Reports on Progress in Physics

PURPOSE-LED PUBLISHING

PAPER • OPEN ACCESS

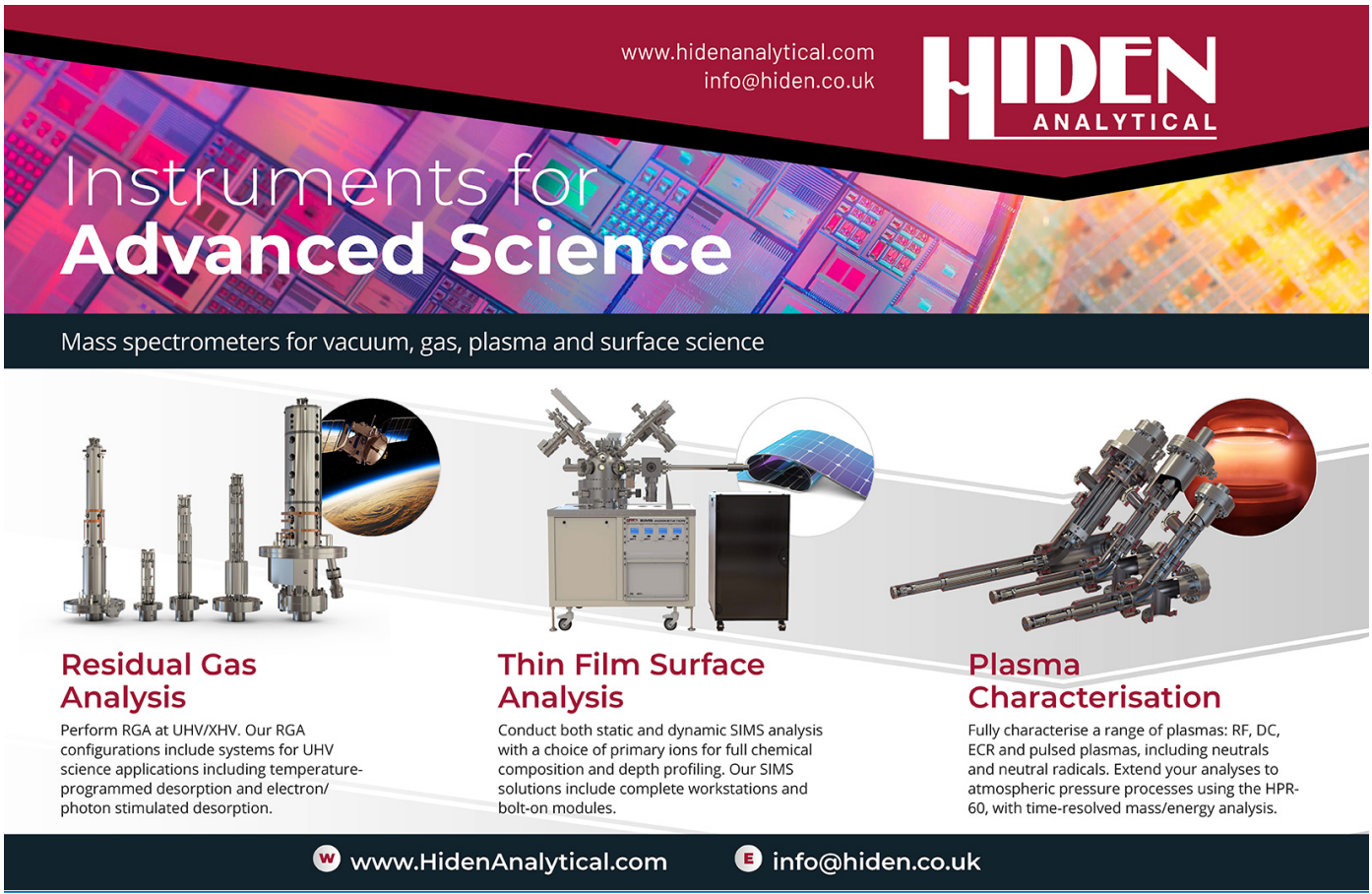
Inherent structural descriptors via machine learning

To cite this article: Emanuele Telari *et al* 2025 *Rep. Prog. Phys.* **88** 068002

View the article online for updates and enhancements.

You may also like

- Enhanced ferromagnetism in monolayer Cr_2Te_3 via topological insulator coupling
Yunbo Ou, Murod Mirzhalilov, Norbert M Nemes et al.
- A review of the fraction of four-coordinated boron in binary borate glasses and melts
Oliver L G Alderman, Nagia S Tagiara, Ian Slagle et al.
- Dynamically generated decoherence-free subspaces and subsystems on superconducting qubits
Gregory Quiroz, Bibek Pokharel, Joseph Boen et al.



This content was downloaded from IP address 92.24.44.104 on 31/10/2025 at 10:40

OPEN ACCESS
IOP Publishing
Reports on Progress in Physics
Rep. Prog. Phys. **88** (2025) 068002 (15pp)
https://doi.org/10.1088/1361-6633/add95b

Inherent structural descriptors via machine learning

Emanuele Telari[1], Antonio Tinti[1,9,*], Manoj Settem[1], Carlo Guardiani[1], Lakshmi Kumar Kunche[1], Morgan Rees[2], Henry Hoddinott[2,3], Malcolm Dearg[4], Bernd von Issendorff[5], Georg Held[3], Thomas J A Slater[4], Richard E Palmer[2], Luca Maragliano[6,7], Riccardo Ferrando[8] and Alberto Giacomello[1]

[1] Dipartimento di Ingegneria Meccanica e Aerospaziale, Sapienza Università di Roma, Roma 00184, Italy
[2] Nanomaterials lab, Mechanical Engineering, Swansea University, Swansea SA1 8EN, United Kingdom
[3] Diamond Light Source, Harwell Science and Innovation Campus, Didcot OX11 0DE, United Kingdom
[4] Cardiff Catalysis Institute, School of Chemistry, Cardiff University, Cardiff CF24 4HQ, United Kingdom
[5] Department of Physics, Albert-Ludwigs-Universität, Freiburg in Breisgau 79098, Germany
[6] Dipartimento di Scienze della Vita e dell'Ambiente, Università Politecnica delle Marche, Ancona 60131, Italy
[7] Center for Synaptic Neuroscience and Technology, Istituto Italiano di Tecnologia, Genova 16132, Italy
[8] Dipartimento di Fisica, Università di Genova, Genova 16146, Italy
[9] Laboratory of Molecular Simulation (LSMO), Institut des Sciences et Ingénierie Chimiques, École Polytechnique Fédérale de Lausanne (EPFL), Sion 1950, Switzerland

E-mail: antonio.tinti@[uniroma1.it/epfl.ch]

Received 22 October 2024, revised 19 April 2025
Accepted for publication 15 May 2025
Published 5 June 2025

Corresponding editor: Dr Lorna Brigham

Abstract

Finding proper collective variables for complex systems and processes is one of the most challenging tasks in simulations, which limits the interpretation of experimental and simulated data and the application of enhanced sampling techniques. Here, we propose a machine learning (ML) approach able to distill few, physically relevant variables by associating instantaneous configurations of the system to their corresponding inherent structures as defined in liquids theory. We apply this approach to the challenging case of structural transitions in nanoclusters, managing to characterize and explore the structural complexity of an experimentally relevant system constituted by 147 gold atoms. Our inherent-structure variables are shown to be effective at computing complex free-energy landscapes, transition rates, and at describing non-equilibrium melting and freezing processes. In addition, we illustrate the generality of this ML strategy by deploying it to understand conformational rearrangements of the bradykinin peptide, indicating its applicability to a vast range of systems, including liquids, glasses, and proteins.

* Author to whom any correspondence should be addressed.

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

© 2025 The Author(s). Published by IOP Publishing Ltd

Supplementary material for this article is available online

Keywords: metal nanoclusters, machine learning, molecular dynamics, collective variables, free energy calculations

1. Introduction

Describing atomic/molecular processes is a notoriously difficult endeavor [1] even in apparently simple cases such as the isomerization of a small molecule [2]. Producing a low-dimensional representation of such processes usually requires the introduction of functions of the system coordinates, called collective variables (CVs). CVs can be exploited in advanced simulation techniques [3–5] for accelerated sampling, free-energy (FE) calculations, and identification of transition mechanisms for a variety of phenomena, including transitions in hard [3], soft, and biological [2] matter, and chemical reactions [6]. More generally, starting from the unpractical description in terms of atomic coordinates, CVs attempt to distil essential physical information about complex processes including non-equilibrium ones [7].

Recently, machine learning (ML) has emerged as an invaluable tool for the discovery of CVs [8–12] in overly complicated systems or when physical intuition fails. In this work, we introduce a generally applicable ML approach for characterizing structural transitions of actual physical systems. We define CVs capable of discriminating structural motifs in noisy finite-temperature configurations based on their zero-temperature counterparts, taking inspiration from the inherent structure concept which we borrow from the theory of liquids [13] (figure 1(a)). In order to devise few, physically informed CVs, we employ a neural network characterized by the convergent-divergent architecture typical of autoencoders. We train the network such that, while taking structural descriptors evaluated on finite-temperature realizations as inputs, it learns to associate them to the zero-temperature counterparts of the original descriptors in the output (figure 1(b)). The resulting latent variables, which we call inherent structure variables (ISVs), can be thus computed on-the-fly during the dynamical evolution of a system. In addition, they offer a unified description of instantaneous configurations belonging to different temperatures, as they refer to the associated inherent configurations. For the same reason, ISVs can be adopted to describe both equilibrium and non-equilibrium conditions. The present approach constitutes therefore a unique mapping between instantaneous configurations and a universal representation that enables, within the same framework, phase space exploration, FE and rate calculations or trajectory analysis, and are general interest for any system in which structural diversity is an issue, e.g. nanoclusters [14, 15], bulk crystals [3, 16], glasses [17, 18], and biomolecules [19, 20].

Here, the ISV approach is applied to structural transitions in metal nanoclusters, a challenging class of experimentally relevant systems characterized by a startling variety of motifs [21]. Indeed, due to their small size, metal nanoclusters can break translational and rotational symmetries, allowing for multiple twinned structures such as icosahedra (Ih) and decahedra (Dh) in addition to standard crystal lattices, as face-centered-cubic (fcc) [21, 22]. Moreover, they support several types of surface and internal defects and overall shapes [23–25]. Additionally, we showcase the generality of the ISV approach by applying it to detect dynamical conformational changes in the bradykinin (BK) peptide.

