncertainty. In this sense, when combined, two distinct descriptors such as, e.g. SOAP and LENS, complement and improve each other. Furthermore, such an approach also allows tracking the origin of local dynamical (LENS) fluctuations occurring on the surface, outlining microscopic structure-dynamics relationships. The off-diagonal entry in the matrix of figure 2(b) representing the transition of atoms from the edge atomic environment (in blue) to the pink (LENS) environment ($\sim$0.1 % probability) reveals that those atoms diffusing with high-speed on the

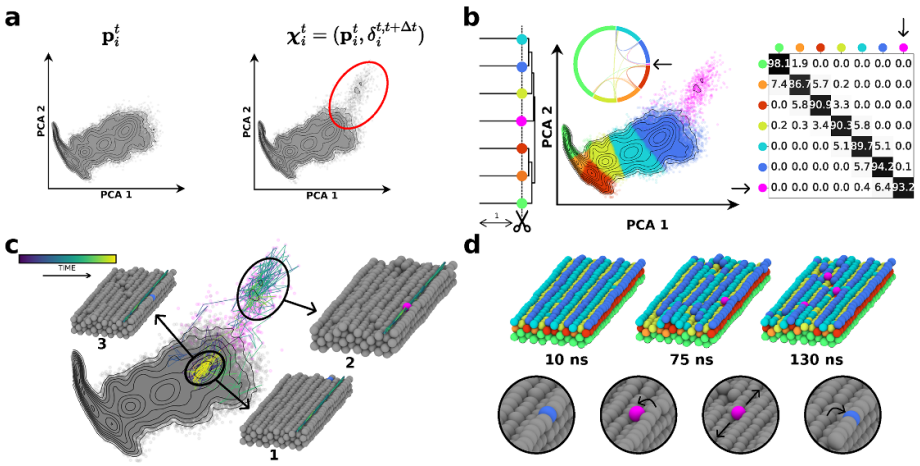

Figure 2. Combined SOAP&LENS analysis of a Cu(211) surface at 600 K. (a) Left: first two PCA components of the SOAP power spectra of the **Cu(211)** system. Right: first two PCA components of the SOAP&LENS combined descriptor: the new cloud of points emerging in the PCA projection of the χ vector is highlighted by the red circle. (b) SOAP&LENS based analysis of **Cu(211)** system. Left: hierarchical clustering based dendrogram (from an HDBSCAN* classification, see figure S2(a)) and dendrogram cutting, defining the merged macro-clusters (accordingly to clusters in figure 1(a), except for a new pink cluster). Middle: PCA of the SOAP&LENS dataset (first two principal components), colored based on the detected macro-clusters and chord diagram (fluxes). Right: transition probability matrix for the dynamical transition between the macro-clusters, highlighting the new cluster in pink. (c) Trajectory of an atom detaching from an edge, running on the surface, and re-entering into the edge. The trajectory is shown both on the PCA plot and on the snapshots, colored from blue to yellow according to the time evolution. (d) Three surface MD snapshots colored based on the classification: bulk atoms in green, sub-surface in orange and red, surface 'valleys' in yellow, faces in cyan, edges in blue and pink atoms detaching from the edges and running on the surface (an example of this process is reported in the zoom below).

metal surface come from the surface edges (see movie S1). After their creation and diffusion, such diffusing pink atoms are then again reabsorbed into the surface edges (∼6.4 % probability). The large imbalance between the probabilities for the creation and annihilation of these LENS diffusing atoms (figure 2(b) right, ∼0.1% vs. ∼6.4%) indicates that the emergence of such fast atoms is a rare event. Yet, it is clear that detecting such diffusing atoms is key for understanding the behavior of the system. Figure 2(c) provides an example of the structural variation of an atom undergoing such transition, following its trajectory both on the PCA plot and along the MD. The atom's trajectory is color-coded based on the MD simulation time, from dark blue to yellow, showing atoms that after residing within the surface edges (dark blue lines, example snapshot 1), detach and diffuse on the surface becoming part of this pink LENS environment (green lines, example snapshot 2), and then being reabsorbed into the edges (yellow lines, example snapshot 3). Figure 2(d) shows a complete representation of the **Cu(211)** surface colored based on corresponding SOAP&LENS environments. In contrast to the snapshots of figures 1(c) and (d), this comprehensive approach captures all the key SOAP as well as LENS environments, providing a more complete characterization of this system.

By combining these two descriptors, it becomes evident that the motion of atoms diffusing on the surface (pink atomic environment) originates from fluctuations within the SOAP environment, which defines the edges of the surface (blue atomic environment).

