

ABSTRACT

Title of Dissertation: **TOWARDS FULLY AUTOMATED
ENHANCED SAMPLING OF NUCLEATION
WITH MACHINE-LEARNING METHODS**

Ziyue Zou
Doctor of Philosophy, 2024

Dissertation Directed by: **Professor Pratyush Tiwary
Department of Chemistry and Biochemistry**

Molecular dynamics (MD) simulation has become a powerful tool to model complex molecular dynamics in physics, materials science, biology, and many other fields of study as it is advantageous in providing temporal and spatial resolutions. However, phenomena of common research interest are often considered rare events, such as nucleation, protein conformational changes, and ligand binding, which occur on timescales far beyond what brute-force all-atom MD simulations can achieve within practical computer time. This makes MD simulation difficult for studying the thermodynamics and kinetics of rare events. Therefore, it is a common practice to employ enhanced sampling techniques to accelerate the sampling of rare events. Many of these methods require performing dimensionality reduction from atomic coordinates to a low-dimensional representation that captures the key information needed to describe such transitions.

To better understand the current challenges in studying crystal nucleation with computer simulations, the goal is to first apply developed dimensionality reduction methods to such sys-

tems. Here, I will present two studies on applying different machine learning (ML) methods to the study of crystal nucleation under different conditions, i.e., in vacuum and in solution. I investigated how such meaningful low-dimensional representations, termed reaction coordinates (RCs), were constructed as linear or non-linear combinations of features. Using these representations along with enhanced sampling methods, I achieved robust state-to-state back-and-forth transitions. In particular, I focused on the case of urea molecules, a small molecule composed of 8 atoms, which can be easily sampled and is commonly used in daily practice as fertilizer in agriculture and as a nitrogen source in organic synthesis. I then analyzed my samples and benchmarked them against other experimental and computational studies.

Given the challenges in studying crystal nucleation using molecular dynamics simulations, I aim to introduce new methods to facilitate research in this field. In the second half of the dissertation, I focused on presenting novel methods to learn low-dimensional representations directly from atomic coordinates without the aid of *a priori* known features, utilizing advanced machine learning techniques. To test my methods, I applied them to several representative model systems, including Lennard Jones 7 clusters, alanine dipeptide, and alanine tetrapeptide. The first system is known for its well-documented dynamics in colloidal rearrangements relevant to materials science studies, while the latter two systems represent problems related to conformational changes in biophysical studies. Beyond model systems, I also applied my methods to more complex physical systems in the field of materials science, specifically iron atoms and glycine molecules. Notably, the enhanced sampling method integrated with my approaches successfully sampled robust state-to-state transitions between allotropes of iron and polymorphs of glycine.

TOWARDS FULLY AUTOMATED ENHANCED SAMPLING OF
NUCLEATION WITH MACHINE-LEARNING METHODS

by

Ziyue Zou

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2024

Advisory Committee:

Professor Pratyush Tiwary, Chair/Advisor
Professor Yifei Mo
Professor Jason D. Kahn
Professor Taylor J. Woehl
Professor Leah G. Dodson

© Copyright by
Ziyue Zou
2024

Acknowledgments

I would like to express my deepest gratitude to everyone who has supported me throughout the journey of completing this dissertation at the University of Maryland. First and foremost, I am grateful to my advisor, Pratyush Tiwary, for his invaluable guidance, continuous support, and unwavering patience. Since my research topics were much different from those other students were working on, I do appreciate that he provided a lot of help, especially with my starting projects. Back in the day as a new PI, I believe it must be a tough decision to make to offer a position to someone with zero computational research knowledge.

I would also like to extend my thanks to the members of my committee, John D. Weeks, Yifei Mo, Jason D. Kahn, Taylor J. Woehl, and Leah G. Dodson for their insightful feedback, constructive criticism, and for generously sharing their time and knowledge.

Special thanks to my colleagues in the Tiwary lab, Yihang Wang, Sun-ting Tsai, Zachary Smith, Luke Evans, and Dedi Wang for creating a collaborative and inspiring environment. As elder graduate students in the lab, they provided tremendous help in my research. I am particularly grateful to Sun-ting for being an amazing mentor and for his guidance in most of my projects at UMD, Eric R Beyerle for our wonderful collaboration, Zack and Dedi for their assistance with various aspects of coding and machine-learning techniques, and for being sounding boards for ideas. Besides the elder members in the group mentioned earlier, I would also like to thank all other group members: Shams Mehdi, Lukas Herron, Akashnathan Aranganathan, Bodhi P Vani,

Suemin Lee, Vanessa J Meraz, Ruiyu Wang, Xinyu Gu, Anjali Verma, Richard John, Venkata Adury, and Mariadelia Argüello-Acuña for being supportive, helpful and consultative during group discussion.

Lastly, I would like to thank my family for their unconditional love and support. Your belief in me has been a driving force throughout this journey, and I could not have done it without you.

Thank you all for being a part of this incredible journey, and thank you UMD for providing me an opportunity to work with these amazing people.

Table of Contents

Acknowledgements	ii
Table of Contents	iv
List of Tables	vii
List of Figures	viii
List of Abbreviations	x
Chapter 1: Introduction	1
1.1 Motivation	1
1.1.1 Nucleation Mechanisms	2
1.1.2 Time-scale Problem	3
1.2 Enhanced Sampling Method: Metadynamics	4
1.2.1 Well-tempered metadynamics	5
1.3 Finding Reaction Coordinate	6
1.3.1 Spectral Gap Optimization of Order Parameters (SGOOP)	6
1.3.2 State Predictive Information Bottleneck (SPIB)	9
1.3.3 Graph and Graph Neural Network	10
1.4 Outline of Dissertation	12
Chapter 2: Towards Automated Sampling of Polymorph Nucleation and Free Energies with SGOOP and Metadynamics	14
2.1 Introduction	14
2.2 Methods	18
2.2.1 Order parameters for nucleation	18
2.2.2 Spectral gap optimization of order parameters (SGOOP)	21
2.2.3 Simulation setup	22
2.3 Results	22
2.3.1 Landmarks for locating polymorphs	22
2.3.2 Reaction coordinate, sampling and converged free energies	23
2.3.3 Comparison with experimental measurements	26
2.4 Conclusion	30
Chapter 3: Driving and characterizing nucleation of urea polymorphs in water	32
3.1 Introduction	32

3.2	Discovering and Biasing Reaction Coordinates for Nucleation	35
3.2.1	Order parameters for nucleation	35
3.2.2	State Predictive Information Bottleneck (SPIB)	41
3.2.3	Estimating free energies with Metadynamics	42
3.3	Results and Discussion	43
3.3.1	Urea Nucleation	43
3.3.2	Urea Polymorphism	44
3.3.3	Comparison with previous studies	47
3.4	Conclusion	48
Chapter 4: Enhanced Sampling of Crystal Nucleation with Graph Representation Learning Variables		53
4.1	Introduction	53
4.2	Methods	56
4.2.1	Graph Neural Networks (GNN) based model	57
4.2.2	Dataset Preparation	60
4.2.3	Well-Tempered Metadynamics	62
4.3	Results and Discussions	63
4.3.1	Phase Transitions in Pure Iron	65
4.3.2	Nucleation of Glycine from Melt	68
4.4	Conclusion	70
Chapter 5: On Graphical Representations and Slow Modes Identifier for Enhanced Sampling Methods		72
5.1	Introduction	72
5.2	Methods	73
5.2.1	GNN-based SPIB	73
5.2.2	Well-Tempered Metadynamics	73
5.3	Results and Discussions	74
5.3.1	Lennard Jones 7	76
5.3.2	Alanine Dipeptide	77
5.3.3	Alanine Tetrapeptide	79
5.4	Conclusion	81
Chapter 6: Conclusion and Outlook		83
Appendix A: Supplementary Materials		87
A.1	Chapter 2: Supporting Material to “Towards Automated Sampling of Polymorph Nucleation and Free Energies with SGOOP and Metadynamics”	87
A.1.1	SGOOP Input and Parameters	87
A.1.2	1-d Reaction Coordinate χ_4	88
A.1.3	1-d Reaction Coordinate χ'_6	88
A.1.4	Further Analyses on χ_6	89
A.2	Chapter 3: Supporting Material to “Driving and characterizing nucleation of urea and glycine polymorphs in water”	90

A.2.1	Sampling Efficiency Comparison	90
A.2.2	Intermolecular Angle Formalization	91
A.2.3	Classifying Urea Polymorphs	92
A.3	Chapter 4: Supporting Material to “Enhanced Sampling of Crystal Nucleation with Graph Representation Learnt Variables”	93
A.3.1	Details on featurization and model training	93
A.3.2	Well-tempered metadynamics simulations setup	94
A.3.3	Classification with trained local decoder in iron nucleation	95
A.3.4	GNN classifier for glycine polymorphs	96
A.3.5	Free energy difference time series of glycine polymorphs	96
A.3.6	Effects of N_c selection on relative stability rankings	97
	Bibliography	108

List of Tables

2.1	Landmarks for stable configurations of urea.	23
3.1	Parameters for WTmetaD	41
4.1	Notations for the machine learning model used in this chapter	58
4.2	WTmetaD Parameters	63
5.1	Well-Tempered Metadynamics Parameters.	74
A.1	Sampling Efficiencies of Expertise-Based (EB) and ML RC in WTmetaD Simulations	91
A.2	Free energy difference between crystalline phases and molten phase of iron on baseline methods	96
A.3	Free energy difference of crystal structures of glycine polymorphs with respect to the initial molten/liquid phase	97

List of Figures

1.1	Schematics to classical and non-classical nucleation mechanisms	3
1.2	A schematic of metadynamics enhanced sampling method	5
1.3	A schematic of spectral gap optimization of order parameters dimensionality reduction method	7
1.4	Schematics of state predictive information bottleneck method	8
1.5	Definition of graph objects and graph neural nets	10
2.1	Definition of characteristic vectors in urea molecules	17
2.2	Reaction coordinates learned by SGOOP for urea in vacuum	24
2.3	The time series of WTmetaD simulation biasing SGOOP optimized RC	27
2.4	Tabulated free energy difference bntween urea polymorphs	28
3.1	Visualization of the feature vectors and corresponding intermolecular angles	39
3.2	Summary to WTmetaD simulations and snapshots of urea polymorph observed	41
3.3	Thermodynamic and RC analyses	51
3.4	Free energy differences among urea polymorphs	52
4.1	Schematic of the pipeline of the graph representation learning architecture	54
4.2	Geometric representations of iron and glycine molecules	60
4.3	Results of WTmetaD simulations of iron nucleation from its melt	64
4.4	Results of WTmetaD simulations of the nucleation of glycine zwitterions	66
5.1	A schematic of training the GNN-SPIB model	74
5.2	Summary of WTmetaD simulation results biasing along machine learnt reaction coordinates	75
5.3	Summary of WTmetaD simulation results in alanine dipeptide system	78
5.4	Summary of WTmetaD simulation results in alanine tetrapeptide system	80
A.1	WTmetaD simulation and thermodynamics analyses to urea nucleation in vacuum	98
A.2	Convergence check to simulation of urea nucleation in vacuum	99
A.3	Details on SGOOP learned RC	99
A.4	Time series biasing individual OPs $\bar{\theta}_1$ and $\bar{\theta}_2$	100
A.5	Spectral gap analyses of SGOOP learned RC	101
A.6	Probability distribution of intermolecular angle OPs	102
A.7	Definitions of urea solution and solid urea in angular space	103

A.8 Examination on local decoder in identifying iron clusters	104
A.9 Comparison between GNN decoder purposed and standard CNA method	104
A.10 $\Delta G_{\ell \rightarrow s}$ as a function of simulation time	105
A.11 Two-dimensional free energy surfaces of cluster size of glycine polymorphs	106
A.12 Robustness check on state definitions	107

List of Abbreviations

CV	Collective variable
OP	Order parameter
RC	Reaction coordinate
ML	Machine learning
NN	Neural network
MLP	Multi-layer perceptron
VAE	Variational autoencoder
AE	Autoencoder
CNT	Classical nucleation theory
DNT	Dynamical nucleation theory
MD	Molecular dynamics
WTmetaD	Well-tempered metadynamics
imetaD	Infrequent metadynamics
GNN	Graph neural network
EGNN	Equivariant graph neural network
GAT	Graph attention net
GMM	Gaussian mixture model
GBF	Gaussian basis function
GLY	Glycine
AD	Alanine dipeptide
Ala ₁	Alanine dipeptide
Ala ₃	Alanine tetrapeptide
LJ7	Lennard Jones 7
RAVE	Reweighted autoencoded variational Bayes for enhanced sampling
SPIB	Information bottleneck
MFPT	Mean first passage time
MaxCal	Maximum Caliber
MSM	Markov state model

Chapter 1: Introduction

1.1 Motivation

Crystal nucleation is known to be the first step of crystallization, a process ubiquitous across biological and chemical systems. [1, 2] However, studies of this process in a variety of species, whether in experiments or *in silico*, remain difficult due to various factors. While experiments provide a macroscopic view of the nucleation mechanism, the onset of the nucleation event may reach the limit of the spatial and/or temporal resolution of the instruments. [3] On the other hand, computational simulations have become powerful tools over the past decades, free from these limitations, as the time step for integrating Newton's equations of motion is set small enough to capture most important atomic motions like hydrogen atoms'. Standard MD simulations have been applied to the study of many complex dynamics such as self-assembly, protein conformational changes, ligand binding, and more. [4–13] These problems are not only of great importance to research studies, but they are also closely related to industry, particularly in the development of active pharmaceutical ingredients (APIs), [14] the discovery of novel materials, [15] and other fields. Moreover, MD simulations allow researchers to track individual particles in many-body systems, offering a unique perspective that is difficult to achieve with experimental instruments. Another advantage of MD simulations is the ability to prepare systems under extreme conditions. For example, homogeneous nucleation can be hard to study experimentally because it requires a

pure solution without any impurities, a condition that is simple to achieve using computer simulations.

1.1.1 Nucleation Mechanisms

For many years, the picture painted by classical nucleation theory (CNT) [16–19] was seen as an accurate method for describing nucleation processes and, indeed, the robustness of this method can be seen in its continued use for describing both the thermodynamics and kinetics of many nucleation processes. [20–22] However, it has also become clear that not all nucleation events follow a purely classical pathway where the necessary and sufficient reaction coordinate (RC) is the size of the spherical crystal nucleus. [14, 18, 23–33] In Fig. 1.1, a schematic summarizes both of the nucleation schemes under the description of classical nucleation theory (subplot a)) where the crystallization process depends solely on the size of cluster size. The competition between cluster surface tension and free energy preference in the formation of a more stable solid phase leads to an energy barrier when the cluster reaches its critical size. In Fig. 1.1b), my observations on non-classical nucleation behavior from computer simulations of urea which I will discuss in detail in the upcoming chapters. Furthermore, even if N , the size of the nascent crystal nucleus analogous to r the radius of nucleus, can be shown to be a sufficiently good RC, CNT also stipulates that there is a single energy barrier along the RC corresponding to the barrier required to form the surface of the spherical critical nucleus of size N^* . However, it is well-known that for many nucleation processes the Ostwald step mechanism, [34] whereby the most kinetically accessible crystalline structure forms first, followed by the more thermodynamically stable one, is the preferred nucleation pathway. This mechanism that requires the presence of at least two

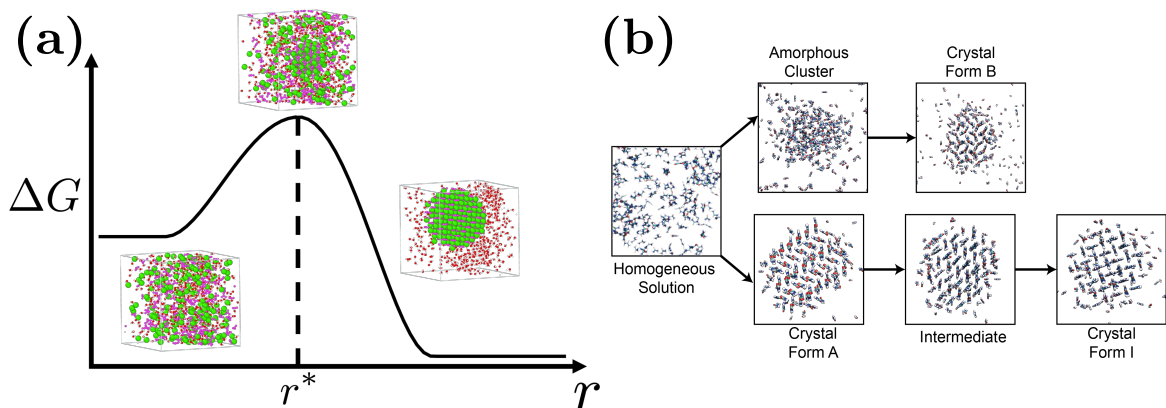


Figure 1.1: Schematics of mechanisms described by a) classical nucleation theory where the nucleation process is described by the size of the cluster (snapshots are rendered from seeded simulation of calcium carbonate), and b) non-classical nucleation behavior simulated and observed in the homogeneous nucleation of urea. Snapshots are generated from Ovito [35] and VMD [36].

energetic barriers, one to go from the liquid state to the kinetically stabilized crystal structure and a second to go from the kinetically stabilized crystal structure to the thermodynamically stable crystal structure. In this dissertation, I present several attempts to probe the nucleation mechanism of simple organic molecules with enhanced sampling methods, which will be introduced in later sections.

1.1.2 Time-scale Problem

At the microscopic scale studied in computer simulations, nucleation is a rare event occurring over typical simulation timescales. Even with accurate force fields, whether built using ML methods or not, nucleation events are difficult to sample computationally without using an enhanced sampling method or applying a biasing force to the nucleation RCs. [33, 37–40] In the standard MD scheme, the time step of the integration equation of motion is typically 5 fs at most, to capture the fast fluctuations of hydrogens. However, rare events of interest, such as crystal

nucleation and ligand unbinding, may occur on timescales ranging from milliseconds to minutes, varying by system and condition. This means that sampling one such event can take up to years of computer time on a CPU, not to mention measuring thermodynamic or kinetic properties.

1.2 Enhanced Sampling Method: Metadynamics

To overcome the aforementioned timescale limitation in standard brute-force molecular dynamics simulations, enhanced sampling methods are required to speed up the occurrence of rare events. Here in this dissertation, I focused on one of these methods, called metadynamics. [41] In metadynamics, Gaussian biases are frequently deposited along the defined reaction coordinates, allowing the system to escape from where it is initially trapped (see Fig. 1.2). The instantaneous Gaussian bias, $V(s(R), t)$ along n -dimensional collective variables, $s(R)$ of Cartesian coordinates space of all atoms R , at time, t , can be expressed as,

$$V(s, t) = \omega \exp \left(- \sum_i^n \frac{[s_i(R) - s_i(R, t)]^2}{2\sigma_i^2} \right) \quad (1.1)$$

where ω is the height of the Gaussian potential, and σ_i is the width of the bias. Although with metadynamics systems may easily escape from their initial state, the thermodynamic measurements often oscillate instead of converging to a well-defined value given that Gaussian potentials are being deposited even when the energy wells are filled. Therefore, in this dissertation, I applied the well-tempered variant of metadynamics to systems of interest, which leads to better convergence.

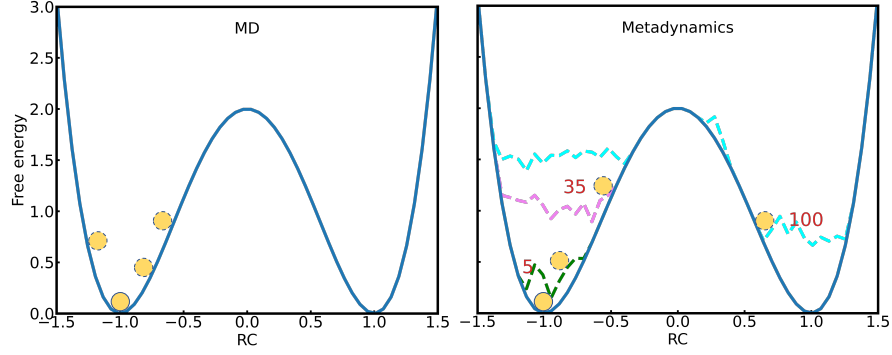


Figure 1.2: To studies of rare events, system may commonly be trapped in its initial states (left). With metadynamics, Gaussian potentials are frequently deposited as suggested by the colored lines with numbers indicating numbers of bias addition and therefore the system can escape from the initial state (right).

1.2.1 Well-tempered metadynamics

In WTmetaD, a variant of the metadynamics method, history-dependent Gaussians are deposited along pre-defined biasing variables to facilitate state-to-state back-and-forth transitions between different metastable states where the system would normally be trapped. [40] In particular, the bias is rescaled as a factor of pre-defined bias factor as system revisiting the same location in RC space, with this, the height of Gaussian potential in this variant, $\omega(s(R), t)$, becomes position dependent and can be expressed as,

$$\omega(s(R), t) = \omega_0 \exp - \frac{V_g(s(R), t)}{k_B \Delta T} \quad (1.2)$$

where ω_0 is the initial bias height, k_B is the Boltzmann constant, V_g is the total biases already deposited prior to time frame t , and ΔT is a parameter homogeneous to temperature and it is often written as the bias factor $\gamma = (T + \Delta T)/T$. With this formalism, WTmetaD can converge

and provide an estimate of the free energy surface, $F(s)$:

$$F(s) = -\frac{T + \Delta T}{\Delta T} V_g(s) + c(t) \quad (1.3)$$

where T is the temperature of the system and $c(t)$ is a correction term (see Ref. [42] for details on $c(t)$). With the WTmetaD method, one can, in principle, sample rare events, but as introduced above, the Gaussian biases need to be deposited onto an n -dimensional CV space. In the next section, I introduce how I define these variables for biasing.

1.3 Finding Reaction Coordinate

One main focus of this dissertation is on finding (sub)optimal reaction coordinates in low dimensions, which allow enhanced sampling methods to sample larger configuration spaces while preserving the mechanism and physics. With metadynamics, the biasing CV space is normally restricted to less than or equal to 3 dimensions for computational efficiency. Many dimensionality reduction methods have been introduced and are being developed. Here I introduce three methods that were applied to systems discussed in later chapters. Throughout this thesis, I used the terms collective variables (CVs) and order parameters (OPs) interchangeably to refer to approximations of the reaction coordinate, and these variables (or parameters) are defined by complex functions of atomic coordinates.

1.3.1 Spectral Gap Optimization of Order Parameters (SGOOP)

SGOOP utilizes the principle of Maximum caliber (MaxCal) (see Ref. [43–45] for details of MaxCal) to construct estimates of transition probability matrices along several trial RCs

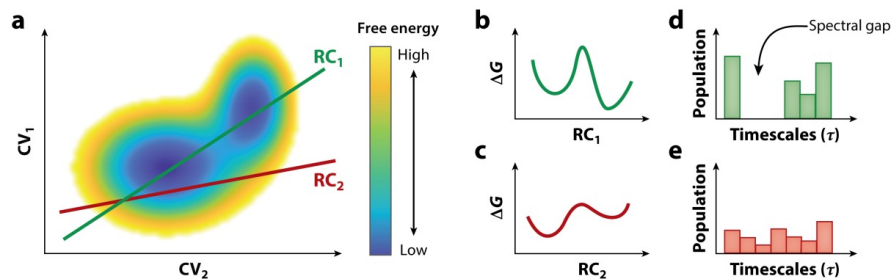


Figure 1.3: An example complex free energy surface in CVs space with two possible reaction coordinates RC_1 and RC_2 in a). The projections of free energy surface on individual RC space is shown in b) and c). As indicated in d), RC_1 as good CV leads to a larger spectral gap and herein larger time scale separation between fast and slow modes, while e) suggests RC_2 does not have similar properties. Figure reprinted from Ref. [43] with permission. Copyright 2020 Annual Reviews.

(see schematics of SGOOP methods in Fig. 1.3). This is done by combining stationary density information from biased simulations along trial RCs and dynamical information from short unbiased simulations. Ref. [46] illustrated through numerical examples how the RC learnt from SGOOP provides a very good approximation to those from more accurate but expensive methods. I now summarize the key concepts behind SGOOP. Let π_m denote the stationary distribution along a spatially discretized trial RC with m denoting the grid index. This can be obtained from reweighting a biased simulation performed along any preliminary RC approximation [47]. In addition to π_m , I also introduce $\langle N \rangle$ as the average number of nearest-neighbor transitions along any spatially discretized trial RC per unit time, which provides dynamical constraints to be used in the MaxCal framework. $\langle N \rangle$ can be estimated by directly counting from short unbiased MD trajectories. With these two pieces of stationary and dynamical information, MaxCal provides a tractable approximation for the transition matrix K along any trial RC:

$$K_{mn} = -\frac{\langle N \rangle}{\sum \sqrt{\pi_m \pi_n}} \sqrt{\frac{\pi_n}{\pi_m}} \quad (1.4)$$

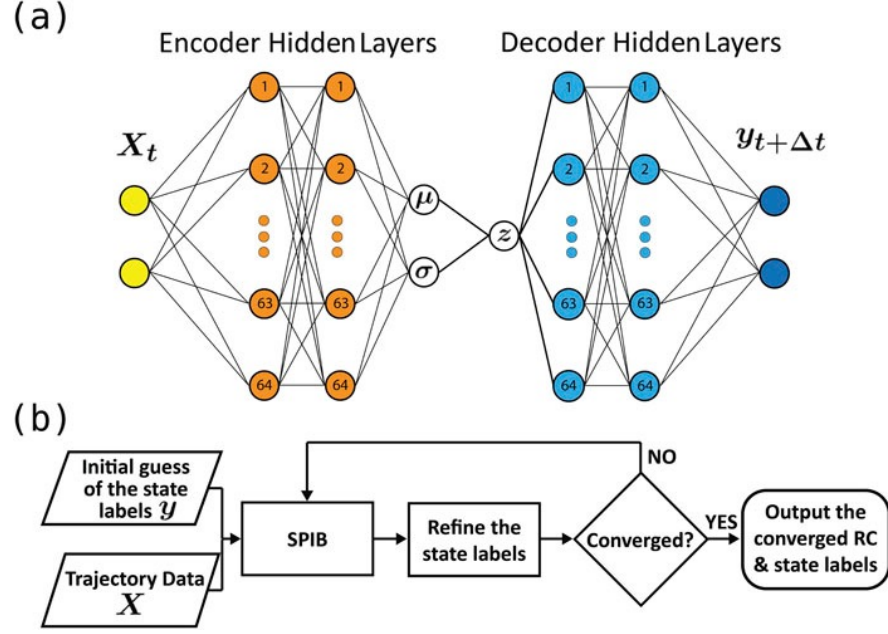


Figure 1.4: a) The architecture of the State Predictive Information Bottleneck (SPIB) method. The input features at time t , X_t are first compressed by passing through encoder MLPs to yield probability distribution parameters. Samples drawn from the distribution are then decoded by the decoder for constructing the state label in future time $t + \Delta t$, $y_{t+\Delta t}$. b) The flowchart of the overall optimization scheme is shown and SPIB allows for iteratively improving its latent variables with better sampled input data. Figure reprinted from Ref. [48] with permission. Copyright 2021 AIP.

For any putative RC, one can construct such a transition probability matrix and diagonalize it to get eigenvalues $\lambda_0 = 1 > \lambda_1 > \lambda_2 > \dots > 0$. These eigenvalues directly provide information about the timescales of different transition processes when projected on the corresponding RC along which the matrix was built. I can then find an optimal RC χ by selecting from the trial RCs the one with maximal spectral gap which is defined as the maximal difference between the visible slow and hidden fast processes. Such differences can be easily quantified as $\lambda_n - \lambda_{n+1}$ where n is the number of visible slow processes that I really care about.

1.3.2 State Predictive Information Bottleneck (SPIB)

The state predictive information bottleneck (SPIB), which takes the form of a variational autoencoder (VAE) [49, 50], has been applied to various systems. [48, 51–54] In standard autoencoder (AE) architecture, the model is composed of an encoder where the input features are compressed (by convolutional operations) and a decoder where the latent variables are expanded to high-dimensional representations (also by convolutions). Alternatively, VAE incorporates variational inference which allows for the sampling of new data from the latent space, in particular, this is achieved by mapping input features into parameters of a probability distribution to the latent variable space. SPIB differs from the traditional VAE because instead of attempting to maximize the model’s ability to re-construct the input from the low-dimensional latent space, it instead maximizes the quality of reproducing the population of the system’s metastable states after a given lagtime τ in the future. (see Fig. 1.4 for its architecture) That is, the SPIB finds the latent space, i.e., the output of the encoder, that best reproduces the metastable state populations at time $t + \tau$ given the values of the input OPs at time t [48]. This approach allows SPIB to simultaneously perform dimensionality reduction and accurate future state prediction by minimizing a loss function that is inspired by the variational information bottleneck formalism [48, 55, 56].

Conceptually, this approach is similar to performing a Markov state model and clustering analysis [57, 58] with a continuous basis set. In SPIB the number and neighborhood of each metastable state is adjusted on-the-fly during the training of the model. The final learned model is output once the neighborhoods spanned by each metastable state have converged to below a pre-defined threshold, a tunable hyperparameter of the model. As a variant to the Reweighted autoencoded variational Bayes for enhanced sampling (RAVE) family [59], SPIB RC can be

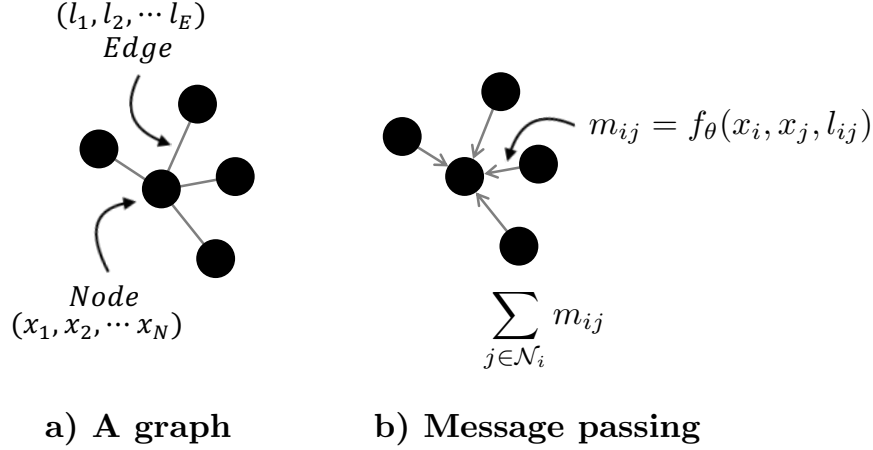


Figure 1.5: a) shows a simple graph data defined by node feature, X_n , connectivity, $a_{ij} \in \{0, 1\}$ and edge feature, L_e . To this dissertation, all graphs are computed with the atomic coordinates from particles in simulation cell. b) presents the exchange of information is standard graph neural nets via message passing. The message of node i is a learnable function of the features from itself and from its neighboring nodes j in $\mathcal{N}_i$.

improved iteratively with better-sampled data as the workflow shown in Fig. 1.4 b).

1.3.3 Graph and Graph Neural Network

Different from the previous two methods, graph neural networks provide detailed descriptions of individual nodes and their neighborhoods with message passing functions on graph data. Here, I provide a brief introduction to graph objects: a graph $G = (V, E)$ has two primary components: vertices (or nodes) V and edges E . The node features $\{X_n\}$ attributed to individual nodes relate to particles' properties such as atom type in my studies. While edge features $\{L_e\}$ are generally defined by interatomic distances, but higher-order representations like angles and spherical harmonics may boost the performance of the graph model. The neighborhood is defined by an $n \times n$ adjacency matrix where each entry $a_{ij} \in \{0, 1\}$ indicates whether node i is connected to node j or not (see Fig. 1.5 a)). The canonical property of graph data allows the nodes in graph

data to shuffle randomly, and this will result in rearrangement of the adjacency matrix, but does not necessarily affect the graph data itself. Taking this into consideration, graph neural networks generally incorporate permutation symmetry. In later chapters, I will show that this feature can be useful for problems where not only permutation symmetry, but also rotational and translational symmetries are important, such as crystal nucleation.