Navigating the structural complexity of clusters has been a long-standing challenge [26]. The fact that dozens of structural families can be identified in metal nanoclusters [27] makes the question about the kinetics and mechanisms of transitions between them even more urgent: how do the atoms of an fcc nanocluster rearrange into a decahedral one? At what rates does such a process take place? Previous studies [14, 28] of structural transitions in metal nanoclusters have relied on carefully identified CVs tailored for a specific transition. However, considering the structural complexity of metal clusters, CVs capable of capturing the fine structural details and navigating the variety of structural motifs is crucial and non-trivial. Parallel tempering (PT) has proven an effective means to explore the structural landscape of coinage metals [29, 30] exploiting configuration exchanges between replicas at different temperatures to overcome FE barriers without the need of specifying CVs. Building upon this large database of structures, we recently used ML to construct a low dimensional representation of such a landscape that is both physically meaningful and capable of discriminating fine structural details [27]. The key idea was using a translationally and rotationally invariant representation of the cluster, the radial distribution function (RDF), and reduce its dimensionality using a convolutional autoencoder. This approach allowed us to classify locally minimized structures into tens of structural families.

The CVs we defined in [27] are efficient at classifying structures after the removal of thermal noise by means of energy minimization. However, in order to study the finite-temperature evolution of nanoclusters and employ enhanced simulation approaches, CVs able to deal with noisy, finite-temperature configurations are needed. This is the goal of the present work that was achieved by the ML approach introduced above (figure 1) using RDFs as suitable structural descriptors. In fact, the ISV concept can be implemented using different descriptors depending on the system of interest and on the required level of detail, as demonstrated in the additional biomolecule case we considered, where an interatomic distance matrix is employed.

The first system considered in this work is a gold nanocluster of 147 atoms (Au_{147}), which is a magic number

for the formation of Ih, Dh, and fcc clusters. This cluster was specifically selected because it is characterized by the coexistence, over a wide range of temperatures, of a much widervariety of structural motifs with respect to other elemental clusters of similar size [30]. This structural wealth was confirmed by scanning transmission electron microscopy (STEM) of size-selected nanoclusters, which is reported below. As a consequence of the competition of many structural motifs, Au_{147} represents a vastly challenging system despite the relatively small number of atoms and the consequent affordability of computer simulations [29, 31].

To demonstrate the generality of the proposed approach, we additionally considered the challenging case of tracking in real-time the dynamical conformational changes of a small and flexible biomolecule in water, a BK-derived decapeptide [32]. BK is a hormone involved in inflammation, blood vessel dilation and pain perception that binds to specific G-protein coupled receptors (GPCR) [33]. To circumvent the scarce availability of structural data on GPCRs BK and other natural agonists have been used as templates in the design of novel drugs targeting this receptor family [34]. Additionally, BK is a curvature-sensing peptide that has been engineered to recognize and purify nanoscale vesicles such as liposomes and small exosomes from the blood serum [32]. Structural characterization of BK remains elusive, as existing studies report conflicting conformational data [34, 35], leaving its precise structure in water unresolved. We thus use this challenging case to further put the ISV framework to the test. For BK, the chosen descriptor was the distance matrix between the α-carbons, while the training set was generated using Hamiltonian replica exchange. These different implementation choices further underscore the flexibility of the ISV approach and of the proposed neural network architecture.

2. Results and discussions

2.1. Inherent structural variables by ML

In this section, we devise an approach to build a low-dimensional structural description that enables on-the-fly structural analysis and biasing of molecular simulations. In order to achieve this goal, instantaneous configurations are used as an input, differently from similar approaches aiming at the static classification of [27] which rely solely on locally minimised structures. The proposed approach for obtaining descriptors with a general structural meaning from instantaneous configurations draws inspiration from the inherent structure idea pioneered by Stillinger and Weber for liquids [13]: each instantaneous configuration is thought as a fluctuation around the closest local minimum on the potential energy surface [36]; by quenching, one can refer each dynamical configuration to its *inherent structure* which does not depend on the particular way the original configuration was obtained (equilibrium or non-equilibrium simulations, different temperatures, protocols, etc), see figure 1(a).

Similarly to our previous attempts, we leverage RDFs as a convenient translationally and rotationally invariant description of the atomic structures. From a structural point of view, RDFs are particularly convenient descriptors as they contain extensive information about the structure of clusters [14, 27]—both long and short range order—regardless of whether they are instantaneous configurations or locally minimized ones. The major drawback of RDFs, which has so far limited their use as CVs, is related to their relatively high dimensionality. This limitation can be alleviated by means of ML. Alternative descriptors, such as SOAP [37] or ACE [38], could be used when a detailed description of the local environment of individual atoms is required. However, for the nanocluster system hereby considered, RDFs are both computationally more efficient and more sensitive to discriminate structural motifs in clusters [39], as they are able to also provide abundant information regarding the overall cluster shape.

We propose to couple and compress the high-dimensional information contained in RDFs relative to instantaneous and inherent configurations by means of a conveniently modified convolutional autoencoder. Autoencoders are neural networks with a convergent-divergent architecture as sketched in figure 1(b), well-suited for reducing the dimensionality of data [40]. In our implementation (figure 1(b)), the RDF computed from a specific instantaneous atomic configuration is fed to the autoencoder. In parallel, the same atomic configuration is subjected to a short energy minimization to quench it to the local minimum, resulting in a noise-free RDF. The network is then taught to minimize the mean squared error loss between the output and the inherent-structure RDF, at variance with the usual autoencoder strategy of matching identical inputs and outputs. Thus, in a single step, our network is capable of analyzing instantaneous atomic configurations and match them with their inherent structure, while producing a low dimensional representation. In data science terms, the strategy can be summarized as a classification task where each instantaneous-configuration RDF is labeled according to its inherent counterpart. In summary, by non-linearly combining information from the input and the output, the bottleneck obtained by such an approach provides a limited number of descriptors, the ISVs, capable of assigning similar values to different instantaneous configurations which share the same inherent structure.

As an application, we considered a real-world example, a gold nanocluster consisting of 147 gold atoms, Au_{147}. Interactions are modeled via the many-body 'Gupta' second-moment tight-binding QEq potential [41]. This potential is known to capture well the of variety structural motifs of gold at this size [27], which correspond to those experimentally observed in our STEM data shown in the following. We note that other approaches, such as DFT calculations, could in principle be used in conjunction with the ISVs, although it is outside the scope of this work which is focused on demonstrating the generality of the approach. The training set was taken from [29], using Au_{147} configurations generated by PT at different temperatures. Details about dataset and training are reported in the Methods section and in the supplementary text.

The most important hyperparameter of the autoencoder is the bottleneck size; here we found that the optimal compromise between information compression and

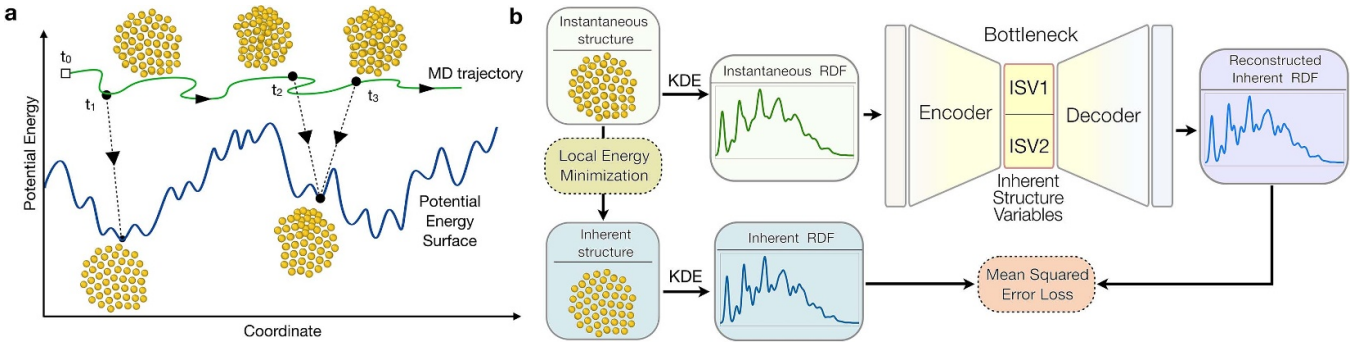


Figure 1. Proposed approach to distill inherent structural variables. (a) Schematic representation of the relation between instantaneous configurations at finite temperature and the related inherent structures on the potential energy surface. Instantaneous configurations are affected by significant thermal noise. A local minimisation removes the thermal contribution and allows obtaining of the corresponding inherent structure. (b) Sketch the of working principle behind the ISV encoder-decoder neural network. Structural descriptors of the system, such as the radial distribution function (RDF), constitute the encoder's input, whose output is used by the decoder to reconstruct the inherent state counterpart of the original descriptors. The mean square error (MSE) loss function is used to measure the performance of the network in validation and training.

reconstruction performance was achieved for a bottleneck of size 2 (see figure S3). By comparing the generated space against the structural classification of [27], the two ISVs were found to be expressive enough to encode the fine structural details of the Au nanoclusters (see figure S4).

The possibility to compute structural CVs directly from instantaneous structures opens the way to use them on-the-fly in atomistic simulations, e.g. for analyzing the dynamical structural evolution of the system and to bias trajectories exploiting the intrinsic differentiability of neural networks. Results below indeed show that the ISVs are suitable for FE and rate calculations, as well as for analyzing non-equilibrium processes in complex and realistic systems. We were able to handle these diverse applications by training the network only once, as the description conveniently unifies information from different temperatures contained in the dataset.

We further remark that the proposed strategy is general in several ways: 1) it can be used in conjunction with simulations of different kinds, including DFT; 2) due to the flexiblity of neural networks, it can be used on a variety of physical inputs, notably different kinds of descriptors; 3) our inherent-structure approach to CVs could prove beneficial for dynamical analysis in other fields,including liquids/glasses [18], and proteins [19].

In this regard section 2.5 showcases the performance attained by ISVs in describing of the complex structural behavior of Bradichinin, a notable oligopepeptide, when solvated in water.. In addition, in order to demonstrate that the ISV method can be suited also for the study of larger systems, the approach was also tested on a larger metal nanocluster: Au_{309}. Also in this case the resulting variables were able to effectively discriminate different structural motifs and making it possible to easily distinguish different regions corresponding to fcc, Dh and amorphous configurations, namely the main structural families that can be observed in the larger Au cluster (see figure S25). Our full results for Au309 clusters are reported in the supplementary materials.

2.2. Free-energy landscape

We computed the FE landscape of Au_{147} at 400 K in the 2D space defined by the ISVs obtained by the strategy illustrated in figure 1. We used Monte Carlo Umbrella Sampling simulations in combination with the weighted histogram analysis method (WHAM) algorithm to unbias the probabilities of approximately 15 000 restrained simulations spanning all relevant regions of the ISV space. This procedure allowed for the reconstruction the FE landscape reported in figure 2. Further details are offered in the Methods section and supplementary text. Previous attempts to reconstruct the structural FE landscape of metal nanoclusters were limited to clusters of few atoms [42] or to selected structural transformations [14] due to lack of sufficiently informative and low-dimensional CVs capable of comprehensively describing the structural wealth of nanoclusters.