We further test our approach on different systems. We carried out a second test on a 309 atoms icosahedral gold nanoparticle (**Au-NP**) model, simulated for 2 ns at $T = 200$ K using the Gupta potential [10, 40], (see Methods section for details). In these conditions, this **Au-NP** was demonstrated to have non-trivial dynamics [10, 28]. In figure 3(a), a SOAP-based analysis of the MD trajectory reveals the dominant structural environments within the nanoparticle vertices in blue, surface in lime, sub-surfaces atomic environments in orange, bulk atoms in red and also surface defects in yellow and rosette in light-green. The dynamics of these SOAP atomic environments is quantified by the exchange chord diagram and in the transition probability of figure 3(a) (right). At the same time, analysis of the LENS time series unveils a crucial insight, overlooked by a pure SOAP analysis (figure 3(b)). After ∼180 ns of MD simulation, the nanoparticle undergoes a sharp local structural transition involving one vertex, which penetrates the surface generating a distinctive structure known as a rosette (figures 3(a) and (d)): in light-green). Notably, the creation of a rosette (six symmetrical neighbors around an intruded center) from a vertex (five symmetrical neighbors) is an event that is known to happen in such icosahedral nanoparticles and that can be observed experimentally [10, 41]. The LENS analysis shows the emergence of strong signals when the vertex intrudes and triggers the formation of the rosette (figure 3(b), left). In particular, the magenta colors in figure 3(b) reveal, after such local transition, the presence of a highly dynamic liquid-like' region surrounding the

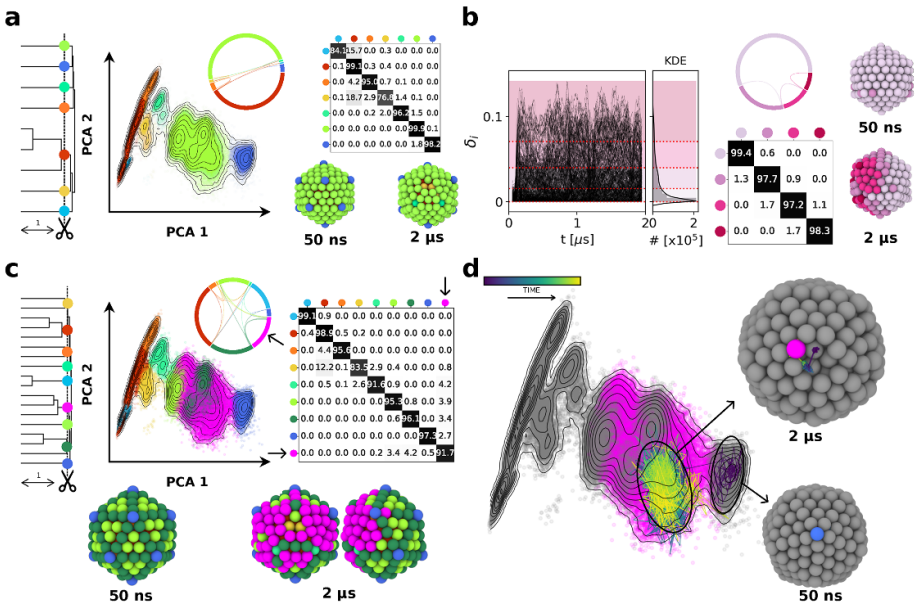

Figure 3. Combined SOAP&LENS analysis of a cold Au-NP surface at 200 K. (a) SOAP-based analysis of the **Au-NP** system. Left: hierarchical clustering based dendrogram (from an HDBSCAN* classification, see figure S1(b)) and dendrogram cutting, defining the merged macro-clusters. Middle: PCA of the SOAP dataset (first two principal components), colored based on the detected macro-clusters, and chord diagram (fluxes). Right: transition probability matrix for the dynamical transitions between the macro-clusters (SOAP environments). Bottom: two nanoparticle MD snapshots where atoms are colored based on the classification: vertices in blue, surface atoms in lime, sub-surface in light-green, yellow and orange, and bulk atoms in red. (b) LENS analysis of **Au-NP**. Left: LENS time-series and classification. Middle: chord diagram (fluxes) and transition probability matrix. Right: MD snapshots with atoms colored based on the LENS clusters: more/less dynamic atoms in brighter/lighter colors. (c) SOAP&LENS based analysis of the **Au-NP**. Left: hierarchical clustering based dendrogram (from an HDBSCAN* classification, see figure S2(b)) and dendrogram cutting, defining the merged macro-clusters, colored according to cluster classification in figure 3(a), except for a new pink cluster. Middle: PCA of the SOAP dataset (first two principal components), colored based on the detected macro-clusters, and chord diagram (fluxes). Right: transition probability matrix for the dynamical transitions between the macro-clusters, highlighting the new cluster in pink. Bottom: three nanoparticle MD snapshots colored according to the classification: vertices in blue, surface atoms in lime, sub-surface in light-green, yellow and orange, bulk atoms in red and 'liquid-like' region in pink. (d) Trajectory of an atom detaching from a vertex and entering the surface of the nanoparticle and giving rise to the rosette environment. The trajectory is shown both on the PCA plot and on the snapshots, colored from blue to yellow according to the time evolution.

