Machine learning for molecular simulations of crystal nucleation and growth

Sapna Sarupria[1*], Steven W. Hall[2] and Jutta Rogal[3,4*]

[1]Department of Chemistry, University of Minnesota, Minneapolis, 55455, MN, USA.
[2]Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, 55455, MN, USA.
[3]Department of Chemistry, New York University, New York, 10003, NY, USA.
[4]Fachbereich Physik, Freie Universität Berlin, Berlin, 14195, Germany.

*Corresponding author(s). E-mail(s): sarupria@umn.edu; jutta.rogal@nyu.edu;

Abstract

Molecular simulations are a powerful tool in the study of crystallization and polymorphic transitions yielding detailed information of transformation mechanisms with high spatiotemporal resolution. However, characterizing various crystalline and amorphous phases as well as sampling nucleation events and structural transitions remain extremely challenging tasks. The integration of machine learning with molecular simulations has the potential of unprecedented advancement in the area of crystal nucleation and growth. In this article, we discuss recent progress in the analysis and sampling of structural transformations aided by machine learning and the resulting potential future directions opening in this area.

Keywords: nucleation, crystal, machine learning, molecular simulations

Introduction

The properties of materials depend not only on their composition but inherently also on the details of their crystalline or amorphous structure. In condensed phase systems, different crystalline phases composed of the same elements or molecules, also known as polymorphs, can exhibit vastly different materials properties. Prominent examples include the many different polymorphs of ice (with ice XIX as the most recent experimentally characterized one [1]) or the versatility of carbon forming diamond, graphite, graphene, or fullerenes. The stability of different polymorphs and transformations between polymorphs are of importance in a wide range of application areas such as pharmaceuticals (impacting, for example, bioavailability), organic electronics (affecting properties such as charge transport [2]), or metals and alloys for high-performance materials in the energy and transportation sector. Controlling the formation of specific polymorphs and possibly stabilizing metastable forms is key in the design of novel materials with tailor-made properties. Molecular simulations of nucleation and growth during crystallization and of polymorphic transitions can provide a fundamental understanding of the underlying molecular processes and mechanisms. In recent years, machine learning (ML) has become a valuable tool in many areas of molecular simulations [3]. In this article, we focus particularly on how ML approaches can aid in the analysis

of simulation data as well as the sampling of crystallization processes and polymorphic transitions. In the next section, we discuss the application of ML to the analysis of local and global structural features, the evaluation of transformation mechanisms, and the identification of suitable reaction coordinates for crystallization processes. The section on ML aided sampling of nucleation and growth illustrates recent ideas on how low-dimensional representations of transition paths extracted from ML models can be combined with enhanced sampling approaches to explore nucleation and polymorphic transitions with molecular simulations. We close with a brief perspective on how the rapid developments in the field of ML may further benefit the analysis and sampling of crystallization processes.

ML aided analysis of nucleation and growth

In inferring mechanisms of crystallization from molecular simulations, a key task is the characterisation of structural environments to be able to distinguish between liquid, amorphous, and different crystalline phases. This is typically done through the use of collective variables (CV) that are functions of the phase space coordinates. Traditionally, these CVs are physically motivated and, for example, based on the symmetry of the local structure such as Steinhardt bond order parameters [4], coordination numbers, or tetrahedrality. Moreover, the analysis of transformation mechanisms is often linked to the definition of a reaction coordinate (RC) that describes the progress along the structural transition. The RC is usually given by a linear or nonlinear combination of suitable CVs. Given the complexity of crystallization processes, the identification of an appropriate RC can, however, be highly non-trivial and may require a non-intuitive combination of CVs. In the following, we illustrate the application of ML approaches in structure classification and RC determination.

Local and global structure identification

Deriving traditional CVs for structure classification often requires *a priori* knowledge of the structural features that characterize the different phases to be distinguished. A single CV might not be sufficient and usually a combination of CVs is necessary, in particular, when differentiating between several crystalline structures simultaneously. Furthermore, the identification of any new structure requires the development of new CVs which can be difficult and time-consuming. As an alternative, ML-based classification has emerged as a powerful tool for structure identification.

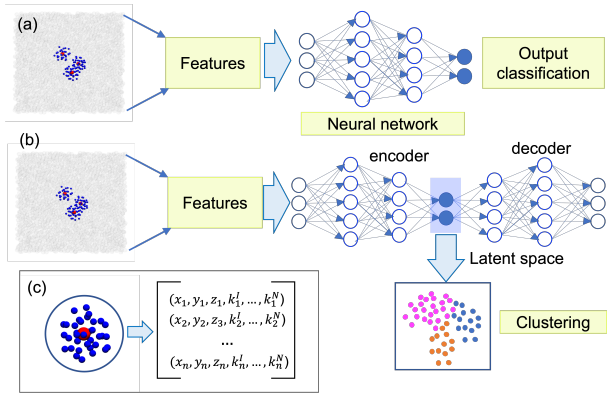

Fig. 1 Schematic of (a) supervised and (b) unsupervised learning for crystal structure identification. (c) Illustration of point cloud which is input to PointNet-based structure classification [5]. The input features include the x-, y-, and z-coordinates of the neighbors of the atom to be classified, and additional features k can be added. For example, k can be the atomic identity.