The FE landscape in figure 2(a) offers a high-resolution picture of Au_{147} structures at 0.8 times the melting temperature. At this temperature, the prevailing structure is Dh, followed by fcc and Ih; amorphous clusters, which occupy the upper left corner, have large free energies (see the red isolines). As a first approximation, data show that the FE landscape consists of three main basins: fcc, Dh, and Ih. Interestingly, the three basins are connected by two kinetic bottlenecks (corresponding to the FE saddle points) separating fcc from Dh and Dh from Ih. The Dh basin constitutes a central hub through which all structural transitions at 400 K are expected to pass. Additionally, although not relevant at this temperature, the mildest slope leading to the amorphous region is found close to the Dh–Ih bottleneck on the Ih side. As we will see, the topology and connectivity between these basins, being based on inherent structural descriptors, offers a general and clear-cut picture of equilibrium and non-equilibrium transitions for Au_{147}, including the melting and freezing processes discussed in section 2.4.

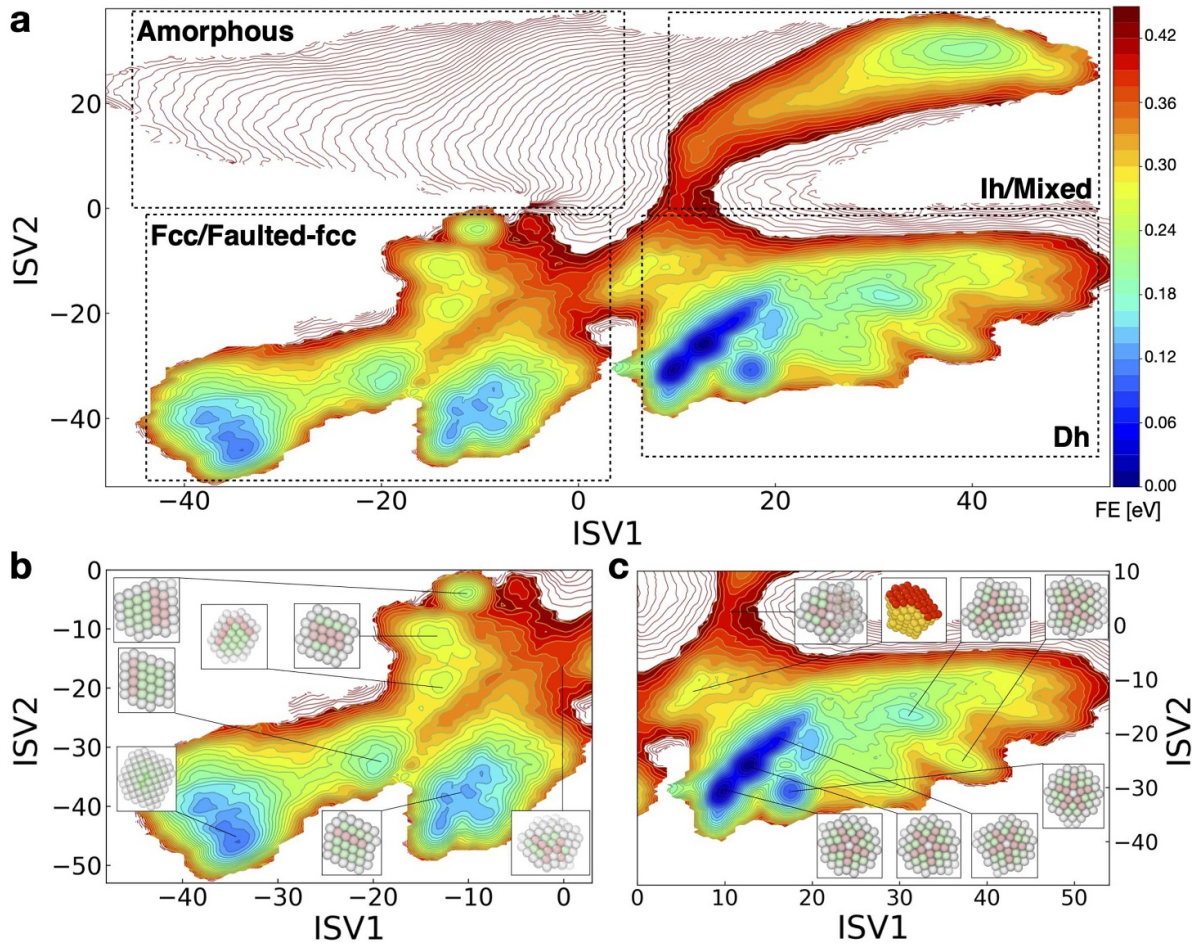


Figure 2. FE landscape of Au_{147}. (a) Contour plot of the FE landscape obtained by US simulations at 396 K, after WHAM calculations. Color coding of the contours is associated to FE values, as reported in the horizontal colorbar. For values of the free energies above 0.44 eV are displayed only the isolines. The landscape can be divided, according to the FE, in three main regions: fcc and faulted-fcc structures region in the bottom left, Ih and mixed structures region in top right and Dh region in the bottom right. Dh region is connected via a saddle-point to twin and fcc region and via another saddle point to Ih and mixed structure region, while the other two regions are not directly communicating on the landscape. In the top left is where amorphous structures are located, which are associated with very high free energies at this specific temperature. (b) Detailed enlargement of the FE contour plot shown in panel (a), showing the fcc and faulted-fcc region together with the representative structures associated with the local minima and the bottleneck linking the fcc region with the Dh basin. (c) Detailed enlargement of the FE contour plot shown in panel (a), showing the Dh region and the representative structures associated with the local minima and the bottleneck connecting the Dh region to Ih and Mixed structures. Atoms colored in green, pink, and white have fcc, hcp, and undefined coordination, respectively.

The kinetic bottleneck separating the fcc basin from Dh is characterized by structures with surface defects. These defected nanoclusters are characterized by the convergence of two hcp planes (which are the typical feature of twin structures) that give rise to the first seed of a local five-fold axis [43], i.e. the distinguishing feature of the decahedral geometry (figure 2(b)). The saddle point between Dh and Ih features the formation of an hcp island at the surface of an otherwise decahedral cluster (figure 2(c)).

A closer look to the three main basins highlights the presence of multiple local minima in the fcc and Dh basins, which correspond to metastable structures. The former basin is populated by fcc and various defected structures thereof, chiefly characterized by twinning plane(s) (figure 2(b)). Perfect fcc occupies a rather broad FE minimum at the extreme left. Immediately close to it, a minimum corresponding to a twin cluster with the hcp plane immediately below the surface is found. The basin then forks into a sub-basin on the lower right, which gathers clusters with a single twinning plane in different central positions, and one on the upper right, with multiple minima corresponding to different arrangements of two hcp planes.

In the Dh basin, the most populated sub-basin corresponds to a central five-fold axis, with multiple local minima pertaining to different kinds of surface defects (figure 2(b)). As expected the absolute FE minimum coincides with the perfect Dh structure. Importantly, in between the two main saddle points separating fcc from Dh and Dh from Ih, a local minimum is present characterized by the presence of an hcp island; although characterized by a relatively large FE, this

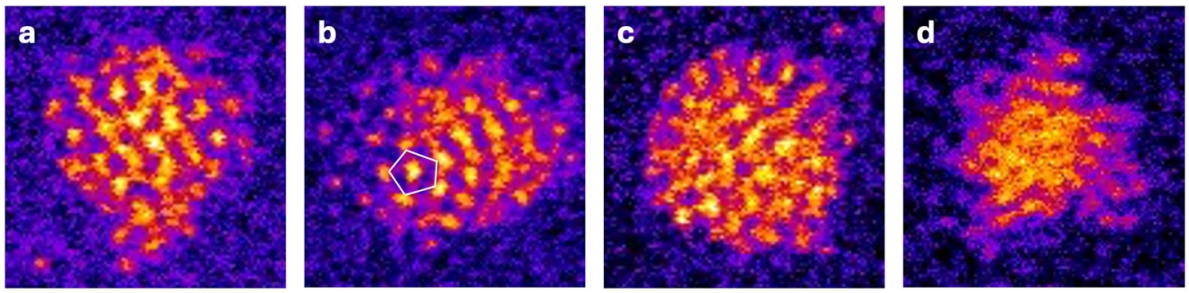


Figure 3. Experimental HAADF-STEM micrographs of Au_{147}. Gold nanoclusters are size-selected to be composed of 147 ± 3 atoms and scans are performed at different temperatures. Representative structures are shown: (a) fcc (300 °C), (b) decahedron with the local pentagonal arrangement marked (350 °C), (c) icosahedron (300 °C), and (d) amorphous (200 °C).

structure occupies a pivotal point for many transitions. Other local minima exist on the far right in which a groove is formed at the cluster's surface. At this temperature, only one Ih minimum is present, corresponding to the perfect structure. However, a pseudoplateau is present close to the bottleneck, corresponding to mixed structures with mainly amorphous and Ih features [29] that play a major role in melting and freezing.

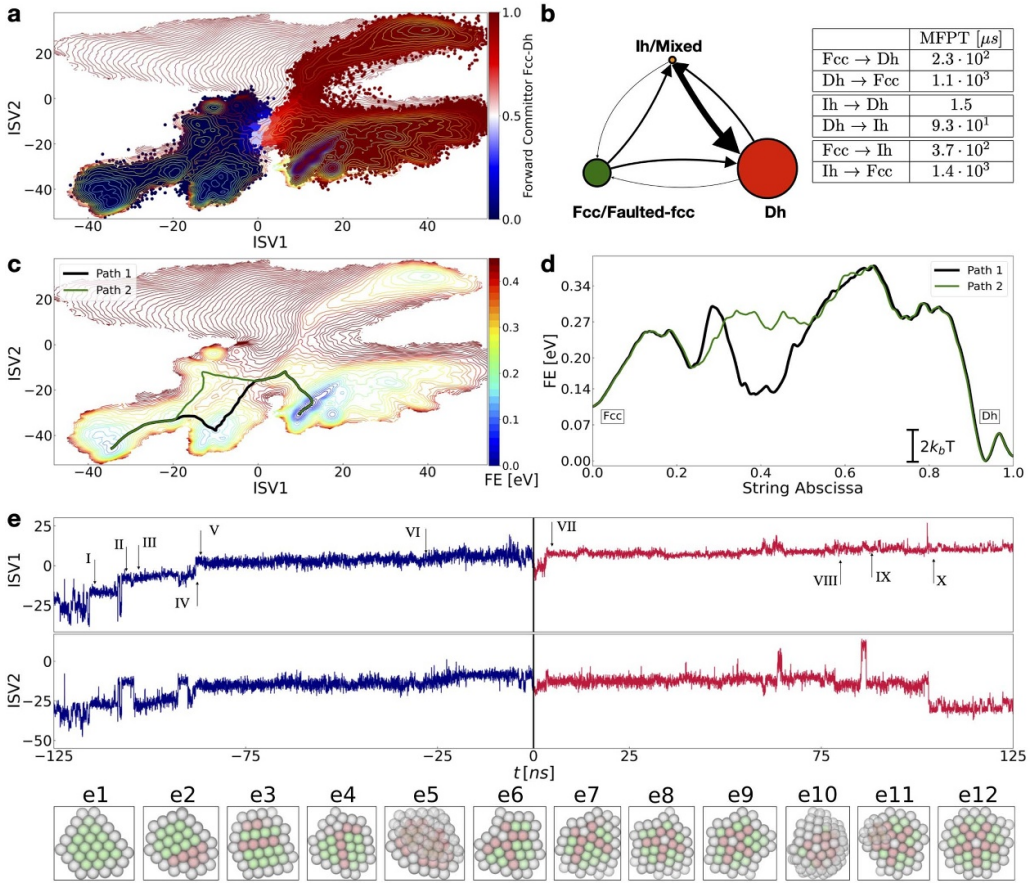
To explore the validity of the simulations, we imaged experimentally size-selected gold clusters, soft-landed onto a carbon support, with the aberration-corrected high-angle annular dark field (HAADF) STEM [44]. Imaging at such small sizes poses intrinsic difficulties, due to the fast structural transitions (see next section) and to the interactions of the electron beam with the cluster. Nonetheless, results for Au_{147} confirm the main structural families found by simulation: fcc features are clearly visible in several clusters (figure 3(a)); the fivefold axis characteristic of Dh was also detected (figure 3(b)); regular Ih-like structures with arc-like surface features could also be imaged, especially at higher temperatures (figure 3(c)). Finally a proportion of amorphous clusters are seen at all temperatures (figure 3(d)).