In graph neural nets (GNNs), messages are exchanged only between neighboring nodes which is defined by the adjacency matrix (i.e., edges) and this subsequently affect the nodes' embeddings. However, with multiple message-passing layers and update functions, information about the nodes and edges from hops away (outside the neighborhood defined by the adjacency matrix) can be exchanged. [60] Here I break down how embeddings of a node are updated and this involves functions which are called message passing functions and update functions. The messages is first computed by a learnable function, f_θ , between the node of interest i and all nodes in its neighborhood $\mathcal{N}_i$ (see Fig. 1.5 b)). These messages are then aggregated by permutation-invariant operator, which typical are sum, mean, and max operations. Finally, the embeddings of particular nodes are updated incorporating both messages from previous steps and the representation of the node itself through a learnable update function. Depending on the learning scheme and the task of the model, a readout function following the last message-passing layer maps the embeddings to graph-level, node-level, or edge-level predictions. In my case, I focus on graph-level predictions to provide a macroscopic view of the studied system and, to some extent, they are equivalent to order parameters one would generally consider in many enhanced sampling methods.

1.4 Outline of Dissertation

Here I briefly summarize the following chapters:

- In Chapter 2, I applied the SGOOP method to study urea nucleation in vacuum. A library of order parameters (OPs) for the study of nucleation was introduced to SGOOP, resulting in a linear combination of these variables after fully accounting for the static and dynamic information. All sampled metastable states were compared with existing MD studies, and thermodynamic measurements in terms of free energy differences with respect to the initial liquid phase were performed. I then benchmarked my results against experimental observations, and the computed transition pressure between the ambient and high-pressure products was in agreement with each other.
- In Chapter 3, I applied the SPIB method to study the case of homogeneous nucleation of urea in water. Here, I found completely different nucleating behaviors compared to those observed in vacuum, and none of them could be described or explained by classical nucleation theory. Moreover, not just one, but all polymorphs observed in previous studies of urea in vacuum were sampled and characterized. I also provided explanations from the perspectives of the learned reaction coordinates and attempted to measure the thermodynamics of those mechanisms.
- In Chapter 4, I introduced a GNN-based dimensionality reduction method that allows one to learn low-dimensional representations directly from atomic coordinates without requiring prior knowledge about the input features. The input graphs are constructed with only pairwise distances as edge features. Latent representations were trained to distinguish struc-

tures in the training data at both graph-level and node-level. This protocol was then examined in two systems: allotropes of iron particles and polymorphs of glycine molecules. With the learned graph-based RC, state-to-state transitions were sampled between all crystal structures synthesized in the crystallography database. These configurations were then ranked thermodynamically, and the ranking aligned with those measured in experiments.

- In Chapter 5, I combined GNN with the SPIB method. Separately, GNN provides atomistic information to the system of interest, while SPIB learns the dynamics of the underlying transitions. This resulted in a complete picture of learning meaningful representations in dimensionality reduction methods where both configurational and dynamical information is taken into account. I then applied this pipeline to three model systems: Lennard Jones 7 clusters, alanine dipeptide, and alanine tetrapeptide. With only type-0 representations (i.e., distances), I obtained comparable results to those with higher-order features like angles and spherical harmonics. The comparisons were again drawn from the thermodynamic properties with WTmetaD simulations using typical good CVs.

Chapter 2: Towards Automated Sampling of Polymorph Nucleation and Free Energies with SGOOP and Metadynamics

2.1 Introduction

Understanding how different polymorphs can form has practical importance in material science [1, 61] and the pharmaceutical industry. [62–64] However, experimental measurements suffer resolution problems, as nucleation events involve the sub-nanometer lengthscale and sub-picosecond timescale. [65] In principle, computational methods such as Molecular Dynamics (MD) simulations can provide detailed information about nucleation process since by construction they follow all-atom and femtosecond resolution. Unfortunately, this femtosecond temporal resolution limits the time scale that MD simulation can reach to around milliseconds, preventing it from sampling nucleation events which typically occur at seconds and much slower timescales. [3] Moreover, liquid-solid nucleation is further complicated by the presence of multiple competing crystal structures or complicated polymorphism, [66] which is shown to be important in the pharmaceutical industry, material science and elsewhere. [1, 61–64] Understanding the nucleation of these polymorphs is therefore essential since different polymorphs lead to different physical and chemical properties. [3]

In order to overcome the timescale limitation faced by MD simulations in the study of nu-

cleation and more generally, several enhanced sampling methods have been developed. [41, 67–71] A large fraction of these methods needs a prior estimate of a biasing or progress coordinate that can differentiate relevant metastable states from each other, and can also quantify the progress of the slow process being studied. In principle, many approximate choices of this coordinate are sufficient for reliably enhanced sampling. In practice however, the enhancement in sampling is tremendously benefited if the biasing or progress coordinate approximates the true reaction coordinate (RC) for the process being studied. For the study of nucleation, the first such coordinate one can think of comes from classical nucleation theory (CNT) [72] which posits the size of a nucleus assumed spherical as a progress coordinate. CNT can however give wrong nucleation rates even for liquid Argon that can be off by almost twenty orders of magnitude. [29] This has led to a large effort over the decades aimed at developing better variables for quantifying the onset of nucleation. [73–78] At the same time, a large number of techniques have been developed for the more general problem of learning approximate reaction coordinates for rare event systems given limited sampling. [79–87] The problem of constructing good biasing or progress coordinates is especially accentuated when it comes to studying polymorphism because such crystal structures all fall into a narrow range of energies [71] and very likely have similar orientations.

In this chapter, we demonstrate how one such method “Spectral gap optimization of order parameters (SGOOP)” [81] can be used in conjunction with well-tempered metadynamics [88] for sampling different polymorphs and accurately capturing their relative free energies. While the methods are described in greater detail in Sec. 2.2.2, here we give an overview. Metadynamics involves constructing a bias potential by depositing history-dependent Gaussian kernels along the chosen biasing coordinate, which is typically low-dimensional, to encourage state-to-state transitions. SGOOP helps construct such a biasing coordinate as a linear combination of several

order parameters (OPs). The central idea behind SGOOP is that the optimal RC will maximize the spectral gap of the dynamics projected along it, and the spectral gap itself can be learnt efficiently from biased metadynamics runs along trial coordinates, through a maximum path entropy or caliber framework. [43, 81]

In order to construct the RC as a combination thereof, we have many possible OPs at our disposal. Broadly speaking these fall into two classes. The first type of OPs require pre-knowledge of the target crystal structure, and are based upon comparing the instantaneous configuration of a system with a specific reference structure. These include the Steinhardt bond order parameters [89] that compare systems with different spherical harmonics. Another example is the SMAC order parameter introduced by Giberti et al. [40] which describes the systems by selected angles between feature vectors. The second type of OPs are more generic and involve observables such as mean distance, mean dihedrals, and thermodynamics variables such as entropy and enthalpy. The first class of OPs is not very well suited for the study of polymorphism because of the loss in generality when specifying orientations. Thus in this chapter, we mainly focus on the second type and the detailed definitions are given in Sec. 2.2.1. An optimized RC is constructed as a function of those order parameters through SGOOP.

We illustrate the framework by studying urea polymorph nucleation from the melt, which is a small molecule commonly used as fertilizer and active pharmaceutical ingredient (API). The urea system has been widely studied both experimentally [90–97] and computationally. [40, 98, 99] For consistency, we used the same notations of polymorphs of urea as in Ref. [98]. In particular, form I and IV have been synthesized in lab and studied in detailed, whereas form A and B have been found only in simulations so far. Ref. [99] suggests that order parameters such as orientational entropy play an important role in urea nucleation and in exploring the configura-

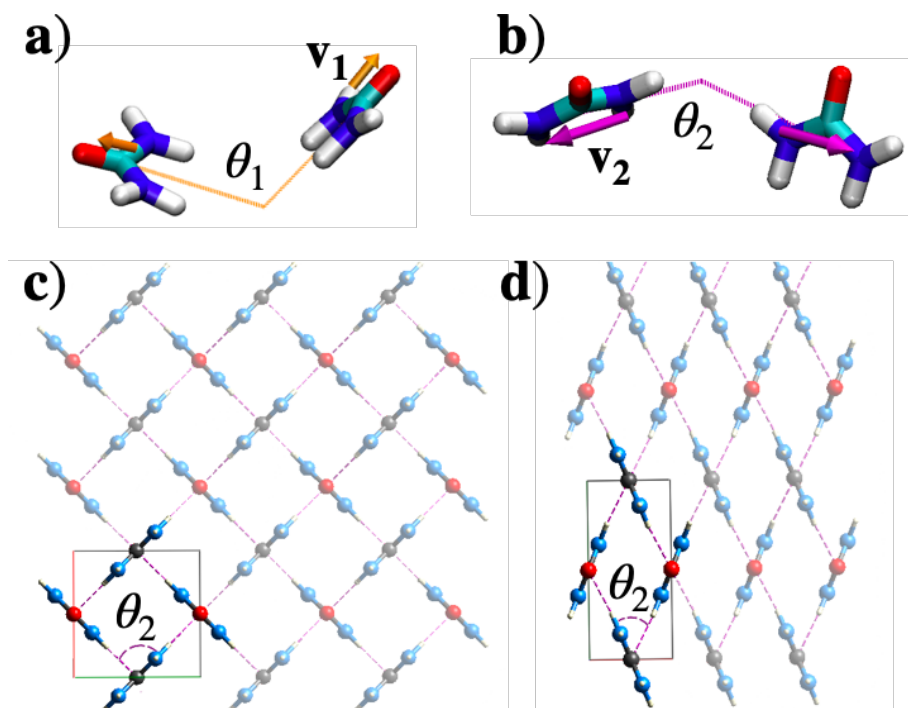


Figure 2.1: Definition of characteristic vectors, $\mathbf{v}_1$ and $\mathbf{v}_2$ in the upper panel (a and b, respectively). Configurations of forms I ($P\bar{4}2_1m$) and IV ($P2_12_12$) with $\bar{\theta}_2$ angle along [001] in the lower panel (c and d). $\mathbf{v}_1$ is defined as the dipole moment of urea molecules and $\mathbf{v}_2$ connects the two nitrogen atoms. Direct measure of $\bar{\theta}_2$ order parameter allows us to distinguish these two forms. Crystallographic data of forms I and IV were obtained from Cambridge Structural Database (CCDC 731958, 731961). [93] Structures were generated with Visual Molecular Dynamics (VMD) [36] and CrystalExplorer 17.5. [100]

tion space relevant to the processes at work. We apply SGOOP to build a 1-d RC as a function of these entropy variables as well as other OPs that we describe in Sec.2.2.1. By performing well-tempered metadynamics along this optimized 1-d RC, we obtain correct ranking of thermodynamic stabilities of different polymorphs. The use of such an optimized 1-d RC for enhanced sampling lowers the computational cost relative to non-optimized or higher-dimensional biasing variables. Finally, by extrapolating on the basis of the free energy difference obtained from our simulations, we obtain good agreement with experiments regarding the pressure at which one of the metastable polymorph becomes the more stable form.

2.2 Methods

2.2.1 Order parameters for nucleation

Many order parameters have been developed to study the nucleation process. [22, 74, 75, 77, 101–104] Here we consider four types of order parameters. which are then used to construct an optimized 1-d RC as a function thereof. These are (i) pair orientational entropy, [105] (ii) enthalpy, (iii) averaged intermolecular angles, and (iv) mean coordination number. We now describe them systematically:

- Pair orientational entropy: Piaggi et al. [99] recently introduced an approximation to the pair entropy as a biasing variable for metadynamics. Referring to liquid state theory, [106] the excess entropy with respect to the entropy of ideal gas can be expressed in an infinite sum of correlation functions, and has been shown as a successful variable for analysis by considering only the first term. [69, 107–109] The excess pair orientational entropy is then defined as:

$$S_{\theta} = -\pi\rho k_B \int_0^{\infty} \int_0^{\pi} [g(r, \theta) \ln g(r, \theta) - g(r, \theta) + 1] r^2 \sin \theta dr d\theta \quad (2.1)$$

where ρ is the density of the system, k_B is Boltzmann constant, r is the intermolecular distance, θ is the angle formed between characteristic vectors, and $g(r, \theta)$ is a mollified correlation function to ensure its differentiability for enhanced sampling methods. Given the polarity of urea molecules, two characteristic vectors are introduced (Fig. 2.1 upper

panel.). The first vector $\mathbf{v}_1$ is along the direction of the carbonyl group, which is also the direction of its dipole moment. The second vector $\mathbf{v}_2$ connects the two nitrogen atoms. Angles that are formed between a pair of two $\mathbf{v}_1$ vectors on two adjacent urea molecules are denoted θ_1 while those between $\mathbf{v}_2$ are denoted as θ_2 . Using these two angles in Eq. 2.1, we define two representations of entropy denoting them S_{θ_1} and S_{θ_2} .

- Enthalpy: To complete the picture of thermodynamics related variables, enthalpy was introduced as an order parameter in Ref. [105]. Unlike entropy which requires approximations to be calculated, enthalpy can be computed directly from its definition:

$$H = E + PV \quad (2.2)$$

where E is the instantaneous total potential energy, V is the volume and P is the pressure. Given the simulation is performed at isothermal-isobaric ensemble, pressure is a constant.

- Mean coordination number: We consider the mean coordination number for urea molecules in light of several studies on gas/liquid nucleation [29, 110, 111] and we calculate this as [110]

$$N = \frac{1}{n} \sum_{ij} \frac{1 - (r_{ij}/r_c)^6}{1 - (r_{ij}/r_c)^{12}} \quad (2.3)$$

where n is the total number of molecules, r_c is a cutoff and j denotes molecules besides molecule i . For molecules with distance r_{ij} within this threshold distance, they are identified as neighbors. Similarly, we defined the cutoff distance as the radius of first shell, 0.64 nm. While we do not expect that N will be able to distinguish different polymorphs from

each other, it is well-documented to be useful for differentiating liquid from solid phases and thus we consider it here as one of the possible constituents of the reaction coordinate.

- Averaged intermolecular angles: To further distinguish polymorphs from each other, here we introduce another order parameter that is based on the averaged values of the angles θ_1 and θ_2 introduced earlier (Fig. 2.1 upper panel). For either of the two angles, this variable is defined as follows:

$$\bar{\theta} = \frac{\sum_{ij} \sigma(r_{ij}) f(\theta_{ij})}{\sum_{ij} \sigma(r_{ij})} \quad (2.4)$$

where a step function of angle θ is defined in a form of hyperbolic tangent that ensures continuous differentiability

$$f(\theta) = \frac{1}{2}[(\pi - 2\theta) \tanh(5\theta - 9.25) + \pi] \quad (2.5)$$

and $\sigma(r_{ij})$ is a smooth switching function of intermolecular distance r_{ij} . This converts θ to its supplementary angle when such angle is larger than a cutoff θ_{ij} . Considering the C_{2v} symmetry of urea molecule, permutation of $\mathbf{v}_2$ is unavoidable by directly calculating average angles. One alternative is to define a step function to introduce rotational invariance as achieved through the function f . This order parameter is developed for the purpose of differentiating form I and form IV along the direction of $\mathbf{v}_2$, where form IV ($P2_12_12$) is known to be a high-pressure product of urea in experiment and form I ($\bar{P}42_1m$) is the most stable structure found at ambient conditions. [92] The main difference of these two polymorphs are illustrated by θ_2 , as it equals 50° for form IV and 90° for form I (shown in

Fig. 2.1 lower panel). In this chapter, the threshold of the step function of measured angles is set to be 100° , which shows better performance than a cutoff of 90° in distinguishing form I from its collapsed version, and a cutoff of distance is set to be $0.64nm$ corresponding to the range of first neighbor shell.

2.2.2 Spectral gap optimization of order parameters (SGOOP)

While the different order parameters (OPs) described in Sec. 2.2.1 can collectively differentiate polymorphs from each other, it is not obvious how to use these OPs to perform enhanced sampling of polymorph nucleation. Considering all 6 OPs collectively as biasing or progress coordinates in a protocol such as metadynamics [41], umbrella sampling [67] or forward flux sampling [68, 69, 74, 112] is possible in principle, but not in practice. [113] Thus, our next objective is to compress these OPs into a lower-dimensional approximate reaction coordinate (RC) which best encapsulates the relevant slow degrees of freedom. In this chapter, we use the method “Spectral Gap Optimization of order parameters (SGOOP)” [81] to find such an optimal 1-d RC as a linear combination of the OPs we defined in Sec. 2.2.1 (see Ch. 1.3.1). Note that here we restrict ourselves to a 1-d RC as a trade-off between achieving accurate description of the slow processes while staying computationally efficient. For more complicated systems, recent developments using SGOOP allow us to systematically quantify if more dimensions need to be added to the RC. [114]

2.2.3 Simulation setup

The system consists of 108 urea molecules described using generalized Amber force field (GAFF). [115] The partial charges of urea were adopted from Amber03 database [116]. Biased and unbiased MD simulations were performed with a modified version of GROMACS-2020.2 [117] patched with PLUMED 2.6.1 [101, 118] in constant number, pressure and temperature (NPT) ensemble with an integration time step of 2 fs. The system was coupled with a thermostat using stochastic velocity rescaling method [119] at 450 K with 0.1 ps relaxation time. The pressure was controlled using the Parrinello-Rahman barostat [120] at 1 bar with 10 ps relaxation time. The long-range electrostatic interaction in reciprocal space was calculated with Particle-mesh Ewald method. [121] The cutoffs for both electrostatic and Van der Waals interactions in real space were set to 0.9 nm. The hydrogen bonds were constrained with LINCS algorithm. [122] We used well-tempered metadynamics [42, 88] to enhance the sampling, where Gaussian kernels with width and height equal to 0.2 RC units and 7.48 kJ/mol ($\approx 2k_B T$) were deposited every 1 ps. The bias factor for the tempering was 50 and the production runs were performed for 400 ns.

2.3 Results

2.3.1 Landmarks for locating polymorphs

In order to locate existing polymorphs during our enhanced sampling simulations, we first establish their corresponding landmarks in the order parameter space. The landmark for each polymorph can be found by the average of each of the 6 order parameters obtained in 2 ns long

unbiased simulation initiated from their individual configurations. These landmark values in CV space are then used as the fingerprints for tracing the formations of different polymorphs. Our primary interest in this chapter is in the polymorphs I and IV which have been synthesized in experiments. These forms can be identified most cleanly simply using the two averaged angles $(\bar{\theta}_1, \bar{\theta}_2)$ introduced in Sec. 2.2.1. We have also shown the landmark values for forms A and B in Table 2.1 even though they have not been seen in experiments yet, but were reported in past computational work. [99] Forms A and B can be easily identified by using $(S_{\theta_1}, S_{\theta_2})$.

2.3.2 Reaction coordinate, sampling and converged free energies

By maximizing the spectral gap value using SGOOP (Sec. 2.2.2) using a basin-hopping algorithm [123], we find the optimized RC as $\chi_6 = 0.885S_{\theta_1} + 0.330S_{\theta_2} + 0.906\bar{\theta}_1 + 1.0\bar{\theta}_2 - 0.011N - 0.017H$. See Appendix for further analysis of SGOOP results. By directly looking at the coefficients of the different OPs in this RC (Fig. 2.2 (a)), it can be seen that all 4 OPs except the mean coordination number N and the enthalpy H contribute to the RC. Yet another way to compare the contributions of different OPs is after accounting for their different scales, which can be done by dividing out their respective standard deviations from the coefficients in

Table 2.1: Landmarks for stable configurations of urea.

Phase	S_{θ_1}	S_{θ_2}	$\bar{\theta}_1$	$\bar{\theta}_2$	N	H
I	-6.11	-2.12	0.257	1.22	11.609	-766.93
IV	-4.95	-3.57	0.248	0.428	11.516	-764.61
A*	-4.20	-5.79	0.389	0.411	12.288	-763.25
B*	-1.76	-3.39	1.272	0.666	11.543	-762.99
Liquid	-1.13	-1.14	1.040	0.936	11.650	-760.12

*Polymorphs not seen when biasing χ_6 or in experiments but were seen in Ref. [99].

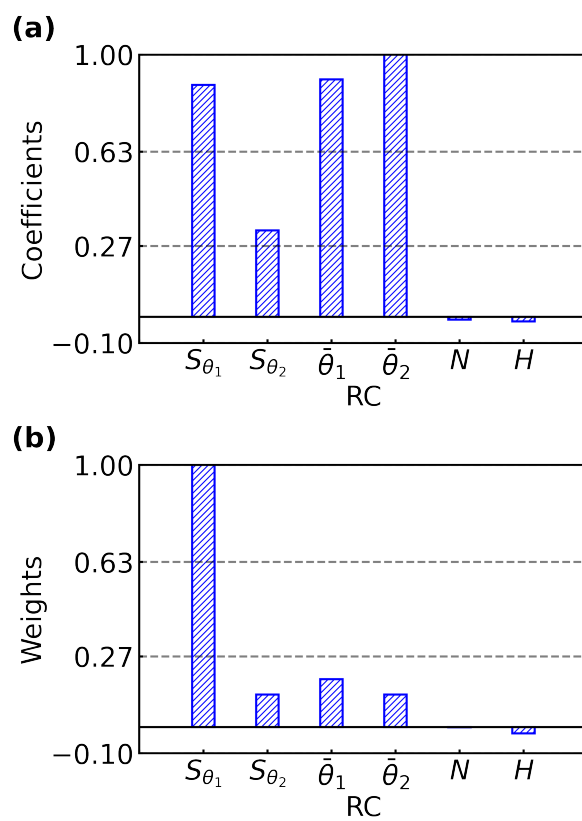


Figure 2.2: Reaction coordinates learned by SGOOP for urea in vacuum. Bar plots of a) coefficients of 1-d reaction coordinate optimized by SGOOP and b) same coefficients, but scaled by their relative standard deviations. The weights in (b) are rescaled properly with respect to corresponding reweighted standard deviations observed in the different OPs of 9 individual 100 *ns* long WTmetaD simulations for SGOOP optimization (details can be found in Appendix) and thus represent relative importance of each order parameters involved in the 1-d reaction coordinate.

Fig. 2.2 (a). These standard deviations were computed from reweighting 9 individual 100 ns long biased MD simulations along S_{θ_1} and S_{θ_2} starting from the melt, which are the same input trajectories for SGOOP algorithm (details given in Appendix). The subsequent weights are shown in Fig. 2.2(b). Here we can see that N and H still have minimal contributions to the RC, but the pair orientational entropy ($S_{\theta_1}, S_{\theta_2}$), especially S_{θ_1} , now has the most dominant weight. This is reminiscent of Argon liquid-droplet nucleation where second moment of coordination numbers have significantly higher weights.

Interestingly, even though the weight for H is small, the SGOOP RC χ_6 captures a negative correlation between entropy and enthalpy in agreement with the definition $G = H - TS$ of the free energy. The mean coordination number N which is a measure of number of neighboring molecules, useful primarily for driving phase transitions from liquid to solid, does not necessarily have the ability to classify different polymorphs as solid urea are mainly face center cubic (FCC). As such it is not entirely surprising that it does not have a high weight in the SGOOP derived RC. In the Appendix we also report another RC found as a local maximum solution to the spectral gap optimization procedure but with lower spectral gap than the RC reported here. As shown in Appendix, the accuracy of the sampling performed using this sub-optimal RC is found to be poorer than the one shown here, reflecting the predictive power of the spectral gap metric of SGOOP. Finally, given the very low weights for H and N in χ_6 , we performed SGOOP to obtain a RC composed of only the other 4 OPs, namely S_{θ_1} , S_{θ_2} , $\bar{\theta}_1$, and $\bar{\theta}_2$. Enhanced sampling performed along this RC also shows slightly worse than with χ_6 but still relatively decent convergence, consistent with our observation that N and H are relatively, though not entirely, unimportant for polymorph nucleation.

Once we have the optimized RC χ_6 described in the previous two paragraphs, we perform

well-tempered metadynamics simulation biasing this RC starting from liquid state. The time series of the RC χ_6 obtained from WTmetad is shown in Fig.2.3(a), where frequent state-to-state transitions occur and no evidence of substantial hysteresis, i.e. system getting trapped in a state is found suggesting χ_6 as a good approximation to the RC. Fig. 2.3(b) shows the free energy profile along χ_6 obtained from averaging over four independent WTmetaD simulations. Fig.2.3(b) shows a distinctive barrier of height $45 k_B T$ (i.e. $168 kJ/mol$) between liquid (black dotted line in Fig. 2.3(a)) and solid states. It is interesting to note that this 1-d profile along χ_6 does not show a discernible barrier between forms I and IV, even though our simulations did not get trapped in these states. However this is evidence that even for urea nucleation in vacuum one should be able to improve the RC further most likely by considering further order parameters that could contribute to the RC.

2.3.3 Comparison with experimental measurements

We now evaluate the quality of our RC and of the sampling so obtained by comparing with experimental observations. While it is difficult to make quantitative comparisons to experiments due to finite-size effects, [111, 124] we expect that the stable polymorph will have lower free energy than the unstable polymorphs. Here we define the free energy difference between any 2 phases α and β as

$$\Delta G_{\alpha \rightarrow \beta} \equiv -G_{\alpha} + G_{\beta} \equiv \frac{1}{\beta} \log \frac{P_{\alpha}}{P_{\beta}} \quad (2.6)$$

where β is the inverse temperature and P_{α} and P_{β} are the probabilities of the system being in phase α and β respectively. We obtain these probabilities by reweighting the histogram obtained

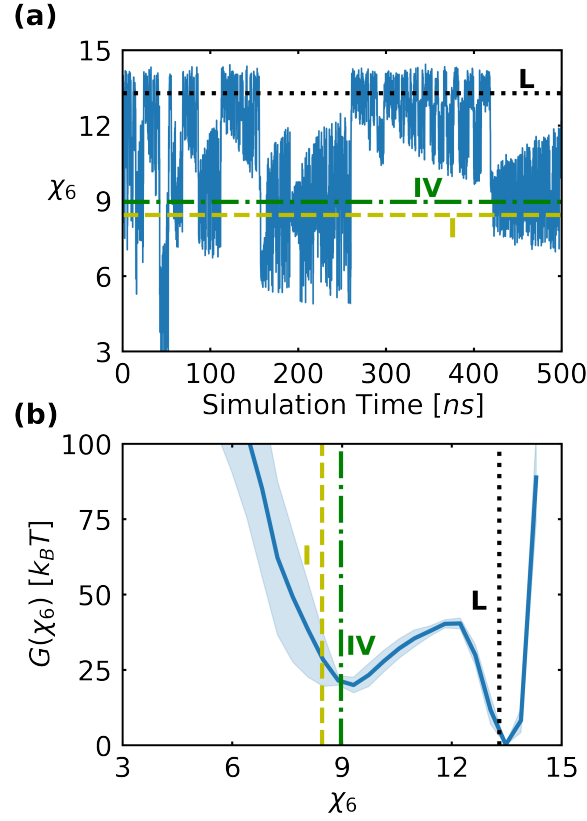


Figure 2.3: The time series of WTmetaD simulation biasing 1-d SGOOP optimized RC χ_6 (a) and its associated reweighted free energy surface averaged over four individual 1-d WTmetaD simulations. (b) Forms I (in yellow) and IV (in green) were frequently visited in the trajectory and corresponding phases have been labeled in free energy profile. Error bars for the free energy are shown through shaded blue color. An unstable structure in which all dipoles are along the same direction is also sampled once throughout the simulation at 42 ns .

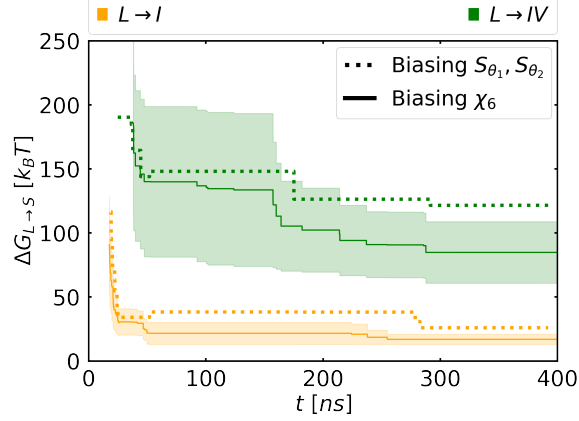


Figure 2.4: The free energy difference ΔG of forms I (in orange) and IV (in green) with respect to liquid phase over time from 2-d WTmetaD simulation biasing S_{θ_1} and S_{θ_2} (dotted lines) and 1-d WTmetaD biasing χ_6 (solid lines). A 400 ns trajectory of 2-d WTmetaD was reproduced from Ref. [99]. Four independent simulations were collected and averaged over for the computation of free energy differences and error bars are shaded with corresponding colors. Form I has much lower free energy than that of form IV from both simulations and this agrees with experimental observations. [93]

from biased metadynamics simulations. [47] Fig. 2.4 shows the free energy differences for the different experimentally relevant phases through the use of Eq. 3.8 by averaging over four independent WTmetaD simulations biasing χ_6 . Fig. 2.4 also shows the results from 2-dimensional WTmetaD biasing both S_{θ_1} and S_{θ_2} as reported in Ref. [99]. The free energy of each polymorph was estimated in a range of $(\bar{\theta}_1, \bar{\theta}_2)$ centered around the landmarks defined in Table. 2.1 which allows us to account for thermal fluctuations. The set $(\bar{\theta}_1, \bar{\theta}_2)$ was deliberately selected for the computation of free energies since the averaged angle OPs do a better job than the optimized RC χ_6 of distinguishing the different polymorphs (time series along average angle OPs are shown in Appendix). It can be seen in Fig. 2.4 that our 1-d RC performs as well if not better in terms of accuracy and convergence speed relative to the 2-d metadynamics.

From simulations biasing χ_6 , the free energy difference between liquid state and phase I is 63 ± 15 kJ/mol, while the difference between liquid state and phase IV is 317 ± 90

kJ/mol . This is qualitatively consistent to experimental observations that form I is the most stable configuration of urea molecules relative to the high pressure product form IV. The other polymorphs which we do not find biasing χ_6 have to the best of our knowledge never been synthesized experimentally, thus reflecting their extremely low probabilities. We would like to first justify the validity of our free energy difference calculation. Since biases were deposited along χ_6 , the probability distributions for form I and form IV computed directly from χ_6 would be sensitive to numerical parameters such as bin sizes, due to close landmark values (see Fig. 2.3). However, when calculating the free energy differences we histogram in the 2-d $(\bar{\theta}_1, \bar{\theta}_2)$ space whose values are able to separate these two solids cleanly (see Appendix). One possible reason would be OPs values associated to form IV having much larger variances than those of form I (see Appendix for time series corresponding to each form), and this leads to very well-separated probability distributions. We further tested the experimental reliability of the obtained Gibbs free energy differences by doing a back-of-the-envelope calculation of the approximate pressure at which urea undergoes a phase transition, and the form IV becomes more stable relative to form I. [125] To do so we assume a linear model $\Delta G(p, T) \approx \Delta G(p_0, T) + (p - p_0)\Delta V$. This gives us the phase transition pressure is approximated to be $1.27 \pm 0.58 \text{ GPa}$ for I-IV transition, which is in excellent agreement, especially considering that force-field errors, with the value of 0.69 GPa at 433K reported from experimental measurements. [90]

We need to point out configurations sampled in metadynamics simulation performed may have defects such as vacancies or even edge dislocations. These phenomena come from the difference in commensurate ratio in periodic lattice and simulation box. [40] Seemingly good order parameters may not well represent a state when the underlying structure is distorted, and therefore, affect the extrapolation of thermodynamic and kinetic information. Another challenge

is that of finite size effects. The size of simulation box (864 atoms) utilized in this chapter is much smaller than typical system size, of around 10^4 Lennard-Jones particles [126], used for study nucleation process, and definitely much smaller than experimentally relevant sizes. Many works [3, 70, 79, 127–130] have reported strategies for dealing with this which could be useful for making more quantitative comparisons of free energies and eventually kinetics [29, 111] with experiments, which was not the objective of the present work.