ML approaches for structure classification can be categorized into two groups: supervised and unsupervised learning, as schematically shown in Fig. 1(a) and (b). In supervised learning, the data to train the ML model are labeled, that is the input features have to be provided together with the information about the corresponding structure. The ML model then learns the decision boundaries between the different structures in the space of input features. Once trained, the input features computed for any arbitrary structure can be passed to the ML model to assign a structure. Unsupervised learning, on the other hand, does not require any structure assignment to the data which is particularly useful when analysing data with unknown or transient structures that are different from the bulk crystalline phases. Having, for example, trained an autoencoder, the ML model can perform a nonlinear dimensionality reduction of the input features and project the data into a low-dimensional latent space (see

Fig. 1(b)) in which the different structures should be sufficiently separated. The regions in the latent space that correspond to different structures can be identified by applying a clustering algorithm. Any unknown structure can then be projected into the latent space and will be classified by its location on the latent space map.

A key component in ML structure identification in molecular simulations is the design of the input features for the ML model. There is often a trade-off between feature complexity and model complexity. Input features that are relatively complex and highly tuned functions of a particular system typically require simpler ML models than more general input features (e.g., direct Cartesian coordinates). Additionally, the features used for structure classification need to be rotation-, translation-, and permutation-invariant. The majority of efforts in ML structure identification are based on supervised learning. One of the first applications to condensed phase systems is the work by Geiger and Dellago [6] who were motivated by the failure of traditional CVs to distinguish between various polymorphs of ice. They used radial and angular symmetry functions as input features to a neural network (NN) to classify the local structural environment of six different ice polymorphs. One limitation of their approach is that the input features need to be selected and tuned specifically for the structures of interest. As the different phases of ice and water are notoriously difficult to classify, this system has inspired several other ML models. GCIceNet [7] is based on a graph NN that treats each water molecule as a node and each hydrogen bond as an edge in the graph representation. In contrast to the symmetry functions, no significant feature tuning is required in this approach. The model successfully classified nine different phases of ice/water and the graph representation was also demonstrated to work effectively for unsupervised learning. Another model, called DeepIce [8], focused specifically on minimal feature engineering. The input to DeepIce for a given water molecule are the x-, y-, and z-coordinates of the neighboring molecules. This input is transformed into features through a number of sub-networks operated on the Cartesian coordinates, spherical coordinates, spherical harmonics and Fourier transforms of the Cartesian coordinates. DeepIce has, so far, only been applied to distinguish liquid water and hexagonal ice. An even more general approach to derive input features was proposed in connection with the PointNet architecture [9] which was originally developed to process and classify point cloud data. In the context of structure classification, the environment around each atom is treated as a point cloud and only the relative Cartesian coordinates of the neighboring atoms are needed as input [5], as shown in Fig. 1(c). PointNet was applied to a broad range of systems including Lennard-Jones particles, water/ice, water at interfaces, and mesophases.

Unsupervised learning approaches for structure identification vary in the design of the input features, the method chosen for dimensionality reduction, and the approach to cluster the data in the latent space. As in supervised learning, input features ranging from rather structure specific to fairly general have been used, including Steinhardt bond order parameters [10], a (possibly) large number of distances, angles, spherical harmonics etc. [11, 12], a graph representation (GCIceNet, discussed above) [7], a graphlet decomposition of the input [13, 14], and a point cloud representation [15]. Autoencoders are a popular choice for nonlinear dimensionality reductions and have been employed in combination with several of the input features [7, 10, 13, 15]. Other approaches for dimensionality reduction that have been proposed in this context are, for example, uniform manifold approximation and projection for dimension reduction (UMAP) [11, 12, 16] and diffusion maps [14]. Clustering in the latent space is usually performed with standard methods such as Gaussian mixture models or hierarchical density-based spatial clustering of applications with noise (HDBSCAN) [17]. More recent work focused on building a data-centric crystal classifier (DC3) [18]. DC3 utilizes radial symmetry functions in combination with Steinhardt bond order parameters but instead of tuning the parameters of these features, each type of feature was calculated for a wide range of parameters, leading to hundreds of features per atom. The NN automatically determined the features most relevant to accurately classifying the local atomic environments. Their approach performed competitively against more traditional and specialized methods. The model was also designed to identify amorphous versus crystalline structure as well as to recognize a crystal structure that has not been previously identified. It,

therefore, performs a hybrid of supervised and unsupervised learning.