rosette, coexisting with a 'crystalline-like' domain in the remaining portion of the **Au-NP**. It is worth noting how a SOAP analysis alone overlooks such a dynamic surface non-uniformity: for the SOAP descriptor, rich in structural information, this local dynamical change does constitute a relevant effect. In the SOAP-based analysis, such a 'liquid-like' region is classified together with the crystalline region, as a global surface cluster (lime color), even if the dynamic behavior of the two regions is different. Therefore, the SOAP description fails to capture part of the system physics: it incorporates two distinct regions with entirely different dynamical behaviors into one single cluster characterized by an averaged structural representation.

In figure 3(c), we show the results of SOAP&LENS based analysis, where we combined the SOAP spectrum of each atom at every timestep with the LENS signal for the same atom at the subsequent Δt. In this case, the combined analysis reveals that a significant portion of the PCA-reduced data -in particular, that central region referring to the surface of the nanoparticle (in figure 3(a): in lime)- corresponds to a highly dynamic LENS environment (figure 3(c): pink). This allows us to disentangle the 'liquid-like' region from the well-defined crystal-like structural domains on the nanoparticle surface. Furthermore, the results of figure 3(c) demonstrate again how, also in this case, the addition of LENS improves the accuracy in the detection of the SOAP environments. Comparing of figures 3(a) *vs.* (c), it is clear how the analysis robustly distinguishes now the edges (dark green), faces (lime) and vertexes (blue), as well as rosettes (light-green) and defects (yellow) on the icosahedral nanoparticle surface. Similar to the case of **Cu(211)**, a strong correlation arises between the 'liquid-like' dynamical domain and specific structural environments: the LENS (pink) cluster in the transition matrix is found connected to the faces (lime, ∼3.9%), edges (green, ∼3.4%), vertices (blue, ∼2.7%) and especially the rosettes (light-green, ∼4.2%) of the nanoparticle. This is interesting, considering that the pink dynamical region (local 'melting' of the nanoparticle surface) originates from the creation of a first rosette (a defect in the icosahedron).

In figure 3(d), we show an example of a structural variation event that gives rise to the formation of a rosette structure. This transition is depicted both on the PCA plot and on the snapshots, where the trajectory