2.4 Conclusion

In summary, in this chapter we have demonstrated how one can learn a 1-dimensional approximate reaction coordinate for urea polymorph nucleation through the SGOOP approach. [81] The RC is expressed as a linear combination of six order parameters which take into account various thermodynamic and structural information. In addition to previously used order parameters such as entropy, enthalpy, coordination number and others, we also introduced averaged angles $(\bar{\theta}_1, \bar{\theta}_2)$ as order parameters which can be used to distinguish experimentally observed forms I and IV. Entropy was found to have the highest weight in the reaction coordinate, while coordination number, not surprisingly, had the lowest weight as it can only distinguish liquid from solid but not polymorphs from each other. Biasing along our 1-d RC with well-tempered metadynamics, we could obtain frequent back-and-forth transitions between liquid to those experimentally observed polymorphs within shorter simulation time compared to previous work, and not visiting the polymorphs not reported in experiments. We also performed free energy analysis and showed that form I has lower free energy than form IV. This is also in agreement with the experiment fact where form I was seen in the ambient condition while form IV can only be seen in

high pressure condition. Our calculations lead to an accurate estimate of the pressure at which phase transition between these forms happens. We thus believe that this study represents a step forward towards more automated all-atom resolution sampling and understanding of polymorph nucleation in generic, complex molecular systems.

In the next chapter, we will continue our study of crystal nucleation using machine learning methods, and to further challenge ourselves, we solvated the urea system with water molecules and enlarged the system box. These therefore required us to consider not only the solute urea molecules but also taken the solvent molecules into account.

Chapter 3: Driving and characterizing nucleation of urea polymorphs in water

3.1 Introduction

Although traditionally viewed through the lens of classical nucleation theory (CNT) [19], which postulates the only important variable for describing nucleation and subsequent crystal growth is the size of the nascent crystal [18], more recent studies of nucleation for a number of physical systems under a variety of conditions via high-resolution experiments [131–134] and molecular dynamics (MD) simulations [14, 18, 23–32] have shown that CNT alone is not sufficient for an accurate description of the mechanism of crystal formation, which often involves multiple, competing pathways [17, 131, 135]. Thus it is of interest to discover other important variables besides or in addition to the size of the nucleus to accurately describe the nucleation process. One can imagine this situation is especially relevant for substances such as H₂O or organic molecules possessing multiple stable, crystalline structures or polymorphs, where both the lattice structure as well as crystal size are important in distinguishing the solid and the liquid states from each other.

From an experimental perspective, nucleation is a ubiquitous phenomenon, but the microscopic processes driving the nucleation event typically occur over time- and lengthscales that are, generally speaking, below the resolution of current physical apparatuses [131, 132]. Thus, applying a combination of theory and simulation to describe the microscopic nature of nucleation

is required to gain a detailed understanding of the molecular nature of the nucleation process and build a bridge to the time- and lengthscales accessible by computational and experimental approaches.

Even with supercomputers capable of simulating hundreds of microseconds per day [136], the timescale for sampling most nucleation processes, occurring on the timescale of seconds, still remains out-of-reach. Since nucleation is a rare event from the microscopic perspective, nucleation events can generally only be observed in unbiased simulations under conditions of heavy supersaturation or by applying an external driving force. Thus, finding excellent reaction coordinates along which the system can be biased to accelerate the sampling is still very much of importance through a variety of enhanced sampling methods [41, 67–71]. In parallel to the development of more powerful computing machines and sampling algorithms, there has also been an explosion in the use of machine learning (ML) techniques to sift and interpret the results of long simulations [137–144]. Neural networks and ML in general have been implemented to discover good RCs for describing nucleation in a more automatic manner. These neural network derived RCs can then be used as the biasing coordinates in different enhanced sampling methods. In fact, even traditionally RC-free methods, such as path sampling methods (e.g. forward flux sampling) have been shown to benefit from a judicious choice of low-dimensional projections along which to calculate the flux [113, 145].

In principle, the RC can be directly expressed as a function of simple variables such as high-dimensional atomic positions. But, for various computational reasons, generally the RC is expressed as a linear or non-linear function of lower-dimensional order parameters, which are able to distinguish metastable states. These OPs themselves are generally *a priori* designed functions of the atomic positions. Approximating the true RC with smoothly differentiable functions

of the OPs has become one of the main aspects of studying crystal nucleation using enhanced sampling approaches such as well-tempered metadynamics [88], which is our interest in this work, but also more generally for other sampling approaches [3, 28, 29, 37, 40, 74, 75, 79, 126, 144, 146–148, 148–155].

In this chapter, we continue our study on nucleation of urea molecules but now in solution, while the goal was kept the same, i.e., sampling all targeted polymorphic urea clusters and perform thermodynamic measurements among these metastable states. Here we utilize a neural network framework termed the state predictive information bottleneck (SPIB) [48, 51] to extract a two-dimensional RC for describing the nucleation of different polymorphs of urea from solution. SPIB is a variant of RAVE that discovers a set of low-dimensional RCs capable of faithfully predicting the metastable states given the input OPs, which can then be biased in enhanced sampling simulations [51]. The input OPs originate from a library of candidate OPs and are system-specific. For the nucleation processes studied here, the OPs are derived from a set reporting on the global and local molecular orientations and packing of the nucleating species, e.g. coordination numbers [29, 111, 156], radial distribution functions [99, 105] and Steinhardt bond OPs [89], *inter alia*. We describe SPIB in more detail in Ch. 1.3.2.

We find that biasing MD simulations of urea in water using the RCs discovered by SPIB allows for a sufficient enhancement of the sampling such that multiple back-and-forth from homogeneous solution to finite-sized crystal-like clusters are observed over practical compute times on high-performance resources. Furthermore, through the use of a linear architecture during the encoding of the RCs, we are able to directly interpret the RCs discovered. Finally, by reweighting these biased simulations [47] the relative sampling preference, in terms of free-energy difference between the crystal polymorph and the initial solution state, among finite clusters of each poly-

morph embedded in solution is determined.

3.2 Discovering and Biasing Reaction Coordinates for Nucleation

3.2.1 Order parameters for nucleation

As mentioned in the Introduction, in this work as well as in work by others [3, 28, 29, 40, 74, 75, 79, 126, 146–152], it is common to build the reaction coordinate as a function of OPs that collectively distinguish among competing metastable states, which has led to the development of several rich OPs for the study of nucleation and phase transitions in general:

- *Coordination numbers and associated moments:* As introduced above, CNT is based on the size of the nucleus, and the formation of such crystal nuclei is closely related to the formation of a denser phase. A continuous and differentiable coordination number of any molecule i can be defined as:

$$c(i) = \sum_j \frac{1 - (r_{ij}/r_c)^6}{1 - (r_{ij}/r_c)^{12}} \quad (3.1)$$

where r_c is a radial cutoff and r_{ij} denotes the distance between reference sites for two molecules i and j . When averaged over all molecules in the system, the coordination number can serve as an approximation to the true reaction coordinate in the study of gas-liquid transitions [29, 111, 157].

Inspired by this definition, in this manuscript we specifically consider two sets of populations of molecules, with coordination numbers greater than 8 and 11, denoted as N_{8+} , N_{11+} , respectively. We also consider the second moment of all coordination numbers as

an OP, denoted μ_c^2 [156]. However, such coordination number based descriptors are not sufficient to describe the reaction coordinate in events like nucleation from the melt and polymorphism in molecular systems because nucleation is accompanied with the appearance of a space group. Therefore, in addition to the OPs N_{8+} , N_{11+} and μ_c^2 derived from coordination number, we include other OPs into the dictionary, as introduced next.

- *Steinhardt bond OPs*: These OPs, originally introduced in Ref. [89], have been used to distinguish Lennard-Jones solids [156, 158–163] and have been extended to the study of ice polymorphs [164–168]. They map a given local environment onto specific degrees of spherical harmonics [89], defined through :

$$q_{lm}(i) = \frac{\sum_j \sigma(r_{ij}) \mathbf{Y}_{lm}(\mathbf{r}_{ij})}{\sum_j \sigma(r_{ij})} \quad (3.2)$$

where $\mathbf{Y}_{lm}$ is the l^{th} order of spherical harmonics and m ranges from $-l$ to l . As these bond OPs refer to distributions as superpositions of perfect crystalline structure, reliable results have been acquired in classifying face-centered-cubic (fcc), body-centered-cubic (bcc), hexagonal-close-packed (hcp) and liquid-like structures [158]. Here, we add $\bar{q}_4$, $\bar{q}_6$ from Ref. [89] corresponding to the average values of the fourth and sixth orders of spherical harmonics to our OP library. However, as pointed out by Dellago and Lechner [169] thermal fluctuations may destroy the ability in phase identification and, as such, the variants of original bond OPs have been developed for better differentiating ability at the expense of a higher computational cost; these are omitted here.

- *Interfacial water*: The contribution of solvent molecules can hardly be neglected when

studying the nucleation in solutions [30]. Here we introduce a new OP measuring approximately the population of interfacial waters or more generally solvent molecules surrounding solid-like solute molecules by using tunable distance cutoffs. This variable is defined as

$$N_s = \sum_i \frac{1}{2} \xi(i) c_{solvent}(i) \quad (3.3)$$

where

$$\xi(i) = \tanh(c_{solute}(i) - c_o) + 1 \quad (3.4)$$

Here i denotes a sum over solute molecules and c_o is a tunable parameter for the determination of solid-like aggregates which is set to be 5.0 for all systems in this work [156]. Effectively, the OP in Eq. 3.3 is a product of two smoothened Heaviside functions: $\xi(i)$ only counts solute clusters larger than the threshold c_0 , and $c_{solvent}(i)$ is a coordination number between such selected solute atoms and the solvent, as given in Eq. 3.1. Thus, N_s is large when two conditions are simultaneously satisfied: when the solute cluster is sufficiently large and also well-coordinated with solvent. If either one of these conditions is not satisfied, N_s is small. Thus, large values of N_s indicate cluster formation and low values indicate a liquid state or the presence of small, solvent-separated clusters.

- *Averaged intermolecular angles and associated moments:* The estimation of the size of solid-like aggregates and their solvation states is useful and intuitive for the study of molec-

ular systems in different environments [29, 111, 157], but nucleation mechanisms can easily consist of non-classical behavior and polymorphs such that solely distance based OPs fail in categorizing them. In the spirit of Steinhardt bond OPs [89] and environment similarity OPs [170], we introduce a set of OPs which account for the orientation of molecular feature vectors of solutes to help better discriminate polymorphs. For any molecule i , we define:

$$\theta(i) = \frac{\sum_j \sigma(r_{ij}) \frac{1}{2} [(\pi - 2\tilde{\theta}) \tanh(5\tilde{\theta} - 9.25) + \pi]}{\sum_j \sigma(r_{ij})} \quad (3.5)$$

where $\tilde{\theta}$ is the angle formed between characteristic vector on molecule i and molecule j as illustrated in Fig. 3.1, and σ is a switching function of intermolecular distance. The hyperbolic tangent switching function is applied to remove the mirror image symmetry (see Appendix for detailed explanation). We then compute the mean of angular distribution functions, denoted as $\bar{\theta}_1$ and $\bar{\theta}_2$ where 1 and 2 denote the two intramolecular vectors specified for crystalline configuration identification (see Fig. 3.1 (a) and (b) for definitions of selected vectors for urea). In addition to these OPs, we also consider in our OP dictionary the second moments $\mu_{\theta_1}^2$ and $\mu_{\theta_2}^2$ of the collection of θ values defined through Eq. 3.5.

- *Pair entropy*: In Ref. [105] Piaggi and co-workers introduced an OP which approximates radial and orientational entropies from radial different distribution functions. This OP is formulated as:

$$S(r) = -2\pi\rho k_B \int_0^\infty [g(r) \ln g(r) - g(r) + 1] r^2 dr \quad (3.6)$$

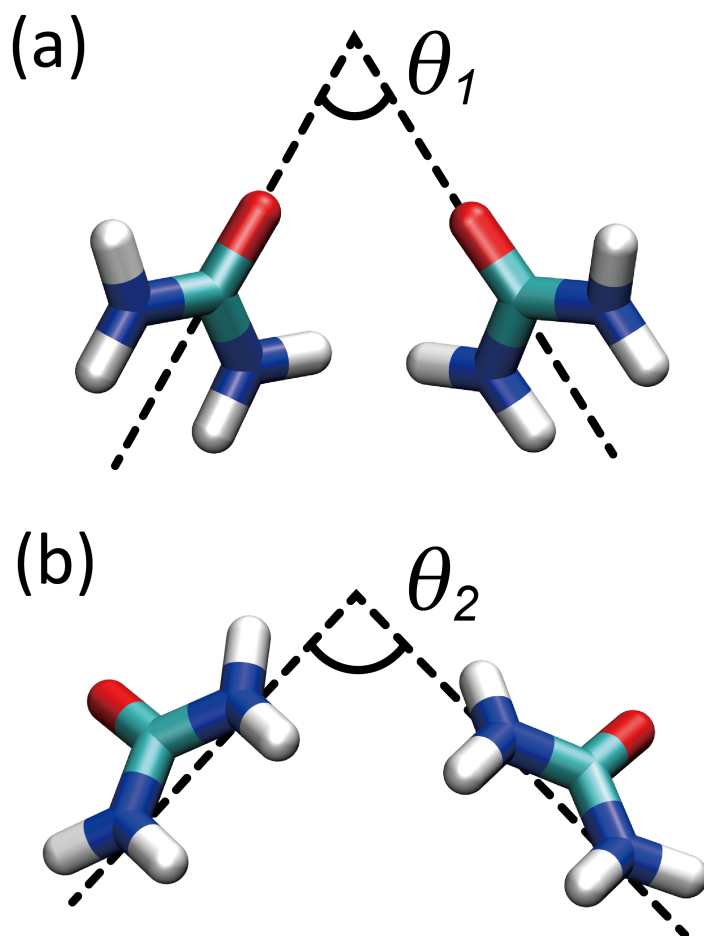


Figure 3.1: Visualization of the feature vectors and corresponding intermolecular angles (a) θ_1 and (b) θ_2 for urea molecules. The feature vectors shown here are the vectors connecting C-O and N-N. Different atom types are indicated with different colors, specifically carbons in cyan, oxygens in red, nitrogens in blue, and hydrogens in white. Snapshots are rendered using Visual Molecular Dynamics (VMD) [36].

where $g(r)$ is the radial distribution function, ρ is the density, and k_B is Boltzmann’s constant. This OP was introduced in order to account for the possible entropy-favored crystalline structures. An advanced version of this OP is developed in Ref. [99]. This pair orientational entropy OP, denoted $S(r, \theta)$, depends on both the intermolecular distance and the relative angle θ between a given intramolecular vector equivalently defined for each solute molecule in the simulation:

$$S(r, \theta) = -\pi\rho k_B \int_0^\infty \int_0^\pi [g(r, \theta) \ln g(r, \theta) - g(r, \theta) + 1] \times r^2 \sin(\theta) dr d\theta \quad (3.7)$$

It was found that such an approximation to the entropy is useful for distinguishing polymorphs in simulations of urea and naphthalene, and a variant of Eq. 3.7 is used in [155] to successfully discover and distinguish polymorphs of 1:1 mixtures of resorcinol and urea. Since Eq. 3.7 only accounts for a single angle, and not the three angles necessary to specify exactly the relative orientation between molecules, this expansion of the entropy, though useful, is necessarily approximate, as are the other OPs previously described in this section.

While individually these different OPs might have limitations, they can supplement each other and provide a palette through which the complex molecular processes governing nucleation can be described. It is by combining them through the State Predictive Information Bottleneck (SPIB) [48] approach that we construct low-dimensional projections that are much closer to the true RC than any of the OPs on their own. We next briefly summarize the SPIB approach, referring to Ref. [48] for further details.

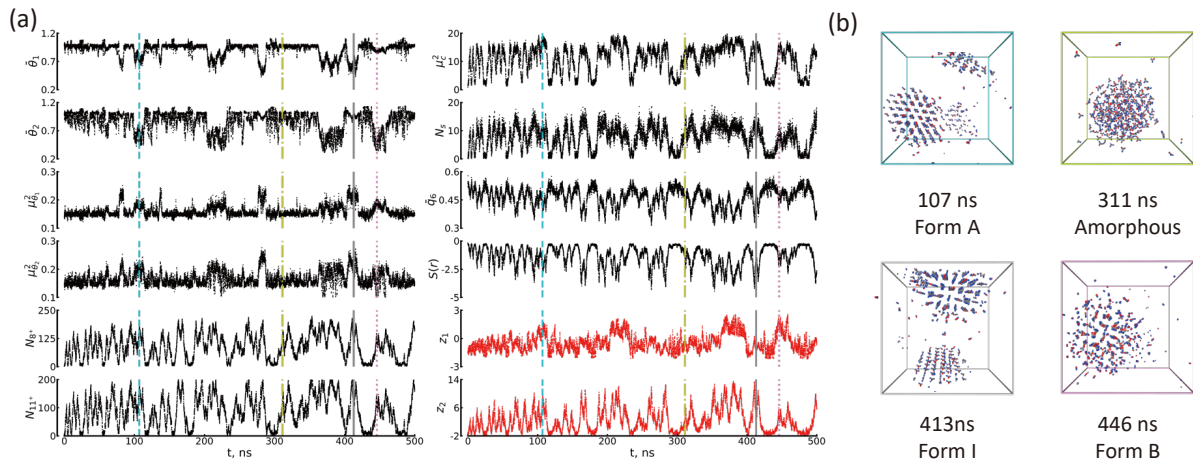


Figure 3.2: Sampling different finite-sized urea clusters in aqueous solution with WTmetaD simulations biasing the SPIB-learned 2-dimensional reaction coordinate. (a) shows a representative time series of OPs (black) and reaction coordinates (red), clearly demonstrating multiple back-and-forth transitions between different phases. Vertical cyan dashed, green dash-dotted, grey solid, and pink dotted lines indicate representative transitions to form A, amorphous, form I, and form B from homogeneous solution. (b) shows snapshots of clusters sampled during trajectory shown in (a), along with the time (vertical lines in (a)) at which they were observed. Ovito 3.3.5 [35] is the visualization tool used for snapshot generation and water molecules are removed.

3.2.2 State Predictive Information Bottleneck (SPIB)

Given the evidence that CNT is insufficient to accurately describe the nucleation event for a number of chemical systems, for instance because knowledge of more than just the solute coordination number OP is required, we need an approach to combine the different possible OPs to develop a lower-dimensional reaction coordinate for describing nucleation in urea and the SPIB method (see Ch.1.3.2 for details) is used to construct a 2-dimensional reaction coordinate with provided OPs introduced above.

Table 3.1: Parameters for WTmetaD

System	ω (kJ/mol)	γ	σ_1 (RC unit)	σ_2 (RC unit)	T (K)
Urea	5.0	100	0.2	0.2	300

3.2.3 Estimating free energies with Metadynamics

In order to learn the SPIB, or the approximate RC, as a combination of different OPs, we need access to a trajectory that has visited different possible conformations. Such a trajectory is generated here through the method Metadynamics, which helps the system escape free energy minima by periodically depositing Gaussian kernels as a function of the variable being biased. For the initial metadynamics runs we bias orientationally informative OPs such as intermolecular angles $\bar{\theta}$ and orientational entropies S_{θ} . After performing SPIB analysis on these runs, we then bias along the SPIB-learned RCs in future iterations. In particular, here WTmetaD is used where kernels are decreased in height for biasing regions being revisited, thus leading to better convergence of the free energy surface [88]. The WTmetaD parameters are reported in Table 5.1. The height of Gaussian deposition, ω , is selected to $2 k_B T$ at 300 K and the bias factor, γ , is set to be 100, in order to overcome the high energy barrier associated with nucleation problems [124]. The width of the Gaussian, σ , is set to the thermal fluctuation of the corresponding approximate reaction coordinate estimated from a short (~ 10 ns) unbiased MD run.

After classifying frames in the biased simulation into either the liquid-like state or a particular crystal polymorph, the free energy difference between two states a and b at temperature T can be calculated as follows:

$$\Delta G_{a \rightarrow b} = G_b - G_a = k_B T \log \frac{P_a}{P_b} \quad (3.8)$$

where k_B is Boltzmann's constant and P_a , P_b denote the Boltzmann probabilities of states a , b obtained from reweighting the WTmetaD simulations [47].

In this study, we find that using a two-dimensional latent space in the SPIB is sufficient to learn the important physical processes driving polymorph nucleation in urea in aqueous solution. However, the SPIB approach is not limited to two-dimensional latent spaces, and the bottleneck approach allows for the use of a latent space of arbitrary dimensionality. Furthermore, it is not mandatory to use WTmetaD to generate biased trajectories for input to the SPIB; unbiased simulations may be used (provided the sampling of the processes of interest is sufficiently good) or biased simulations from another enhanced sampling method. Since the SPIB protocol involves biasing along the discovered latent variables, if a high dimensional latent space (more than three-dimensional) is required, then it will likely be more computationally efficient to select a biasing approach more suited to higher dimensional CVs [171].

3.3 Results and Discussion

We performed classical all-atom MD simulations for all systems using GROMACS version-2022.2 [117] patched with PLUMED 2.6.1 [118, 119] to perform the WTmetaD. Further simulation details are provided in Materials and Methods section. Below we provide detailed results for homogeneous nucleation of urea molecules in water.

3.3.1 Urea Nucleation

Urea is commonly used as fertilizer and as nitrogen feedstock for organic synthesis. Due to its industrial importance, urea has been studied extensively and several crystalline structures have been synthesized and reported. These include polymorphs named I (space group: $P\bar{4}2_1m$), III ($P2_12_12_1$), IV ($P2_12_12$) and V($Pm\bar{c}n$), where polymorph I corresponds to the typical poly-

morph stabilized at ambient conditions and the rest are high-pressure, high-temperature products [91–94, 96, 97, 154]. Moreover, an ever-enriched dictionary of polymorphism in urea has been established with computational methods using classical and quantum approaches [30, 40, 98, 99, 124, 127, 172, 173]. We adopt the same notations as in Ref. [99], where two more polymorphs of urea A ($Pnma$) and B ($P\bar{1}$) are introduced. It remains unknown if these two crystal structures A and B are artifacts due to use of the generalized Amber force field (GAFF) [115] or simply experimentally yet to be observed. Despite the orientational difference in the lattice cell, hydrogen bonding pattern in both forms A and B (all in type III) has been shown to be different from experimental forms (type I and type II) [98, 173].

Chronologically, we start with a preliminary round of well-tempered metadynamics to facilitate state-to-state transitions along trial reaction coordinates. For this, we biased the two averages of the two feature angles (Fig. 3.1 (a)), $\bar{\theta}_1$ and $\bar{\theta}_2$ as they have been considered as OPs in distinguishing crystal structures of urea from each other [149]. Several transitions to and from the solid states of interest are observed in the WTmetaD simulations (data not shown) biasing this two-dimensional coordinate (4 independent runs of 400 ns each, totalling $\sim 1.5 \mu\text{s}$) and trained as input data for SPIB along with other candidate OPs.

3.3.2 Urea Polymorphism

From this preliminary WTmetaD simulation, we performed SPIB to learn a 2-d latent representation for the reaction coordinate (denoted as z_1 and z_2) as linear combinations of OPs discussed in Materials and Methods section. We then perform a second round of WTmetaD biasing along this SPIB learned 2-d RC consisting of 4 independent runs of 500 ns each. The time

series for different OPs and the 2-d RC components are shown in Fig. 3.2(a) in black and red, respectively. This unequivocally shows the presence of numerous back-and-forth state-to-state transitions (see Appendix for detailed sampling efficiency comparison), a hallmark of good enhanced sampling [42]. In particular, three crystal structures, namely polymorphs I, A, and B, are visited during simulations. The corresponding snapshots for these from one are shown in Fig. 3.2(b) along with that of an amorphous bulk phase. The evidence for the formation of these polymorphs can be found in the time series of intermolecular angle OPs (top left two rows in Fig. 3.2(a)), as either one or both of these OPs drop, indicating formation of an ordered phase based on their definitions. Comparing to the known polymorphs introduced above, structures from experimental products III, IV, V [95] and zero-temperature prediction C [98] are missing. It is possible that the former high-pressure structures are less stable as the existence of mediating solvent molecules, and the latter polymorph, could potentially be unstabilized by the addition of entropic effects under finite temperature. As these solid states have not been found in previous studies using MD simulations [40, 127] (except form IV being observed in nucleation from the melt [105]), we therefore believe there is remaining scope here for exploration of the configuration space in future studies from other perspectives such as force-field refinement.

In Fig. 3.3(a), the reweighted free energy surface along the two-dimensional latent representation z_1 and z_2 is plotted [47]. The states of interest, one liquid-like and three crystal-like phases, have been labeled on this surface. However, no minima are observed corresponding to regions marked as polymorph states of urea, which could be caused by multiple factors. One of the main reasons suggested is the effect of finite-size, which can be corrected analytically using CNT or by increasing the size of the simulation box accordingly [124, 127]. However, in this work, we emphasize the construction of ML RCs that explore the configuration space reported in previous

studies rather than conveying detailed, more quantitative, investigations on thermodynamic and kinetic properties, which are left for future studies.

In Fig. 3.3(b), we provide the coefficients for different OPs, indicating how much they contribute to the RC (see **Materials and Methods** for details on calculating these coefficients). From these bar plots, it can be seen that several OPs are weighted heavily in the RC, comprising attributes of intermolecular angle and coordination number. Referring to the definition of these two categories of OPs, the coordination number is a simple but useful OP in the identification of phases, suggestive of the CNT formalism where only the size of the nucleating cluster is sufficient. However Fig. 3.3(b) clearly shows that the coordination number in such complicated molecular systems is not enough for capturing the slow degrees of freedom for crystallization. In addition, our RC indicates that the orientation of feature vectors needs to be taken into account as these vectors are significant when classifying polymorphic urea [173].

To support the above discussion, we project the SPIB-learned RCs z_1 and z_2 used to bias onto orientational informative OP space $(\bar{\theta}_1, \bar{\theta}_2)$ in Fig. 3.3(c) and (d) respectively. Here it can be seen that the driving forces behind the two representations are clearly different. z_1 specifically enforces the transitions from the initial liquid-like state (orange hexagon) to states with lower $\bar{\theta}_2$ values (which are form A (cyan star) and form B (pink circle)) and it shows a strong correlation to $\bar{\theta}_2$ OP itself. On the other hand, z_2 pushes the system away from the initial liquid-like state to all potential solid phases.

3.3.3 Comparison with previous studies

For the purpose of accurate classification of crystal-like clusters for further analyses, the local crystallinity (SMAC) OPs are used (see Appendix for full expressions) [30, 40, 124, 127]. In this way, polymorphs can be screened precisely by properly positioning the centers of various switching functions. The size of the nucleus can then be computed using graph theoretic algorithms [76]. After this classification, the free energy corresponding to each phase can be evaluated, giving the free energy difference between two arbitrary states a and b at temperature T as described in Estimating free energies with Metadynamics section. Fig. 3.4 shows the free energy differences of various finite-sized crystal nuclei of urea formed in aqueous solution with respect to the initial liquid-like state obtained after reweighting the biased WTmetaD simulations [47].

Without systematic assessment of the effects of finite-size limitations one can easily find results, especially for kinetic but also thermodynamic properties, in disagreement with those of million-atoms simulations [130, 174]. Many approaches to overcome these limitations have been introduced [111, 130, 175], and, not surprisingly, accurately accounting for the effects of finite-size still remains an active field of research. Therefore, we draw only qualitative analyses on these visited metastable states in terms of the relative preference of forming each crystalline structure in aqueous solution from the state of homogeneous solution.

As can be seen from this figure (Fig. 3.4), our simulations suggest that form A is the most stable polymorph in the given water model, followed by form I and finally form B. A similar relationship between forms A and I in aqueous solution has been established and studied in detail in previous work: whereas a small nucleus favors crystal structure A, interconversions between A and I occur when cluster grows larger. In particular, Ref. [127] suggests a barrierless A-to-

I transition occurs as clusters larger than ~ 50 molecules when applying corrections to leverage finite-size effect, and Ref. [172] claims such a transition occur at a cluster size of ~ 530 molecules associated to a $\sim 125 k_B T$ energy barrier when performing a seeding method by inserting pre-built crystals into large simulation boxes.

To the best of our knowledge, metastable state B has never been synthesized in aqueous solution of urea, and its very high free energy relative to the other forms may explain why such a structure is difficult to sample. We can explain why we were able to sample this metastable polymorph in our simulations and also in Ref. [99]. Structurally, form B has a completely different orientation among other polymorphs in which its dipole moment in the direction of C-O vector is no longer parallel or anti-parallel to its neighbors. This is likely a result of entropic contributions [99]. By having entropy-related OPs in our dictionary of OPs that comprise the RC, we are able to accelerate the substantial entropic degrees of freedom together with other more energy dominated modes, and therefore both enthalpic- and entropic-favored crystalline structures of urea are formed in aqueous solution with enhanced sampling method biasing along machine learned RC.

3.4 Conclusion

Although nucleation is a frequent event on the experimental and biological timescales, with currently accessible compute resources nucleation is a rare event computationally, even when utilizing enhanced sampling techniques [27, 29, 40, 111, 149, 176]. Thus, even for relatively simple systems such as the aqueous urea systems studied here, sampling polymorph configurations and, more dauntingly, determining relative thermodynamic stabilities of these polymorphs, is difficult

computationally. In this work we have demonstrated a more-or-less automatable protocol to obtain back-and-forth sampling of different polymorphs in aqueous solution. Specifically, we have shown that good sampling of polymorphs is possible when classical molecular dynamics simulations of urea are biased along optimized reaction coordinates found using the machine learning approach State Predictive Information Bottleneck (SPIB) [48] which belongs to the RAVE family of methods [59, 85, 137]. A linear, two-dimensional approximation to the reaction coordinate extracted from SPIB [48] can be used to 1) effectively bias nucleation simulations of aqueous urea to enhance the sampling of polymorph nucleation, and 2) interpret the SPIB RCs to determine which OPs are the most important drivers of nucleation.

When interpreting the RCs derived from the SPIB analysis, we find that for urea, the intermolecular angle and cluster size OPs are the most important contributors to the RC. This observation reflects that size informative OPs suggested by CNT are insufficient in describing slow modes towards nucleation of urea molecules; in addition orientational OPs are needed. Each polymorph is identified using different metrics and then ranked appropriately as per their Boltzmann weights at 300 K [47]. We find, given the specific conditions studied here, for urea, finite-sized crystals of form-A are preferentially formed over form-I, while form-B is the least favorable with the homogeneous solution taken as a reference state.

Since we performed the SPIB analysis on the unreweighted energy surface, we cannot yet offer any kinetic insights regarding the nucleation of either organic molecule studied here. In addition we did not consider the question of finite-size effects, which would especially impact kinetic properties. This will be done in future work using approaches such as Ref. [111]. Secondly, our simulations would have been more representative of supersaturation levels of experiments and avoid solute depletion which commonly exists in the small-scale atomistic simulations with

the use of constant chemical potential ensemble, which is an active area of research for studying nucleation [176].

To summarize, this work makes several significant contributions to the field of nucleation research and enhanced sampling in general. We have applied deep learning based RC construction methods to arguably the most complex set of systems to date and used them to successfully find reaction coordinates that accelerate the nucleation process of multiple polymorphs for urea aqueous solutions. Based on these results, we believe using machine learning methods to construct approximate reaction coordinates can provide novel solutions and information regarding non-classical effects in the nucleation of molecules from aqueous media.

Although in this chapter we successfully sampled many state-to-state and back-and-forth transitions between urea polymorphs, the constructed free energy surface obtained from WT-metaD simulations does not show distinct minima to those solid states. Despite the finite size effect summarized above, our dimensionality reduction methods are largely limited by the library of input OPs. It is difficult to learning optimal latent variables if the model is provided with sub-optimal or bad CVs. Therefore, in the next chapter, we introduce a different machine learning scheme which allows one to directly learn low-dimensional representations from high-dimensional atomic coordinates space.