Since the early work of Geiger and Dellago [6], the field of ML-based structure identification has evolved significantly. Many different ML methods are now available for structural classification. The emerging unsupervised approaches have the potential to discover novel (and transient) structures that could help in our understanding of crystallization processes. The diversity of training data and systems used in ML-based structural classification methods make it difficult to directly compare them. Furthermore, such comparison becomes even harder because the methods were motivated by different purposes such as achieving high accuracy, understanding input features critical for classification, and achieving high computational speeds. Moving forward, developing a broad dataset that can be used as a community standard against which to benchmark the algorithms can help in the development of faster and accurate classification methods.

Analysis of crystallization mechanisms

In molecular simulations, crystallization mechanisms are often deciphered by studying the emergence of crystalline structures from the metastable liquid. A recent example where ML-based structure classification was crucial in analyzing simulation data is the crystallization of the binary colloidal AB_{13} crystal [19]. The clear distinction between the different, somewhat exotic, crystalline phases facilitated the observation of the growing nucleus, thus providing insight into the nucleation mechanism.

A more general approach to describe the progress of a crystallization or transformation process is the committor p_B. For a transition between two (meta)stable states A and B of a system (for example, between the liquid and solid state or between two crystalline phases), the committor measures the probability that a given configuration commits to state B before going back to A. Consequently, the committor increases monotonously from 0 to 1 along the transformation and can be considered as an ideal RC. It is, however, rather costly to compute and the committor itself does not provide any physical insight. But it can be used to evaluate the quality of a proposed RC: any CV or combination of CVs that constitute a good RC have to show a strong correlation with the committor. In fact, the combination of CVs that best correlates with the committor can be considered as an optimal representation of the RC, as illustrated in Fig. 2 for a simple 2-dimensional potential energy surface. For configurations at a single value of the trial RC x in Fig. 2(b), a wide range of p_B values is observed, indicating that x does not capture the progress of the transition between the two metastable states and is therefore not a good RC. In Fig. 2(c), $x+y$ correlates well with p_B, making it a much better approximation of the RC. This combination of CVs is often non-trivial and non-intuitive to identify. An ML model that is trained to predict the committor can thus be beneficial to identify the best combination of CVs corresponding to the RC along which the mechanism and kinetics can be evaluated.

In their pioneering work, Ma and Dinner [20] used genetic NNs (GNNs) to predict the committor and identify the RC. The number of input features was limited to two or three CVs to retain interpretability. For each set of CVs, an NN was trained and a genetic algorithm was employed to efficiently find the CV combination with the best statistical fit. Similar in spirit, recent studies utilized simulation data from transition path sampling (TPS) to train an NN to predict the committor [21, 22]. Instead of only a few CVs, a few hundred molecular features were used as input. A sensitivity analysis of the trained NN subsequently revealed the input features most important in predicting the committor and, thus, the RC. Furthermore, symbolic regression with this reduced number of input features was used to generate a human-interpretable model of the committor that approximates the complicated NN function. The approach was applied to LiCl ion pair dissociation in water which revealed that the optimal RC needs to capture the complex interplay between solvent and counterion environments. Another application was to the nucleation of methane hydrates where a switch in the mechanism with temperature was observed which could be expressed in a simple mathematical expression extracted from the trained NN. Frassek et al. [23] used an extended autoencoder approach (EAE) to identify the CVs that contribute dominantly

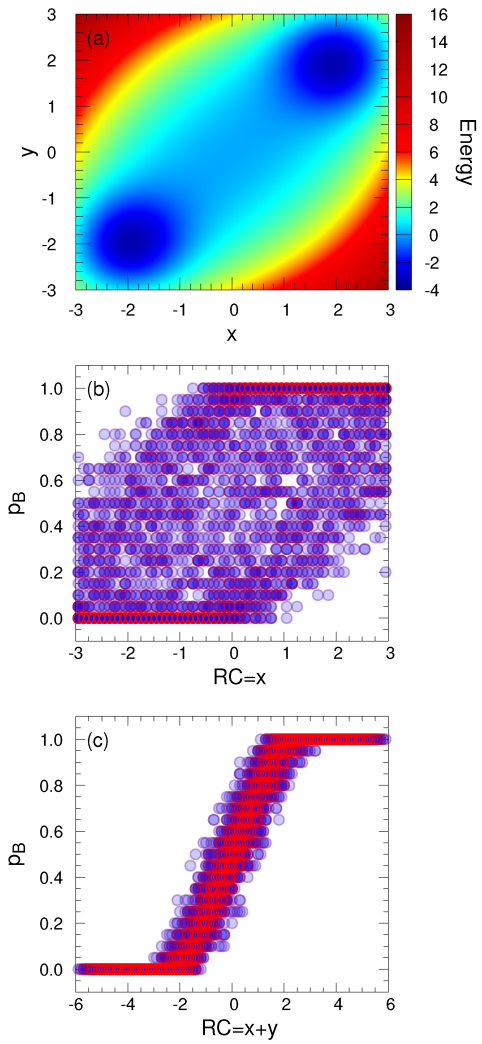

Fig. 2 (a) Model potential energy surface with two metastable states. (b) p_B when x is the trial RC. (c) p_B when $x+y$ is the trial RC. p_B estimates are represented as semi-transparent blue circles with red outlines in (b) and (c). Each p_B estimate for a given point on the surface in (a) was calculated by simulating 20 trajectories with Langevin dynamics starting from that point and monitoring whether the trajectories reached the left or right metastable state first.

to the RC. From the latent space of the autoencoder, both the input data were reconstructed and the committor was predicted simultaneously. The EAE model was applied to analyse TPS data for methane hydrate nucleation and identified important methane and water structural motifs that determined the probability of growing into a hydrate nucleus.