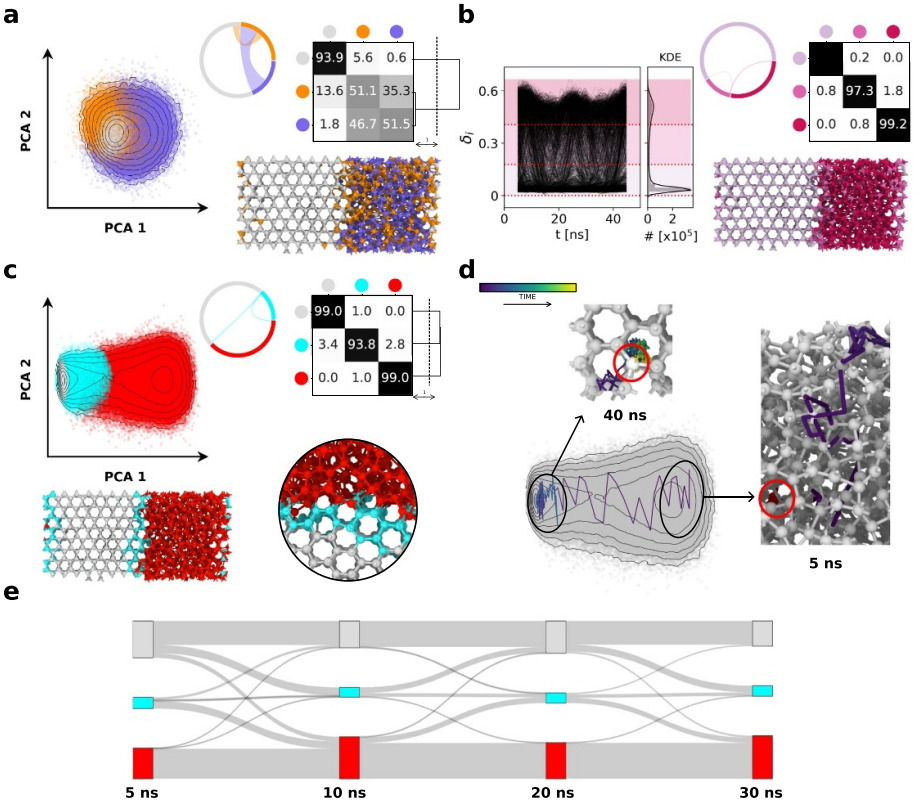

Figure 4. Combined SOAP&LENS analysis of ice-liquid water equilibrium in correspondence of the transition temperature. (a) SOAP-based analysis of the **TIP4P/Ice** system. Left: first two PCA plot, colored according to GMM clustering (see Methods section for details). Right: chord diagram (fluxes) and transition probability matrix of the clusters and its hierarchical clustering based dendrogram showing the relation within them. Bottom: a snapshot along the trajectory colored based on the cluster classification, ice molecules in white, liquid water in orange and purple. (b) LENS-based analysis of **TIP4P/Ice**. Left: LENS time-series and classification. Middle: chord diagram (fluxes) and transition probability matrix. Bottom: an MD snapshot with atoms colored based on the LENS clusters, more/less dynamic atoms in brighter/lighter colors. (c) SOAP&LENS based analysis of the **TIP4P/Ice**. Left: first two PCA plot, colored according to GMM clustering (see Methods section for details). Right: the chord diagram (fluxes) and transition probability matrix of the clusters and its hierarchical clustering based dendrogram showing the relation within them. Bottom: an MD snapshot colored according to the cluster classification: ice molecules in white, liquid water in red and the interface in cyan. Right: zoom of the interface region. (d) Trajectory of a molecule that undergoes a phase transition, from liquid water to ice, crossing the interface. The trajectory is reported both on the PCA plot and on the zoomed snapshots, colored from blue to yellow according to the time evolution. (e) Flow chart of the transitions between the three phases, colored accordingly, at 5 ns, 10 ns, 20 ns and 30 ns.

of the vertex atom (blue, at 50 ns) is color-coded according to time evolution, ranging from dark blue to yellow (2 μs of MD). This demonstrates how the vertex atom entering into the surface, leads to the emergence of a 'liquid-like' region surrounding the rosette (pink, at 2 μs of MD).

As a last case study, we present the effectiveness of our SOAP&LENS analysis in capturing distinct phases within a system where ice and liquid water coexist in correspondence of the solid and liquid transition temperature. We analyzed 50 ns of an atomistic simulation of water modeled with **TIP4P/Ice** force field, containing 2048 molecules at equilibrium between the two phases (∼50% ice and ∼50% liquid water) at the transition temperature [28, 29]. A pure SOAP-based (structural) analysis, reported in figure 4(a), can distinguish the two main phases (ice in white and liquid water in orange and purple). The two clusters in orange and purple in figure 4(a), correspond to tiny variations of the same environment (liquid water). This is clearly shown in the probability matrix and in particular in the Hierarchical Clustering based dendrogram, where the purple and orange atomic environments are very close to one each, and both are in comparison very far from the white one (see figure 4(a) right). However, recently we have demonstrated that a pure LENS (dynamic) analysis can detect easily both the ice and water environments, plus also the interface between them [28]. Figure 4(b) shows the LENS time series, which clearly highlight two distinct statistically relevant environments, with different dynamics, separated by an interface environment where the ice/liquid water molecular transitions occur. The flux chord diagram and the probability transition matrix of figure 4(b) (right) reveal how the ice/liquid phase transition of the molecules takes place through the interface. Figure 4(c) displays the combined SOAP and LENS in a unique dataset, thereby providing a PCA that is significantly distorted compared to the SOAP one of figure 4(a). Two main density peaks are evident (in white and red) corresponding respectively to ice and liquid water. GMM clustering now clearly detects a

distinct area on the PCA corresponding to the ice-water interface (in cyan). In figure 4(d), we highlighted one explicatory trajectory (on the PCA plot and on the snapshot) of a water molecule undergoing phase transition from liquid water to ice, crossing the interface. The flow chart in figure 4(e) provides a qualitative visualization of the transitions between the various environments considering specific time intervals along the trajectory (e.g. at 5 ns, 10 ns, 20 ns and 30 ns). Also in this case, the addition of a LENS component to the SOAP vectors offers a clear advantage over a purely structural analysis (SOAP only). In this specific case, it is interesting to note how LENS retains large part of the information contained in the system trajectory compared to SOAP. This is evident, for example, if we compare the cumulative variance contained in the dataset as a function of the number of principal components of the PCA. In figure S4, we clarify that to reach the 99% of the cumulative variance of the dataset 8 components are needed in a purely structural SOAP dataset, while for example, when LENS is embedded into the dataset, with only three components we largely exceed the 99% of variance. This demonstrates how, in this system, the LENS descriptor might retain more comprehensive information regarding the key features that characterize the system, compared to the SOAP descriptor.