2.3. *Transition rates and mechanisms*

The description offered by ISVs, other than performing FE calculations, allows in general to have an on-the-fly description of the dynamics of the system in the low dimensional space. This feature can be exploited to gather information from unbiased trajectories and their dynamical evolution. We made use of that in order to complete the picture offered by the FE landscape and gather information about the kinetics of the main transitions highlighted by the landscape of figure 2(a). We applied very established methods to obtain such information, namely Markov state models (MSMs) [45] together with transition path theory (TPT) [46]. These methods rely on the analysis of a great collection of relatively short unbiased trajectories, described by an appropriate set of observables, in order to quantify the timescales of the slowest processes of a system. Thanks to the ISVs, we were able to launch a wealth of unbiased trajectories, about 4000, distributed over the most relevant regions of the space and track their dynamical evolution, which then has been fed to the aforementioned analysis tools (see figures S14–S17).

Analysis of unbiased trajectories allowed us to compute the committor for the transitions between the three major structural families, i.e. Ih–Dh and fcc-Dh transition (figures S18–S19). Given two states A and B, simply defined as regions in the CV space, for the $A \to B$ transition the forward committor can be defined as $q^+ = P(\tau_B < \tau_A)$, τ_B and τ_A being the time intervals needed for a trajectory to visit basin B or A, respectively. In a similar way, the backward committor can be defined as $q^- = P(\tau_A < \tau_B) = 1 - q^+$. In a nutshell, the forward committor measures the probability that a trajectory started at a given point in the ISV space ends up in state B ($q = 1$) rather than A ($q = 0$). For instance, the forward committor for the fcc/Dh transition reported in figure 4(a) shows that trajectories started in the Ih or Dh basins most likely fall in the Dh basin, while those initialized in the fcc one will fall in fcc. While this is intuitive, the most important finding is the exact matching between the region where $q = 0.5$ computed from unbiased trajectories and the saddle point in the FE landscape computed independently by US (the fcc/Dh bottleneck). Similarly, the committor for Ih–Dh transition test has been found to be in good agreement with the FE landscape (figure S18(b)). This provides a strong validation of the quality of the ISVs, which describe these processes without the artefacts due to insufficient CVs [1].

In addition to the committor, overall transition rates and mean first passage times (MFPTs) between the three most relevant basins were computed. The rates were estimated by feeding to the MSM the stationary probability distribution computed by means of US simulations (figure 2). The states chosen for this analysis correspond to the three main basins (Dh, fcc, and Ih) which are also those relevant for experiments. The plot in figure 4(b) shows that Ih–Dh is the fastest transition, happening in ca. 1.5 μs; the related FE barrier is $\Delta F = 6\, k_B T$, leading to an estimated prefactor $t_0 = 3$ ns, assuming an Arrhenius kinetics for the MFPT $= t_0 \exp(\Delta F / k_B T)$, with k_B the Boltzmann constant and T the absolute temperature. On the other hand, the fcc-Dh transition, which is characterized by a slightly higher barrier $\Delta F = 8\, k_B T$, takes more than 100 times more, which corresponds to a much larger effective prefactor, $t_0 = 77$ ns. This difference can be understood if one considers



	MFPT $[\mu s]$
Fcc → Dh	$2.3 \cdot 10^2$
Dh → Fcc	$1.1 \cdot 10^3$
Ih → Dh	1.5
Dh → Ih	$9.3 \cdot 10^1$
Fcc → Ih	$3.7 \cdot 10^2$
Ih → Fcc	$1.4 \cdot 10^3$

Figure 4. Rates and mechanisms of structural transitions. (a) Plot of the forward committor q^+ values of the fcc-Dh transition for different regions of the landscape computed with TPT and MSM. (b) Graph plot of the rates between the three macro region of the ISVs space, representing fcc and faulted-fcc structures (green), Ih and mixed structures (orange) and Dh structures (red). Mean first passage times (MFPT) are written in the table. The arrows width is proportional to the log of the ratio between the rates of the transitions (equal to MFPT^{-1}) and the smallest rate (Ih → fcc). The circles areas are proportional to the equilibrium probabilities of finding the system in one of the three states. (c) Countour plot with only isolines of the FE where two different optimal paths to go from the fcc basin to Dh absolute minimum are reported. Colorcode of the contour is the same of figure 1(a) and it is reported in the colorbar on the right. (d) FE profile along the two paths shown in panel (c). The scalebar reports the conversion to $k_\text{B}T$ units of the FE. (e) Plot versus time of the two ISVs along two different MD unbiased trajectories, initialized in the $q^+ = 0.5$ region. The two trajectories share same initial structure and opposite initial velocities and one is plotted in red for positive times, while the other one is plotted reversed for negative times. Continuous thick black line marks the starting time of both trajectories, while arrows point to the most notable transition steps. Below are reported the most significant structures encountered along the transition. Atoms colored in green, pink, and white have fcc, hcp, and undefined coordination, respectively.

the presence of multiple local minima in the fcc/twin basin that effectively slows down diffusion to reach the kinetic bottleneck with Dh.

In figure 4(c) we report the most probable paths joining the fcc minimum with the Dh one, computed by the string method [47] applied on the FE landscape. At least two independent paths are possible, one passing through the simple twin minimum and one through the region corresponding to clusters with multiple hcp planes. Even though both pass through the same transition state, the latter path corresponds to the energetically favored option, as the intermediate barriers are lower (figure 4(d)). While these paths are the statistically most relevant ones in the limit of low thermal noise, the mechanism of the transformation can be observed dynamically and in atomic detail by selecting reactive trajectories for the same transition. This can be done by launching unbiased trajectories from the relevant transition state ($q^+ = 0.5$) with different initial velocities and stitching together two reactive branches ending up in the products ($q^+ = 1$) or in the reactants ($q^+ = 0$).

Figure 4(e) shows two trajectories initialized with opposite initial velocities, sharing the same initial configuration, selected in the transition state region. In such a way, due to the reversibility of the dynamics, the two trajectories can be seen as two portions of the same dynamical evolution single reactive trajectory connecting fcc to Dh, going forward and backward in time. On the ISV plots, we mark the times corresponding to the key structural changes (indicated by the Roman numerals I to X) during the transition.

In the initial phase (up to I), the cluster fluctuates between fcc (e1) and hcp islands. At this point an increase in ISV1 coincides with the development of peripheral stacking fault (SF) with partial {111} surface facet (e2). Very briefly (II to III),

both ISV1 and ISV2 increase resulting in a higher fraction of SF in the cluster and then there is a reduction in ISV2 without much appreciable change in ISV1 (III to IV) leading to the cluster adopting a twin plus hcp island (e3) and to a lesser extent twin structures. In this duration the cluster makes an excursion to the SF again where ISV2 increases and back. The parallel twin (plus hcp island) develops into a peripheral Dh (e4) around IV. A sharp rise in ISV1 around V coincides with the increase of the length of some of the twin planes emanating from the Dh axis at the expense of the longest twin plane. Around VI, ISV2 increases slightly in correspondance to the formation of hcp islands or an additional 5-fold axis when the hcp islands are on the adjacent {111} facets (e5) which persists up to to the transition point at 0 ns. Moving on to the forward branch, around VII, we observe the annihilation of an exiting peripheral 5-fold axis (e6) and creation of another peripheral 5-fold axis (e7). A further rearrangement pushes this 5-fold axis inwards (e8). The cluster remains in this arrangement for a long period (up to VIII) with appearance and disappearance of hcp islands. A spike in ISV2 around 64 ns results in the cluster adopting a mixed structure (Dh and Ih features co-exist) very briefly. A slight dip in ISV2 at VIII results in rearrangement of the surface fcc islands which move away from the Dh axis (e9). This persists up to the beginning of another spike in ISV2 which indicates a transformation into mixed structure (e10) with three 5-fold axes. After this the cluster transforms back (IX) into Dh structure with equi-length twins (e11). A final lowering in ISV2 around X takes the cluster into best Dh minimum where the cluster adopts the global minimum structure (e12).

Overall, the above fcc $\rightarrow$ Dh transition can be summarized as—fcc initially forms faulted structures with twins/SFs which then leads to the formation of a peripheral 5-fold axis. This undergoes further rearrangement with the 5-fold axis moving inwards and a quick excursion to the Ih/mix region leading to Dh with equi-length twins which eventually rearranges to the global minimum Dh. The initial part of the transition from fcc to peripheral Dh was previously observed [48] in Cu_{170} and Ag_{146} nanoclusters. The twin $\rightarrow$ Dh transition in Au_{147} analyzed using disconnectivity graphs [31] suggested that the transition proceeds via disordering with multiple 5-fold axes (i.e. Ih/mix structures). In contrast, our results show that disordering is not necessarily needed to go from twin to Dh. However, we observed that a quick transformation to Ih/mix structures lead to the formation of Dh with equi-length twins which is a feature of the global minimum.

2.4. Venturing into non-equilibrium: melting and freezing

ISVs obtained exploiting the inherent structure description are also feasible to analyze and drive non-equilibrium simulations. To demonstrate this point, we performed freezing and melting simulations for Au_{147}, imposing an increasing or decreasing, respectively, temperature ramp of $1\,\mathrm{K\,ns^{-1}}$ to thousands of independent replicas, see the Methods section for details. Figure 5 shows the distribution of structures observed at different temperatures during the freezing process.