In the examples discussed, the objective is to use ML to decipher and identify the combination of CVs that describe the transition as accurately as possible while retaining interpretability. This advances nucleation studies in multiple ways: (*i*) identification of important CVs yields better mechanistic insights into nucleation and crystallization, which is also crucial for understanding polymorphic transitions; (*ii*) the important CVs can be used in enhanced sampling methods (e.g. metadynamics, umbrella sampling, etc.) to accelerate the sampling of nucleation processes (see also the next section); (*iii*) by focusing on the important CVs, it could be possible to develop high-throughput methods to screen through the effects of solution conditions and additives on nucleation. The discussed methods, however, still depend on a predetermined set of CVs and factors that are not captured by these CVs while contributing to the nucleation mechanism may be missed. Furthermore, interpretation of the latent space is generally not directly possible and can only be approximated through, for example, a sensitivity analysis possibly in combination with symbolic regression as discussed above.

ML aided sampling of nucleation and growth

Nucleation and growth processes in condensed phase systems often take place on timescales that are inaccessible by straightforward molecular dynamics (MD) simulations. These processes can be characterized as transitions between metastable states (for example, the liquid and solid state or different crystalline phases), that is, local minima on the free energy landscape, that are separated by significant energy barriers. The time spent within each of the metastable states is much longer than the transition time and this separation of timescales causes the transitions (and therefore the transformation processes) to be *rare events* on a molecular timescale. Over the years, a number of enhanced sampling approaches have been developed that facilitate the exploration of rare events. Often, these approaches require a suitable, low-dimensional RC along which the sampling is being performed. As discussed in the previous section, ML approaches can be used to analyze and interpret transformation mechanisms

and identify low-dimensional descriptors. However, the data needed for this analysis have to be sampled to begin with. This constitutes a chicken-and-egg problem of data-driven approaches to identify suitable CVs that are then needed in the enhanced sampling to produce the data. In the following, we discuss two ideas: the first one is to construct ML-based CVs using only data from the metastable states without having to sample the rare event explicitly; the second is iterative approaches where the data sampled along a putative CV are used to further improve the CV that is then employed in further enhanced sampling.

Collective variables for enhanced sampling

To sample structural transformations, CVs that can distinguish between the structural environments in the different phases are an obvious choice. For example, to study the nucleation of a crystalline phase from a supercooled liquid, the local environment around each atom or molecule can be characterized as representing either the solid or the liquid and all solid particles can subsequently be clustered. The size of this solid cluster may then be used as a 1-dimensional RC in enhanced sampling.

Structure identification based on ML has been treated mainly as a classification problem (see previous section). Consequently, the decision function separating the classes for different structural environments can be used as CV in enhanced sampling of structural transformations [24, 25]. The main advantage is that to train the ML structure classification, data are only needed within the different metastable states that represent the different phases which can be sampled efficiently with standard MD. An important aspect that needs to be considered if a CV is to be applied in enhanced sampling is that the CV needs to be differentiable with respect to atomic positions to provide the corresponding biasing forces. Since the forces are required in every simulation step the computation of the CV and its derivatives should also be computationally relatively cheap to avoid a significant increase in the computational cost. This requires a careful tuning of the complexity of the ML model as well as the selection of input features derived from the atomic configurations.

One recent example applied to crystallization is the combination of the structure factor (SF) with an NN and linear discriminant analysis (LDA) [26]. The peaks of the full three dimensional structure factor globally characterize the system as being either liquid or solid. However, the number of peaks is too large to be directly used as CVs in an enhanced sampling scheme. Instead, an NN was employed to combine the SF peaks non-linearly and the LDA was performed in this reduced space to separate the solid and liquid phase. The resulting 1-dimensional *deep-LDA* CV was then used in on-the-fly probability enhanced sampling [27] to explore the free energy landscape of crystallization of sodium chloride and carbon dioxide [26].

Another example is the local structure classification where atom-centered symmetry functions are used as input features for a classification NN (see the section on local and global structure identification). In [6], the crystallization of supercooled water was studied using the largest crystalline cluster as CV in umbrella sampling. The identification of water molecules in a crystalline environment was performed with a classification NN but the NN output entered the CV only indirectly, and no force evaluations with respect to the CV and thus the NN output were needed.