In conclusion, this study points out the intrinsic limitations of relying solely on structural descriptors to comprehend the physics of dynamically evolving systems. By integrating a microscopic dynamic descriptor, like LENS, with a structural counterpart (e.g. SOAP), we obtain numerous advantages. First, this integration improves the accuracy of both structural and dynamic classifications, 'cleaning up' the noise and reducing the degeneracy issues intrinsic to both individual analyses. Second, this paves the way for understanding how given structural microscopic environments within the system can generate specific dynamic behavior (fluctuations). This opens new routes to learn microscopic-scale structure-dynamic relationships (e.g. those of figures 2 and 3) that are key to understanding the behaviors and properties of these, and in general of a variety of complex systems. These results are also reminiscent of general concepts in physics. For instance, when studying the behavior of a system, the sole positional information of the objects is insufficient to predict the dynamic behavior of the system at non-zero temperature (e.g. information on velocities is also needed). Similarly, these results demonstrate how coupling a purely structural parameter like SOAP, which provides information only on the relative structural arrangements, with a descriptor that is rich in local dynamic information, offers fascinating insights. We expect that such type of approach, given its abstract nature, will be highly valuable in characterizing the behavior of complex systems across various domains and potentially also beyond the atomistic/molecular scale.

3. Methods

3.1. MD simulations and pre-processing

The atomistic model of **Cu(211)** surface (see figures 1 and 2) is composed of $N_{211} = 2400$ atoms. The MD simulation is conducted at $T = 600$ K via LAMMPS software [42] using a neural network potential built using the DeepMD platform [39], as described in detail in [9]. The sampled trajectories are 150 ns long. A total of 502 frames are extracted every $\Delta t = 0.3$ ns along the MD trajectory and used for the analysis.

The atomistic model for the icosahedral **Au-NP** is composed of $N_{\text{Au-NP}} = 309$ gold atoms (figure 3). The **Au-NP** model is parametrized according to the Gupta potential, [40] and is simulated for 2 μs of MD at $T =$ 200 K using the LAMMPS software [42] as described in detail in [10]. 2000 frames are extracted every $\Delta t =$ 1 ns of the MD trajectory and then used for the analysis.

The atomistic Ice/Water interface model of figure 4 is composed of $N_{\text{TIP4P}} = 2048$ water molecules. The MD simulation is conducted at $T = 268$ K. The **TIP4P/Ice** water model [43] is used to represent both the solid phase of ice and the phase of liquid water [39], as described in detail in [28]. The sampled trajectory is t = 50 ns long, sampled and analyzed every 0.1 ns.

All MD trajectories are firstly pre-processed in order to obtain a `hdf5` database, containing the data needed to extract the SOAP spectra and LENS values by using the software *cpctools*, accessible at: https://github.com/GMPavanLab/cpctools. For the **Cu(211)** surface, we computed the SOAP spectra on both the surface and the bulk ($N_{211} = 2400$ atoms in total), removing most of the deep bulk atoms, thus obtaining the 900-atoms system analyzed herein. We analyzed all the atoms of the **Au-NP** system. In the **TIP4P/Ice** water system of figure 4 we computed the SOAP spectra for all the O atoms considering also the H atoms in the environment, while we did the LENS analysis by considering only the O atoms. In all cases, the analysis is then conducted by building both the local SOAP environments and the LENS values of each unit within a sphere of radius r_{cut} (see 1(b)), equal to 6 Å for the **Cu(211)**, 4.48 Å for the **Au-NP**, and 6 Å for the **TIP4P/Ice** system.