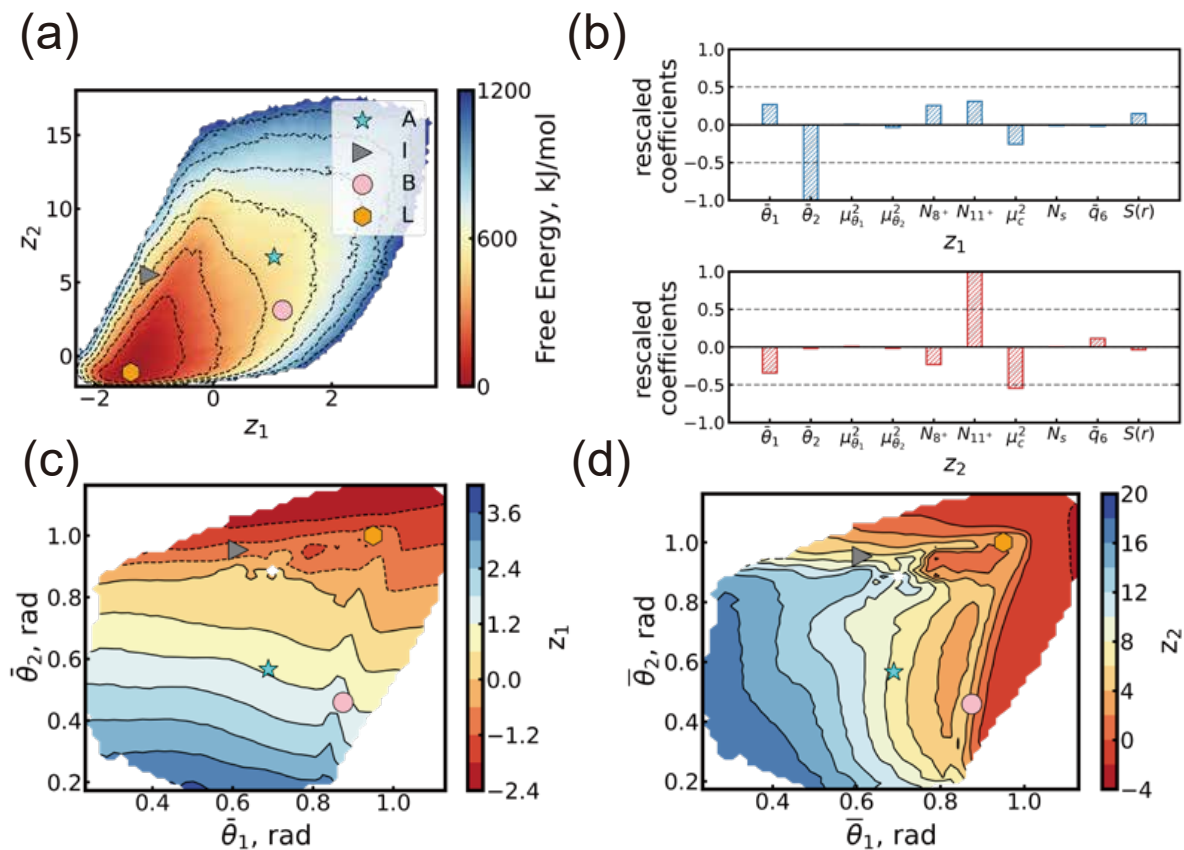


Figure 3.3: (a) Reweighted free energy surface in kJ/mol along SPIB-learned reaction coordinates with landmarks for states of interest (finite-sized crystal nuclei A, I, B in aqueous solution, and the initial homogeneous solution L)

(b) Rescaled coefficients for each OP in construction of SPIB-learned RCs. The coefficients are rescaled with respect to the corresponding fluctuations of components in WTmetaD simulations (see **Materials and Methods** for details). The importance of the urea molecules relative orientation can be seen from the first reaction coordinate while the second coordinate shows the relevance of clustering. (c) z_1 -projection onto $(\bar{\theta}_1, \bar{\theta}_2)$ space; and d) z_2 -projection onto $(\bar{\theta}_1, \bar{\theta}_2)$ space. The markers asterisks (cyan), triangles (grey), circles (pink), and hexagons (orange) indicate crystalline structures A, I, B, and the initial liquid-like phase L, respectively.

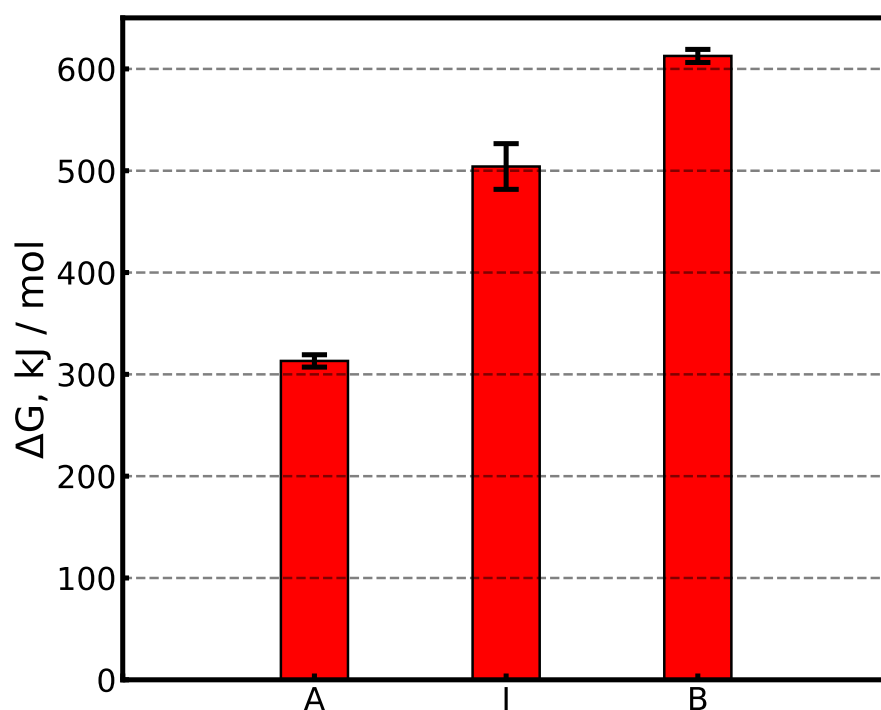


Figure 3.4: Free energy difference in kJ/mol between different crystal nuclei formed in aqueous solution with respect to the initial liquid-like state as calculated from WTmetaD simulations. This indicates that the formation of form A in aqueous solution is preferred to that of form I, and forming polymorph B is the least likely among the three crystal-like states in aqueous solution, comparatively. The error bar is computed over four independent production runs.

Chapter 4: Enhanced Sampling of Crystal Nucleation with Graph Representation Learning Variables

4.1 Introduction

In last chapter, we study crystal nucleation of solvated urea molecules with a 2-dimensional reaction coordinates learned by SPIB. Although state-to-state transitions are sampled, a meaningful free energy surface of the system is remaining to be constructed. One of the limitation would relate to the lack of choice of the candidate OPs to SPIB. Therefore, in this chapter, we introduce another machine learning technique which allows us to construct latent representation from the atomic coordinate space.

Over the years a vast range of such hand-crafted and machine-learned OPs have been proposed for the study of crystal nucleation. These can be split into different classes. A first class includes task-specific OPs whose definitions rely on particular orientations of particles or molecules and their local environments in the corresponding crystalline packings of interest. [27, 30, 77, 89, 170, 177, 178] A second class of OPs is more generic and does not need prior knowledge of the relevant crystalline packings. These rely on the computation of the exact or approximated thermodynamic observables. Examples are approximate entropy, enthalpy [99, 105] and moments of coordination number [28, 29]. Given their generic nature, these OPs can be ap-

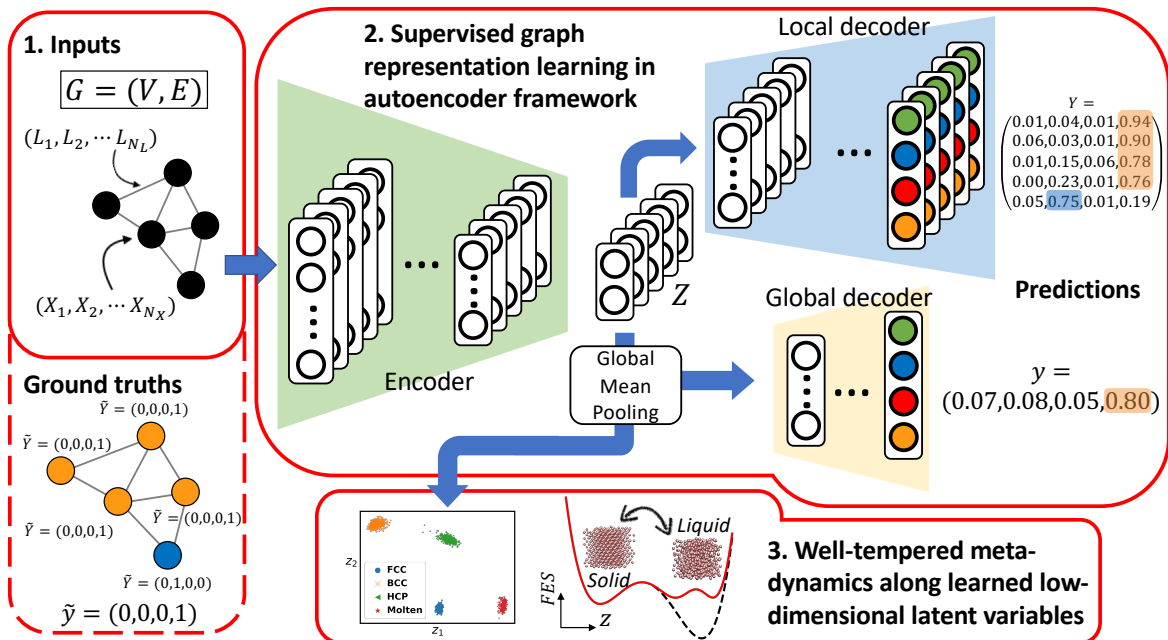


Figure 4.1: Schematic of the pipeline of the graph representation learning architecture. In part 1, atomic coordinates generated from MD simulations are first converted into graphs consist of tensors of node and edge features (X and L respectively; in solid red box) with labels (in dashed box). In part 2, graph data is fed in the GNN model for the training process. Under a supervised learning scheme, ground truths (i.e. labels) are applied for computing and backpropagating the loss. Once the model is trained, the frozen encoder part along with a global mean pooling layer provides latent variables computed on-the-fly as reaction coordinates for metadynamics (part 3).

plied to systems without any prior knowledge for the exploration of the free energy landscape and screening of metastable allotropes or polymorphs. However, it is important to note that this generality may occasionally lead to slow convergence and inefficiency in computing free energy. [40]

Different from the above two hand-crafted classes of OPs, recent breakthroughs in machine learning (ML) techniques have given rise to a range of neural network (NN) based OPs for a variety of problems, including crystal nucleation. The inherent differentiability of these OPs makes them suited for various enhanced sampling methods that involve the modification of a system's Hamiltonian. [179–187] We specifically highlight graph neural networks (GNNs) which

have emerged as powerful tool in the realm of materials science, including but not limited to efficient descriptions of material energetics [188,189], accurate predictions on material properties [190,191], and robust classifiers of crystal structures and defects [102,113,141,192,193].

Early chapter involving graph-theoretic ideas, predating the development of GNNs, was already reported for crystal nucleation. Examples include social permutation invariant (SPRINT) coordinates by Pietrucci and Andreoni [194], smooth overlap of atomic positions (SOAP) fingerprints by Bartók *et al.* [195], and other approaches. [76] In the same vein, carefully designed GNN models can construct an optimal representations of complex molecular systems that are invariant to translational, rotational, and permutational symmetries. Translation and rotation invariance can be achieved by introducing input features like radial distances and angles, which themselves remain invariant to both translation and rotation. An example can be found in Ref. [196], which proposed a pretrained GNN-based representation generator on translation and rotational invariant features with atom-centered symmetry functions for OP construction. On the other hand, in order to classify the overall state of a given structure as a given phase of matter, a natural way is to apply pooling aggregators, where pooling over fully connected layers enforces correspondences between mappings of feature space and classes, and this leads to permutation invariance in GNN models. These properties in combination make GNN models useful in capturing characteristics of the state of highly ordered crystal structures.

While the above approaches are elegant and powerful, there are not yet many approaches where the learned representations from a GNN are biased to enhance the sampling of nucleation processes. The only notable exception is a very recent preprint. [185] In this work, we develop a GNN-based autoencoder framework for acquiring low-dimensional representations that is then used in enhanced sampling of crystal nucleation in two different systems. A GNN model in an

encoder-decoder setup is trained to precisely classify crystalline structures using local environments. It ensures permutation, translation, and rotation invariance in its latent outputs. A global pooling layer achieves invariance to permutational symmetry, while translation and rotation invariances arise from carefully chosen input features. These compact learnt variables can be easily integrated into various enhanced sampling techniques, thanks to the differentiability of machine learning models. To demonstrate the efficacy of our approach, we apply it to two challenging systems, namely iron and glycine, in the context of nucleation from the melt using metadynamics. Our results show the obtained latent variables capture key configurational features from the training dataset and are able to enhance the samplings as evidenced by frequent back-and-forth transitions. Our chapter has complementary aspects to Ref. [185]. While we benchmark directly on input atomic configurations, allowing the GNN model to probe their structural differences, the model from Ref. [185] provides an accurate and efficient method for reconstructing conventional order parameters by including their information in the loss function. We then conduct analysis with the aid of the trained GNN models for computation of relative free energies which exhibit excellent agreements with existing literature regarding the stability rankings of various allotropic and polymorphic structures.

4.2 Methods

As introduced above, simulating rare events like crystal nucleation require both associated reaction coordinates and enhanced sampling methods to increase movement along these coordinates. Here we provide detailed information on both of these aspects in the following three subsections: we first introduce graphs and graph neural nets in Sec. 4.2.1, and we present details

on how these graphs are prepared (Sec. 4.2.2). Then, we summarize the sampling method well-tempered metadynamics, in Sec. 5.2.2. A schematic is presented in Fig. 4.1. We provide details on setting up the simulations in the Appendix and descriptions of the notations in the machine learning model in Tab. 4.1.

4.2.1 Graph Neural Networks (GNN) based model

In this work, we use a k-nearest neighbor (kNN) algorithm [197] to construct the neighborhood $\mathcal{N}_v$ where k is a tunable parameter. In Fig. 4.1 panel 1, a graph with 5 vertices is presented as an example with N_X number of node features and N_L number of edge embeddings, these feature tensors eventually serve as input to GNN models. For simplicity, node features are set to 1 for systems studied in this chapter and edge features are selected to be translation and rotation invariant (which we discuss in the next subsection, Sec. 4.2.2). Models designed in this chapter adopt a supervised learning scheme which therefore requires ground-truth labels of different crystal phases for computing and minimizing the loss function. The atom-wise labels are generated with different baseline classifying methods and the graph-level label is determined by the leading population of fingerprints on nodes.

As shown in Fig. 4.1, the learning scheme is composed of three parts: creating graph data, training GNN models to learn low-dimensional order parameters (OPs), and performing enhanced sampling along a further reduced space. In this chapter this is done through an embedded autoencoder framework that allows one to obtain a low-dimensional representation for generic enhanced sampling methods. Specifically, starting from selected node and edge features (described separately in the next subsection) as input features, the encoder, colored in green in

Table 4.1: Notations for the machine learning model used in this chapter

Notations	Descriptions
V, X	Nodes and node embeddings
E, L	Edges and edge embeddings
N	Number of vertices (nodes)
$\mathcal{N}_v$	Neighborhood set of node v
l	Model layer index
F	Edge convolution network in ECC layers
y, Y	Global and local predictions
$\tilde{y}, \tilde{Y}$	Graph and node labels (one-hot)
Z	Node latent variables
z	Global (pooled) latent variables
$\mathcal{L}$	Loss function
β	Hyperparameter in loss function

Fig. 4.1, takes the input and compresses them into a relatively lower-dimensional local decoder Z . For the applications shown in this work, Z is $(N, 2)$ dimensional, where N is the total number of nodes in a graph. The local decoder (in blue in Fig. 4.1) predicts the structure of individual nodes, which could be atoms or molecules, given the information about their neighborhood. In addition, a global decoder, shown in yellow, classifies the entire input graph by coupling it with a global pooling layer. Here we choose to use the global mean pooling layer which avoids the effect of system size. In other words, this makes the trained model size-agnostic (i.e., transferable to systems of the same species of any size). The output of the global mean pooling operation, z , is generally two-dimensional and is biased in enhanced sampling.

The graph convolutional layers, specifically edge-conditioned convolution (ECC) layers [188, 192, 198], allow message passing of node and edge features of the linked neighboring nodes

$\mathcal{N}_v$ into the individual vertex through convolution operations, as described in Eq.4.1:

$$X^l(i) = \frac{1}{|\mathcal{N}_v|} \sum_{j \in \mathcal{N}_v} F^l(L(j, i); w^l) X^{l-1}(j) + b^l, \quad (4.1)$$

where l is the layer index in the neural network, w and b are learnable weights and biases of the network. A filter network [199] F^l parameterized by weights w outputs an edge-specific weight matrix given edge attributes $L(j, i)$. We keep the same machine learning architecture for the different systems studied in this work. In particular, we keep zero hidden layers in both decoders to optimize the ability of the encoder to classify different crystal structures.

The learning objective $\mathcal{L}$ of this model consists of a sum of two cross-entropy losses provided in Eq. 4.2,

$$\mathcal{L} = - \sum_i^N \sum_{c \in \text{classes}} \tilde{Y}_{i,c} \log Y_{i,c} - \beta \sum_{c \in \text{classes}} \tilde{y}_c \log y_c. \quad (4.2)$$

where classes in this chapter correspond to crystal structures. The first term, which is a local prediction term, computes the cross-entropy loss of the node logits $Y_{i,c}$ of node i class c with respect to the node-level ground truth $\tilde{Y}_{i,c}$ which is intrinsically a binary indicator (0 if the node does not belong to class c and 1 if it does so) and then sums over all classes and nodes. The second term, which is a global prediction term, calculates the cross entropy between the c^{th} class graph prediction (y_c) and the target ($\tilde{y}_c$). A hyperparameter β is introduced to control the relative importance in local and global prediction.

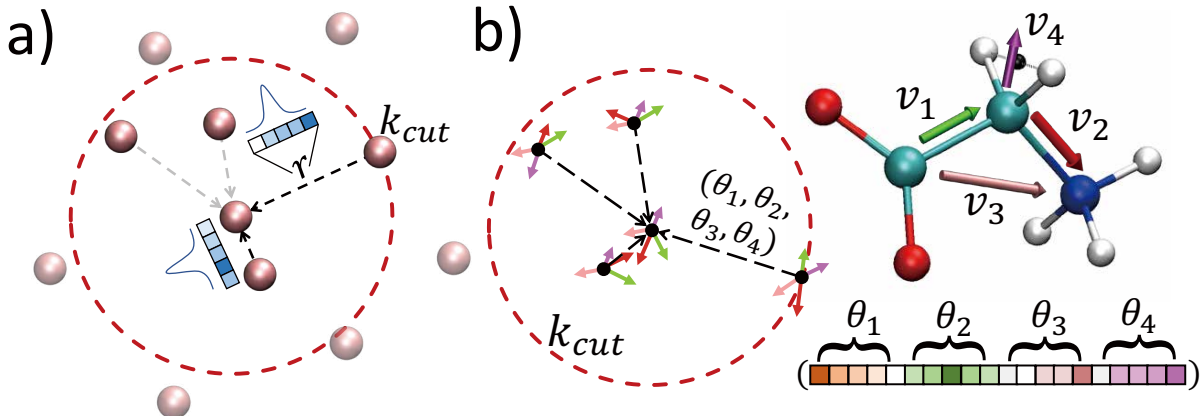


Figure 4.2: Geometric representations of (a) iron particles and (b) glycine molecules. Gaussian basis functions are applied to featurize edge attributes individually and node features are set to be one for both systems. The neighborhood is defined with the k-nearest neighboring algorithm ($k_{cut} = 4$ as an example). v_i 's are intramolecular characteristic vectors and θ_i 's are the corresponding intermolecular angles formed by specific v_i . v_1 is the C-C $_{\alpha}$ vector, v_2 is the N-C $_{\alpha}$ vector, v_3 is the C-N vector and v_4 is the C $_{\alpha}$ -H $_{center}$ vector. Atoms are colored with respect to their species: irons in mauve, oxygens in red, carbons in cyan, nitrogens in blue, and hydrogens in white. Ghost particles (i.e., center of mass of glycine and center of hydrogen atoms) are in black.

4.2.2 Dataset Preparation

Like all data-driven methods, prior information needs to be provided to train our model. In this work, the model uses all-atom coordinates from MD simulations initiated from different perfect crystal structures. The training data for iron allotropes was generated by LAMMPS built-in lattice functional; while the molten phase was prepared by random insertion of iron particles. Separately, four supercells composed of 432 α -Fe, 256 γ -Fe, 180 ϵ -Fe, and 285 molten Fe were constructed and 1 *ns* short MD trajectories were initiated accordingly. These structures were equilibrated at respective temperatures where they are expected to be stable. Specifically, this was 1000 K for α -Fe, 900 K for γ -Fe, 900 K for ϵ -Fe, and 2000 K for molten Fe. A total of 2000 frames of MD snapshots, corresponding to 500 frames for each configuration, were then converted into graph representation and trained via the proposed model discussed above.

The node feature was set to be unity. A sparse adjacency matrix was constructed on k-nearest neighbors (k-NN) metric. [200] The k value for iron is set to be 50 and a detailed discussion is provided in Sec.4.3.1. Each pair of linked nodes was attributed the radial distance as the edge feature, followed by a Gaussian basis function introduced in Schnet shown in Fig.4.2a) . [189] An edge feature, L , can be obtained by expanding distance, r , into t slices as follows:

$$L_t(i, j) = \exp(-\gamma(r_{ij} - \mu_t)^2), \quad (4.3)$$

where μ_t the center of the t -th Gaussian with width γ and $r_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ guarantees translation and rotation invariance of the GNN model. For the iron system, edge features are sliced into 20 dimension (details are provided in Appendix).

Crystal structures of glycine polymorphs were obtained from the Crystallography Open Database. [201] In a similar manner as for iron, different supercells consisting of 128 molecules as α -glycine [202], 128 as β -glycine [203], 108 as γ -glycine [202], and 137 as liquid glycine (denoted as ℓ -glycine) were built using unit cells of glycine polymorphs prepared with the Mercury package. [204] MD simulations of all glycine polymorphs were carried out at 300 K ; while the liquid glycine was obtained at around its melting temperature of 500 K . All MD simulations are of length 2 ns and 2000 frames were obtained from each simulation for training the model. Every glycine molecule is treated as one entity with the corresponding center of mass (i.e., one node per molecule; black dots in Fig.4.2b)) whose feature is 1. The leading 6 nearest neighbors are defined as neighboring molecules referring to the position of the center of mass of each molecule. Different from iron, edge attributes are concatenations of four intermolecular angles under Gaussian basis function into one long feature vector as shown in Fig.4.2b) bottom right.

The angles are chosen for the molecular system in the light of physical intuition and also in light of previous work. [32, 105] They are defined by the intramolecular vectors: ν_1 is the C-C $_{\alpha}$ vector, ν_2 is the N-C $_{\alpha}$ vector, ν_3 is the C-N vector and ν_4 is the vector of C $_{\alpha}$ and the geometric center of the two associated hydrogens (see Fig.4.2b) for illustrations). Similarly, all intermolecular angles are expanded into 10 slices which leads to 40 edge features for each node in total.

4.2.3 Well-Tempered Metadynamics

The low-dimensional latent variables learnt through the global mean pooling operation (Fig. 4.1) serve as low-dimensional descriptors of various competing phases. Due to the loss function in Eq. 4.2 these capture both local and global information, making them well-suited for driving short-range and long-range fluctuations relevant to nucleation. Here we do so by performing well-tempered metadynamics [42] along these variables, while expecting that our protocol should be fully amenable to other enhanced sampling approaches.

In WTmetaD, history-dependent Gaussians are deposited along pre-defined biasing variables reaction coordinates to facilitate state-to-state back-and-forth transitions between different metastable states the system would normally be trapped in. We refer to Ref. [42] for further details of WTmetaD. Here we used the latent variables (z_1, z_2) from Fig. 4.1 as the variables being biased. Other parameters used in performing WTmetaD simulations are reported in Tab. 5.1. Iron nucleation simulations were performed with LAMMPS-23Jun2022 simulator [205] and glycine simulations were performed with GROMACS-2021.6 MD engine. [206] Both packages were patched to PLUMED-2.8.1 with the Pytorch module enabled. [118, 119] Codes for reproducing the simulations in this chapter are available at Github.

Table 4.2: WTmetaD Parameters

System	$\omega(k_bT)$	γ	σ_1 (RC unit)	σ_2 (RC unit)	T (K)	pace
Iron	1.0	50	0.6	0.6	1800	500
Glycine	2.0	100	0.15	0.1	500	500

4.3 Results and Discussions

We evaluate the ability of the GNN-learnt low-dimensional latent representations to enhance sampling by performing well-tempered metadynamics for the two selected representative systems, namely, iron (Sec.4.3.1) and glycine (Sec.4.3.2) initiated from their molten or liquid phases. Iron as one of the most abundant elements on Earth has received significant interest given its importance in steels and alloys and in geology. Many allotropes of pure iron exist, which are the body centered cubic (BCC) α -Fe, the face centered cubic (FCC) γ -Fe, the hexagonal close packing (HCP) ϵ -Fe and the BCC δ -Fe. In addition to iron, we assess the reliability of our protocol on the nucleation of polymorphs of the simplest amino acid glycine. This is an important system as physical properties in different glycine polymorphs vary which can change the effect of glycine as an inhibitory neurotransmitter. [207] Furthermore, the existence of many possible space groups complicates the problem in molecular crystals in general. [208] Three polymorphs, namely α -, β -, and γ -glycine, exist in zwitterionic glycines at ambient conditions. In particular, α -glycine (space group: $P2_1/n$) and β -glycine ($P2_1$) are monoclinic and γ -glycine ($P3_1$) is trigonal. We compute free energy differences between these different metastable allotropes/polymorphs and compare with respective literature.

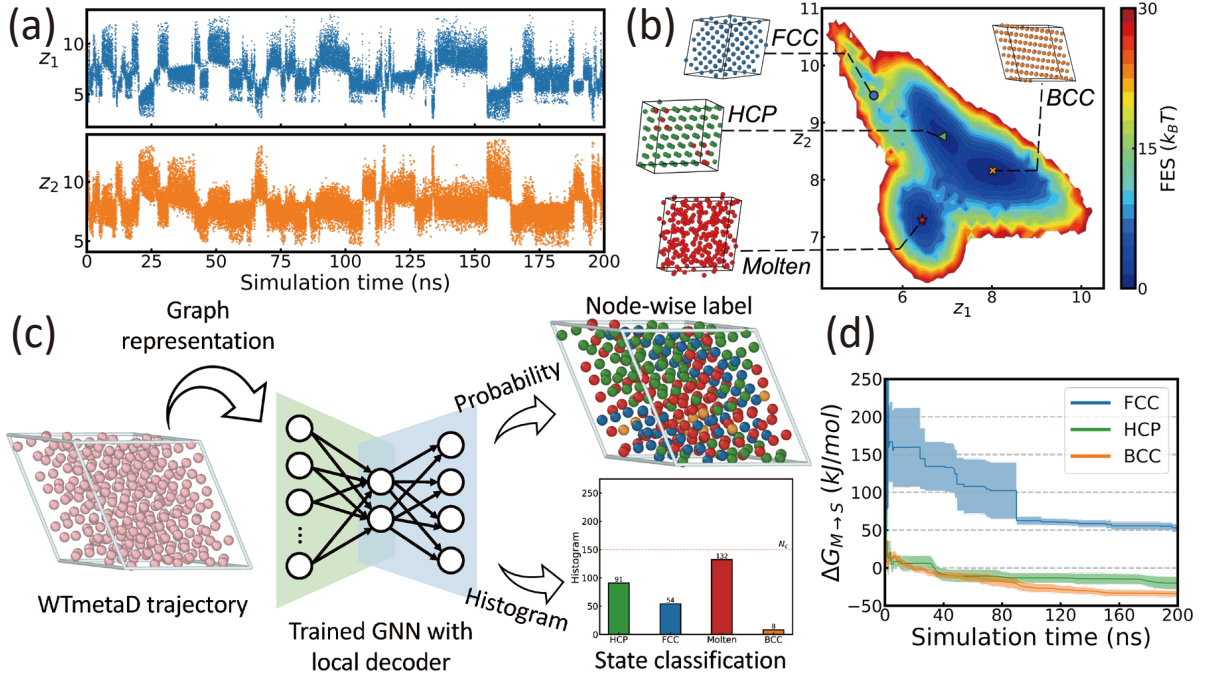


Figure 4.3: Results of WTmetaD simulations of iron nucleation from its melt. Machine learnt latent representations z_1, z_2 as a function of simulation time in panel a) shows frequent state-to-state transitions. Subplot b) shows reweighted FES (free energy surface) in $k_B T$ in the latent variable space (z_1, z_2) with landmarks of sampled configurational snapshots from OVITO package. [35] Iron particles are color-coded with respect to the adaptive common neighbor analysis method. [209] Subplot c) shows the workflow of post-processing the obtained WTmetaD trajectories by passing them through the full GNN model with the local decoder. The state labels are generated from the output prediction and consequently applied for computing the time series of free energy difference of solid states of interest (S) with respect to the molten phase M in subplot d). Computed standard errors over four individual production runs are represented as shaded regions.

4.3.1 Phase Transitions in Pure Iron

Experimental measures suggest α -Fe remains stable at temperature less than 1184 K ; while γ -Fe is stable between 1184 K to 1665 K . When above 1665 K , γ -Fe transforms to another BCC structure, δ -Fe until sublimation. In addition, the HCP ϵ -Fe is stable at a pressure greater than 13 GPa . [210,211] The enrichment of crystal packings makes iron a challenging system and an excellent test case for our proposed protocol.

The graph data for iron particles is shown in Fig.4.2a). Since the system of interest here is in single species, the feature of individual nodes is set to be 1. We set a k-NN cutoff ($k = 50$) that is much larger than the coordination number (number of neighboring particles in the first shell) of close packing structures. This helps gain information from nodes several hops away in order to account for the structural similarities in FCC (abcabc...) and HCP (ababab...) crystals. The edge features are the Gaussian expanded radial distance as introduced in Methods (Sec. 4.2.2). The edge feature is an expanded radius distance in 20 dimensions (details on the slicing of features are provided in Appendix).