A merit of the ISV space is to allow the visualization of the time-evolution of structural populations, which can then be compared against the corresponding equilibrium estimates. These equilibrium distributions were taken from US simulations at 400 K and from the PT data of [29] at other temperatures. The system starts in the amorphous basin at 600 K and explores a region which coincides with equilibrium expectations. At lower temperatures (500 K), the Ih basin, which is the closest one to amorphous, becomes densely populated, with a prevalence of mixed structures; few trajectories also fall in the Dh and fcc basins. At even lower temperatures (400 K), the amorphous population has disappeared, leaving room to the three main structural motifs. Interestingly, when one compares the populations obtained in non-equilibrium and in equilibrium, there is a striking difference concerning Ih and Dh (figure 5(a)): Ih are kinetically trapped in the freezing simulations accounting for ca. 40% of the population, while the equilibrium fraction would be negligible. This happens mainly at the expense of the Dh population that decreases from 80% down to 50% in non-equilibrium (at 400 K). If the cooling is sufficiently fast (and the temperature is then kept low), it is possible to select Ih clusters. More generally, ISVs produce an intuitive map that could be useful for designing controlled freezing protocols capable of selecting specific polymorphs [28] in clusters of different metals and sizes; actually, this approach is expected to be more effective for larger clusters for which the typical transition rates are slower [49].

The melting simulations follow a similar protocol, with the initial configurations being extracted from the three main (meta)stable basins. To achieve melting of Dh and fcc clusters (figures 6(a) and (b)), the system has to traverse the mixed region in the Ih basin which, due to its position, plays a major role in both melting and freezing. Interestingly, if the system is initialized close to a perfect Ih (figure 6(c)), it still has to traverse the same mixed region but it reaches it for the first time at much lower temperatures. The system is thus able to overcome the Ih–Dh barrier and to populate the Dh basin; Dh then melts with the usual mechanism at higher temperatures ($>$500 K). The crucial role played by mixed/Ih structures close to the melting/freezing temperature is further supported by the fact that they can be observed by HAADF-STEM at high temperatures (figure 3(c)).

2.5. Generality of ISVs: the case of the BK peptide

In order to demonstrate the generalisability of the proposed ISV framework, we used it for the real-time classification of the structural motifs of a biomolecule: the BK peptide. To study computationally BK structural motifs, we first generated a dataset of 10 000 structures by running Hamiltonian replica exchange molecular dynamics (H-REMDs). A 2D representation of the conformational landscape is then obtained by training an AE with the same strategy illustrated in figure 1 but using as descriptor of the structures the 10×10 matrix of the pairwise distances between α-carbons. At a later stage, a k-means clustering [50] with 16 centroids was performed to characterize how different structures were distributed in the

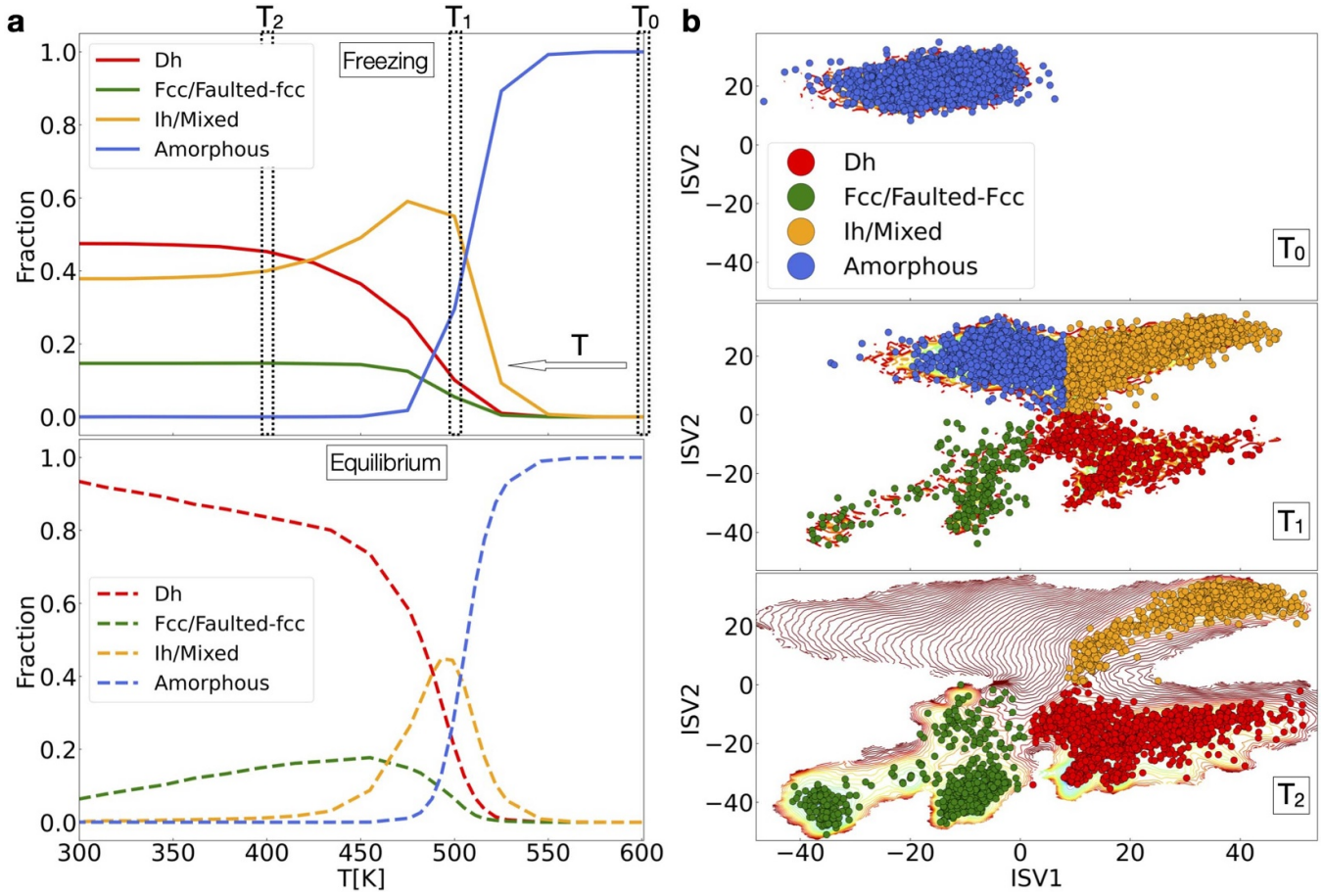


Figure 5. Freezing of Au_{147}. (a) Comparison between fractions of main structural families vs. temperature observed in freezing simulations (top panel) and equilibrium simulations (bottom panel) performed in the work of [29]. The structures have been split in the same 4 structural families shown in figure 2(a). Fractions of amorphous structures are reported by the blue lines, fcc/faulted-fcc by green lines, Dh by red lines and Ih by orange lines. (b) Instantaneous distributions for all freezing trajectories at three specific temperatures, $T_0 = 600$ K, $T_1 = 500$ K, $T_2 = 400$ K, highlighted also in the non-equilibrium fractions plot of panel (a). For the plots of T_0 and T_1 below the points there are the contour plots of the log densities of PT-MD data from [29] at those specific temperatures. For the plot of T_2 the contour plot of the FE shown in figure 2(a) is shown. Points are colored according to their structure type using the same color code as in panel (a).

2D ISV space generated by the network, shown in figure 7(a). The structural analysis performed on the k-means clusters suggests that the points in the ISV space can be broadly grouped in three main structural families corresponding to three different regions of the space, namely distorted S-shaped conformations (upper left region), distorted W-shaped conformations with a N-terminal loop (upper right region), and W-shaped conformations (lower region). For a detailed description of the clustering, see the supporting materials (figures S20 and S21, tables S2 and S3) Long (500 ns), unbiased MD simulations started in each of these regions demonstrate that trajectories remain confined each within its initial regions presumably signalling the presence of large free-energy barriers separating the three structural families (figure 7(b)).

More in detail, the trajectory starting in the upper right basin spends most of its time in cluster 6, corresponding to a distorted W-shape with an N-terminal loop in hairpin-like conformation. The high stability of this conformation is likely due to the persistent interactions (hydrogen bond or salt bridge) between the side-chain of Arg1 and the COO– terminal group of Lys10 or, alternatively, the carbonyl group of Arg9 (figure 7(c)). These interactions are present in all W-shaped conformations, but the structures in this basin also exhibits a hydrogen bond between the NH_3+ terminal group of Arg1 and the side-chain hydroxyl or the backbone carbonyl group of Ser6. Due to its remarkable stability, cluster 6 lies at the centre of all conformational transitions in this basin with cluster 14 and 15 at the periphery (figure 7(c)). This picture is consistent with the relative positions of these clusters in the ISV space (figure 7(a)), and with the ISV-based analysis of a long unbiased trajectory initialized by a configuration in cluster 6.

Figure 7(f) shows that the trajectory originating from cluster 5 spends most of the time in cluster 7. In this case, the path goes back and forth along the sequence $C5 \rightarrow C13 \rightarrow C2 \rightarrow C7$ spanning asymmetric hairpin-like (clusters 5 and 13), S-shaped (2), and distorted S-shaped conformations (7) all belonging to the same basin. Finally, the trajectory starting in cluster 0 and visiting the lower basin of W-shaped motifs has the highest structural variability and is described in the supplementary material (figure S22).

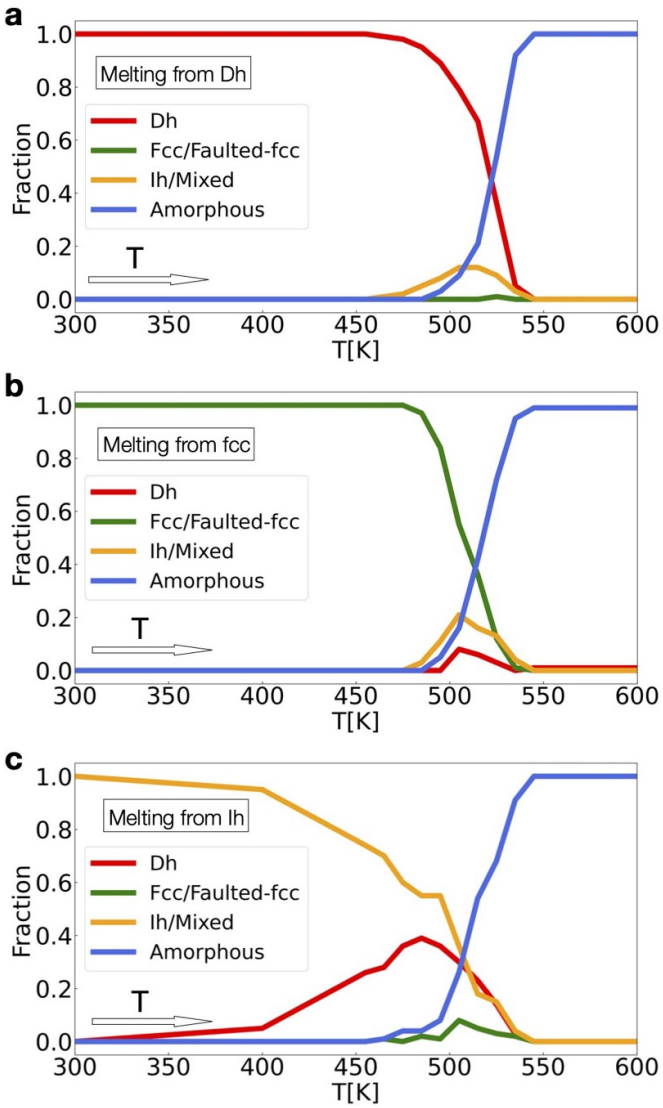


Figure 6. Melting of Au_{147}. (a) Fractions of main structural families vs temperature observed in melting simulations initialized from a Dh structure. The structures have been split in the same 4 structural families shown in figure 2(a). Fractions of amorphous structures are reported by the blue lines, fcc/faulted-fcc by green lines, Dh by red lines and Ih by orange lines. (b) Fractions of main structural families vs temperature observed in melting simulations initialized from an fcc structure. Same color code of panel (a). (c) Fractions of main structural families vs temperature observed in melting simulations initialized from an Ih structure. Same color code of panel (a).