3.2. SOAP analysis

To describe the structural environment surrounding each particle within the simulations, we use the SOAP descriptor. We compute the SOAP spectrum $\mathbf{p}_i^t$ representing the local structural environment of each particle i at every timestep t within a cut-off radius r_{cut} (6 Å for the **Cu(211)**, 4.48 Å for the **Au-NP**, and 6 Å for the **TIP4P/Ice** system) through the software *cpctools*, accessible at: https://github.com/GMPavanLab/cpctools. The SOAP vectors are generated using *describe* [44], and both l_{max} and n_{max} parameters for spherical harmonics, and number of radial basis functions are set to 8. The results in a 576-component vector represent the environments of one particle at a certain timestep for the single species systems in (**Cu(211)** and **Au-NP**), while in a 1728-component vector for the ice/water interface. Then, we applied the PCA algorithm to each dataset (as implemented in the SciPy python package [45]), reducing the dimensionality of the representation to the first n-components, in order to reach a certain cumulative variance within each system, as reported in table S1. To analyze the reduced data of the **Cu(211)** and **Au-NP** systems, we applied the HDBSCAN* [27] clustering algorithm set up with `min_cluster_size = 80` for the former and `min_cluster_size = 150` for the latter, obtaining 7 and 9 environments, respectively. We used soft-clustering to assign the point classified as noise to their closer cluster. From the cluster transition probability matrix (see figures S1(a) and (b)), we found the relations within the environments via hierarchical clustering algorithm. Then, merging the ones closer than 1 in terms of the chosen metrics (`correlation`) and linkage (`average`), we obtained 6 and 7 macro-clusters respectively for **Cu(211)** and **Au-NP** systems. Regarding the **TIP4P/Ice** system, we followed a slightly different procedure: indeed, as clear from the PCA of the SOAP spectra reported in figure 4(a), there are no clear density-based patterns, and HDBSCAN* failed to assign meaningful clusters, as shown in figure S3. Thus, instead of HDBSCAN* clustering algorithm, we employed a GMM [25] setting the number of clusters to three, without merging clusters *a posteriori* bust still applying Hierarchical Clustering being interested in cluster relations. Then, for all the systems, we compute the clusters' fluxes, i.e. the number of particles going from one cluster to another, following the cluster assignment along the trajectory. The fluxes are visualized as chord diagrams in figures 1(c), 3(a) and 4(a). The width of the arcs represents the total number of transitions experienced by each cluster during the simulation, including both self-transitions and those to other clusters. The chords linking the clusters depict their interconnections, with the extension of the chord's base indicating the amount of particles exchanging between connected clusters. The color of the chords indicates the dominant direction of particle transfer between clusters. Then, normalizing the flux matrices on each row, we obtained the transition probability, reported in figures 1(c), 3(a) and 4(a).

3.3. LENS analysis

We compute the $\delta_i(t)$ signals for all the systems following a similar procedure reported in Crippa *et al* [28], by using the *cpctools* software accessible at: https://github.com/GMPavanLab/cpctools, and reducing the noise by using a Savitzky and Golay [46] filter (as implemented in the SciPy python package [45]). Each δ_i(t) signal is smoothed using a common polynomial order parameter of $p = 2$ and a time-window of 20 frames in the crystalline **Cu(211)** surface, 100 frames for both the water/ice interface and the **Au-NP** system. After the noise reduction, the clustering of the δ_i data is performed: in the case of **Cu(211)**, the clustering thresholds are set as previously [28] while for both the **Au-NP** and the **TIP4P/Ice** systems are set by means of kmeans algorithm [24] implemented in SciPy python package [45]. The kmeans algorithm requires the definition of the number of clusters as an input: in both cases of gold nanoparticle and ice/water interface, we set four and three clusters respectively, according to the number of macro clusters previously found [28]. Knowing the cluster assignment, we compute the cluster fluxes, i.e. the number of particles going from one cluster to another, for each system. The fluxes are reported as chord diagrams of figures 1(d), 3(b) and 4(b), representing the data as reported above. Then, normalizing the flux matrices on each row, we obtain the transition probability, reported in figures 1(d), 3(b) and 4(b).

3.4. SOAP&LENS combined analysis

The combined SOAP&LENS descriptor is obtained by following the procedure illustrated in figure 1(e) and explained in detail in the Results section. The SOAP power spectrum of each particle i at every time step t ($\mathbf{p}_i^t$) is combined with the subsequent LENS scalar value ($\delta_i^{t+\Delta t}$), obtaining a new vector $\boldsymbol{\chi}_i^t = (\mathbf{p}_i^t, \delta_i^{t+\Delta t})$. Each SOAP power spectra are normalized on their norm, while the LENS scalar is intrinsically normalized within zero (no neighborhood changes) and one (the whole neighborhood changes). In this way, while retaining different information and having two distinct mathematical forms (a high dimensional vector and a scalar), the two components have the same 'weight' in the dataset. This procedure, when iterated throughout the entire trajectory, results in a new dataset including $N = N_{\text{particle}} \times N_{\text{frames}}$ vectors. Each vector contains $n + 1$ components: n components representing the SOAP power spectrum and 1 component representing the LENS value. Starting from this $\boldsymbol{\chi}_i^t$ representation of the particle local environments, we