We find that well-tempered metadynamics biasing along the 2-dimensional latent variable learnt from GNN leads to robust, multiple state-to-state transitions without any hysteresis, as can be seen from the time series in Fig.4.3a). The associated free energy surface is then constructed following the appropriate reweighting scheme. [47] Three distinct energy basins correspond to the initial molten state, HCP iron, and BCC iron (Fig.4.3b)). On the contrary, no distinct basin is sampled for the FCC state of iron, which suggests such a configuration is thermodynamically less stable at the simulation temperature of 1800 K . This is expected because FCC is reported to be the least stable allotrope among other forms from both experiments [212] and zero temperature

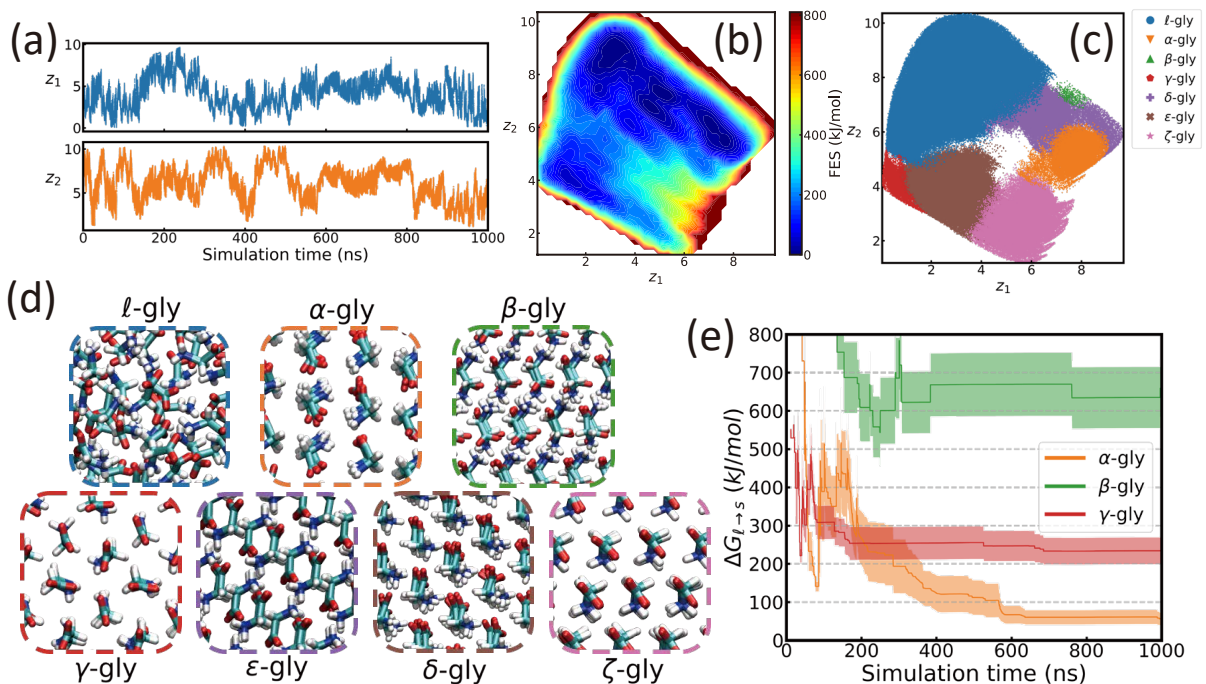


Figure 4.4: Results of WTmetaD simulations of the nucleation of glycine zwitterions. a) shows machine learnt latent representations z_1, z_2 vs. simulation time. Subplot b) shows reweighted FES (free energy surface) in the latent variable space (z_1, z_2) with multiple basins observed. Subplot c) is the scatter plot of liquid and solid glycines in (z_1, z_2) space classified by post-trained GNN model, and d) consists of snapshots to each corresponding class rendered from Visual Molecular Dynamics (VMD). [36] Free energy difference between states of interest vs. simulation time in e) shows convergence of the production runs. Standard errors are represented as shaded regions.

calculations [213, 214].

An advantage of the proposed ML model is that its local decoder (Fig.4.1) provides an estimation, in a probabilistic sense, of the class or crystalline structure of individual nodes belonging to. [89, 215, 216] The populations of different crystal packings can then be computed by the summation of individual weights from the output Y of the local decoder, $N_{class} = \sum_i^N Y_{i,class}$ as shown schematically in Fig.4.3c). Additionally, we present results in the form of a 2-dimensional free energy surface in the Appendix for indicators of potential solid-solid transitions. The free energy difference ΔG can be computed between phases and, as an example, the equation between

liquid, ℓ , and solid, s , is shown as follows:

$$\Delta G_{\ell \rightarrow s} = -k_B T \ln \left(\frac{\langle H(N_s - N_c) \rangle}{\langle H(N_\ell - N_c) \rangle} \right), \quad (4.4)$$

where $\langle \cdot \rangle$ is the reweighted unbiased average and H is the Heaviside step function with a size cutoff N_c . [170] The value of N_c should be properly tuned in order to avoid assigning coexistence states (i.e., a mixtures of multiple configurations) as perfect crystals. The tabulated free energy differences between metastable crystalline states and the starting molten state are plotted in Fig.4.3d) with a threshold value of 150 in Eq.4.4, which means frames are categorized into the corresponding state for free energy computation when $N > 150$ (within a total of 285 Fe irons in the box). We also report values of free energy differences under larger N_c thresholds in Appendix. In addition, Fig.4.3d) suggests that the WTmetaD simulations performed are well-converged. The values were averaged over four independent production runs of 200 ns . The thermodynamically most stable configuration of iron is its BCC form, and the HCP form is shown to be less stable than BCC. The least stable form of iron at its melting temperature is the FCC configuration. This is validated in Ref. [212], the experimental measured free energy difference between BCC and FCC iron, $\Delta G_{BCC-FCC}$, is 0.082 eV and $\Delta G_{FCC-HCP}$ is -0.023 eV at room temperature. In addition, zero temperature calculations [213, 214] also show similar measures which are in good agreement with the stability rankings obtained in this work, BCC > HCP > FCC iron in decreasing order.

4.3.2 Nucleation of Glycine from Melt

The developed model is then assessed to a more complicated molecular system, glycine. Geometric data for zwitterionic glycine is slightly more intricate than that of iron (see [Methods](#) for detailed information). As more degrees of freedom are incorporated into the graphs of glycine molecules, a smaller neighborhood of individual molecules is defined with only 6 closest neighbors ($k = 6$). Intermolecular angles of characteristic vectors, ν_1 , ν_2 , ν_3 , and ν_4 are again expanded under basis functions which leads to 40-dimensional edge features between linked nodes.

[Fig. 4.4](#) summarizes results from WTmetaD simulations biasing latent representations z_1, z_2 learnt by the GNN model. Several transitions can be identified by evaluating the time series in subplot a) along with multiple distinct wells in the reweighted free energy surface of (z_1, z_2) space ([Fig. 4.4b](#)). However, after closely examining the obtained trajectories with visualization tools, we found that more polymorphic glycines were sampled even though the model was trained only on solid glycine in its three well-studied ambient products synthesized experimentally. These new polymorphs that our simulations visit have been however reported previously in Ref. [\[202, 217\]](#) as high-pressure structures. We thereby trained a new GNN model as a classifier on liquid glycine and all associated crystals including structures found under the effect of different pressures, and this leads to in total of six polymorphs namely α -, β -, γ -, δ -, ϵ -, and ζ -glycine. The notations for the solid glycines are adopted from Ref. [\[217\]](#). Here, we only briefly describe the three additional glycine polymorphs since they are not the main focus of this chapter and their relative stabilities remain unclear: α -glycine remains stable to pressures up to 23 *GPa*, meanwhile, β -glycine undergoes phase transition to δ -glycine ($P2_1/c$) at 0.8 *GPa*, γ -glycine transforms to ϵ -glycine (Pn) under application of pressure and the process complete at between

4 to 5 *GPa*, and decompression of ϵ -glycine leads to ζ -glycine (*I1*). [217, 218]

The distributions of individual glycine polymorphs are shown as scatter points in (z_1, z_2) space in Fig. 4.4c) with cutoff value $N_c = 65$, and the corresponding snapshots are provided in Fig. 4.4d). All observations above suggest the trained GNN latent representations capture the configurational information among various polymorphic structures of glycine molecules. Structures which are close to each other in configuration space (i.e. the associated high-pressure components) are sampled at ease with robust WTmetaD simulations.

The free energy difference $\Delta G_{\ell \rightarrow s}$ as a function of simulation time shown in Fig. 4.4e) is computed from Eq. 4.4 with a value of cutoff 65 which is set slightly larger than half of the population of glycine in the simulation cell (free energy differences with other N_c values are reported in the Appendix). The figure is averaged over 9 independent runs of 1000 *ns* (see the Appendix for full $\Delta G_{\ell \rightarrow s}$ vs. simulation time plot). This indicates the thermodynamic stability of ambient glycine polymorphs ranks in α -gly > γ -gly > β -gly at 500 *K* which is in consistent with the stability ranking reported in the literature which γ -gly is the most thermodynamically stable at ambient temperature and the densest α -gly becomes the most stable at temperatures above 440 *K*. [218] In addition, we also identified a transition pathway of γ -gly to ϵ -gly to ζ -gly, while no direct transitions from γ -gly to ζ -gly, supported by 2-dimensional free energy analyses (see Appendix for details), which was also observed and reported experimentally. [219] Overall, this shows that the trained model is robust in learning structural properties for classifying configurations with simple graph convolution operations and readily computable features. Benefiting from the WTmetaD method, introducing biases along these low-dimensional latent representations validates the feasibility of obtaining relative free energy differences between competing allotropes/polymorphs starting only from their chemical identity and possible target structures.

4.4 Conclusion

Computational methods for investigating crystal nucleation have recently shown their strength in providing high temporal- and spatial-resolution descriptions. [3, 126] Due to the timescale limitations resulting from the rare event nature of nucleation, it is however necessary to perform enhanced sampling molecular dynamics. Most enhanced sampling methods involve biasing or following selected low-dimensional descriptors, and methods of constructing these descriptors remains to be an active field of research. Recent advancements in machine learning, particularly graph neural networks, have made it possible to achieve a better understanding of crystal nucleation from a perspective of learning relative slow modes [220], or on efficient computation of order parameters [185].

In this work, we have introduced a data-driven GNN-based representation learning model within an autoencoder framework to extract low-dimensional variables from configurational features found in experimental crystal structures. These variables serve as the key inputs for performing enhanced sampling methods. Our method employs straightforward convolutions and pooling techniques. To validate the usefulness of our machine learning pipeline, we studied nucleation of different allotropes and polymorphs of iron and glycine respectively from the melt. We biased the GNN based latent variables in well-tempered metadynamics and were able to achieve multiple back-and-forth state-to-state transitions and converged free energy estimates, both hallmarks of reliable sampling. This proves the robustness and potential of our graph neural network learnt variables for enhanced sampling. The thermodynamic stability rankings among allotropes or polymorphs are in agreement with experimental measures.

The protocol proposed here can be further improved in many ways, and here we highlight

some possible avenues for future research. The current network only consists of convolution and pooling operations and it should be possible to introduce advanced manipulations on graph data, such as the attention mechanism, to better interpolate and even extrapolate configurational information of complex species. [141, 190] Secondly, an obstacle to drawing quantitative comparisons to literature is due to finite size effects, which have been investigated under a variety of systems including atomic and molecular. For a more comprehensive summary in finite size effects, the readers is referred to Ref. [3]. To address this issue, various methods have been developed, including the introduction of constant chemical potential ensemble [221] and analytical corrections to single-species [222, 223] and solvated systems [111, 127, 130]. In future work, we hope to combine data-driven approaches with those from statistical mechanics [224] to not just automate biasing variable discovery, as was done in this work, but also address finite size effects.

With GNN-based model, high-dimensional graph data is directly taken as input and used to construct latent representations which are able to distinguish states of interest. However, graph data by construction only considers configurations properties and does not provide dynamic information into account. Therefore, we the next chapter, we combine the GNN model developed in this chapter onto the existing SPIB model, which allows the model to learn both conformational and dynamical information while learning.

Chapter 5: On Graphical Representations and Slow Modes Identifier for Enhanced Sampling Methods

5.1 Introduction

In this chapter, we continue our investigation on finding optimal collective variables for enhanced sampling methods using GNN. Although GNNs alone can capture key conformational information of systems of interest and yield good thermodynamic measurements, the GNN model itself cannot capture any dynamical information given its construction. To circumvent this obstacle, we improve upon our model from the previous chapter by combining the existing GNN model with methods that learn kinetic information. The model is the SPIB model which are introduced in Sec. [3.2.2](#) in Introduction.

While SPIB requires the input to be in low-dimensional CVs space – which is exactly what the GNN model output after its pooling operators – the combination of these two methods, should therefore fulfill the requirements of an optimal reaction coordinate. In this chapter, we will present our studies on three model systems and examine the quality of samplings when biasing along the variables learned by this combined model.

5.2 Methods

5.2.1 GNN-based SPIB

While both the GNN model and the SPIB model are introduced in previous chapters, our goal in this chapter is to combine the two methods. This integration enables the new model to learn information from both the structural changes in the system and the dynamics of the transitions. This results in a direct concatenation of the two models, meaning the system will first be converted into its geometric representations with an assigned state label for the GNN model. The GNN model extracts key orientational information from high-dimensional graph inputs and then provides graph-level features in relatively low dimensions to the SPIB model. The SPIB model can then extract dynamical information using the output embeddings from the GNN model. For reaction coordinates in enhanced sampling methods, the decoder is discarded, leaving the model as a combination of the GNN and the encoder of the SPIB model. A schematic is provided in Fig. 5.1.

5.2.2 Well-Tempered Metadynamics

Again, we evaluated our dimensionality reduction method by performing well-tempered metadynamics simulation. The method itself has been thoroughly introduced in earlier chapters and here we provides the parameters in Tab. 5.1 for performing the simulations instead.

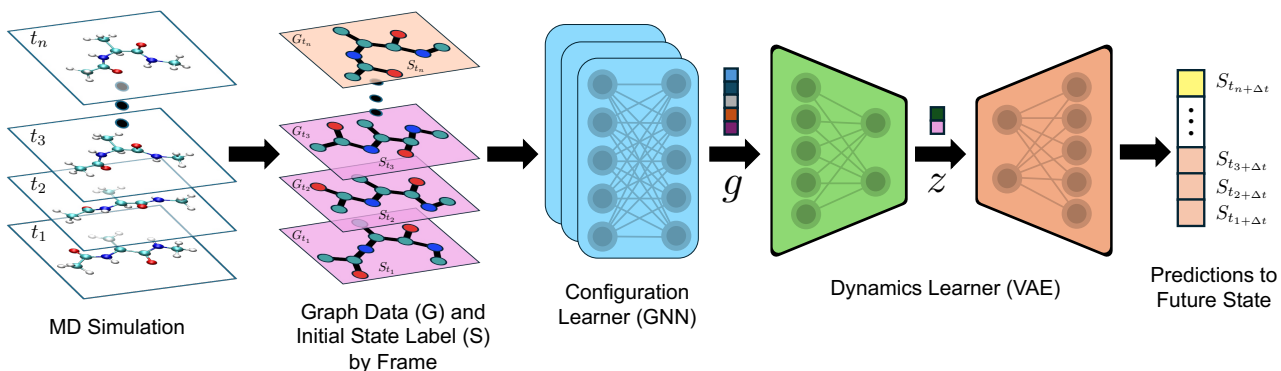


Figure 5.1: A schematic of training the proposed GNN-SPIB model. The input trajectory is first converted into graph objects by frames and then fed into the GNN model. Through multiple message passing, aggregation, and update functions, a graph-level embedding in low-dimension is obtained after the global pooling layer. The following SPIB model can then learn the dynamics under past future information bottleneck framework. The entire framework is used for training and the decoder part of the SPIB model will be removed when using in enhanced sampling methods.

5.3 Results and Discussions

We evaluate the ability of the machine-learnt low-dimensional latent representations to enhance sampling by performing well-tempered metadynamics for the three model systems, they are Lennard Jones 7 (LJ7, in Sec.5.3.1), alanine dipeptide (AD, in Sec.5.3.2) and alanine tetrapeptide (Ala₃, in Sec.5.3.3). All three systems have been studied in detail which therefore provide

Table 5.1: Well-Tempered Metadynamics Parameters.

System	$\omega(k_B T)$	γ	σ_1	σ_2	$T(K)$	pace
LJ7*	0.1	5	0.5	0.5	0.1	500
Ala ₁	0.5	8	0.5	0.5	300	500
Ala ₃	0.5	5	0.2	0.2	350	500

*LJ7 system is in reduced Lennord Jones units.

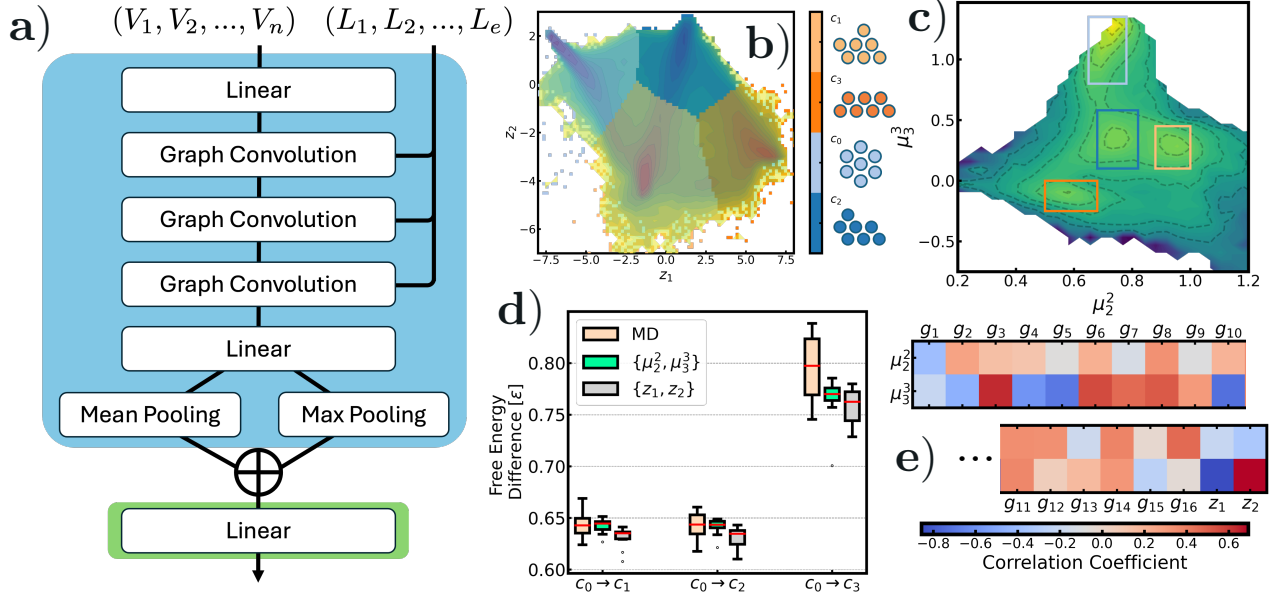


Figure 5.2: Summary of WTmetaD simulation results biasing along machine learnt reaction coordinates, z_1 and z_2 , in Lennard Jones 7 system: a) a schematic of how the 2-d reaction coordinate is computed with node features $\{V_n\}$ and edge features $\{L_e\}$ while $\oplus$ denotes for concatenate operation; b) state labels predicted by the trained model in RC space; c) reweighted free energy surface projected onto good CV space, μ_2^2 and μ_3^3 with state definitions in colored boxes; d) box plots of free energy differences from c) between sampled metastable states comparing among conventional long MDs and WTmetaDs biasing good CVs; and e) correlation coefficients of model embeddings to the moment variables.

sufficient information while we compare sampling results under various schemes. As a general pipeline to the three systems, we first train the models with data collected from short MD simulations at higher temperatures in which all targeted metastable states are sampled but likely not converged. We then perform WTmetaD simulations biasing along the learned embeddings of the network, compute free energy difference between states, and compare with results from long MD simulations and WTmetaD simulations biasing conventional good CVs.

5.3.1 Lennard Jones 7

As suggested by its name, Lennard Jones 7 (LJ7) consists of a cluster of 7 Lennard Jones particles in 2 dimension and there exist 4 configurations namely c_0 (hexagon), c_1 (capped parallelogram 1), c_2 (capped parallelogram 2), and c_3 (trapezoid) (see Fig.5.2c) for particles' arrangements). Given its simple construction, the dynamics and landscape of LJ7 are well-studied which makes it an excellent model system for benchmarking enhanced sampling methods of rare events. [225] The good CVs which are usually considered as reaction coordinates to this system are the second moment, $\mu_2^2 = \frac{1}{N} \sum_{i=1}^N (c_i - \bar{c})^2$ and third moment, $\mu_3^3 = \frac{1}{N} \sum_{i=1}^N (c_i - \bar{c})^3$, of coordination number, c , where $\bar{c}$ is the mean of c and N is the number of particles. With these definitions, one may see the key to classifying metastable states in LJ7 system is the fluctuation and distribution of coordination numbers of individual LJ particles, rather than the ensemble average, which is a typical candidate collective variable to consider in colloidal systems.

Fig.5.2 summarizes the results when biasing along the 2-dimensional machine learnt embeddings, z_1 and z_2 . The input graphs are composed of identical node attributes $\{V_n\} = \{1, 1 \dots 1\}$ and pair distance edge features $\{L_e\}$. The graph convolution layers are directly adopted from Ref. [226] and graph embeddings are pooled with mean and max operators (see Fig.5.2a)). The resulting features are then fed into the SPIB model for extrapolating dynamic information (in green). Fig.5.2b) shows the predicted state labels in colored regions of input training data and four distinct local minima are observed. The training data is a trajectory of 1×10^7 steps at a higher temperature $k_B T = 0.2\epsilon$.

WTmetaD simulations were performed biasing these 2-d reaction coordinate, z_1 and z_2 , and resulting free energy surfaces are shown in Fig.5.2c). For better evaluation practice to the

sampling quality, we projected the free energy surface onto the more meaningful moments space. Four distinct local minima were sampled which is always an indicator of good samplings. To further evaluate the sampling quality, we tabulated the free energy differences between the sampled configurations to the initial state, c_0 . To benchmark, we performed additional 10 individual production runs in long MD simulations and 10 WTmetaD simulations biasing good CV set $\{\mu_2^2, \mu_3^3\}$ along with 10 WTmetaD simulations biasing the machine learnt 2-d CV $\{z_1, z_2\}$, while MDs last for 1×10^9 steps and WTmetaDs last for 1×10^8 steps. As reported in Fig.5.2d), the machine-learned CV shows comparable results to the conventional CVs, noticing that no information about the coordination was directly fed into the model as inputs but only pairwise distances of neighboring nodes. However, several latent variables (e.g. g_3, g_5, g_6 , and g_{10}) seem to correlate strongly to especially μ_3^3 as suggested by the correlation coefficient plot on Fig.5.2e). Not surprisingly, the 2-d RC, z_1 and z_2 constructed in forms of linear combination of graph embeddings $\{g_i\}$ correlates to the two known good CVs as well.

5.3.2 Alanine Dipeptide

If the rearrangement of LJ7 cluster is considered a simple model to the dynamics of colloidal system, alanine dipeptide (AD) molecule is an excellent toy model to biopolymer conformational changes. Here we focus on transitions in vacuum between three conformers, $C5$, $C7_{eq}$, and $C7_{ax}$. The set of good CVs commonly used in enhanced sampling methods is dihedral angles ϕ and ψ (Fig. 5.3a)). [39, 227] To our model, different from the previous example, the three elements (excluding non-heavy hydrogens) in alanine dipeptide are one-hot encoded as in their node features, and again we use as edge features the inter-atomic distances as in LJ7 model.

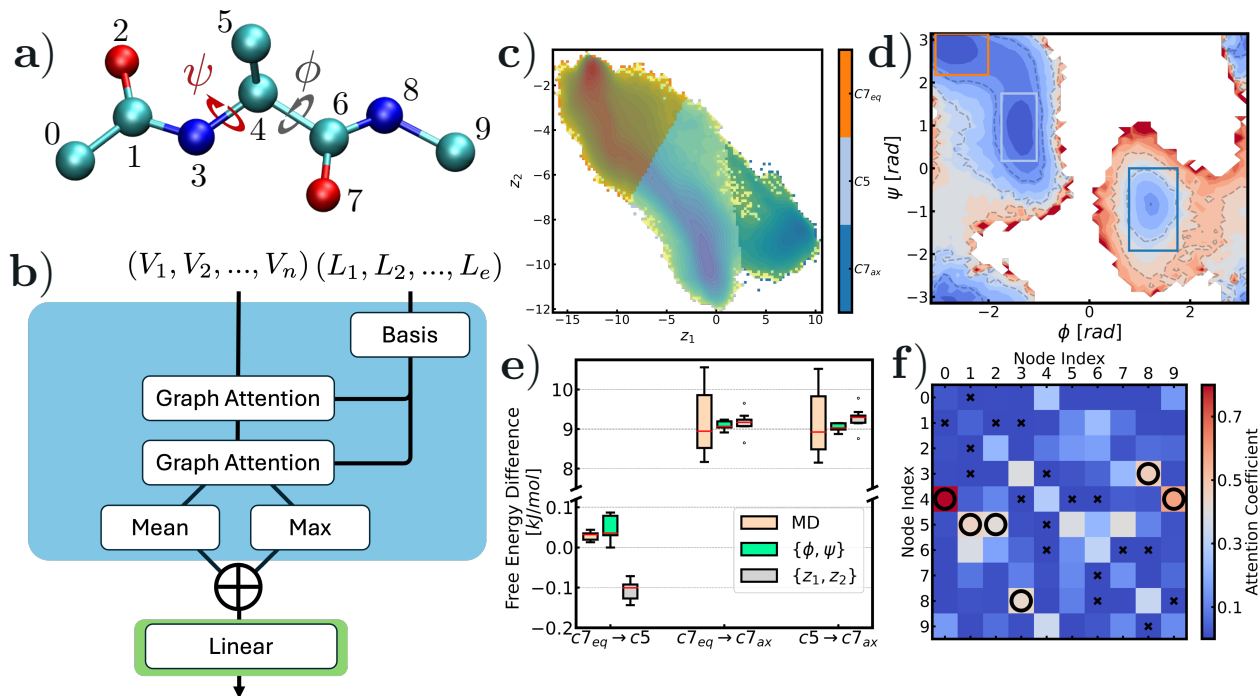


Figure 5.3: Summary of WTmetaD simulation results in alanine dipeptide system: a) representation of alanine dipeptide molecule with definition to good CV, ϕ and ψ . Graph representation is constructed only with heavy atoms and atomic labels are followed by assigned node index to graphs; b) reweighted free energy surface biasing the machine learnt RC projected onto (ϕ, ψ) space with boxed state definitions; c) a schematic of the reaction coordinate with node $\{V_n\}$ and edge $\{L_e\}$ features; d) state label prediction in different colors from the model decoder; e) free energy differences with state defined in c) under different sampling schemes; and f) attention coefficient retrieved from the last graph attention network convolutional layer with $\times$ denotes for bonded nodes and $\circ$ highlights the top 6 edge pairs in attention weights.

However, to increase the expressiveness of the model, we applied a basis function to slice the edge feature. [189]

The overall architecture remains the same to the study of LJ cluster, while we swapped the graph convolution layers to graph attention network (GAT) convolution layers. [228, 229] Three distinct minima shown when projected onto the learned latent space with input data sampled at $T = 400K$ (Fig. 5.3c)) and reweighted free energy surface on (ϕ, ψ) space from WTmetaD simulations using z_1 and z_2 variables is shown in Fig. 5.3d) which all targeted states are well-sampled. The free energy differences between individual conformers were tabulated and benchmarked with

brute-force MD simulations and WTmetaD simulations using (ϕ, ψ) dihedrals (see Fig.5.3e)). The results are in good agreement with discrepancies less than 1 kJ/mol . Additionally, we retrieve the attention coefficients from the model (see heatmap in Fig.5.3f)) and the important edge pairs identified by the network correspond to atoms in the second shell of the radial distribution functions.

5.3.3 Alanine Tetrapeptide

In the third example, we study the conformational changes in the alanine tetrapeptide (Ala_3) in vacuum. As a more complicated model system to alanine dipeptide, there exists 8 metastable states in total while we follow the notations in Ref. [114] and these states of interest are defined as $s_i, i = 1 \dots 8$. In particular, 6 dihedral angles are considered important and they are $\phi_1, \phi_2, \phi_3, \psi_1, \psi_2$, and ψ_3 (see Fig. 5.4c) for their definitions). With a similar architecture to previous studies, the Ala_3 molecule is first converted into graph objects and all hydrogens are removed while building the graphs. Nodes inside the graph are defined to be fully connected and edge features are interatomic distances. We adopted a skip connection scheme which allows information from all graph layers to flow to pooling operator which therefore can possibly improve the expressiveness of the model in capturing features from all metastable states. Additionally, as shown in Fig. 5.4a) we considered all three typical pooling operations, namely mean, max and sum operations in this example.

The input training data is projected onto the trained latent space (z_1, z_2) in Fig. 5.4b). As presented by the color coding, the model was successfully optimized to learn 7 out of 8 metastable states, while s_6 and s_7 states remain barely distinguishable from each other. This is due to the

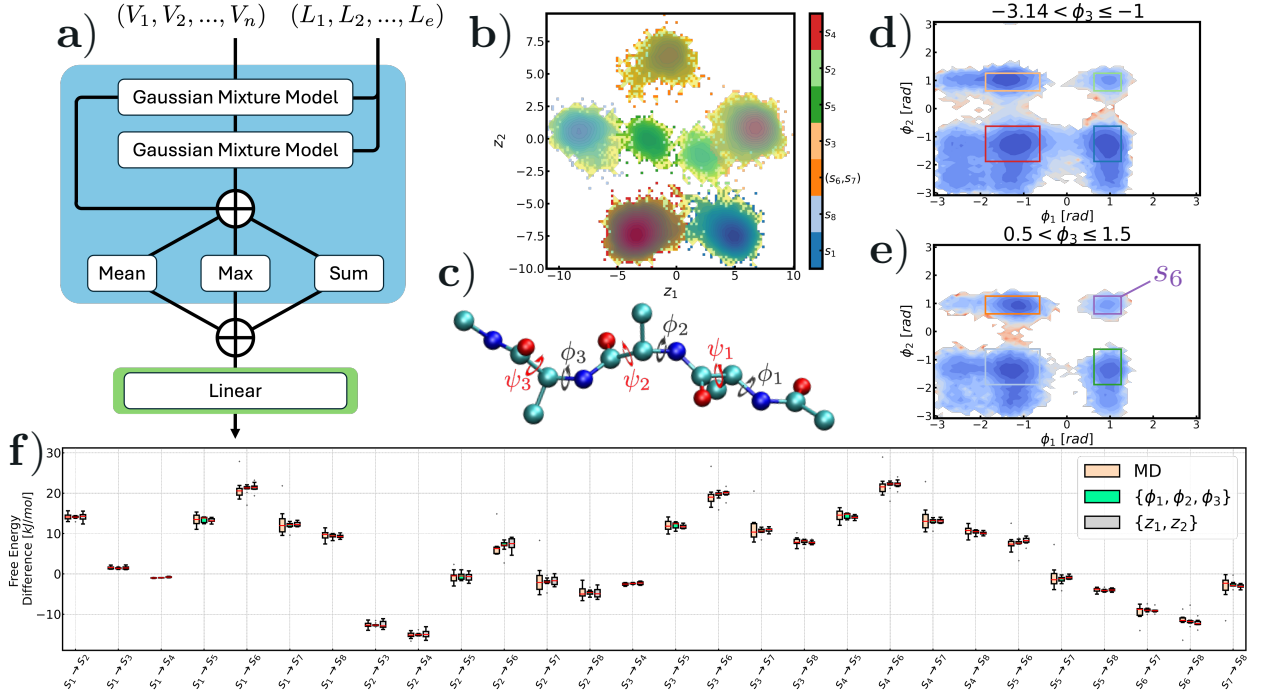


Figure 5.4: Summary of WTmetaD simulation results in alanine tetrapeptide system: a) a schematic of the reaction coordinate and embeddings of each graph convolution layers are collected for graph-level pooling via skip connections; b) the learnt latent variable space (z_1, z_2) with state labels predicted by the model; c) representation of alanine tetrapeptide molecule with definition to characteristic dihedral angles, $\phi_1, \phi_2, \phi_3, \psi_1, \psi_2$, and ψ_3 and only heavy atoms are considered during graph construction; d) reweighted free energy surface of WTmetaD simulations using 2-d (z_1, z_2) variables projected onto (ϕ_1, ϕ_2, ϕ_3) space; and f) tabulated free energy differences between all metastable state from brute force MD simulations, WTmetaD simulations biasing (ϕ_1, ϕ_2, ϕ_3) , and WTmetaD simulations biasing (z_1, z_2) .

lack of samples in s_6 provided while training the model. Our input training data is a $1\mu s$ long MD simulation at $400K$ where s_6 is rarely sampled. Nevertheless, performing WTmetaD simulations using (z_1, z_2) for $200 ns$ does drive the system to visit all 8 metastable states in alanine tetrapeptide. To show the sampling quality of our simulations, we project the trajectory onto (ϕ_1, ϕ_2, ϕ_3) space and label all target states in Fig. 5.4d). The free energy differences between all 8 states are computed in box plot in Fig. 5.4e) along with two benchmarks: standard long MDs and 3-d WTmetaDs using (ϕ_1, ϕ_2, ϕ_3) variables. The values of free energy differences are converged to well-defined values align to the those from benchmark methods.