In summary, ISVs have enabled the characterization of the complex structural landscape of the BK peptides and the dynamical interconversions between motifs, highlighting their potential for the study of diverse molecular systems. Furthermore, these results also demonstrate the flexibility of ISVs in working with various descriptors, which can thus be tailored to the specific system or computational needs. In the BK case, the distance matrix was used that is computationally more efficient than the RDF.

3. Conclusions

In conclusion, we devised a ML approach to obtain few general and yet informative CVs that enable the dynamical analysis of structural transitions. We coupled information from instantaneous atomic configurations and the related inherent structures using autoencoders. This approach distilled a small set of inherent structural variables capable of finely describing the structural landscape, evolution, and transitions of metal nanoclusters, both in equilibrium and in non-equilibrium. Our ISVs, in conjunction with umbrella sampling, allowed us to compute a high-resolution two-dimensional FE landscape of Au_{147} nanoclusters, revealing that the topology of the structural space comprises three major FE basins: fcc, Dh, and Ih. STEM experiments confirmed the existence of these structures in Au_{147}. In addition, our simulations shown that the minima are connected by two kinetic bottlenecks with Dh at the center. The basins are populated by several local FE minima, accounting, at finite temperature, for a wealth of metastable states and structural transitions among them. Transition rates between the three main FE basins were computed by means of MSMs, which allowed to validate quality of the ISVs by means of committor evaluations. In addition, the two ISVs were capable of tracking the structural evolution of non-equilibrium melting and freezing simulations, rationalizing routes to polymorph selection and recurring melting patterns. The scalability of the network was checked up to Au clusters of 309 atoms The generality of the strategy was tested by using ISVs to describe in real-time the structural rearrangements of the BK peptide, showing the existence of three basins corresponding to different motifs and punctuated by local minima. The successful analysis of this elusive biological case and the different descriptors used showcase the versatility of ISVs, suggesting their applicability in different contexts, including the field where the idea of inherent structures originated, i.e. liquids [13] but also glasses [18], colloids [51], and biomolecules [19].

4. Materials and methods

4.1. Deep learning

ISVs are obtained by training a modified autoencoder neural network, which associates to the descriptors of a non-minimized structure its inherent counterpart, being for Au_{147} the descriptors the RDFs of the structures, while for BK the distance matrices of alpha carbons. The network has the typical convergent-divergent architecture of autoencoders, where the first half, i.e. the encoder, is composed of convolutional layers, while the second half, i.e. the decoder mirrors the encoder and is composed of deconvolutional layers. The ISVs for Au_{147} are obtained as output of the trained encoder. The dataset for the network training is a collection of $613,872$ Au_{147} structures generated by means of PT. Reducing by 90% the dataset yielded satisfactory results for the fcc showing that the quality rather than the

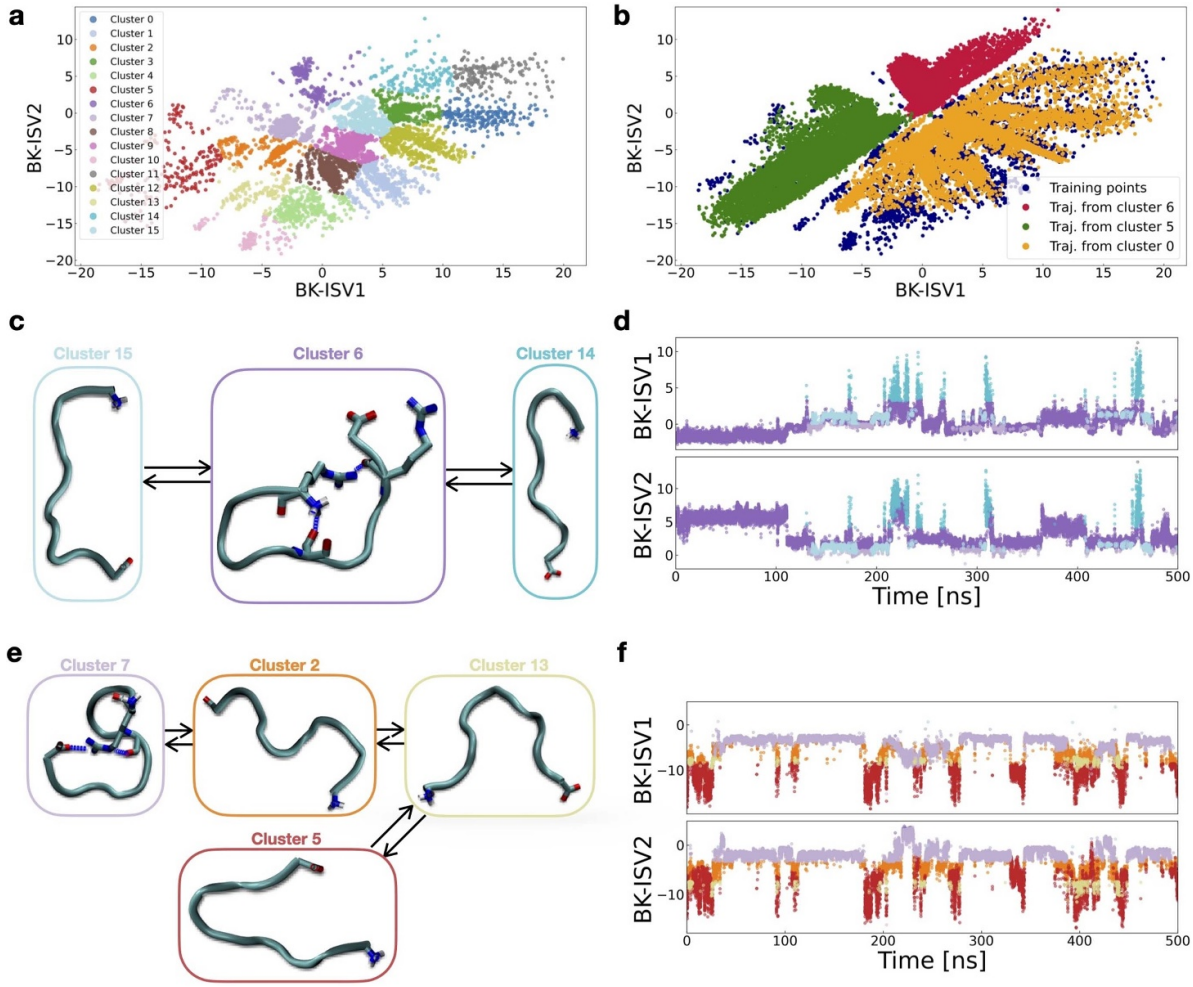


Figure 7. Real-time structural classification of bradykinin peptide dynamics via ISVs. (a) 2D ISV space generated by training the network using as descriptors the distance matrices of the different BK structures. The variables have been named BK-ISV to distinguish them from the Au_{147} ISVs. In the plot, the points are colored according to the results of k-means clustering with 16 centroids. (b) Plot of three 500 ns unbiased trajectories initialized in different regions of the ISV space: from cluster 6 (red), cluster 5 (green), and cluster 0 (yellow). The dark blue points below represent the training data. Structural diagram depicting the main structural transitions (c) and time history of the ISVs (d) for the trajectory initialized in cluster 6. Points are coloured according the cluster label in panel (a). Structural diagram depicting the main structural transitions (e) and time history of the ISVs (f) for the trajectory initialized in cluster 5.

quantity of data matters, see figures S23 and S24. Every structure was then minimized leading to the computation of the RDF for both instantaneous and inherent structures. Details about the BK dataset can be found in the BK simulations section.

The network has then been trained with a bottleneck size equal to 2 for both systems, feeding it the non-minimized structures descriptors and comparing the outputs with the associated inherent structures descriptors via a mean square error loss function.

For a more detailed description of the dataset [52], network architecture and training, see supplementary text.

4.2. Umbrella sampling Monte Carlo simulations

Umbrella sampling simulations [53] were performed using a Metropolis Monte Carlo code which was custom written in C++ for this purpose (supplementary text). A total of 15 022 simulations, distributed all over the ISV landscape have been performed. Simulations have been initialized using the thermalized structure in the training dataset that is closest to the restraining value. After careful tuning the harmonic spring constant of the umbrellas has been set equal to 0.1 eV. The simulations consisted in a total of $20 \cdot 10^6$ MC moves. CVs values have been sampled every $5 \cdot 10^3$ moves for a total of $4 \cdot 10^3$ samples for every simulation. After discarding the first $1/4$ of samples, random sampling with replacement was used to generate 10 different samplings from the original populations. FE was then reconstructed using each of these samplings allowing for the statistical error estimation via the boostrapping method. For the reconstruction of the FE landscape for each boostrap realization the WHAM algorithm [54] has been used in the implementation by Grossfield [55]. A more detailed description of the simulations procedure with information on

the convergence of the FE reconstruction is provided in the supplementary text.

4.3. Molecular dynamics and MSMs

MD simulations were performed using the LAMMPS code [56] augmented with custom Python code to perform on-the-fly estimation of the ISVs. Simulations have been thermalized using a Langevin thermostat with a time constant of 1.0 ps. A 5 fs timestep was used during integration. Each integration was carried for 0.5 μs (corresponding to $100 \cdot 10^6$ timesteps), during which the sampling of the ISVs was performed every 50 ps resulting in a total of 10^4 samples for every simulation. A grand total of 4 448 of these simulations have been performed, starting from initial configurations distributed all over the most relevant regions of the FE landscape (figure S14(a)). These simulations amounted to a total sampling time of 2.2 ms. Again simulations have been initialized by picking the thermalized structure of the training dataset closest to the selected starting point. MSM [45] calculations have been performed using the DeepTime [57] library. This analysis has been conducted in the ISV space, leveraging information on the stationary distribution obtained by the US calculations. Committor was estimated using TPT [46, 58] as implemented in DeepTime library [57]. Additional information regarding MSM is reported in the supplementary text.