follow the same *bottom–up* procedure, described above, applied to the pure SOAP dataset. To highlight the real effect of the LENS component, avoiding biased results, we performed the *bottom–up* analysis by using the same parameters. Indeed, upon applying PCA to the SOAP&LENS dataset of each system, we considered the first n-PCA components to match the PCA variance retained in the SOAP analysis, as reported in table S1. We apply the clustering algorithm (both HDBSCAN* and GMM) to this new reduced dataset, by using the same parameters (`min_cluster_size = 80` for the **Cu(211)** and `min_cluster_size = 150` for the **Au-NP** and `n_component` = 3 for the **TIP4P/Ice**), and then the Hierarchical Clustering dendrogram cutting under the same conditions i.e. closer than 1 in terms of the chosen metrics (`correlation`) and linkage (`average`). Details regarding the computational cost of the SOAP and LENS analyses for each system are reported in table S2 in the supporting information. The data shown in table S2 indicate that the LENS analysis is approximately one order of magnitude less expensive than the SOAP computation (with a slight variability depending on the system of interest), when performed under comparable conditions. Thus, the SOAP&LENS analysis computational cost is comparable to a pure SOAP calculation. While our method is general and owns the advantage of transferability to diverse systems, some limitations may concern the size of the system, namely, the number of individuals and frames taken along the simulation trajectory that can be effectively analyzed. Increasing too much the size (in terms of number of units) and the trajectory sampling produces an increase in terms of computational cost. Parallelization of this analysis code will help dealing with this limitation in the next future.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://zenodo.org/records/10213827[47].

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).

Conflict of interest

The authors declare no competing interests.

ORCID iDs

Martina Crippa https://orcid.org/0000-0002-6682-0015
Annalisa Cardellini https://orcid.org/0000-0002-6359-6118
Matteo Cioni https://orcid.org/0000-0003-4391-2096
Gábor Csányi https://orcid.org/0000-0002-8180-2034
Giovanni M Pavan https://orcid.org/0000-0002-3473-8471

References

[1] Andrews J, Gkountouna O and Blaisten-Barojas E 2022 Forecasting molecular dynamics energetics of polymers in solution from supervised machine learning *Chem. Sci.* **13** 7021
[2] Gasparotto P and Ceriotti M 2014 Recognizing molecular patterns by machine learning: an agnostic structural definition of the hydrogen bond *J. Chem. Phys.* **141** 174110
[3] Davies M B, Fitzner M and Michaelides A 2022 Accurate prediction of ice nucleation from room temperature water *Proc. Natl Acad. Sci. USA* **119** e2205347119
[4] Noé F, Olsson S, Köhler J and Wu H 2019 Boltzmann generators: sampling equilibrium states of many-body systems with deep learning *Science* **365** eaaw1147
[5] Gardin A, Perego C, Doni G and Pavan G M 2022 Classifying soft self-assembled materials via unsupervised machine learning of defects *Commun. Chem.* **5** 82
[6] Cardellini A, Crippa M, Lionello C, Afrose S P, Das D and Pavan G M 2023 Unsupervised data-driven reconstruction of molecular motifs in simple to complex dynamic micelles *J. Phys. Chem.* B **127** 2595–608
[7] Capelli R, Gardin A, Empereur-Mot C, Doni G and Pavan G M 2021 A data-driven dimensionality reduction approach to compare and classify lipid force fields *J. Phys. Chem.* B **125** 7785–96
[8] Lionello C, Perego C, Gardin A, Klajn R and Pavan G M 2023 Supramolecular semiconductivity through emerging ionic gates in ion-nanoparticle superlattices *ACS Nano* **17** 275–87
[9] Cioni M, Polino D, Rapetti D, Pesce L, Delle Piane M and Pavan G M 2023 Innate dynamics and identity crisis of a metal surface unveiled by machine learning of atomic environments *J. Chem. Phys.* **158** 124701
[10] Rapetti D, Delle Piane M, Cioni M, Polino D, Ferrando R and Pavan G M 2023 Machine learning of atomic dynamics and statistical surface identities in gold nanoparticles *Commun. Chem.* **6** 143