To study solid-solid transformation mechanisms between two crystalline phases in molybdenum, the NN classification of the local structure was directly used to apply a biasing force and drive the structural transformation with metadynamics and driven adiabatic free energy dynamics [28, 29]. The output of the classification NN for each atom was combined into global classifiers that represent the fraction of each crystalline phase in the entire system. Within this space of global classifiers a path CV was defined along which the fraction of each phase changes, thus driving the transformation [28, 29]. The nonlinear combination of these global classifiers into a path CV is much more efficient than using the global classifiers directly as CVs since the phase fractions are not independent (as one phase grows another one must shrink). In this system, the transformation proceeds via local changes through a disordered/amorphous region at the phase boundary (Fig. 3(c)). The biasing forces entering the enhanced sampling are large for atoms close to the interface and almost zero

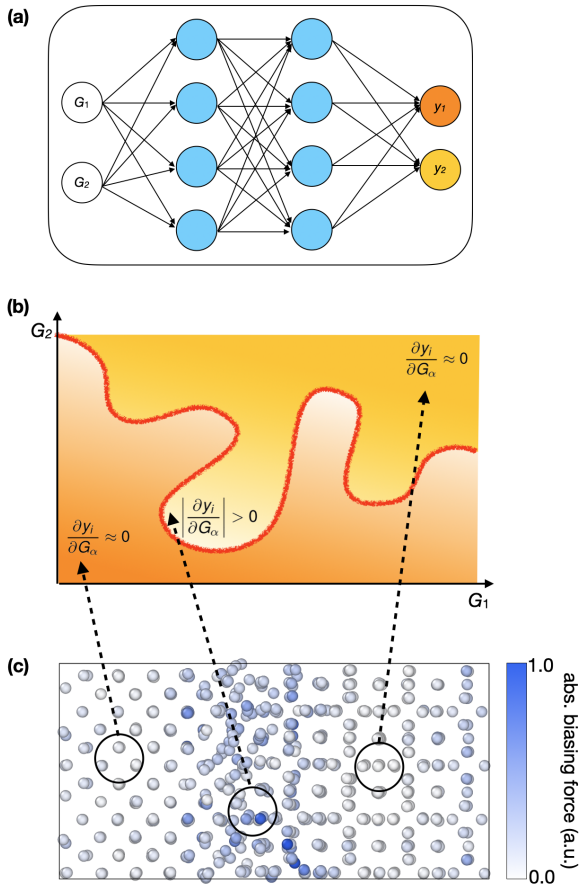

Fig. 3 (a) Schematic representation of a classification neural network with features G_α in the input layer and classes y_i in the output layer. (b) Schematic representation of the learnt decision function in the space of input features; the derivative of the NN output with respect to the input features is large close to the decision boundary (red line) and small far away from it. (c) Snapshot of an interface between a body-centred cubic (left) and topologically closed-packed (right) crystal structure. The color gradient indicates the magnitude of the biasing force in the enhanced sampling using a NN-based path CV (see [28, 29] for details). Biasing forces are large for local environments that are identified as close to the decision boundary and small for local environments that resemble the bulk crystalline phases.

in bulk regions (Fig. 3(c)) which is directly connected with the derivatives of the classification NN output: forces are small if the local environment is well within one of the classes and large if the local environment is close to a decision boundary (Fig. 3(b)). The enhanced sampling therefore promotes the local structural changes at the interface that drive the phase transformation.

A slightly different idea of combining a set of CVs into a 1-dimensional RC is the spectral gap optimization of order parameters (SGOOP) [30]. Here, the objective is to identify a linear combination of CVs that best separates the slow visible and fast hidden dynamics projected along the resulting 1-dimensional RC. This approach was recently used to study crystallization in urea [31] by combining six global descriptors to distinguish between the liquid and crystalline phases with SGOOP. Metadynamics was employed to explore the free energy along the SGOOP RC and sample the crystallization process. In contrast to the other approaches discussed in this section, SGOOP does require sampling of data along the transition as it evaluates the dynamics rather than structure classification. Possibly, a combination of SGOOP with ML approaches could enable a nonlinear combination of the trial CVs providing additional flexibility in the description of complex crystallization processes.

Iterative sampling and collective variable identification

Several iterative approaches have been introduced in the last few years that combine enhanced sampling and ML-aided identification of optimal, low-dimensional RCs. The basic idea that these approaches have in common is to start with data from an unbiased simulation or from enhanced sampling along non-optimal CVs, employ ML to propose improved CVs, continue the enhanced sampling with this new set of CVs, and subsequently include the new data in the training of the ML model. As this iterative process continues, the enhanced sampling should become more and more efficient and the ML model of the RC more accurate.