4.4. Non equilibrium simulations

Non equilibrium MD simulations (freezing and melting) were performed using the LAMMPS code using a Langevin thermostat with the same time settings described in the previous section. The freezing simulations start from a highly disordered liquid configuration which is equilibrated at 600 K for 1 ns. The temperature is then decreased at a rate of 1 K ns^{-1} to a final temperature of 300 K. In the case of melting we considered four different initial configurations—fcc, twin, Ih, and Dh. The initial configurations are equilibrated at 300 K for 1 ns and then the temperature is raised to 600 K at a rate of 1 K ns^{-1}. In both the cases, configurations were sampled every 5 ps in the temperature range of 450–550 K and 50 ps at other temperatures. A total of 4200 freezing simulations and 300 melting simulations were performed.

4.5. BK simulations

The BK decapeptide by Gori *et al* [32] (RPPGFSPFRK) was built using the xleap tool of the AmberTools23 [59] package. All simulations have been performed with the Amber18 suite [60] of programs using the force field ff14SB [61] and the TIP3P water model [62]. Seven sodium and ten chloride ions were added to neutralize the charge of the cationic peptide and reach a 0.15 M concentration of NaCl. The system first underwent 1000 steps of steepest descent minimization followed by 9000 steps of conjugate gradient minimisation. The system was then equilibrated for 2 ns in the NPT ensemble keeping the temperature at 300 K with the Langevin thermostat with collision frequency of 1 ps^{-1}, and the pressure at 1 atm using the Berendsen barostat with pressure relaxation time of 2 ps. The equilibration continued for other 2 ns in the NVT ensemble(simulated via Langevin Thermostat) keeping the temperature fixed at 300 K. In Hamiltonian replica exchange simulations several replicas of the system were run in parallel at the same temperature. The replicas share the same features, yet they are characterized by different potential energy functions. In our case, the torsional term of the force field had been altered by scaling factors c_i. We used 32 replicas with exponentially decreasing c_i, from 1.0 in the unperturbed replica to 0.10 in the last replica (1.0, 0.93, 0.86, 0.80, 0.74, 0.69, 0.64, 0.60,0.56, 0.52, 0.48, 0.45, 0.42, 0.39, 0.36, 0.33, 0.31, 0.29, 0.27, 0.25, 0.23, 0.22, 0.20, 0.19, 0.17, 0.16, 0.15, 0.14, 0.13, 0.12, 0.11, 0.10). We checked that this settings allow for an acceptance probability of the exchanges of 70%. Each replica was equilibrated for 2 ns in the NVT ensemble with its specific scaled potential. The H-REMD simulation was run for 100 ns using the same settings as the NVT simulation, with exchange attempts every 1 ps. The classification obtained by k-means clustering in the ISV space was validated against a purely structural classification obtained by the quality threshold method [63]. In the latter case, the metrics employed was the maximum difference between distances of corresponding pairs of carbon atoms. As summarized in tables S2 and S3 some clear correlations emerge in the two classifications. For instance, ISV-clusters 6, 8, and 9 (W-shaped conformations) correspond to QT-clusters 2, 1, and 5 respectively also representing the same structural motif. Similarly, ISV-cluster 2, corresponding to S-shaped conformations, is mapped into QT-cluster 6 populated by a majority of S-shaped conformations. In the case of hairpin-like and extended conformations, the correspondence between clusters is more complex. However, also in this case ISV-clusters tend to be mapped into QT-clusters exhibiting the same structural motif.

4.6. Experimental methodology

The Swansea University Nanocluster Source, located at the B07 beamline of the Diamond Light Source synchrotron, was used to produce and deposit gold clusters for experimental STEM imaging and thus structure comparison with the theoretical results via the Simulation Atlas approach [44, 64, 65]. Size-selected Au147 clusters ($N = 147 \pm 3$ atoms) were deposited onto silicon nitride heating chips (DENS Solutions) using this DC magnetron-sputtering, inert-gas condensation cluster beam source coupled with a lateral time-of-flight mass selector and deposition stage [66, 67]. The mass filter (resolution $M/\Delta M = 25$) was calibrated with a beam of Ar+ ions. To reduce cluster agglomeration, the cluster beam was rastered across the support to deposit a uniform coverage (approximately 1% by projected surface area) on the silicon nitride imaging window. Clusters were soft-landed [68] at a kinetic energy of 1 eV atom^{-1} and allowed to diffuse and immobilise at pre-formed defect sites created in advance by sputtering of the window with an Ar+ beam at 500 V for 10 min [69]. The agglomeration observed could be associated with harmonics of the incidence Au147 clusters (see below).

HAADF-STEM images were acquired with a JEOL ARM300F (GRAND-ARM) microscope at the electron Physical Sciences Imaging Centre at Diamond Light Source. The electron beam energy was 300 kV and beam current was approximately 30 pA. The probe semi-angle was approximately 23 mrad and the HAADF detector had an inner collection angle of approximately 58 mrad (outer angle approximately 215 mrad). A DENS Solutions Wildfire holder was used to heat the samples to a range of temperatures (100 °C, 150 °C, 200 °C, 250 °C, 300 °C and 350 °C consecutively). Temperature is monitored by a 4-point probe and is typically stable to within ±1 °C; all samples were measured within a central window to ensure accuracy. Videos were acquired using a plug-in for Digital Micrograph, with a frame acquisition time of 1.31 s.

The cluster structure typically fluctuates from frame to frame. The structural assignment of each frame in each cluster video was accomplished by comparison with a Simulation Atlas generated using the abTEM Python package [70]. The PRISM algorithm [71] was used to simulate images (electron energy of 200 keV), a convergence semi-angle of 28 mrad, an interpolation factor of 4 and 10 frozen phonon iterations. Poisson noise was added to the simulated data to approximate an electron fluence of $1 \cdot 10^5$ $e^-/Å^{-2}$ (which is on the order of the fluence used in our imaging). We note that the electron energy and convergence semi-angle of the simulations do not exactly match those of the experiment, but the relevant structural elements used to assign the cluster structures do not depend on the microscope parameters and so the isomers can be assigned regardless.

The Au_{147} clusters were identified as the smallest clusters in each video frame; also found on the surface were larger clusters—being multiples of 147 atoms, as judged by their integrated intensities [72], presumably formed by surface agglomeration. The illustrative example images shown in figure 3 are low-noise Au_{147} clusters. The images shown were chosen to illustrate the principal structural motifs observed in the experiments. These images were processed by application of a high frequency filter to suppress noise and adjustment of brightness and contrast. A color gradient was also mapped onto the greyscale images to better highlight the structural features. The processed frames are compared with the best fits in the simulation atlases for icosahedral, decahedral and fcc structures of an Au_{147} cluster. The atlases cover the full range of polar and azimuthal orientations. Recent examples of this approach are [22, 73] and [74]. Key to the manual best matching process are the patterns and symmetries in the core region of the nanoparticle, where the signal is highest.

Data availability statement

The data that support the findings of this study will be openly available following an embargo at the following URL/DOI: https://doi.org/10.5281/zenodo.15564106.

Acknowledgments

We thank John Russo for useful discussions.

Conflict of interest

There are no competing interests to declare.

Funding

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 2.2.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union—NextGenerationEU-Project Title PINENUT—CUP D53D23002340006 - Grant Assignment Decree No. 957 adopted on 30/06/2023 by the Italian Ministry of University and Research (MUR). We thank Diamond Light Source for access to and support in use of the electron Physical Science Imaging Centre (Instrument E02, Proposal No.: MG28449), and gratefully acknowledge EPSRC Grant EP/V029797/2 for support of the electron microscopy. MR is grateful to EPSRC (via Swansea University) and Johnson Matthey for a PhD scholarship. HH is grateful to EPSRC (via the M2A CDT at Swansea University) and Diamond Light Source for an EngD scholarship.

Author contributions

E T and A T designed the ML algorithms and programmed the MD/MC simulation codes. E T coded the Neural Networks, performed equilibrium simulations and analyzed simulation data. A T supervised simulation work and data analysis. C G and L K K performed BK simulations. C G performed BK structural analysis. M S performed structural analysis and non-equilibrium simulations. E T, A T, R F, and A G conceptualized the work.

H H produced the clusters under the supervision of B v I, G H and R E P. M D performed ac-STEM imaging of clusters under the supervision of T S; T S produced the Simulation Atlas. M R matched experimental images to the simulation atlas under the supervision of R E P; H H, B v I, G H, T S, M R, and R E P wrote the experimental section of the manuscript under the coordination of R E P.

L M supervised rare event methodology, R F and A G supervised the work. A G wrote the original draft, which was reviewed and approved by all authors.

ORCID iDs

Emanuele Telari https://orcid.org/0009-0009-3296-959X
Antonio Tinti https://orcid.org/0000-0002-6750-6503
Carlo Guardiani https://orcid.org/0000-0002-8914-9260

Thomas J A Slater https://orcid.org/0000-0003-0372-1551
Richard E Palmer https://orcid.org/0000-0001-8728-8083
Luca Maragliano https://orcid.org/0000-0002-5705-6967
Riccardo Ferrando https://orcid.org/0000-0003-2750-9061
Alberto Giacomello https://orcid.org/0000-0003-2735-6982