5.4 Conclusion

Even though derivation of various enhanced sampling methods enable molecular dynamics simulations to overcome their timescale limit, many still suffer from finding the optimal coordinates and this remains to be an active field of study. In this chapter, we combine two powerful machine learning techniques, namely graph neural networks and variational autoencoder to learn meaningful low-dimension representation of complex system. With this architecture, it allows the model to not only learn configurations of the studied system from the graph layers, but dynamics from the past future information bottleneck.

We benchmarked this approach on three model systems across the field of material science and biophysics, and they are the Lennard Jones 7 cluster, alanine dipeptide, and alanine tetrapeptide. As shown in Sec.5.3, initiating WTmetaD simulations using these machine learnt reaction coordinates present comparable results to those when basing conventional CVs. Given its simple construction using input data, namely pairwise distances, we believe this protocol can potentially be useful when applied to complex systems where the optimal reaction coordinates for enhanced sampling methods are not *a priori* known.

As one member of the RAVE family, [59,85] the learnt reaction coordinate can be improved iteratively with better sampled input data as shown in Ref. [230] especially when targeting complex systems this becomes crucial. On the other hand, while in this chapter we limit ourselves to only type-0 representations (i.e. pairwise distances) when constructing graphs, high-order representations like angles, torsions, and even spherical harmonics should be adopted. [60] Moreover, inputs to model training are not limited to unbiased/biased simulations with frequent state-to-state transitions. Static discrete data from metastable states drawn from databases can also lead

to meaningful latent representations. [\[231–234\]](#)

Chapter 6: Conclusion and Outlook

Crystal nucleation, a common phenomenon in nature, significantly impacts the fields of materials and pharmaceutical industries. However, many underlying mechanisms in crystal nucleation remain unexplained, often limited by the lack of resolution in experimental instruments. Although computer simulations may offer the necessary resolutions, the inherent many-body characteristic and chaotic, high-dimensional dynamics in nucleation make this process challenging to be sufficiently sampled and therefore difficult in performing meaningful measurements using standard molecular dynamics simulations. Therefore, numerous enhanced sampling methods have been developed, allowing researchers to enhance their sampling of rare events, such as nucleations, with metadynamics being one such method. In this dissertation, I combined the enhanced sampling method, metadynamics, with machine learning methods to study various systems under different conditions. I successfully drove the system to sample the target solid states and performed thermodynamic measurements of these systems. These measures were then compared with the literature and showed qualitative agreements. Additionally, I attempted to explain some of the observed nucleation mechanisms with the learned reaction coordinate, hopefully providing insights into the nucleation processes.

In the first part of the dissertation, I applied two machine learning methods, SGOOP and SPIB, to study urea nucleation in vacuum and water, respectively. In vacuum, I managed to sam-

ple multiple transitions between the ambient product of urea molecules and its high-pressure form. With an optimal reaction coordinate separating the relatively slow modes learned by SGOOP, the constructed free energy surface exhibited a distinct energy barrier, enabling computation of free energy differences and transition pressure. After comparing these values to experimental measures, they were found to be in agreement, suggesting that the pipeline introduced in Chapter 2 can indeed be applied to the study of crystal nucleation, where energy barriers are commonly larger than hundreds of kJ/mol. Subsequently, I shifted focus to the study of urea nucleation in water. Here, I successfully sampled all polymorphs of urea studied in vacuum with robust back-and-forth transitions between them. This serves as strong evidence that the machine learning method captured slow modes of the homogeneous nucleation of urea in water. However, limited by the size of my simulation cell, thermodynamic measures experienced strong finite-size effects, leading to the computed free energy surface showing no minima besides the initial solvated state.

Later on, I also considered the possibility of surpassing the selections of hand-crafted collective variables, which are essential to dimensionality reduction methods like SGOOP and SPIB. Even though many hand-crafted variables reveal underlying physics of processes of interest, their low-dimensional nature makes it difficult to apply them to different systems. Therefore, in Chapter 4, I introduced another machine learning method that directly takes high-dimensional features into account and learns meaningful low-dimensional representations as a nonlinear combination thereof. This method is achieved with the use of graph neural networks (GNNs). GNNs allow us to consider properties from individual nodes, and the exchange of information occurs only within the defined local neighborhood. This construction is very similar to those of hand-crafted collective variables in the study of crystal nucleation mathematically. I therefore examined my method

with two systems: nucleation of iron allotropes and glycine polymorphs. The results in Chapter 4 show that my simulations were nicely converged, and distinct local minima were observed. I then computed their free energy differences and benchmarked them against their thermodynamic rankings from other studies, showing excellent agreements.

Although GNN-based methods provided reliable latent representations of the system of interest, one key piece of information is missing, i.e., the model can only consider static input, and there is no dynamical information taken into account. Therefore, a combination of GNN models with SPIB can perfectly resolve this issue, as GNN models are trained to learn local configurations from high-dimensional input graph objects, and then SPIB is trained to learn the relatively slow modes with readouts from GNN models. I examined my pipeline with three well-studied model systems: rearrangement in Lennard Jones 7 clusters, conformational changes in alanine dipeptide, and alanine tetrapeptide. To evaluate the performance of the learned latent variables, I performed additional WTmetaD simulations using the known good collective variables and compared the results from these two sets of simulations. The sampling results were then compared in their free energy differences between states. While all results were in agreement with the benchmark methods, I therefore concluded that my proposed protocol can indeed learn meaningful representations.

In summary, I first applied existing machine learning methods to the problem of crystal nucleation at an extreme, where very strong biases are required. Through these studies, although I was able to drive rare events, the quality of the sampling still hindered us from obtaining meaningful thermodynamic (or kinetic) measurements. I therefore sought methods to improve the proposed algorithm, and one of the key limitations of existing methods is their high reliance on the selection of input features. On that note, I introduced a GNN-based model to tackle this prob-

lem, which directly constructs low-dimensional representations from high-dimensional atomic coordinate space.

However, several improvements can be made in the future, and some perspectives could be considered. Firstly, the existing framework results in a brute-force concatenation of two machine learning methods, making the overall method somewhat cumbersome. To better merge the two methods in an elegant manner, insights from existing graph variational autoencoder architecture can be considered. [235] Secondly, in order to study complex macromolecules with larger size, graphs with all heavy atoms become unnecessary and computationally inefficient; instead, a better definition of these large proteins should be taken into account. For example, the nodes can be defined by only the $C\alpha$'s, and their residues become the node attributes or edge features. To an extreme, coarse-grain methods might be helpful in constructing informative graphs. Finally, one limitation of the current proposed framework requires the information of the final state, which is not always available when studying new systems. Therefore, it can be beneficial to combine the model with generative models and allow it to learn from similar configurations/graphs. Alternatively, the transferability of the current pipeline into similar systems should be examined.

Appendix A: Supplementary Materials

In this chapter, the supporting information of individual chapters is provided by section.

A.1 Chapter 2: Supporting Material to “Towards Automated Sampling of Polymorph Nucleation and Free Energies with SGOOP and Metadynamics”

A.1.1 SGOOP Input and Parameters

SGOOP requires a stationary probability estimate of the system along with dynamical constraint for Maximum Caliber principle. In general, the stationary probability is obtained from long unbiased or biased simulations, in which state-to-state transitions are sampled at least once. In this work, the input trajectory consists of 9 individual metadynamics simulations biasing S_{θ_1} and S_{θ_2} for 100 ns , concatenated with each other. The stationary probability was then obtained with appropriate reweighting protocol [47]. On the other hand, the dynamical information of the system is obtained from a 500 ns unbiased simulation initiated from liquid state. The probability distribution was constructed using between 20 to 40 bins and the probabilities smaller than a cutoff of e^{-50} , i.e., barriers higher than 50 $k_B T$ in free energy, are trimmed to avoid numerical issues. Thus the transition matrices constructed have the same size as the stationary probability distribution.

A.1.2 1-d Reaction Coordinate χ_4

As mentioned in main text, we also calculated a 1-d SGOOP RC χ_4 . The basinhopping algorithm [236] is applied to search for the maximum of spectral gap space. This optimized RC consisting of four OPs was found as $\chi_4 = 0.5S_{\theta_1} + 0.938S_{\theta_2} - 0.13\bar{\theta}_1 + 1.00\bar{\theta}_2$. Fig. A.2 shows time series from WTmetaD simulation along χ_4 . Frequent state-to-state transitions are observed at the first 200 *ns*. However the system is trapped in solid state for 200 *ns* and escapes the well at 480 *ns*. Fig. A.2 (b) is the reweighted free energy surface along RC χ_4 , where the two basins correspond to liquid and solid states of urea as labeled by the benchmark lines. Fig. A.2(c) shows an unreweighted exploration plot in averaged intermolecular angles $\bar{\theta}_1, \bar{\theta}_2$ space demonstrating that four regions are highly explored and they correspond to all four solid states, including initial liquid state (both phases A and IV are located at lower left region (refer landmark in main text for details)). In particular, the upper right region associated with a value of (0.39, 0.93) of $(\bar{\theta}_1, \bar{\theta}_2)$, is urea in phase I. Fig. A.2(d) is the reweighted free energy surface and only two wells are observed which are identified as form I and liquid. In summary, biasing χ_4 which excludes H and N does a slightly worse job of sampling than the χ_6 reported in main text.

A.1.3 1-d Reaction Coordinate χ'_6

Here we present another 1-d reaction coordinate, χ'_6 optimized using SGOOP with basin-hopping algorithm. Similarly, this 1-d reaction coordinate consists of six order parameters, S_{θ_1} , S_{θ_2} , $\bar{\theta}_1$, $\bar{\theta}_2$, N , H , and is given by $\chi'_6 = -0.787S_{\theta_1} - 0.461S_{\theta_2} - 0.813\bar{\theta}_1 + 1.0\bar{\theta}_2 - 0.383N + 0.076H$. This reaction coordinate has a spectral gap value of 0.2307 under 3-well assumption as shown in Fig. A.3 (a), which is slightly lower than the 0.2332 value for χ_6 reported in main

text and in next section. Actual coefficients for different order parameters are shown as a bar plot in Fig. A.3 (b) and to judge the importance of each component, in Fig. A.3 (c) coefficients are rescaled by reweighted standard deviations of input trajectories. Again, entropy order parameters are shown to be the most important features in driving the nucleation process of urea from the melt among other order parameters in the library. Time series of 1-d WTmetaD simulation biasing χ'_6 is shown in Fig. A.3 (d) in which transitions from the melt to form I and form IV are observed. From reweighted free energy surface averaged over 3 individual runs shown in Fig. A.3 (e), one distinctive barrier between liquid and solid state is sampled and there is no discernible barrier between forms I and IV. Fig. A.3 (f) is free energy difference computed with associated landmarks over time. One can also conclude that biasing this slightly sub-optimal SGOOP RC shows correct ranking of polymorph under experimental observations, whereas form I has low free energy than form IV. However, the large discrepancy (in terms of error bars) between the three individual trajectories leads to large uncertainty in Fig. A.3 (f), which suggests χ'_6 is a slightly poorer approximation to the reaction coordinate than χ_6 reported main text. Moreover, a large amount of defective structures are sampled while biasing along this reaction coordinate.

A.1.4 Further Analyses on χ_6

We rationalize the need for the introduction of averaged angle order parameters $\bar{\theta}_1$ and $\bar{\theta}_2$ with Fig. A.4. The time series of the two averaged angle order parameters utilized in the construction of SGOOP-based RC are shown in Fig. A.4(a) and (b). Followed by their associated free energies (Fig. A.4(c) and (d)). One can easily identify the form I and form IV by evaluating either sets of the analyses (time series or free energies). But, with only one of these order

parameters. Therefore, the free energy calculation applied in the main text are along $\bar{\theta}_1$ and $\bar{\theta}_2$.

The free energy surface along χ_6 through the input trajectories is shown in Fig. A.5 (a). Spectral profile of χ_6 is shown in Fig. A.5 (b) and the spectral gap equals 0.2332. In Fig. A.5 (a), three distinctive wells are identified. By appropriately assigning state labels from landmarks of polymorphs described in main text, the leftmost basin corresponds to form A, the middle one consists of form I and IV, and liquid state locates at the rightmost well. In addition, form B seems to located around transition state between the liquid state and collectively forms I and IV. As shown in Fig. A.3 (a), values of spectral gap for χ_6 and χ'_6 are close to each other. But their performances are different. While both gave correct polymorph ranking, simulations biasing χ_6 with higher spectral gap converged faster than those biasing χ'_6 . The 0^{th} and 1^{st} eigenvectors of the transition matrix are shown in Fig. A.5 (c) and (d), respectively, for further analyses. The 0^{th} eigenvector has no node and the 1^{st} eigenvector has one node located at the transition state of the deepest well, which further suggests χ_6 is a good approximation to the reaction coordinate.

A.2 Chapter 3: Supporting Material to “Driving and characterizing nucleation of urea and glycine polymorphs in water”

A.2.1 Sampling Efficiency Comparison

The sampling efficiencies of expertise-based and machine-learned (ML) reaction coordinates are listed in Tab. A.1. In particular, for the choice of expertise-based biasing variables, we used $(\bar{\theta}_1, \bar{\theta}_2)$ for the aqueous urea system and $(S(r, \theta_1), S(r, \theta_2))$ for the aqueous glycine system. Overall, the sampling efficiency of all polymorphs except the β form of glycine has been improved, demonstrating the advantage of the SPIB framework to construct a set of linear com-

binations of order parameters optimizing the distinguishability of covered states.

Table A.1: Sampling Efficiencies of Expertise-Based (EB) and ML RC in WTmetaD Simulations

System	Transition	EB (per 100 ns)	ML (per 100 ns)
Urea	L→I	0.2	0.8
	L→A	0.5	0.6
	L→B	0.2	0.4

* Solid solid transitions are counted separately, e.g. a transition of urea solution (L) → solid form 1 → solid form 2 is counted as L → S1 and L → S2.

A.2.2 Intermolecular Angle Formalization

Here we illustrate the reason why a switching function is used in the main text while defining intermolecular angle OPs, $\bar{\theta}_1$ and $\bar{\theta}_2$, in the **Order parameters for nucleation** section. The probability distributions of intermolecular angles of C-O and N-N vectors of urea are plotted with dashed lines in Fig.A.6. Since these angles are related to the local environment and are calculated over a large population, they can not directly serve as OPs in classifying states. Instead, we consider their average values as a global representation of the simulation box. As marked by the vertical lines, one is not able to identify urea solution and solid state on basis of the mean value of raw angles if directly used. Instead by the use of a switching function which we quantified in **Order parameters for nucleation** Eq. 5 in the main text we find that the resulting mean values are fairly distinct between urea solution and solid phases (solid lines in Fig.A.6.), thereby leading to much better demarcating OPs.

A.2.3 Classifying Urea Polymorphs

Here we describe the SMAC OPs mentioned in **Urea - Comparison with previous studies** section of the main text [30]. This set of OPs expressed by Eq.A.1 maps specific angles with kernel functions using tunable cutoffs for distance r_{ij} between the i - and j -th molecules, and the number of neighbors n_i of the i -th molecule.

$$\Delta\Gamma_i = \frac{\rho_i}{n_i} \sum_{j=1}^N f_{ij} \Theta_{ij} \quad (\text{A.1})$$

where density of molecule i is defined by a switching function with tunable parameter b and cutoff n_{cut} ,

$$\rho_i = \frac{1}{1 + e^{b(n_{cut} - n_i)}} \quad (\text{A.2})$$

number of neighbors n_i is the summation of individual local coordination number f_i with a tuning on the sharpness of the switching function and threshold r_{cut} ,

$$n_i = \sum_{j=1}^N f_{ij} \quad (\text{A.3})$$

$$f_i = \frac{1}{1 + e^{a(r_{cut} - r_i)}} \quad (\text{A.4})$$

and Θ_i describes the global orientation of molecules within the simulation box in reference of characteristic orientation θ_{ij} ,

$$\Theta_{ij} = \sum_{k=1}^{k_{max}} e^{-\frac{(\theta_{ij}-\theta_k)^2}{2*\sigma_k^2}} \quad (\text{A.5})$$

where σ_k tunes widths of Gaussians. Given $\Delta\Gamma_i$ represents the crystallinity of molecule i with respect to target structure, we then compute the size of nuclei of crystalline structures (denoted as $N_{crystal}$) incorporating with depth-first search algorithm. As shown in Fig. A.7(a), homogeneous solution and solid phases (besides amorphous structures) possess completely different angular distribution functions, which makes it reasonable of the application of SMAC OPs. All configurations for these computations are drawn from WTmetaD simulations biasing SPIB-learnt RCs and each trajectory is 12 *ns* long with a time step of 1 *ps*. By using these distributions as references, the population of corresponding polymorphs can be computed in combination of SMAC OPs by proper placement of switching functions and depth-first search method as suggested in Ref. [76]. Fig. A.7(b) shows estimates of the size of polymorphic urea sampled, and the system is labeled by a specific phase when the cluster size, N , is greater than 50 (red dashed line).

A.3 Chapter 4: Supporting Material to “Enhanced Sampling of Crystal Nucleation with Graph Representation Learnt Variables”

A.3.1 Details on featurization and model training

All graphs in this work had their node feature set to be 1 as introduced in the main text. The edge features were expanded under Gaussian basis functions. Here we provided details on featurization and

how the model was trained for both systems.

For iron nucleation from the melt, the centers of the t -th Gaussian are located at $[0nm, 0.8nm]$ every $0.05 nm$ with $\gamma=400 nm^{-2}$. The GNN model for the iron system consists of one hidden layer of 18 channels in the encoder and zero hidden layers in the decoders. All layers are followed by a ReLU layer except the output layer in the decoders. The model was trained using stochastic gradient descent (SGD) optimizer [237] for 200 epochs. The batch size is 36 and the initial learning rate is 0.01 with 0.1 decreasing every 80 epochs. Hyperparameter β is set to be 0.5.

For glycine nucleation from the melt, the centers of all angular basis functions are in $[0, \pi]$ every 0.1π step with $\gamma = 10.13 rad^{-2}$. The model has zero hidden layers in the encoder which is the same as the decoders to yield efficient computations when performing enhanced sampling without diminishing the ability to classify provided crystalline structures. The model was then trained with SGD optimizer for 65 epochs. The batch size is 256 and the initial learning rate is 0.01 with decrements of 0.5 every 35 epochs. Hyperparameter β is set to be 0.2.

A.3.2 Well-tempered metadynamics simulations setup

285 iron particles were randomly inserted triclinic simulation cell of size $1.6 \times 1.6 \times 1.6 nm^3$. The modified embedded-atom (MEAM) interatomic potential introduced by Etesami and Asadi [213] was applied to resemble the optimal physical properties of iron transitions near its melting temperature. All simulations carried out in this work had the equations of motion integrated with a step size of $2 fs$. The box was first relaxed with energy minimization, followed by equilibrations in NVT and NPT ensembles separately. The thermostat was selected to be the velocity rescaling thermostat at around experimental melting temperature $1800 K$ [119] with a coupling time of $0.1 ps$. The relaxation time of the barostat [238] was set to be $10 ps$ and the ambient pressure was set to be the target value. To avoid the formation of unphysical artificial aggregates in atomic simulation [105], changing shape of the simulation cell was enabled.

The simulations of glycine nucleation were performed in a cubic box of size $2.2 \times 2.2 \times 2.2 \text{ nm}^3$ with 120 zwitterionic glycine molecules via random insertion. The forcefield parameters are obtained from the general amber forcefield (GAFF) with mB3LYP point charges. [115,239] The stochastic velocity rescaling thermostat [119] and Parrinello-Rahman barostat [120] were employed with a relaxation time of 0.1 and 1.0 *ps*, respectively. The reference values were 500 *K* to the thermostat and 1.0 *bar* to the barostat. Particle-mesh Ewald method [121] was applied for the computation of long-range electrostatic interaction and hydrogen bonds were constrained with the LINCS algorithm [122]. The cutoffs of electrostatic and Van der Waals interaction in real space were 0.9 *nm*.

A.3.3 Classification with trained local decoder in iron nucleation

The local decoder predicts the probability of individual nodes in pre-defined class and therefore can be used as an estimator to crystal size. Fig.A.8 is the 2-dimensional free energy surface of the size of different crystal packing and the liquid state. The set of subplots suggests it is feasible to form HCP structure from molten iron (subplot (b)) and possible solid-solid transitions may be observed between BCC and HCP iron (subplot (f)).

With a size cutoff of 150, we estimate the free energy differences between the target and initial states as reported in the main text. To examine the reliability of the trained GNN decoder, we benchmark it to the conventional classifying method, adaptive common neighbor analysis. [209] Scatter plots in Fig.A.9 suggest the decoder has comparable classification ability in distinguishing crystal structures. Here, we present values of free energy differences computed for comparison in Tab.A.2, in addition to a simple local minima approximation method. The results indicate the overall thermodynamic rankings among the three configurations are preserved. The free energy of FCC iron is of large differences which may be due to poor sampling.

Table A.2: Free energy difference between crystalline phases and molten phase of iron on baseline methods

System	Transition	Computed free energy difference (kJ/mol)		
		GNN decoder (this work)	Local minima	aCNA
Iron	M→FCC	52.6 ± 5.6	188.9 ± 33.8	148.3 ± 38.5
	M→HCP	-20.2 ± 8.5	-8.5 ± 2.6	-15.7 ± 6.8
	M→BCC	-34.1 ± 4.2	-9.2 ± 4.7	-33.5 ± 4.2

A.3.4 GNN classifier for glycine polymorphs

As mentioned in the main text, a new GNN model is trained in order to identify all crystalline structures in zwitterionic glycine. Crystallography data including ζ -gly [217], ϵ -gly [202], and δ -gly [202] are used for constructing supercells. Different from previous GNN model (used for biasing), short MD simulations of 2 ns are carried out in NVT ensemble to maintain the crystalline structures especially high pressure products without deforming at between 200 to 300 K . 1000 frames of the system are obtained at interval of 0.001 ps .

A.3.5 Free energy difference time series of glycine polymorphs

Here we plot the free energy difference of all sampled polymorphs of glycine with respect to the liquid phase vs. simulation time in Fig. A.10 and the values are reported in Tab. A.3. The high pressure polymorphs were identified using the same metric as for the ambient products. The size cutoff was set to be 65 over a total of 120 glycine molecules.

Fig. A.11 provides a way of evaluating pathway of underlining solid-solid transitions, where 2-dimensional contour plots were made on the space of size of individual polymorphs identified by the GNN model. Clear evidence of ϵ -gly to ζ -gly and ϵ -gly to γ -gly transitions can be found in Fig. A.11(21) and conversely γ -gly is not likely to form ζ -gly directly, this is in agreement to experimental observations as γ -

Table A.3: Free energy difference of crystal structures of glycine polymorphs with respect to the initial molten/liquid phase

Transition	Free energy difference (kJ/mol)
$\ell \rightarrow \alpha$	56.4 ± 16.9
$\ell \rightarrow \beta$	659 ± 75
$\ell \rightarrow \gamma$	234 ± 35
$\ell \rightarrow \delta$	12.1 ± 3.4
$\ell \rightarrow \epsilon$	69.3 ± 13.4
$\ell \rightarrow \zeta$	228.1 ± 34.6

gly to ϵ -gly transitions occur at pressures around 4 GPa and ϵ -gly transforms into ζ -gly as pressure drops to 0.8 GPa . The remaining transitions observed experimentally were not identified in our simulations. [217]

A.3.6 Effects of N_c selection on relative stability rankings

In this section, we evaluated the computations of free energy differences under different values of size threshold. Several values of N_c are chosen for the determination of a particular (meta)stable state, i.e., the system is classified as crystal (or liquid) only when the number of solid- (or liquid-) like particles is larger than N_c . In Fig. A.12, three additional N_c values are chosen arbitrarily to compare with the value used in the main text. It is clear that the samples from crystals become less as N_c get larger and this leads to larger free energy differences with respect to the liquid phase. However, the metric itself in identifying (meta)stable states does not seem affected as the thermodynamic rankings among crystals preserve.

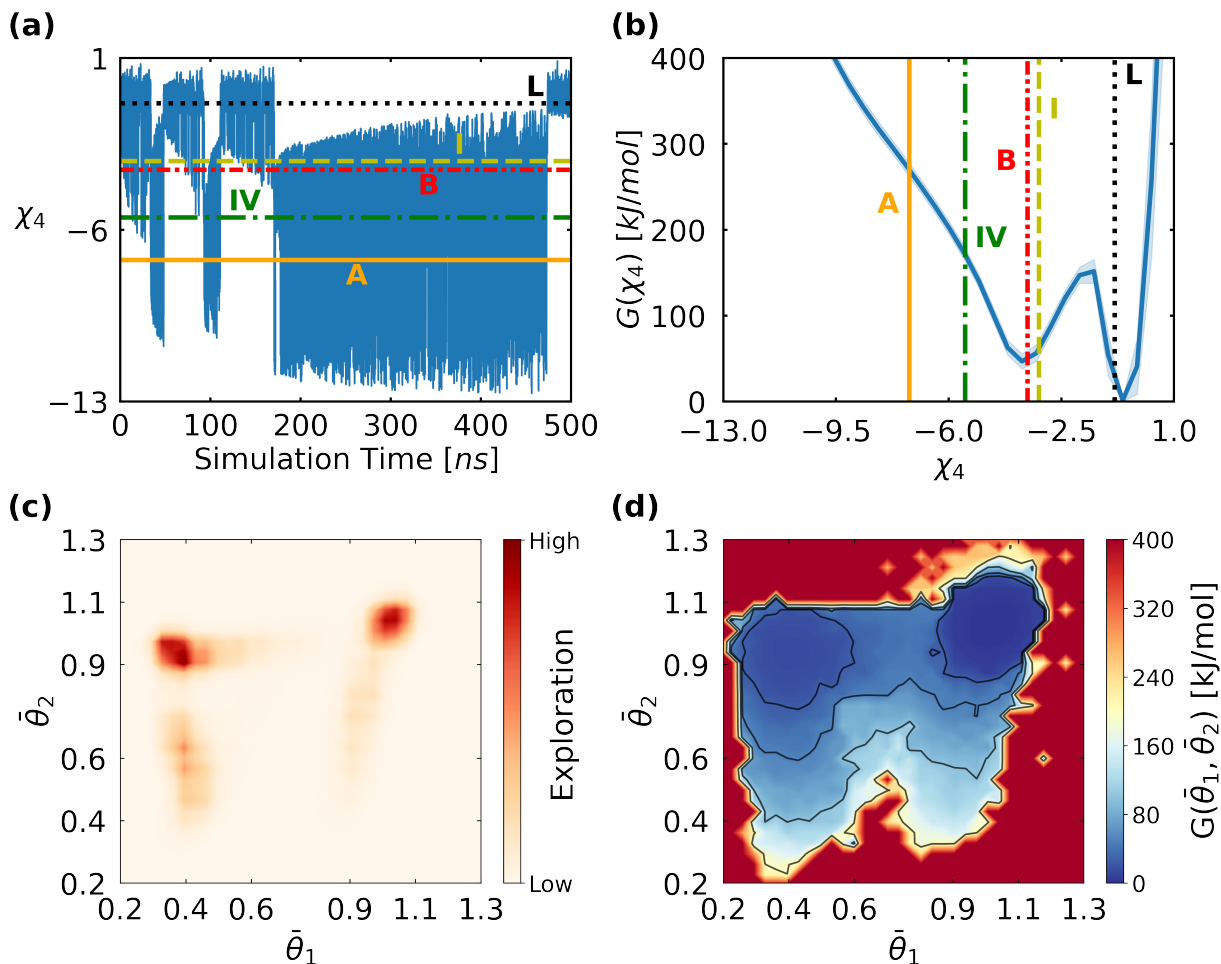


Figure A.1: WTmetaD simulation and thermodynamics analyses: (a) Time series of 1-d WT-metaD biasing χ_4 ; (b) reweighted free energy surface along χ_4 averaged over 3 individual WT-metaD simulations; (c) unweighted exploration plot; and (d) reweighted free energy surface in $(\bar{\theta}_1, \bar{\theta}_2)$ space. The benchmark lines of all polymorphs (sampled in Ref. [99]) are plotted w.r.t corresponding landmarks (see main text for landmark values). The regions between errorbars are filled with corresponding colors in (b). The exploration plot (c) shows configurations which were frequently visited, and they are form I, IV, A, B and liquid. Reweighted free energy surface (d) presents only the basins of form I and liquid.

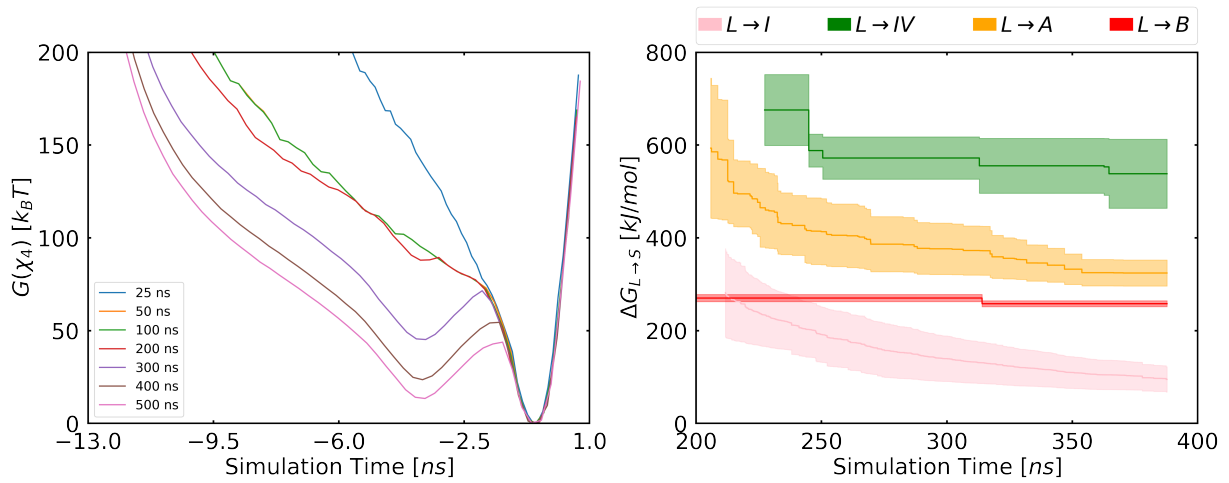


Figure A.2: Convergence check with (a) free energy surface along RC, χ_4 , over simulation time and (b) free energy differences of phases averaged over 3 independent trajectories. Error bars are shaded in corresponding colors in (b). Both methods show slow convergence of χ_4 .