So far, these approaches have been applied to simple low-dimensional model potential energy surfaces and conformational changes in small biomolecules but not yet to crystallization or polymorphic transitions. However, we include a brief discussion here as they have the potential to improve the sampling of complex structural transformations and nucleation processes that require less intuitive and more involved combinations of CVs. Even for seemingly simple particle systems, the competition between

different nucleation pathways may require structure specific CVs that can distinguish between different polymorphs. One example are charged colloidal particles that nucleate in two different crystalline structures with different charge ordering [32]. Another example is nucleation in Ni-Al alloys where, in addition to the size of the solid cluster, the crystallinity and the chemical short range order also need to be included in a suitable RC [33]. In more complex systems, such as the nucleation of methane hydrates, a large number of possible CVs can be proposed that need to be combined in a meaningful way [23, 34]. Similarly, the definition of suitable RCs for nucleation and polymorphic transitions in molecular crystals is non-trivial.

Iterative approaches are also attractive because they promise an automated way in tackling the rare event sampling problem. Ideally, starting with only little information about a system, the exploration of the transition mechanism and the optimization of the RC is performed by the algorithm without much user interference. The nonlinear dimensionality reduction to identify the RC is, for example, performed using autoencoders and then combined with enhanced sampling [35, 36]. Variational autoencoders have also been used, aiming to learn the mapping of molecular simulation data onto the correct probability distribution in the latent space rather than the latent space variables themselves [37]. The RC is then defined as a linear combination of various CVs along which the probability distribution projected from the simulation data best matches the one learnt by the variational autoencoder. In addition, time-lagged autoencoders have been proposed [38] where the data at a time t projected into the latent space are used to predict data at a time $t + \Delta t$. The learnt latent space representation (also called predictive information bottleneck) corresponds to the coordinate that is maximally predictive of a system's future evolution based on its current state [39]. Similar to the variational autoencoders, the latent space can be associated with an RC along which biasing is performed in the enhanced sampling [39].

A related idea of iterative sampling was already introduced in the section on analysis of crystallization mechanisms where the ML model is trained to predict the committor providing a nonlinear combination of possible CVs [21, 22]. The required data are produced with TPS simulations and the reliable prediction of committor values for any arbitrary configuration is, in turn, extremely valuable to accelerate the sampling of trajectories in the TPS algorithm. Again, as more data become available the ML model for committor prediction is further improved and the path sampling becomes more efficient.

Although these iterative approaches are very promising, they are not entirely automated and convergence of the iterative process as well as an ergodic sampling is not guaranteed. For instance, it is discussed in [39] that an insufficient choice of trial CVs to construct the RC might be heuristically spotted by a lack of enhancement in the sampling as the simulation proceeds but there is no proof of completeness. The convergence has, so far, mainly been assessed by an evaluation of the free energy surface. For the simple model potentials and small biomolecules studied in [35, 37, 39] $10 - 20$ iterations were typically performed with $10^6 - 10^7$ steps of biased sampling in between. The number of required iterations as well as sampling steps will, however, depend on the complexity of the studied system. Another challenge is the preprocessing of the simulation data before they can be used as input to the ML model. If the configurations are directly represented, translational, rotational, and permutational symmetries have to be considered and often the representation is reduced to a set of internal coordinates [35, 36]. Alternatively, a possibly large set of trial CVs is computed for each configuration and the RC is determined as either a linear or nonlinear combination of these trial CVs [21, 22, 37, 39]. It will be interesting to explore the performance of these iterative approaches in the sampling of nucleation mechanisms and polymorphic transitions in complex condensed phase systems.

The adjacent possible for computational studies of nucleation

The extensive territory of nucleation and growth is still vastly uncharted by molecular simulations. Larger and more complex systems have just become feasible to probe with growing computational resources and recent years have witnessed

reports of first extensive sampling of nucleation in systems such as pure water and gas hydrates. Still, large free energy barriers associated with the transformation processes prevent the sampling of sufficient transition pathways at reasonable computational cost. The interplay of different interactions especially in multicomponent systems, and the possibility of transient structures governing nucleation makes the identification of the reaction pathways rather difficult. The addition of ML to the arsenal of computational tools has immense potential to help overcome these challenges and expanding the adjacent possible[40, 41] of nucleation studies, thereby ushering the field into a new era of studies.

An immediate application is the analysis of simulation data with ML-supported structure classification and committor predictions. The rapid increase and maturity of ML-based structure identification in condensed phase systems is a testament to the promise of these approaches. As it becomes possible to generate more and more transition pathways, ML-enabled structure identification will help to discover transient structures that potentially drive nucleation. Combined with emerging ML-methods to identify critical parameters that determine the RC, the basic physics of nucleation in these systems will become clearer. While growing computational resources and software developments address some of the challenges in obtaining sufficient simulation data for an ML-based analysis, further developments in the area of ML-enabled enhanced sampling will also be necessary. This might also be complemented by integrating generative ML-models, such as Boltzmann generators [42], that can efficiently propose configurations distributed according to their probability density in the respective ensemble without performing expensive MD simulations. In addition, ML techniques should be explored to address some existing challenges in simulation studies of nucleation such as constant chemical potential simulations, and finite size effects.