[11] Cheng B et al 2020 Mapping materials and molecules Acc. Chem. Res. 53 1981–91
[12] Errington J R and Debenedetti P G 2001 Relationship between structural order and the anomalies of liquid water Nature 409 318–21
[13] Behler J and Parrinello M 2007 Generalized neural-network representation of high-dimensional potential-energy surfaces Phys. Rev. Lett. 98 146401
[14] Rossi K, Pavan L, Soon Y and Baletto F 2018 The effect of size and composition on structural transitions in monometallic nanoparticles Eur. Phys. J. B 91 33
[15] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments Phys. Rev. B 87 184115
[16] Pietrucci F and Martoňák R 2015 Systematic comparison of crystalline and amorphous phases: charting the landscape of water structures and transformations J. Chem. Phys. 142 104704
[17] Behler J 2011 Atom-centered symmetry functions for constructing high-dimensional neural network potentials J. Chem. Phys. 134 074106
[18] Drautz R 2019 Atomic cluster expansion for accurate and transferable interatomic potentials Phys. Rev. B 99 014104
[19] Faber F, Lindmaa A, von Lilienfeld O A and Armiento R 2015 Crystal structure representations for machine learning models of formation energies Int. J. Quantum Chem. 115 1094–101
[20] Gasparotto P, Bochicchio D, Ceriotti M and Pavan G M 2020 Identifying and tracking defects in dynamic supramolecular polymers J. Phys. Chem. B 124 589–99
[21] Musil F, Grisafi A, Bartók A P, Ortner C, Csányi G and Ceriotti M 2021 Physics-inspired structural representations for molecules and materials Chem. Rev. 121 9759–815
[22] Schölkopf B, Smola A and Müller K-R 1998 Nonlinear component analysis as a kernel eigenvalue problem Neural Comput. 10 1299–319
[23] van der Maaten L and Hinton G 2008 Visualizing data using t-SNE J. Mach. Learn. Res. 9 2579–605 (available at: www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf)
[24] Lloyd S 1982 Least squares quantization in pcm IEEE Trans. Inf. Theory 28 129–37
[25] Reynolds D 2009 Gaussian Mixture Models (Springer) pp 659–63
[26] Schubert E, Sander J, Ester M, Kriegel H P and Xu X 2017 DBSCAN revisited, revisited: why and how you should (still) use DBSCAN ACM Trans. Database Syst. 42 1–21
[27] McInnes L, Healy J and Astels S 2017 HDBSCAN: hierarchical density based clustering J. Open Source Softw. 2 205
[28] Crippa M, Cardellini A, Caruso C and Pavan G M 2023 Detecting dynamic domains and local fluctuations in complex molecular systems via timelapse neighbors shuffling Proc. Natl Acad. Sci. USA 120 e2300565120
[29] Caruso C, Cardellini A, Crippa M, Rapetti D and Pavan G M 2023 TimeSOAP: tracking high-dimensional fluctuations in complex molecular systems via time variations of SOAP spectra J. Chem. Phys. 158 214302
[30] Spencer M S 1986 Stable and metastable metal surfaces in heterogeneous catalysis Nature 323 685–7
[31] Jayanthi C S, Tosatti E and Pietronero L 1985 Surface melting of copper Phys. Rev. B 31 3456–9
[32] Yamakov V, Wolf D, Phillpot S, Mukherjee A and Gleiter H 2004 Deformation-mechanism map for nanocrystalline metals by molecular-dynamics simulation Nat. Mater. 3 43–47
[33] Zepeda-Ruiz L A, Stukowski A, Oppelstrup T and Bulatov V V 2017 Probing the limits of metal plasticity with molecular dynamics simulations Nature 550 492–5
[34] Wang X, Zheng S, Shinzato S, Fang Z, He Y, Zhong Li, Wang C, Ogata S and Mao S X 2021 Atomistic processes of surface-diffusion-induced abnormal softening in nanoscale metallic crystals Nat. Commun. 12 5237
[35] Koch R, Borbonus M, Haase O and Rieder K H 1992 Reconstruction behaviour of fcc(110) transition metal surfaces and their vicinals Appl. Phys. A 55 417–29
[36] Wang X-Q 1991 Phases of the au(100) surface reconstruction Phys. Rev. Lett. 67 3547–50
[37] Antczak G and Ehrlich G 2010 Surface Diffusion: Metals, Metal Atoms and Clusters (Cambridge University Press)
[38] Gazzarrini E, Rossi K and Baletto F 2021 Born to be different: the formation process of Cu nanoparticles tunes the size trend of the activity for CO_2 to CH_4 conversion Nanoscale 13 5857–67
[39] Wang H, Zhang L, Han J and Weinan E 2018 DeePMD-kit: a deep learning package for many-body potential energy representation and molecular dynamics Comput. Phys. Commun. 228 178–84
[40] Gupta R P 1981 Lattice relaxation at a metal surface Phys. Rev. B 23 6265–70
[41] Aprà E, Baletto F, Ferrando R and Fortunelli A 2004 Amorphization mechanism of icosahedral metal nanoclusters Phys. Rev. Lett. 93 065502
[42] Thompson A P et al 2022 LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso and continuum scales Comput. Phys. Commun. 271 108171
[43] Abascal J L F, Sanz E, García Fernández R and Vega C 2005 A potential model for the study of ices and amorphous water: TIP4P/Ice J. Chem. Phys. 122 234511
[44] Himanen L, Jäger M O J, Morooka E V, Federici Canova F, Ranawat Y S, Gao D Z, Rinke P and Foster A S 2020 DScribe: library of descriptors for machine learning in materials science Comput. Phys. Commun. 247 106949
[45] Virtanen P et al 2020 SciPy 1.0: fundamental algorithms for scientific computing in python Nat. Methods 17 261–72
[46] Savitzky A and Golay M J E 1964 Smoothing and differentiation of data by simplified least squares procedures Anal. Chem. 36 1627–39
[47] Crippa M, Cardellini A, Cioni M and Pavan G M 2023 Research data supporting: "machine learning of microscopic structure-dynamics relationships in complex molecular systems" Zenodo (https://zenodo.org/records/10213827)