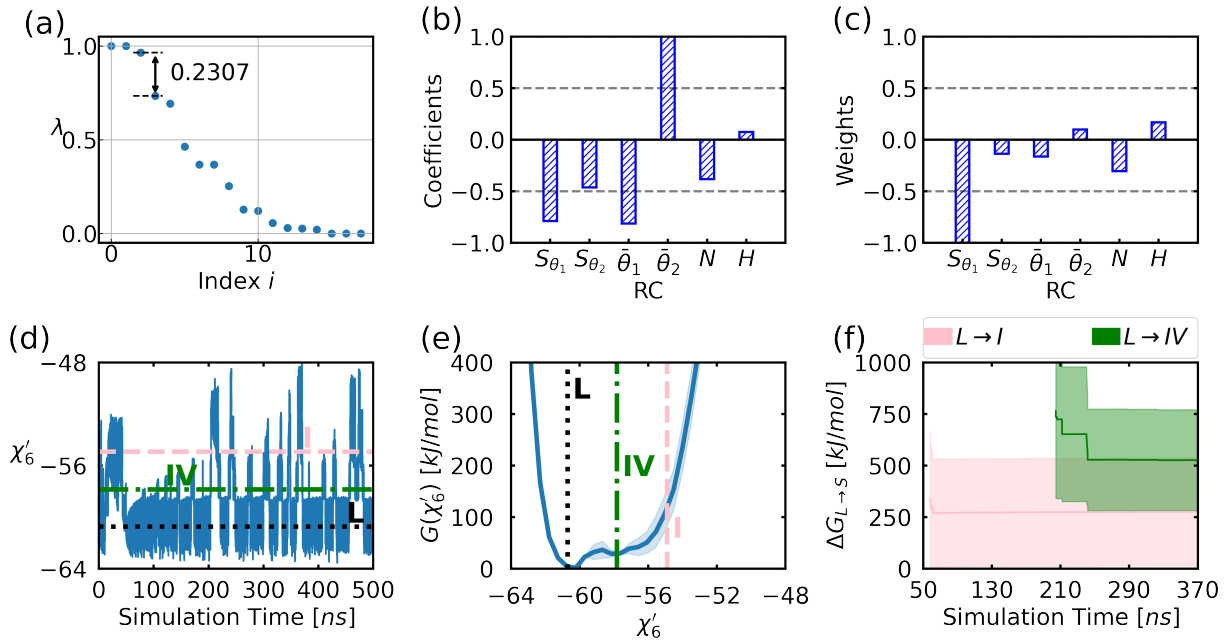


Figure A.3: Details of 1-d SGOOP reaction coordinate χ'_6 : a) spectral gap plot of χ'_6 discretized into 20 wells under 3 wells assumption; bar plots for coefficients b) and rescaled weights c) of each order parameters composed in χ'_6 ; d) time series of χ'_6 with benchmark lines of liquid state (L), form I, and form IV observed with visualization tools; e) reweighted free energy surface along χ'_6 ; and f) free energy difference of form I and form IV with respect to liquid state over time. Shaded area corresponds to error bars.

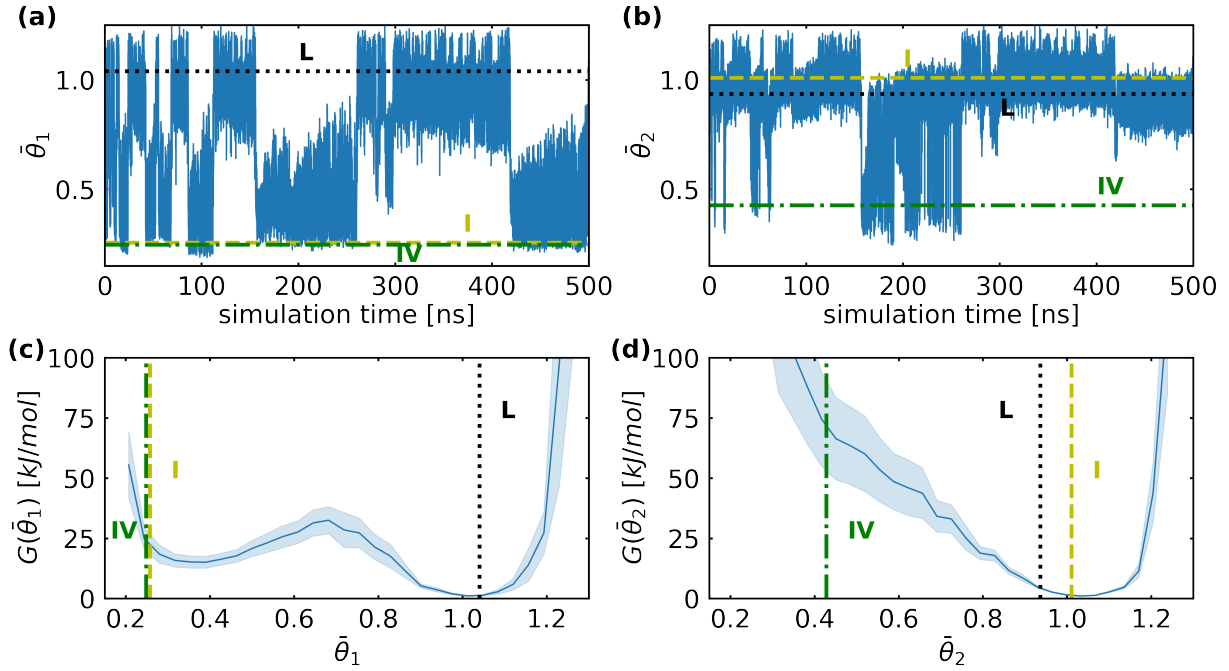


Figure A.4: Time series of $\bar{\theta}_1$ and $\bar{\theta}_2$ for one replica are shown in (a) and (b). Free energy surfaces (c) and (d) are constructed along $\bar{\theta}_1$ and $\bar{\theta}_2$, respectively. Landmarks of metastable states are highlighted with colored lines. As discussed in the main text, SGOOP-optimized reaction coordinate, χ_6 lacks the ability of distinguishing polymorphs I and IV. We performed latter free energy analyses along developed averaged angle order parameters. Evaluations on either time series or free energy plots shows that, by jointly considering $\bar{\theta}_1$ and $\bar{\theta}_2$, one is able to identify form I, form IV and liquid from each other.

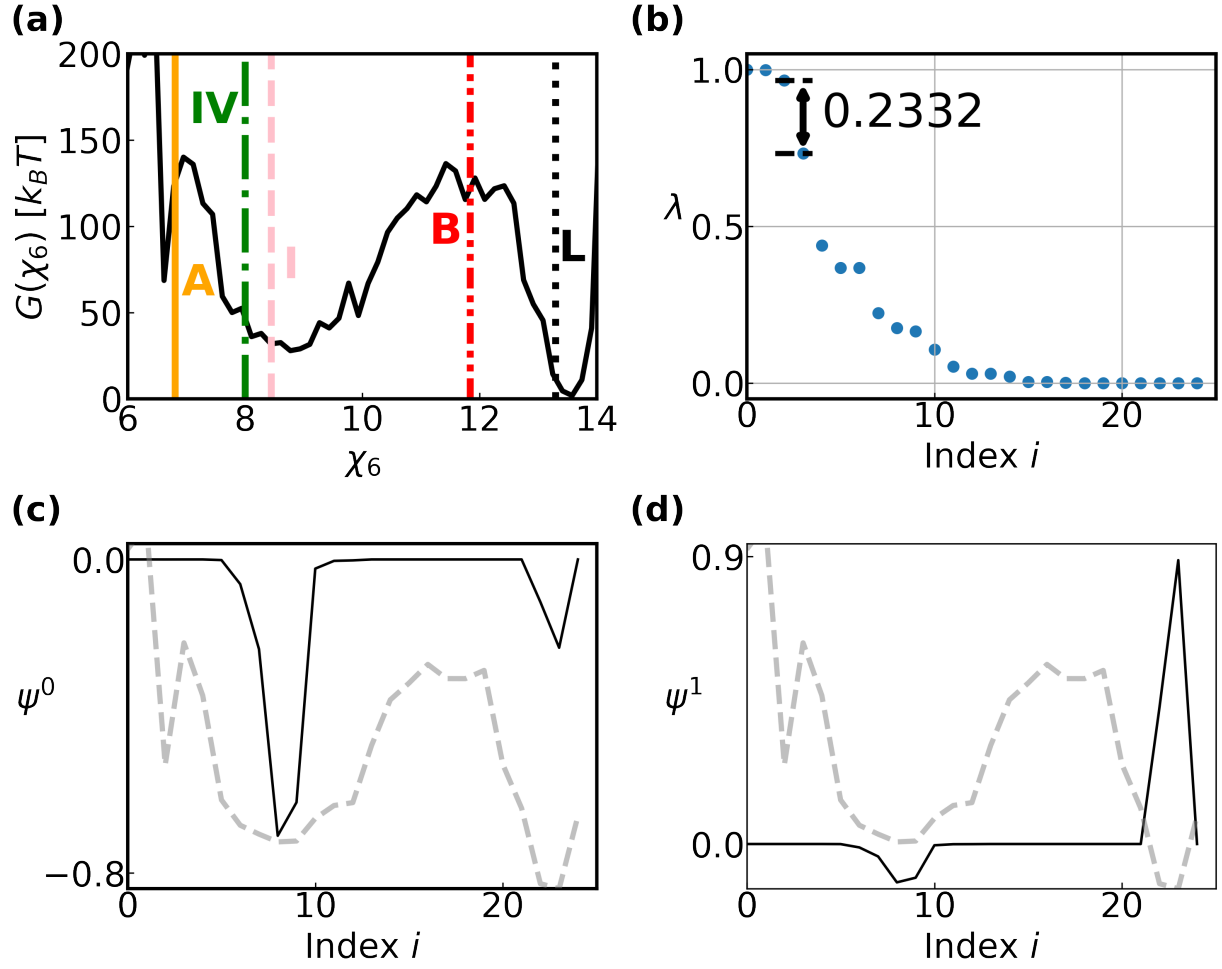


Figure A.5: Spectral gap analyses of χ_6 under 3-well assumption: a) 3 distinctive wells optimized by SGOOP with input data (refer Sec. A.1.1 for details); b) spectral gap plot constructed from exponential of solved eigenvalues of transition matrix; c) the zeroth and d) the first eigenvectors of obtained transition matrix.

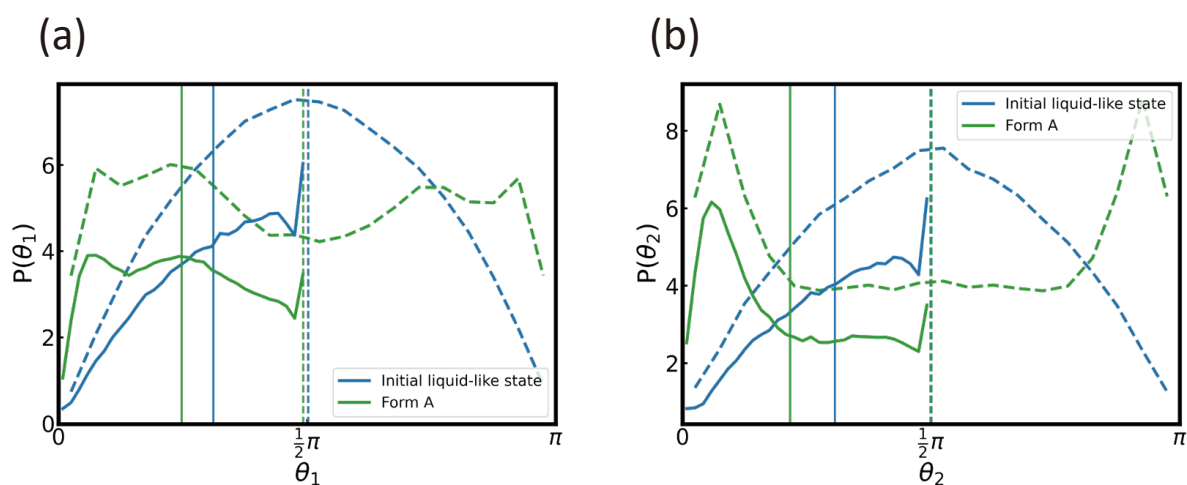


Figure A.6: Probability distribution of intermolecular angle, θ_1 and θ_2 for the initial homogeneous urea solution (in blue) and solid (in green) urea. Geometric information is obtained from snapshots of WTmetaD simulations basing SPIB-learned RCs discussed in the main text. Dashed lines are the distributions the original OP values, and solid lines correspond to the distribution of angle values after using switching function (see main text for details). Vertical lines indicate the mean values of corresponding distributions.

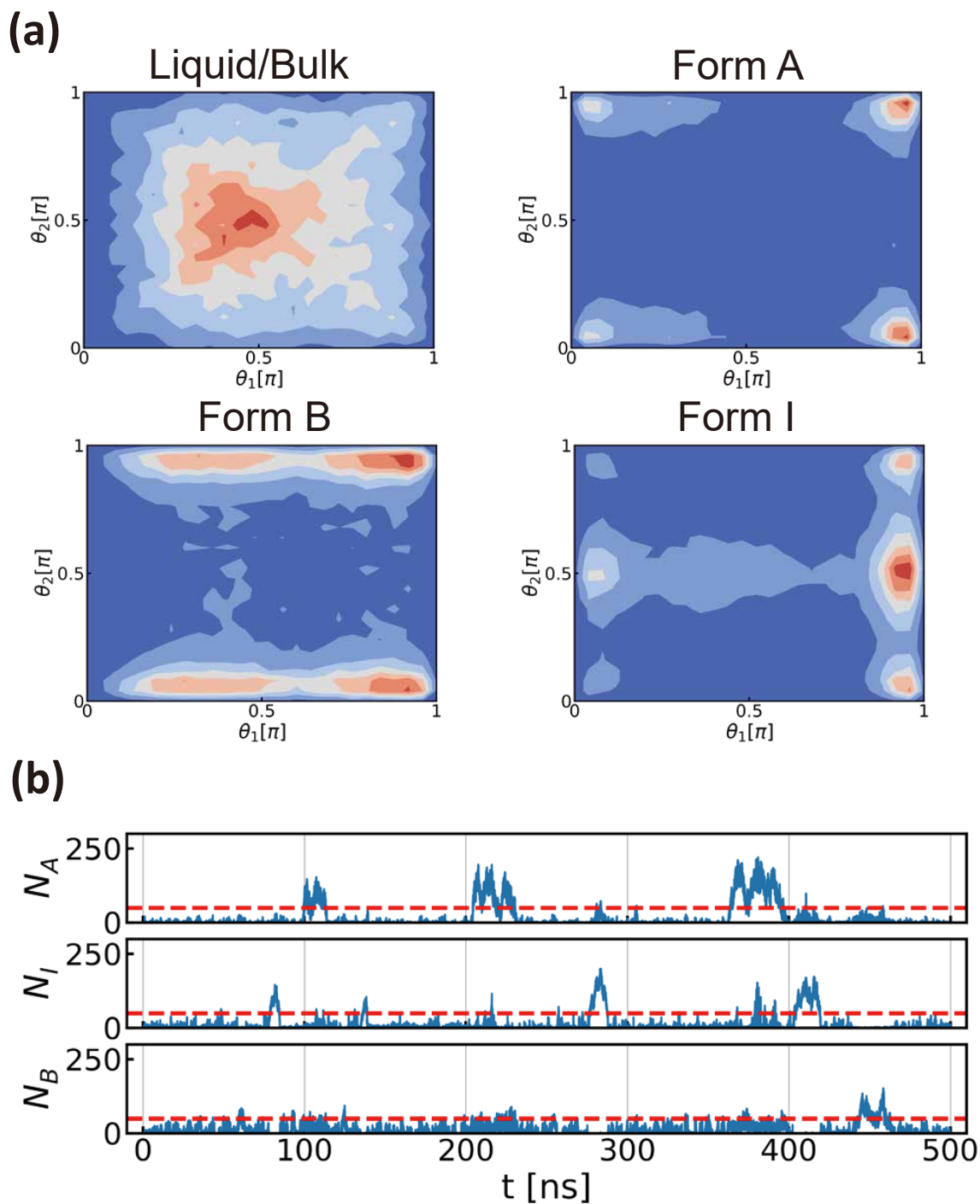


Figure A.7: Definitions of urea solution and solid urea in angular space (panel a) and post-processed time series of crystal urea in terms of cluster size (panel b). Red dashed lines are the threshold in state identification for the SMAC OPs described in **Urea - Comparison with previous studies** section of the main text. Values below this threshold (here 50) are considered noise and classified as the homogeneous solution phase.

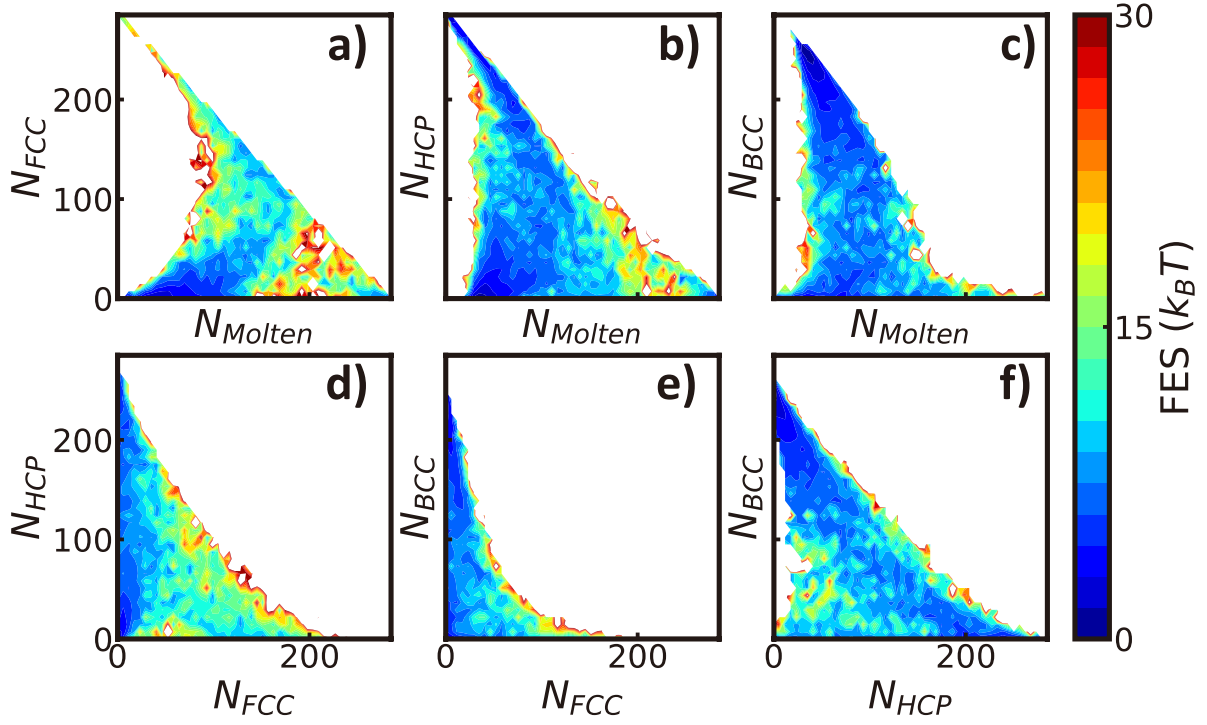


Figure A.8: Cluster size of different phases of iron crystals measured by the trained local decoder.

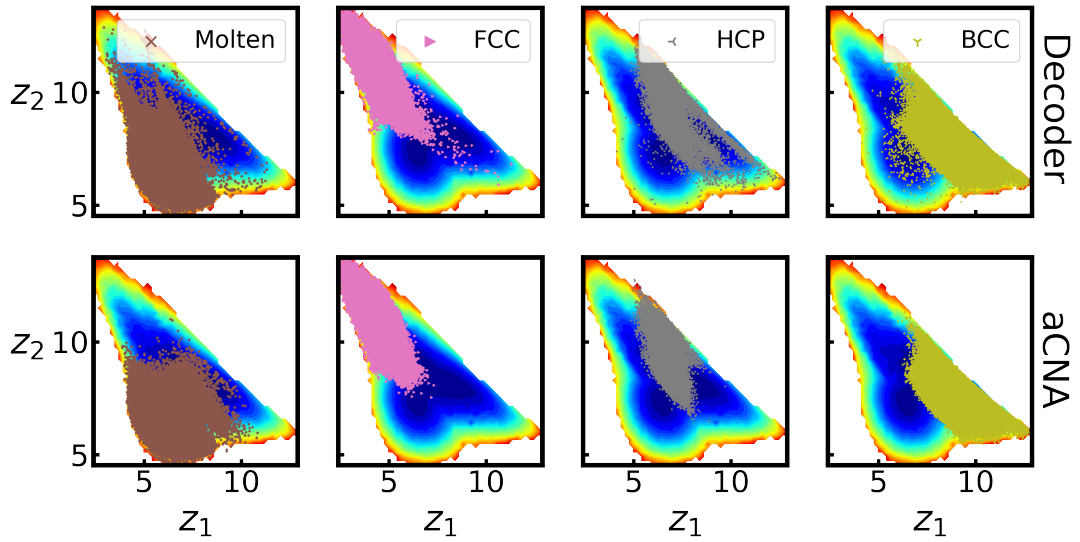


Figure A.9: Scatter plots of two classification methods (GNN decoder, upper panel and adaptive common neighbor analysis [209], lower panel) superimpose to reweighted free energy surface from iron nucleation from melts.

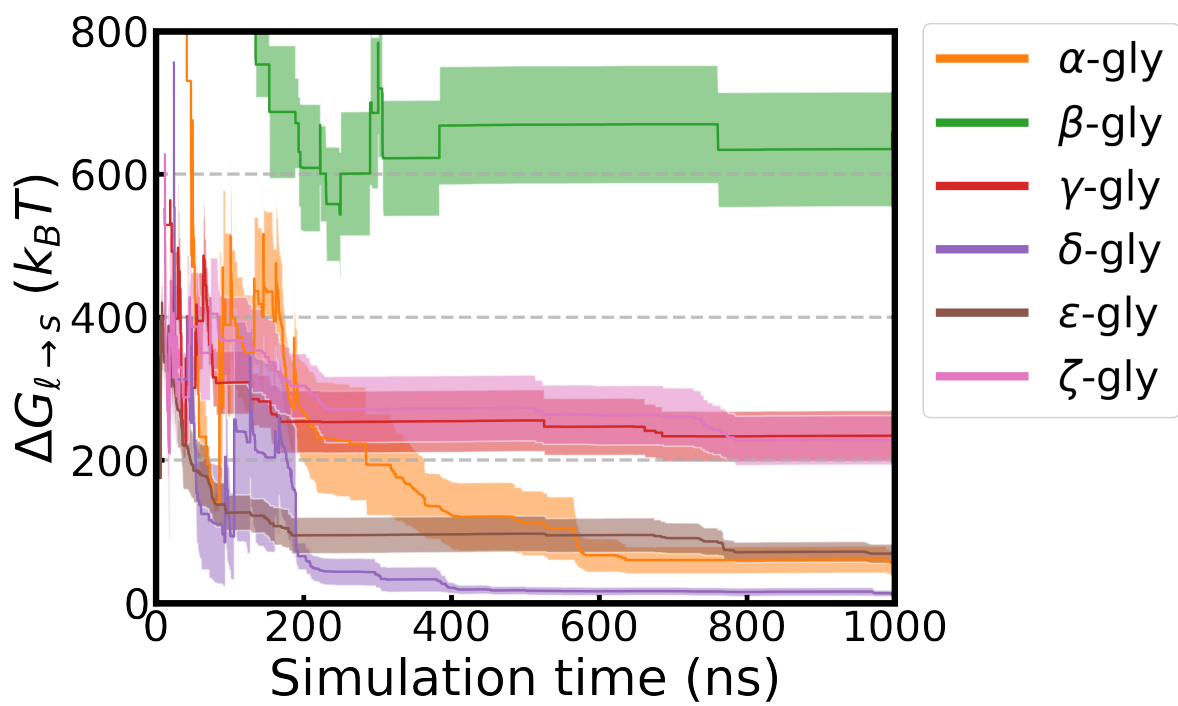


Figure A.10: $\Delta G_{\ell \rightarrow s}$ as a function of simulation time. All high pressure product of glycine are shown to be more stable than glycine polymorphs at ambient condition.

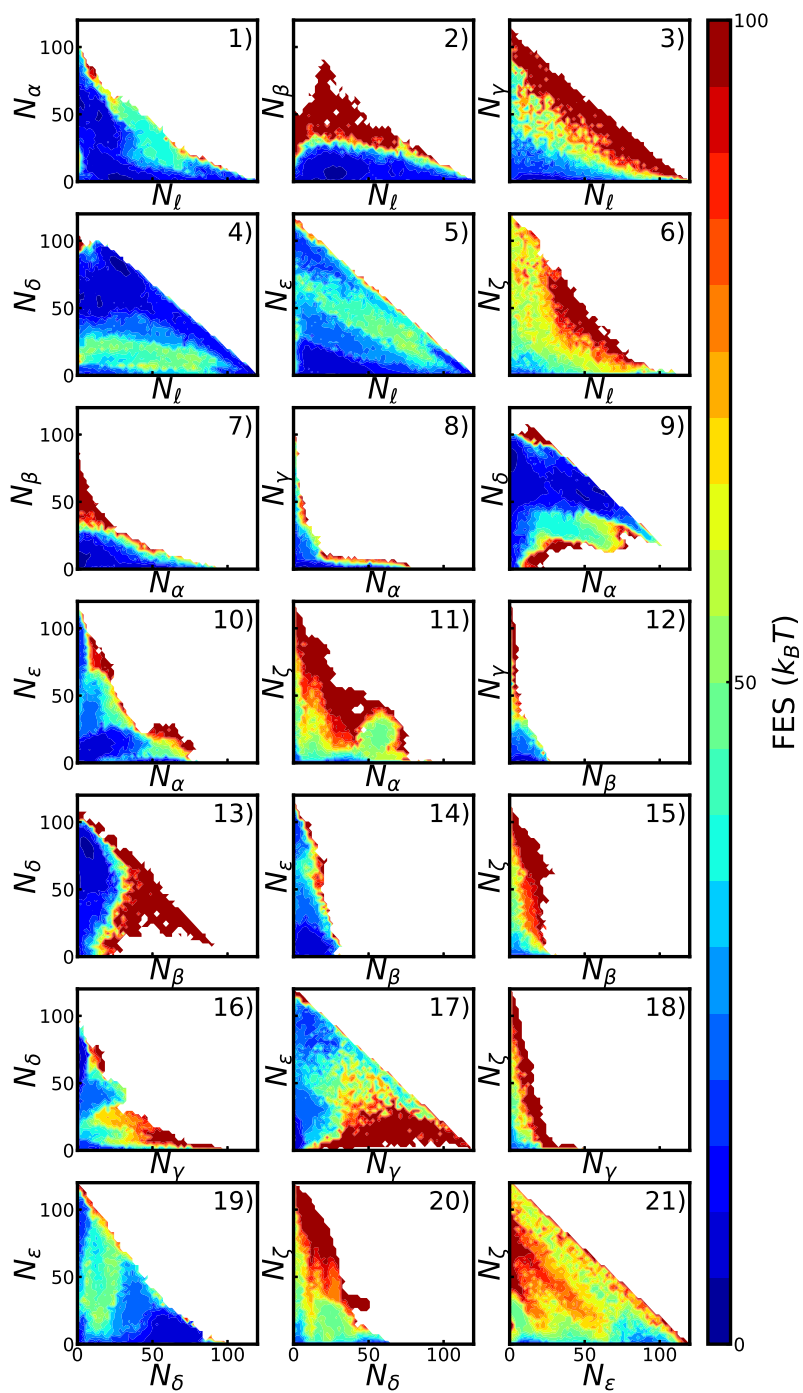


Figure A.11: Two-dimensional free energy surfaces of cluster size of different polymorphs of glycine measured by the trained local decoder.

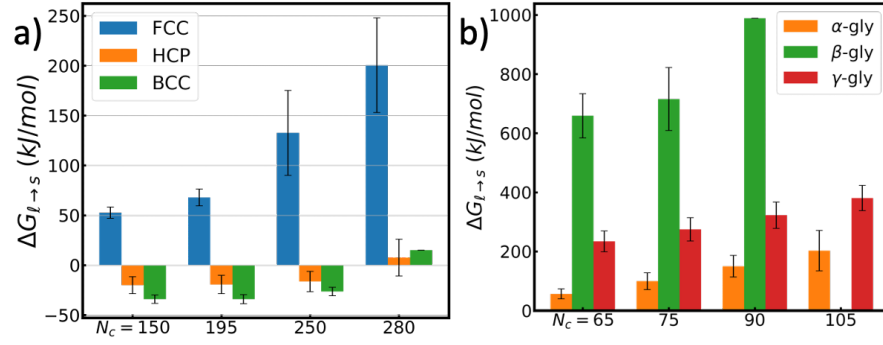


Figure A.12: The default values of N_c were set to be 65 and 150 for iron and glycine systems respectively. Bars without error bar suggest only one run samples the corresponding phase and, similarly, a free energy difference with value zero suggests no such phase is identified in all simulations.

Bibliography

- [1] Denis Gentili, Massimo Gazzano, Manuela Melucci, Derek Jones, and Massimiliano Cavallini. Polymorphism as an additional functionality of materials for technological applications at surfaces and interfaces. *Chemical Society Reviews*, 48:2502–2517, 2019.
- [2] Maria Kuhnert-Brandstätter. Thermomicroscopy in the analysis of pharmaceuticals. Pergamon Press, 1971.
- [3] Gabriele C Sosso, Tianshu Li, Davide Donadio, Gareth A Tribello, and Angelos Michaelides. Microscopic mechanism and kinetics of ice formation at complex interfaces: Zooming in on kaolinite. *The Journal of Physical Chemistry Letters*, 7:2350–2355, 2016.
- [4] Scott A. Hollingsworth and Ron O. Dror. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, 2018.
- [5] Haiyang Niu, Pablo M Piaggi, Michele Invernizzi, and Michele Parrinello. Molecular dynamics simulations of liquid silica crystallization. *Proceedings of the National Academy of Sciences*, 115:5348–5352, 2018.
- [6] Thorsten Bartels-Rausch. Ten things we need to know about ice and snow. *Nature*, 494(7435):27–29, Feb 2013.
- [7] Benjamin J. Murray, Theodore W. Wilson, Steven Dobbie, Zhiqiang Cui, Sardar M. R. K. Al-Jumur, Ottmar Möhler, Martin Schnaiter, Robert Wagner, Stefan Benz, Monika Niemand, Harald Saathoff, Volker Ebert, Steven Wagner, and Bernd Kärcher. Heterogeneous nucleation of ice particles on glassy aerosols under cirrus conditions. *Nature Geoscience*, 3(4):233–237, Apr 2010.
- [8] Peter Mazur. Cryobiology: The freezing of biological systems. *Science*, 168(3934):939–949, 1970.
- [9] James D. Harper, Charles M. Lieber, and Peter T. Lansbury. Atomic force microscopic imaging of seeded fibril formation and fibril branching by the Alzheimer’s disease amyloid- β protein. *Chemistry Biology*, 4(12):951–959, 1997.

- [10] Samuel I. A. Cohen, Sara Linse, Leila M. Luheshi, Erik Hellstrand, Duncan A. White, Luke Rajah, Daniel E. Otzen, Michele Vendruscolo, Christopher M. Dobson, and Tuomas P. J. Knowles. Proliferation of amyloid-42 aggregates occurs through a secondary nucleation mechanism. *Proceedings of the National Academy of Sciences*, 110(24):9758–9763, 2013.
- [11] Martin Karplus and John Kuriyan. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences*, 102(19):6679–6685, 2005.
- [12] Harold A. Scheraga, Mey Khalili, and Adam Liwo. Protein-folding dynamics: Overview of molecular simulation techniques. *Annual Review of Physical Chemistry*, 58:57–83, 2007.
- [13] Pratyush Tiwary, Jagannath Mondal, and B. J. Berne. How and when does an anticancer drug leave its binding site? *Science Advances*, 3(5):e1700014, 2017.
- [14] Alexander G Shtukenberg, Melissa Tan, Leslie Vogt-Maranto, Eric J Chan, Wenqian Xu, Jingxiang Yang, Mark E Tuckerman, Chunhua T Hu, and Bart Kahr. Melt crystallization for paracetamol polymorphism. *Crystal Growth & Design*, 19:4070–4080, 2019.
- [15] Hiroki Nada. Unknown crystal-like phases formed in an imidazolium ionic liquid: A metadynamics simulation study. *The Journal of Chemical Physics*, 160:204501, 6 2024.
- [16] James P Sethna. *Statistical Mechanics: Entropy, Order Parameters, and Complexity*. Oxford University Press, 2021.
- [17] Jim De Yoreo. *A Perspective on Multistep Pathways of Nucleation*, pages 1–17. 2020.
- [18] S Karthika, TK Radhakrishnan, and P Kalaichelvi. A review of classical and nonclassical nucleation theories. *Crystal Growth & Design*, 16(11):6663–6681, 2016.
- [19] Baron Peters. *Reaction Rate Theory and Rare Events Simulations*. Elsevier, 2017.
- [20] Mike Sleutel, Jim Lutsko, Alexander ES Van Driessche, Miguel A Durán-Olivencia, and Dominique Maes. Observing classical nucleation theory at work by monitoring phase transitions with molecular precision. *Nature Communications*, 5(1):5598, 2014.
- [21] Conor Parks, Andy Koswara, Hsien-Hsin Tung, Nandkishor K Nere, Shailendra Bordawekar, Zoltan K Nagy, and Doraiswami Ramkrishna. Nanocrystal dissolution kinetics and solubility increase prediction from molecular dynamics: the case of α -, β -, and γ -glycine. *Molecular Pharmaceutics*, 14:1023–1032, 2017.
- [22] Nathan Duff and Baron Peters. Polymorph specific rmsd local order parameters for molecular crystals and nuclei: α -, β -, and γ -glycine. *The Journal of Chemical Physics*, 135:134101, 2011.
- [23] Pablo M Piaggi, Jack Weis, Athanassios Z Panagiotopoulos, Pablo G Debenedetti, and Roberto Car. Homogeneous ice nucleation in an ab initio machine-learning model of water. *Proceedings of the National Academy of Sciences*, 119:e2207294119, 2022.