The combination of traditional sampling methods with ML-based analysis and sampling tools opens a path towards the study of nucleation in even more exciting and challenging systems such as molecular organic frameworks (MOFs) and zeolites. Another frontier in nucleation studies is the ability to explore the effects of solvent conditions and additives on nucleation. These parameters are routinely used in experiments to control nucleation and polymorph selection. However, very little is understood about the underlying physics from a molecular point of view and, consequently, the choice of these parameters is largely driven by experience. ML-enhanced analysis of nucleation pathways could provide the tools to discover correlations between solution conditions or additives and the emergence of structural features key to the nucleation process. This could also enable high-throughput screening and inverse design of solution conditions for desired nucleation outcomes. To study changes in nucleation mechanisms as solution conditions or additives change, the computational cost to generate the corresponding simulation data still needs to be significantly reduced which requires further progress in ML-supported enhanced sampling. The ultimate goal facilitated by such progress is the inverse design of nucleation processes. This involves the design of additives that not only yield desired polymorph outcomes but also allow us to control competing nucleation pathways. Furthermore, if the metastable structures that emerge along the nucleation pathway can be controlled it becomes possible to engineer exciting responsiveness to solution conditions into the materials. Such control of engineering crystallization kinetics has the potential of unlocking immense novel properties of materials for a vast range of applications.

Acknowledgement

JR acknowledges financial support from the Deutsche Forschungsgemeinschaft (DFG) through the Heisenberg Programme project 428315600. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Award Number DE-SC0015448. S.S. acknowledges support from the National Science Foundation CAREER grant award No. 1653352, and Ruhr University Cluster of Excellence RESOLV for support to travel to and stay in Germany. S.S. acknowledges start-up funds from Department of Chemistry, University of Minnesota.

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest related to the topics discussed in this paper.

References

[1] Gasser, T. M., Thoeny, A. V., Fortes, A. D. & Loerting, T. Structural characterization of ice XIX as the second polymorph related to ice VI. *Nat. Commun.* **12** (1), 1128 (2021) .

[2] Chung, H. & Diao, Y. Polymorphism as an emerging design strategy for high performance organic electronics. *J. Mater. Chem. C* **4** (18), 3915–3933 (2016) .

[3] Noé, F., Tkatchenko, A., Müller, K.-R. & Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **71** (1), 361–390 (2020) .

[4] Steinhardt, P. J., Nelson, D. R. & Ronchetti, M. Bond-orientational order in liquids and glasses. *Phys. Rev. B* **28** (2), 784–805 (1983) .

[5] DeFever, R. S., Targonski, C., Hall, S. W., Smith, M. C. & Sarupria, S. A generalized deep learning approach for local structure identification in molecular simulations. *Chem. Sci.* **10** (32), 7503–7515 (2019) .

[6] Geiger, P. & Dellago, C. Neural networks for local structure detection in polymorphic systems. *J. Chem. Phys.* **139** (16), 164105 (2013) .

[7] Kim, Q., Ko, J.-H., Kim, S. & Jhe, W. GCIceNet: a graph convolutional network for accurate classification of water phases. *Phys. Chem. Chem. Phys.* **22** (45), 26340–26350 (2020) .

[8] Fulford, M., Salvalaglio, M. & Molteni, C. DeepIce: A Deep Neural Network Approach To Identify Ice and Water Molecules. *J. Chem. Inf. Model.* **59** (5), 2141–2149 (2019) .

[9] Qi, C. R., Su, H., Mo, K. & Guibas, L. J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *arXiv:1612.00593 [cs]* (2017). ArXiv: 1612.00593 .

[10] Boattini, E., Dijkstra, M. & Filion, L. Unsupervised learning for local structure detection in colloidal systems. *J. Chem. Phys.* **151** (15), 154901 (2019) .

[11] Adorf, C. S., Moore, T. C., Melle, Y. J. U. & Glotzer, S. C. Analysis of Self-Assembly Pathways with Unsupervised Machine Learning Algorithms. *J. Phys. Chem. B* **124** (1), 69–78 (2020) .

[12] Reinhart, W. F. Unsupervised learning of atomic environments from simple features. *Comput. Mater. Sci.* **196**, 110511 (2021) .

[13] O'Leary, J. *et al.* Deep learning for characterizing the self-assembly of three-dimensional colloidal systems. *Soft Matter* **17** (4), 989–999 (2021) .

[14] Reinhart, W. F. & Panagiotopoulos, A. Z. Automated crystal characterization with a fast neighborhood graph analysis method. *Soft Matter* **14** (29), 6083–6089 (2018) .

[15] Wang, Y., Deng, W., Huang, Z. & Li, S. Descriptor-free unsupervised learning method for local structure identification in particle packings. *J. Chem. Phys.* **156** (15), 154504 (2022) .

[16] McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]* (2020). ArXiv: 1802.03426 .

[17] McInnes, L. & Healy, J. *Accelerated Hierarchical Density Based Clustering*, 33–42 (IEEE, New Orleans, LA, 2017).

[18] Chung, H. W., Freitas, R., Cheon, G. & Reed, E. J. Data-centric framework for crystal structure identification in atomistic simulations using machine learning. *Phys. Rev. Materials* **6**, 043801 (2022) .