References

[1] Bolhuis P G, Chandler D, Dellago C and Geissler P L 2002 Transition path sampling: throwing ropes over rough mountain passes, in the dark *Annu. Rev. Phys. Chem.* **53** 291–318
[2] Bolhuis P G, Dellago C and Chandler D 2000 Reaction coordinates of biomolecular isomerization *Proc. Natl Acad. Sci.* **97** 5877–82
[3] Frenkel D and Ladd A J 1984 New Monte Carlo method to compute the free energy of arbitrary solids. Application to the fcc and hcp phases of hard spheres *J. Chem. Phys.* **81** 3188–93
[4] Laio A and Parrinello M 2002 Escaping free-energy minima *Proc. Natl Acad. Sci.* **99** 12562–6
[5] Maragliano L and Vanden-Eijnden E 2006 A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations *Chem. Phys. Lett.* **426** 168–75
[6] Geissler P L, Dellago C, Chandler D, Hutter J and Parrinello M 2001 Autoionization in liquid water *Science* **291** 2121–4
[7] Allen R J, Valeriani C and Ten Wolde P R 2009 Forward flux sampling for rare event simulations *J. Phys.: Condens. Matter* **21** 463102
[8] Sultan M M and Pande V S 2018 Automated design of collective variables using supervised machine learning *J. Chem. Phys.* **149** 094106
[9] Wehmeyer C and Noé F 2018 Time-lagged autoencoders: deep learning of slow collective variables for molecular kinetics *J. Chem. Phys.* **148** 241703
[10] Sidky H, Chen W and Ferguson A L 2020 Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation *Mol. Phys.* **118** e1737742
[11] Jung H, Covino R, Arjun A, Leitold C, Dellago C, Bolhuis P G and Hummer G 2023 Machine-guided path sampling to discover mechanisms of molecular self-organization *Nat. Comput. Sci.* **3** 334–45
[12] Zeni C, Rossi K, Pavloudis T, Kioseoglou J, de Gironcoli S, Palmer R E and Baletto F 2021 Data-driven simulation and characterisation of gold nanoparticle melting *Nat. Commun.* **12** 6056
[13] Stillinger F H and Weber T A 1983 Inherent structure in water *J. Phys. Chem.* **87** 2833–40
[14] Pavan L, Rossi K and Baletto F 2015 Metallic nanoparticles meet metadynamics *J. Chem. Phys.* **143** 184304
[15] Tribello G A, Giberti F, Sosso G C, Salvalaglio M and Parrinello M 2017 Analyzing and driving cluster formation in atomistic simulations *J. Chem. Theory Comput.* **13** 1317–27
[16] Pipolo S, Salanne M, Ferlat G, Klotz S, Saitta A M and Pietrucci F 2017 Navigating at will on the water phase diagram *Phys. Rev. Lett.* **119** 245701
[17] Fan Y, Iwashita T and Egami T 2017 Energy landscape-driven non-equilibrium evolution of inherent structure in disordered material *Nat. Commun.* **8** 15417
[18] Tanaka H, Tong H, Shi R and Russo J 2019 Revealing key structural features hidden in liquids and glasses *Nat. Rev. Phys.* **1** 333–48
[19] Nakagawa N and Peyrard M 2006 The inherent structure landscape of a protein *Proc. Natl Acad. Sci.* **103** 5279–84
[20] Rao F and Karplus M 2010 Protein dynamics investigated by inherent structure analysis *Proc. Natl Acad. Sci.* **107** 9152–7
[21] Baletto F and Ferrando R 2005 Structural properties of nanoclusters: energetic, thermodynamic and kinetic effects *Rev. Mod. Phys.* **77** 371–423
[22] Foster D M, Ferrando R and Palmer R E 2018 Experimental determination of the energy difference between competing isomers of deposited, size-selected gold nanoclusters *Nat. Commun.* **9** 1323
[23] Mottet C, Tréglia G and Legrand B 1997 New magic numbers in metallic clusters: an unexpected metal dependence *Surf. Sci.* **383** L719–27
[24] Apra E, Baletto F, Ferrando R and Fortunelli A 2004 Amorphization mechanism of icosahedral metal nanoclusters *Phys. Rev. Lett.* **93** 065502
[25] Xia Y, Nelli D, Ferrando R, Yuan J and Li Z Y 2021 Shape control of size-selected naked platinum nanocrystals *Nat. Commun.* **12** 3019
[26] Wales D J, Miller M A and Walsh T R 1998 Archetypal energy landscapes *Nature* **394** 758–60
[27] Telari E, Tinti A, Settem M, Maragliano L, Ferrando R and Giacomello A 2023 Charting nanocluster structures via convolutional neural networks *ACS Nano* **17** 21287–96
[28] Amodeo J, Pietrucci F and Lam J 2020 Out-of-equilibrium polymorph selection in nanoparticle freezing *J. Phys. Chem. Lett.* **11** 8060–6
[29] Settem M, Ferrando R and Giacomello A 2022 Tempering of Au nanoclusters: capturing the temperature-dependent competition among structural motifs *Nanoscale* **14** 939–52
[30] Settem M, Roncaglia C, Ferrando R and Giacomello A 2023 Structural transformations in Cu, Ag and Au metal nanoclusters *J. Chem. Phys.* **159** 094303
[31] Schebarchov D, Baletto F and Wales D J 2018 Structure, thermodynamics and rearrangement mechanisms in gold clusters-insights from the energy landscapes framework *Nanoscale* **10** 2004–16
[32] Gori A *et al* 2020 Membrane-binding peptides for extracellular vesicles on-chip analysis *J. Extracell. Vesicles.* **9** 1751428
[33] Chen J and Johnson E 2007 Targeting the bradykinin B1 receptor to reduce pain *Expert Opin. Ther. Targets* **11** 21–35
[34] Lopez J, Shukla A, Reinhart C, Schwalbe H, Michel H and Glaubitz C 2008 The structure of the neuropeptide bradykinin bound to the human G-protein coupled receptor bradykinin B2 as determined by solid-state NMR spectroscopy *Angew. Chem. Int. Ed.* **47** 1668–71
[35] Bonechi C, Ristori S, Martini G, Martini S and Rossi C 2009 Study of bradykinin conformation in the presence of model membrane by nuclear magnetic resonance and molecular modelling *Biochim. Biophys. Acta* **1788** 708–16
[36] Calvo F, Doye J and Wales D 2002 Equilibrium properties of clusters in the harmonic superposition approximation *Chem. Phys. Lett.* **366** 176–83
[37] De S, Bartók A P, Csányi G and Ceriotti M 2016 Comparing molecules and solids across structural and alchemical space *Phys. Chem. Chem. Phys.* **18** 13754–69
[38] Drautz R 2019 Atomic cluster expansion for accurate and transferable interatomic potentials *Phys. Rev. B* **99** 014104
[39] 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
[40] Hinton G E and Salakhutdinov R R 2006 Reducing the dimensionality of data with neural networks *Science* **313** 504–7

[41] Rosato V, Guillope M and Legrand B 1989 Thermodynamical and structural properties of fcc transition metals using a simple tight-binding model *Phil. Mag. A* **59** 321–36
[42] Santarossa G, Vargas A, Iannuzzi M and Baiker A 2010 Free energy surface of two-and three-dimensional transitions of Au 12 nanoclusters obtained by ab initio metadynamics *Phys. Rev. B* **81** 174205
[43] El koraychy E y, Roncaglia C, Nelli D, Cerbelaud M and Ferrando R 2022 Growth mechanisms from tetrahedral seeds to multiply twinned au nanoparticles revealed by atomistic simulations *Nanoscale Horiz.* **7** 883–9
[44] Wang Z W and Palmer R E 2012 Determination of the ground-state atomic structures of size-selected au nanoclusters by electron-beam-induced transformation *Phys. Rev. Lett.* **108** 245502
[45] Bowman G R, Pande V S and Noé F 2013 *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* vol 797 (Springer Science & Business Media)
[46] Metzner P, Schütte C and Vanden-Eijnden E 2009 Transition path theory for markov jump processes *Multiscale Model. Simul.* **7** 1192–219
[47] Weinan W, Ren W and Vanden-Eijnden E 2002 String method for the study of rare events *Phys. Rev. B* **66** 052301
[48] Huang R, Wen Y, Voter A F and Perez D 2018 Direct observations of shape fluctuation in long-time atomistic simulations of metallic nanoclusters *Phys. Rev. Mater.* **2** 126002
[49] Dearg M *et al* 2024 Frame-by-frame observations of structure fluctuations in single mass-selected Au clusters using aberration-corrected electron microscopy *Nanoscale Horiz.* **9** 143–7
[50] Lloyd S 1982 Least squares quantization in PCM *IEEE Trans. Inf. Theory* **28** 129–37
[51] De Nijs B, Dussi S, Smallenburg F, Meeldijk J D, Groenendijk D J, Filion L, Imhof A, Van Blaaderen A and Dijkstra M 2015 Entropy-driven formation of large icosahedral colloidal clusters by spherical confinement *Nat. Mater.* **14** 56–60
[52] Telari E *et al* 2025 Data of inherent structural descriptors via machine learning *Zenodo* (https://doi.org/10.5281/zenodo.15564106)
[53] Torrie G M and Valleau J P 1977 Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling *J. Comput. Phys.* **23** 187–99
[54] Ferrenberg A M and Swendsen R H 1989 Optimized Monte Carlo data analysis *Phys. Rev. Lett.* **63** 1195
[55] Grossfield A WHAM: the weighted histogram analysis method, version 2.0.11
[56] 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
[57] Hoffmann M *et al* 2021 Deeptime: a python library for machine learning dynamical models from time series data *Mach. Learn.: Sci. Technol.* **3** 015009
[58] Noé F, Schütte C, Vanden-Eijnden E, Reich L and Weikl T R 2009 Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations *Proc. Natl Acad. Sci.* **106** 19011–6
[59] Case D *et al* 2023 Ambertools *J. Chem. Inf. Model.* **63** 6183–91
[60] Case D A *et al* 2018 AMBER 2018
[61] Maier J, Martinez C, Kasavajhala K, Wickstrom L, Hauser K and Simmerling C 2015 ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB *J. Chem. Theory Comput.* **11** 3696–713
[62] Jorgensen W, Chandrasekhar J, Madura J and Klein M 1983 Comparison of simple potential functions for simulating liquid water *J. Chem. Phys.* **79** 926–35
[63] Heyer L, Kruglyak S and Yooseph S 1999 Exploring expression data: identification and analysis of coexpressed genes *Genome Res.* **9** 1106–15
[64] Wang Z W and Palmer R E 2012 Direct atomic imaging and dynamical fluctuations of the tetrahedral Au20 cluster *Nanoscale* **4** 4947–9
[65] Wang Z W and Palmer R E 2012 Experimental evidence for fluctuating, chiral-type Au_{55} clusters by direct atomic imaging *Nano Lett.* **12** 5510–4
[66] Pratontep S, Carroll S J, Xirouchaki C, Streun M and Palmer R E 2005 Size-selected cluster beam source based on radio frequency magnetron plasma sputtering and gas condensation *Rev. Sci. Instrum.* **76** 045103
[67] Issendorff B v and Palmer R E 1999 A new high transmission infinite range mass selector for cluster and nanoparticle beams *Rev. Sci. Instrum.* **70** 4497–501
[68] Di Vece M, Palomba S and Palmer R E 2005 Pinning of size-selected gold and nickel nanoclusters on graphite *Phys. Rev. B* **72** 073407
[69] Claeyssens F, Pratontep S, Xirouchaki C and Palmer R E 2006 Immobilization of large size-selected silver clusters on graphite *Nanotechnology* **17** 805–7
[70] Madsen J and Susi T 2021 The abTEM code: transmission electron microscopy from first principles [version 2; peer review: 2 approved] *Open Res. Eur.* **1** 24
[71] Ophus C 2017 A fast image simulation algorithm for scanning transmission electron microscopy *Adv. Struct. Chem. Imaging* **3** 13
[72] Wang Z W, Li Z Y, Park S J, Abdela A, Tang D and Palmer R E 2011 Quantitative Z-contrast imaging in the scanning transmission electron microscope with size-selected clusters *Phys. Rev. B* **84** 073408
[73] Dearg M, Lethbridge S, McCormack J, Palmer R E and Slater T J A 2024 Characterisation of the morphology of surface-assembled au nanoclusters on amorphous carbon *Nanoscale* **16** 10827–32
[74] Lethbridge S, Pavloudis T M, Slater T J, Kioseoglou J and Palmer R E 2024 Stabilization of 2D raft structures of au nanoclusters with up to 60 atoms by a carbon support *Small Sci.* **4** 2400093