[19] Coli, G. M. & Dijkstra, M. An Artificial Neural Network Reveals the Nucleation Mechanism of a Binary Colloidal AB $_{13}$ Crystal. *ACS Nano* **15** (3), 4335–4346 (2021) .

[20] Ma, A. & Dinner, A. R. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B* **109**, 6769 (2005) .

[21] Jung, H., Covino, R. & Hummer, G. Artificial Intelligence Assists Discovery of Reaction Coordinates and Mechanisms from Molecular Dynamics Simulations. *arXiv:1901.04595 [physics.chem-ph]* (2019) .

[22] Jung, H., Covino, R., Arjun, A., Bolhuis, P. G. & Hummer, G. Autonomous artificial intelligence discovers mechanisms of molecular self-organization in virtual experiments. *arXiv:2105.06673 [physics.chem-ph]* (2021) .

[23] Frassek, M., Arjun, A. & Bolhuis, P. G. An extended autoencoder model for reaction coordinate discovery in rare event molecular dynamics datasets. *J. Chem. Phys.* **155** (6), 064103 (2021) .

[24] Sultan, M. M. & Pande, V. S. Automated design of collective variables using supervised machine learning. *J. Chem. Phys.* **149** (9), 094106 (2018) .

[25] Mendels, D., Piccini, G. & Parrinello, M. Collective Variables from Local Fluctuations. *J. Phys. Chem. Lett.* **9** (11), 2776–2781 (2018) .

[26] Karmakar, T., Invernizzi, M., Rizzi, V. & Parrinello, M. Collective variables for the study of crystallisation. *Mol. Phys.* **119** (19-20), e1893848 (2021) .

[27] Invernizzi, M. & Parrinello, M. Rethinking Metadynamics: From Bias Potentials to Probability Distributions. *J. Phys. Chem. Lett.* **11** (7), 2731–2736 (2020) .

[28] Rogal, J., Schneider, E. & Tuckerman, M. E. Neural-Network-Based Path Collective Variables for Enhanced Sampling of Phase Transformations. *Phys. Rev. Lett.* **123** (24), 245701 (2019) .

[29] Rogal, J. & Tuckerman, M. E. in *Chapter 11. Pathways in Classification Space: Machine Learning as a Route to Predicting Kinetics of Structural Transitions in Atomic Crystals* (eds Salahub, D. R. & Wei, D.) *Theoretical and Computational Chemistry Series* 312–348 (Royal Society of Chemistry, Cambridge, 2021).

[30] Tiwary, P. & Berne, B. J. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proc. Natl. Acad. Sci. U.S.A.* **113** (11), 2839–2844 (2016) .

[31] Zou, Z., Tsai, S.-T. & Tiwary, P. Toward Automated Sampling of Polymorph Nucleation and Free Energies with the SGOOP and Metadynamics. *J. Phys. Chem. B* **125** (47), 13049–13056 (2021) .

[32] Peters, B. Competing nucleation pathways in a mixture of oppositely charged colloids: Out-of-equilibrium nucleation revisited. *J. Chem. Phys.* **131** (24), 244103 (2009) .

[33] Liang, Y., Díaz Leines, G., Drautz, R. & Rogal, J. Identification of a multi-dimensional reaction coordinate for crystal nucleation in Ni_3Al. *J. Chem. Phys.* **152** (22), 224504 (2020) .

[34] DeFever, R. S. & Sarupria, S. Nucleation mechanism of clathrate hydrates of water-soluble guest molecules. *J. Chem. Phys.* **147** (20), 204503 (2017) .

[35] Chen, W. & Ferguson, A. L. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *J. Comput. Chem.* **39** (25), 2079–2102 (2018) .

[36] Belkacemi, Z., Gkeka, P., Lelièvre, T. & Stoltz, G. Chasing Collective Variables Using Autoencoders and Biased Trajectories. *J. Chem. Theory Comput.* **18** (1), 59–78 (2022) .

[37] Ribeiro, J. M. L., Bravo, P., Wang, Y. & Tiwary, P. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *J. Chem. Phys.* **149** (7), 072301 (2018) .

[38] Wehmeyer, C. & Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.* **148** (24), 241703 (2018) .

[39] Wang, Y., Ribeiro, J. M. L. & Tiwary, P. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nat. Commun.* **10** (1), 3573 (2019) .

[40] Kauffman, S. & Kauffman, M. *The Origins of Order: Self-organization and Selection in Evolution* The Origins of Order: Self-organization and Selection in Evolution (Oxford University Press, 1993). URL https://books.google.co.in/books?id=lZcSpRJz0dgC.

[41] Johnson, S. *Where Good Ideas Come From: The Natural History of Innovation* (Penguin Books Limited, 2010). URL https://books.google.co.in/books?id=eOfUiUNby3cC.

[42] Noé, F., Olsson, S., Köhler, J. & Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **365** (6457), eaaw1147 (2019) .