- [24] R Allen LaCour, Timothy C Moore, and Sharon C Glotzer. Tuning stoichiometry to promote formation of binary colloidal superlattices. *Physical Review Letters*, 128(18):188001, 2022.
- [25] Andressa A Bertolazzo, Debdas Dhabal, Laura J S Lopes, Samantha K Walker, and Valeria Molinero. Unstable and metastable mesophases can assist in the nucleation of porous crystals. *The Journal of Physical Chemistry C*, 126:3776–3786, 2022.
- [26] Liam C Jacobson, Waldemar Hujo, and Valeria Molinero. Amorphous precursors in the nucleation of clathrate hydrates. *Journal of the American Chemical Society*, 132:11806–11811, 2010.
- [27] Federico Giberti, Gareth A Tribello, and Michele Parrinello. Transient polymorphism in NaCl. *Journal of Chemical Theory and Computation*, 9:2526–2530, 2013.
- [28] Aaron R Finney and Matteo Salvalaglio. Multiple pathways in NaCl homogeneous crystal nucleation. *Faraday Discussions*, 2022.
- [29] Sun-Ting Tsai, Zachary Smith, and Pratyush Tiwary. Reaction coordinates and rate constants for liquid droplet nucleation: Quantifying the interplay between driving force and memory. *The Journal of Chemical Physics*, 151:154106, 2019.
- [30] Matteo Salvalaglio, Thomas Vetter, Federico Giberti, Marco Mazzotti, and Michele Parrinello. Uncovering molecular details of urea crystal growth in the presence of additives. *Journal of the American Chemical Society*, 134:17221–17233, 2012.
- [31] Haiyang Niu, Yi Isaac Yang, and Michele Parrinello. Temperature dependence of homogeneous nucleation in ice. *Physical Review Letters*, 122(24):245501, 2019.
- [32] Gianpaolo Gobbo, Michael A Bellucci, Gareth A Tribello, Giovanni Ciccotti, and Bernhard L Trout. Nucleation of molecular crystals driven by relative information entropy. *Journal of Chemical Theory and Computation*, 14:959–972, 2018.
- [33] Ziyue Zou, Eric R Beyerle, Sun-Ting Tsai, and Pratyush Tiwary. Driving and characterizing nucleation of urea and glycine polymorphs in water. *Proceedings of the National Academy of Sciences*, 120:e2216099120, 2023.
- [34] R A Van Santen. The Ostwald step rule. *Journal of Physical Chemistry*, 88:5768–5769, 1984.
- [35] Alexander Stukowski. Visualization and analysis of atomistic simulation data with ovito—the open visualization tool. *Modelling and Simulation in Materials Science and Engineering*, 18(1):015012, 2009.
- [36] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.
- [37] Amit Samanta, Mark E Tuckerman, Tang-Qing Yu, and Weinan E. Microscopic mechanisms of equilibrium melting of a solid. *Science*, 346:729–732, 2014.

- [38] Elia Schneider, Luke Dai, Robert Q Topper, Christof Drechsel-Grau, and Mark E Tuckerman. Stochastic neural network approach for learning high-dimensional free energy surfaces. *Physical Review Letters*, 119(15):150601, 2017.
- [39] Luigi Bonati, Valerio Rizzi, and Michele Parrinello. Data-driven collective variables for enhanced sampling. *The Journal of Physical Chemistry Letters*, 11:2998–3004, 2020.
- [40] Federico Giberti, Matteo Salvalaglio, and Michele Parrinello. Metadynamics studies of crystal nucleation. *IUCrJ*, 2:256–266, 2015.
- [41] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99:12562–12566, 2002.
- [42] Omar Valsson, Pratyush Tiwary, and Michele Parrinello. Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint. *Annual Review of Physical Chemistry*, 67:159–184, 2016.
- [43] Kingshuk Ghosh, Purushottam D Dixit, Luca Agozzino, and Ken A Dill. The maximum caliber variational principle for nonequilibria. *Annual Review of Physical Chemistry*, 71:213–238, 2020.
- [44] Steve Pressé, Kingshuk Ghosh, Julian Lee, and Ken A Dill. Principles of maximum entropy and maximum caliber in statistical physics. *Reviews of Modern Physics*, 85(3):1115–1141, 2013.
- [45] Purushottam D Dixit, Abhinav Jain, Gerhard Stock, and Ken A Dill. Inferring transition rates of networks from populations in continuous-time Markov processes. *Journal of Chemical Theory and Computation*, 11:5464–5472, 2015.
- [46] Pratyush Tiwary and B J Berne. Predicting reaction coordinates in energy landscapes with diffusion anisotropy. *The Journal of Chemical Physics*, 147:152701, 2017.
- [47] Pratyush Tiwary and Michele Parrinello. A time-independent free energy estimator for metadynamics. *The Journal of Physical Chemistry B*, 119:736–742, 2015.
- [48] Dedi Wang and Pratyush Tiwary. State predictive information bottleneck. *The Journal of Chemical Physics*, 154:134111, 2021.
- [49] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, pages 1–14, 2014.
- [50] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*, 2017.

- [51] Shams Mehdi, Dedi Wang, Shashank Pant, and Pratyush Tiwary. Accelerating all-atom simulations and gaining mechanistic understanding of biophysical systems through state predictive information bottleneck. *Journal of Chemical Theory and Computation*, 18:3231–3238, 2022.
- [52] Eric R Beyerle, Shams Mehdi, and Pratyush Tiwary. Quantifying energetic and entropic pathways in molecular systems. *The Journal of Physical Chemistry B*, 126:3950–3960, 2022.
- [53] Dedi Wang, Yunrui Qiu, Eric R. Beyerle, Xuhui Huang, and Pratyush Tiwary. Information bottleneck approach for Markov model construction. *Journal of Chemical Theory and Computation*, 20(12):5352–5367, 2024.
- [54] Dedi Wang, Yihang Wang, Luke Evans, and Pratyush Tiwary. From latent dynamics to meaningful representations. *Journal of Chemical Theory and Computation*, 20(9):3503–3513, 2024.
- [55] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.
- [56] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.
- [57] Frank Noé, Illia Horenko, Christof Schütte, and Jeremy C Smith. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *Journal of Chemical Physics*, 126:155102, 2007.
- [58] P Deuffhard, W Huisinga, A Fischer, and Ch Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra and Its Applications*, 315:39–59, 2000.
- [59] Yihang Wang, João Marcelo Lamim Ribeiro, and Pratyush Tiwary. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nature Communications*, 10(1):3573, Aug 2019.
- [60] Alexandre Duval, Simon V Mathis, Chaitanya K Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D Malliaros, Taco Cohen, Pietro Liò, Yoshua Bengio, and Michael Bronstein. A hitchhiker’s guide to geometric gnns for 3d atomic systems, 2024.
- [61] Pradip Mondal and Deepak Chopra. Role of polymorphism in materials science. *Material Science Research India*, 11:43–50, 2014.
- [62] Antonio Llinàs and Jonathan M Goodman. Polymorph control: past, present and future. *Drug Discovery Today*, 13:198–210, 2008.
- [63] Harry G Brittain. *Polymorphism in pharmaceutical solids*. CRC Press, 2018.
- [64] Jie Lu, Zhen Li, and Xiaolin Jiang. Polymorphism of pharmaceutical molecules: perspectives on nucleation. *Frontiers of Chemical Engineering in China*, 4:37–44, 2010.

- [65] S. Karthika, T. K. Radhakrishnan, and P. Kalaichelvi. A review of classical and nonclassical nucleation theories. *Crystal Growth & Design*, 16(11):6663–6681, 2016.
- [66] Aurora J Cruz-Cabeza and Joel Bernstein. Conformational polymorphism. *Chemical Reviews*, 114:2170–2191, 2014.
- [67] G M Torrie and J P Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23:187–199, 1977.
- [68] Rosalind J Allen, Patrick B Warren, and Pieter Rein ten Wolde. Sampling rare switching events in biochemical networks. *Physical Review Letters*, 94:18104, 6 2005.
- [69] Themis Lazaridis and Martin Karplus. Orientational correlations and entropy in liquid water. *The Journal of Chemical Physics*, 105:4294–4316, 1996.
- [70] P Rosales-Pelaez, I Sanchez-Burgos, C Valeriani, C Vega, and E Sanz. Seeding approach to nucleation in the NVT ensemble: The case of bubble cavitation in overstretched Lennard Jones fluids. *Physical Review E*, 101:22611, 6 2020.
- [71] Victor E Bazterra, Marta B Ferraro, and Julio C Facelli. Modified genetic algorithm to model crystal structures. I. benzene, naphthalene and anthracene. *The Journal of Chemical Physics*, 116:5984–5991, 2002.
- [72] Josiah W Gibbs. *The Collected Works.: Elementary Principles in Statistical Mechanics*. Laongmans, 1928.
- [73] Gregory K Schenter, Shawn M Kathmann, and Bruce C Garrett. Dynamical nucleation theory: A new molecular approach to vapor-liquid nucleation. *Physical Review Letters*, 82:3484–3487, 6 1999.
- [74] Ryan S DeFever and Sapna Sarupria. Nucleation mechanism of clathrate hydrates of water-soluble guest molecules. *The Journal of Chemical Physics*, 147:204503, 2017.
- [75] Tarak Karmakar, Michele Invernizzi, Valerio Rizzi, and Michele Parrinello. Collective variables for the study of crystallisation. *Molecular Physics*, 119(19-20):e1893848, 2021.
- [76] Gareth A Tribello, Federico Giberti, Gabriele C Sosso, Matteo Salvalaglio, and Michele Parrinello. Analyzing and driving cluster formation in atomistic simulations. *Journal of Chemical Theory and Computation*, 13:1317–1327, 2017.
- [77] Erik E Santiso and Bernhardt L Trout. A general set of order parameters for molecular crystals. *The Journal of Chemical Physics*, 134:64109, 2011.
- [78] Caroline Desgranges and Jerome Delhommelle. Crystal nucleation along an entropic pathway: Teaching liquids how to transition. *Physical Review E*, 98:63307, 6 2018.
- [79] Baron Peters and Bernhardt L Trout. Obtaining reaction coordinates by likelihood maximization. *The Journal of Chemical Physics*, 125:54108, 2006.

- [80] Jordane Preto and Cecilia Clementi. Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. *Physical Chemistry Chemical Physics*, 16:19181–19191, 2014.
- [81] Pratyush Tiwary and B J Berne. How wet should be the reaction coordinate for ligand unbinding? *The Journal of Chemical Physics*, 145:54113, 2016.
- [82] Mohammad M Sultan and Vijay S Pande. tica-metadynamics: accelerating metadynamics by using kinetically selected collective variables. *Journal of Chemical Theory and Computation*, 13:2440–2447, 2017.
- [83] Eliodoro Chiavazzo, Roberto Covino, Ronald R Coifman, C William Gear, Anastasia S Georgiou, Gerhard Hummer, and Ioannis G Kevrekidis. Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. *Proceedings of the National Academy of Sciences*, 114:E5494–E5503, 2017.
- [84] Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. Vampnets for deep learning of molecular kinetics. *Nature Communications*, 9:1–11, 2018.
- [85] João Marcelo Lamim Ribeiro, Pablo Bravo, Yihang Wang, and Pratyush Tiwary. Reweighted autoencoded variational bayes for enhanced sampling (rave). *The Journal of Chemical Physics*, 149:72301, 2018.
- [86] Zahra Shamsi, Kevin J Cheng, and Diwakar Shukla. Reinforcement learning based adaptive sampling: Reaping rewards by exploring protein conformational landscapes. *The Journal of Physical Chemistry B*, 122:8386–8395, 2018.
- [87] Jiang Wang, Simon Olsson, Christoph Wehmeyer, Adrià Pérez, Nicholas E Charron, Gianni de Fabritiis, Frank Noé, and Cecilia Clementi. Machine learning of coarse-grained molecular dynamics force fields. *ACS Central Science*, 5:755–767, 2019.
- [88] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Physical Review Letters*, 100:20603, 2008.
- [89] Paul J Steinhardt, David R Nelson, and Marco Ronchetti. Bond-orientational order in liquids and glasses. *Physical Review B*, 28:784–791, 1983.
- [90] P W Bridgman. Polymorphism at high pressures. *Proceedings of the American Academy of Arts and Sciences*, 52:91–187, 1916.
- [91] H P Weber, W G Marshall, and V Dmitriev. High-pressure polymorphism in deuterated urea. *Acta Crystallographica Section A*, 58:c174, 6 2002.
- [92] F J Lamelas, Z A Dreger, and Y M Gupta. Raman and x-ray scattering studies of high-pressure phases of urea. *The Journal of Physical Chemistry B*, 109:8206–8215, 2005.
- [93] Anna Olejniczak, Kinga Ostrowska, and Andrzej Katrusiak. H-bond breaking in high-pressure urea. *The Journal of Physical Chemistry C*, 113:15761–15767, 2009.

- [94] M Donnelly, C L Bull, R J Husband, A D Frantzana, S Klotz, and J S Loveday. Urea and deuterium mixtures at high pressures. *The Journal of Chemical Physics*, 142:124503, 2015.
- [95] Kamil Dziubek, Margherita Citroni, Samuele Fanetti, Andrew B Cairns, and Roberto Bini. High-pressure high-temperature structural properties of urea. *The Journal of Physical Chemistry C*, 121:2380–2387, 2017.
- [96] Kinga Roszak and Andrzej Katrusiak. Giant anomalous strain between high-pressure phases and the mesomers of urea. *The Journal of Physical Chemistry C*, 121:778–784, 2017.
- [97] F Safari, M Tkacz, and A Katrusiak. High-pressure sorption of hydrogen in urea. *The Journal of Physical Chemistry C*, 125:7756–7762, 2021.
- [98] Cheng Shang, Xiao-Jie Zhang, and Zhi-Pan Liu. Crystal phase transition of urea: what governs the reaction kinetics in molecular crystal phase transitions. *Physical Chemistry Chemical Physics*, 19:32125–32131, 2017.
- [99] Pablo M Piaggi and Michele Parrinello. Predicting polymorphism in molecular crystals using orientational entropy. *Proceedings of the National Academy of Sciences*, 115:10251–10256, 2018.
- [100] Peter R Spackman, Michael J Turner, Joshua J McKinnon, Stephen K Wolff, Daniel J Grimwood, Dylan Jayatilaka, and Mark A Spackman. Crystalexplorer: a program for Hirshfeld surface analysis, visualization and quantitative analysis of molecular crystals. *Journal of Applied Crystallography*, 54, 6 2021.
- [101] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. Plumed 2: New feathers for an old bird. *Computer Physics Communications*, 185:604–613, 2014.
- [102] Maxwell Fulford, Matteo Salvalaglio, and Carla Molteni. Deepice: A deep neural network approach to identify ice and water molecules. *Journal of Chemical Information and Modeling*, 59:2141–2149, 2019.
- [103] Elise Duboué-Dijon and Damien Laage. Characterization of the local structure in liquid water by various order parameters. *The Journal of Physical Chemistry B*, 119:8406–8418, 2015.
- [104] Hideo Doi, Kazuaki Z Takahashi, and Takeshi Aoyagi. Mining of effective local order parameters for classifying crystal structures: A machine learning study. *The Journal of Chemical Physics*, 152:214501, 2020.
- [105] Pablo M Piaggi, Omar Valsson, and Michele Parrinello. Enhancing entropy and enthalpy fluctuations to drive crystallization in atomistic simulations. *Physical Review Letters*, 119:15701, 2017.

- [106] R E Nettleton and M S Green. Expression in terms of molecular distribution functions for the entropy density in an infinite system. *The Journal of Chemical Physics*, 29:1365–1370, 1958.
- [107] T M Truskett, S Torquato, and P G Debenedetti. Towards a quantification of disorder in materials: Distinguishing equilibrium and glassy sphere packings. *Physical Review E*, 62:993–1001, 6 2000.
- [108] Emanuela Giuffr , Santi Prestipino, Franz Saija, A Marco Saitta, and Paolo V Giaquinta. Entropy from correlations in tip4p water. *Journal of Chemical Theory and Computation*, 6:625–636, 2010.
- [109] Cui Zhang, Leonardo Spanu, and Giulia Galli. Entropy of liquid water from ab initio molecular dynamics. *The Journal of Physical Chemistry B*, 115:14190–14195, 2011.
- [110] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *WIREs Computational Molecular Science*, 1:826–843, 2011.
- [111] Matteo Salvalaglio, Pratyush Tiwary, Giovanni Maria Maggioni, Marco Mazzotti, and Michele Parrinello. Overcoming time scale and finite size limitations to compute nucleation rates from small scale well tempered metadynamics simulations. *The Journal of Chemical Physics*, 145:211925, 2016.
- [112] Amir Haji-Akbari and Pablo G Debenedetti. Direct calculation of ice homogeneous nucleation rate for a molecular model of water. *Proceedings of the National Academy of Sciences*, 112:10582–10588, 2015.
- [113] Ryan S DeFever, Colin Targonski, Steven W Hall, Melissa C Smith, and Sapna Sarupria. A generalized deep learning approach for local structure identification in molecular simulations. *Chemical Science*, 10:7503–7515, 2019.
- [114] Sun-Ting Tsai, Zachary Smith, and Pratyush Tiwary. SGOOP-d: Estimating kinetic distances and reaction coordinate dimensionality for rare event systems from biased/unbiased simulations. *Journal of Chemical Theory and Computation*, 17:6757–6765, 2021.
- [115] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25:1157–1174, 2004.
- [116] David A Case, Thomas E Cheatham III, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz Jr., Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The Amber biomolecular simulation programs. *Journal of computational chemistry*, 26:1668–1688, 2005.
- [117] Mark James Abraham, Teemu Murtola, Roland Schulz, Szil rd P ll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

- [118] Massimiliano Bonomi, Giovanni Bussi, and Carlo Camilloni Camilloni. Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods*, 16:670–673, 2019.
- [119] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126:14101, 2007.
- [120] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52:7182–7190, 1981.
- [121] Ulrich Essmann, Lalith Perera, and Max L Berkowitz. A smooth particle mesh ewald method. *The Journal of Chemical Physics*, 103:8577–8593, 1995.
- [122] Berk Hess. P-lincs: a parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation*, 4:116–122, 2007.
- [123] David J Wales and Jonathan P K Doye. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101:5111–5116, 1997.
- [124] Matteo Salvalaglio, Marco Mazzotti, and Michele Parrinello. Urea homogeneous nucleation mechanism is solvent dependent. *Faraday Discuss.*, 179:291–307, 2015.
- [125] Eun Hee Lee. A practical guide to pharmaceutical polymorph screening selection. *Asian Journal of Pharmaceutical Sciences*, 9:163–175, 2014.
- [126] Katarina E Blow, David Quigley, and Gabriele C Soso. The seven deadly sins: When computing crystal nucleation rates, the devil is in the details. *The Journal of Chemical Physics*, 155:40901, 2021.
- [127] Matteo Salvalaglio, Claudio Perego, Federico Giberti, Marco Mazzotti, and Michele Parrinello. Molecular-dynamics simulations of urea nucleation from aqueous solution. *Proceedings of the National Academy of Sciences*, 112(1):E6–E14, 2015.
- [128] Mark N Joswiak, Michael F Doherty, and Baron Peters. Ion dissolution mechanism and kinetics at kink sites on NaCl surfaces. *Proceedings of the National Academy of Sciences*, 115:656–661, 2018.
- [129] Rosanne N W Zeiler and Peter G Bolhuis. Numerical study of the effect of thiol–disulfide exchange in the cluster phase of β -lactoglobulin aggregation. *Faraday Discuss.*, 158:461–477, 2012.
- [130] Sarwar Hussain and Amir Haji-Akbari. How to quantify and avoid finite size effects in computational studies of crystal nucleation: The case of homogeneous crystal nucleation. *The Journal of Chemical Physics*, 156:54503, 2022.
- [131] Michael H Nielsen, Shaul Aloni, and James J De Yoreo. In situ tem imaging of CaCO₃ nucleation reveals coexistence of direct and indirect pathways. *Science*, 345:1158–1162, 2014.

- [132] Takayuki Nakamuro, Masaya Sakakibara, Hiroki Nada, Koji Harano, and Eiichi Nakamura. Capturing the moment of emergence of crystal nucleus from disorder. *Journal of the American Chemical Society*, 143:1763–1767, 2021.
- [133] James J De Yoreo, Pupa U P A Gilbert, Nico A J M Sommerdijk, R Lee Penn, Stephen Whitelam, Derk Joester, Hengzhong Zhang, Jeffrey D Rimer, Alexandra Navrotsky, Lilian F Banfield, Adam F Wallace, F Marc Michel, Fiona C Meldrum, Helmut Cölfen, and Patricia M Dove. Crystallization by particle attachment in synthetic, biogenic, and geologic environments. *Science*, 349:aaa6760, 2015.
- [134] Benjamin Abécassis, Matthew W Greenberg, Vivekananda Bal, Brandon M McMurtry, Michael P Campos, Lilian Guillemeney, Benoit Mahler, Sylvain Prevost, Lewis Sharpnack, Mark P Hendricks, Daniel DeRossa, Ellie Bennett, Natalie Saenz, Baron Peters, and Jonathan S Owen. Persistent nucleation and size dependent attachment kinetics produce monodisperse PbS nanocrystals. *Chemical Science*, 13:4977–4983, 2022.
- [135] Jim De Yoreo. More than one pathway. *Nature Materials*, 12:284–285, 6 2013.
- [136] David E Shaw, Peter J Adams, Asaph Azaria, Joseph A Bank, Brannon Batson, Alistair Bell, Michael Bergdorf, Jhanvi Bhatt, J Adam Butts, Timothy Correia, Robert M Dirks, Ron O Dror, Michael P Eastwood, Bruce Edwards, Amos Even, Peter Feldmann, Michael Fenn, Christopher H Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Maria Gorlatova, Brian Greskamp, J P Grossman, Justin Gullingsrud, Anissa Harper, William Hasenplaugh, Mark Heily, Benjamin Colin Heshmat, Jeremy Hunt, Douglas J Ierardi, Lev Islerovich, Bryan L Jackson, Nick P Johnson, Mollie M Kirk, John L Klepeis, Jeffrey S Kuskin, Kenneth M Mackenzie, Roy J Mader, Richard McGowen, Adam McLaughlin, Mark A Moraes, Mohamed H Nasr, Lawrence J Nociolo, Lief O’Donnell, Andrew Parker, Jon L Peticolas, Goran Pocina, Cristian Predescu, Terry Quan, John K Salmon, Carl Schwinck, Keun Sup Shim, Naseer Siddique, Jochen Spengler, Tamas Szalay, Raymond Tabladillo, Reinhard Tartler, Andrew G Taube, Michael Theobald, Brian Towles, William Vick, Stanley C Wang, Michael Wozniakowski, Madeleine J Weingarten, John M Williams, and Kevin A Yuh. Anton 3: Twenty microseconds of molecular dynamics simulation before lunch. Association for Computing Machinery, 2021.
- [137] Yihang Wang, Joao Marcelo Lamim Ribeiro, and Pratyush Tiwary. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Current Opinion in Structural Biology*, 61:139–145, 2020.
- [138] Michele Ceriotti. Unsupervised machine learning in atomistic simulations, between predictions and understanding. *The Journal of Chemical Physics*, 150:150901, 2019.
- [139] Moritz Hoffmann, Martin Scherer, Tim Hempel, Andreas Mardt, Brian de Silva, Brooke E Husic, Stefan Klus, Hao Wu, Nathan Kutz, Steven L Brunton, and Frank Noé. Deeptime: a python library for machine learning dynamical models from time series data. *Machine Learning: Science and Technology*, 3:15009, 2021.
- [140] Mahdi Ghorbani, Samarjeet Prasad, Jeffery B Klauda, and Bernard R Brooks. Graphvampnet, using graph neural networks and variational approach to Markov processes for

- dynamical modeling of biomolecules. *The Journal of Chemical Physics*, 156:184103, 2022.
- [141] Suvo Banik, Debdas Dhabal, Henry Chan, Sukriti Manna, Mathew Cherukara, Valeria Molinero, and Subramanian K R S Sankaranarayanan. CEGANN: Crystal edge graph attention neural network for multiscale classification of materials environment. *npj Computational Materials*, 9:23, 2 2023.
 - [142] Richard S Hong, Eric J Chan, Leslie Vogt-Maranto, Alessandra Mattei, Ahmad Y Sheikh, and Mark E Tuckerman. Insights into the polymorphic structures and enantiotropic layer-slip transition in paracetamol form III from enhanced molecular dynamics. *Crystal Growth & Design*, 21:886–896, 2021.
 - [143] Yusuke Mori, Kei ichi Okazaki, Toshifumi Mori, Kang Kim, and Nobuyuki Matubayasi. Learning reaction coordinates via cross-entropy minimization: Application to alanine dipeptide. *The Journal of Chemical Physics*, 153:54115, 2020.
 - [144] Jutta Rogal, Elia Schneider, and Mark E Tuckerman. Neural-network-based path collective variables for enhanced sampling of phase transformations. *Physical Review Letters*, 123:245701, 6 2019.
 - [145] Camilo Velez-Vega, Ernesto E Borrero, and Fernando A Escobedo. Kinetics and mechanism of the unfolding native-to-loop transition of Trp-cage in explicit solvent via optimized forward flux sampling simulations. *The Journal of Chemical Physics*, 133:105103, 2010.
 - [146] Sarwar Hussain and Amir Haji-Akbari. Studying rare events using forward-flux sampling: Recent breakthroughs and future outlook. *The Journal of Chemical Physics*, 152:60901, 2020.
 - [147] Ao Ma and Aaron R Dinner. Automatic method for identifying reaction coordinates in complex systems. *The Journal of Physical Chemistry B*, 109:6769–6779, 2005.
 - [148] Ernesto E Borrero and Fernando A Escobedo. Reaction coordinates and transition pathways of rare events via forward flux sampling. *The Journal of Chemical Physics*, 127:164101, 2007.
 - [149] Ziyue Zou, Sun-Ting Tsai, and Pratyush Tiwary. Toward automated sampling of polymorph nucleation and free energies with the SGOOP and metadynamics. *The Journal of Physical Chemistry B*, 125:13049–13056, 2021.
 - [150] Christian R Schwantes and Vijay S Pande. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *Journal of Chemical Theory and Computation*, 9:2000–2009, 2013.
 - [151] Nils E R Zimmermann, Bart Vorselaars, David Quigley, and Baron Peters. Nucleation of NaCl from aqueous solution: Critical sizes, ion-attachment kinetics, and rates. *Journal of the American Chemical Society*, 137:13352–13361, 2015.

- [152] A Arjun and P G Bolhuis. Molecular understanding of homogeneous nucleation of CO₂ hydrates using transition path sampling. *The Journal of Physical Chemistry B*, 125:338–349, 2021.
- [153] Liam C Jacobson, Masakazu Matsumoto, and Valeria Molinero. Order parameters for the multistep crystallization of clathrate hydrates. *The Journal of Chemical Physics*, 135(7), 2011.
- [154] Matej Badin and Roman Martoňák. Nucleating a different coordination in a crystal under pressure: A study of the b1-b2 transition in NaCl by metadynamics. *Physical Review Letters*, 127:105701, 6 2021.
- [155] Hongxing Song, Leslie Vogt-Maranto, Ren Wiscons, Adam J Matzger, and Mark E Tuckerman. Generating cocrystal polymorphs with information entropy driven by molecular dynamics-based enhanced sampling. *The Journal of Physical Chemistry Letters*, 11:9751–9758, 2020.
- [156] Pieter Rein ten Wolde and Daan Frenkel. Computer simulation study of gas–liquid nucleation in a Lennard-Jones system. *The Journal of Chemical Physics*, 109:9901–9918, 1998.
- [157] Kristof M Bal. Nucleation rates from small scale atomistic simulations and transition state theory. *The Journal of Chemical Physics*, 155:144111, 2021.
- [158] Daniele Moroni, Pieter Rein ten Wolde, and Peter G Bolhuis. Interplay between structure and size in a critical crystal nucleus. *Physical Review Letters*, 94:235703, 2005.
- [159] Pieter ten Wolde, Maria J Ruiz-Montero, and Daan Frenkel. Numerical calculation of the rate of crystal nucleation in a Lennard-Jones system at moderate undercooling. *The Journal of Chemical Physics*, 104:9932–9947, 1996.
- [160] Igor Volkov, Marek Cieplak, Joel Koplik, and Jayanth R Banavar. Molecular dynamics simulations of crystallization of hard spheres. *Physical Review E*, 66:61401, 2002.
- [161] Stefan Auer and Daan Frenkel. Prediction of absolute crystal-nucleation rate in hard-sphere colloids. *Nature*, 409:1020–1023, 2001.
- [162] Manan Chopra, Marcus Müller, and J J de Pablo. Order-parameter-based Monte Carlo simulation of crystallization. *The Journal of Chemical Physics*, 124:134102, 2006.
- [163] Caroline Desgranges and Jerome Delhommelle. Molecular mechanism for the cross-nucleation between polymorphs. *Journal of the American Chemical Society*, 128:10368–10369, 2006.
- [164] Ravi Radhakrishnan and Bernhardt L. Trout. Nucleation of hexagonal ice (ih) in liquid water. *Journal of the American Chemical Society*, 125(25):7743–7747, 2003.
- [165] Ravi Radhakrishnan and Bernhardt L Trout. Nucleation of crystalline phases of water in homogeneous and inhomogeneous environments. *Physical Review Letters*, 90:158301, 6 2003.

- [166] D Quigley and P M Rodger. Metadynamics simulations of ice nucleation and growth. *The Journal of Chemical Physics*, 128:154518, 2008.
- [167] Andrey V Brukhno, Jamshed Anwar, Ruslan Davidchack, and Richard Handel. Challenges in molecular simulation of homogeneous ice nucleation. *Journal of Physics: Condensed Matter*, 20:494243, 6 2008.
- [168] Aleks Reinhardt, Jonathan P K Doye, Eva G Noya, and Carlos Vega. Local order parameters for use in driving homogeneous ice nucleation with all-atom models of water. *The Journal of Chemical Physics*, 137:194504, 2012.
- [169] Wolfgang Lechner and Christoph Dellago. Accurate determination of crystal structures based on averaged local bond order parameters. *The Journal of Chemical Physics*, 129:114707, 2008.
- [170] Pablo M Piaggi and Michele Parrinello. Calculation of phase diagrams in the multithermal-multibaric ensemble. *The Journal of Chemical Physics*, 150, 2019.
- [171] Jérôme Hénin, Tony Lelièvre, Michael R Shirts, Omar Valsson, and Lucie Delemotte. Enhanced sampling methods for molecular dynamics simulations [article v1.0]. *Living Journal of Computational Molecular Science*, 4:1583, 6 2022.
- [172] Taraknath Mandal and Ronald G Larson. Nucleation of urea from aqueous solution: Structure, critical size, and rate. *The Journal of Chemical Physics*, 146:134501, 2017.
- [173] Nicholas F Francia, Louise S Price, Jonas Nyman, Sarah L Price, and Matteo Salvalaglio. Systematic finite-temperature reduction of crystal energy landscapes. *Crystal Growth & Design*, 20:6847–6862, 2020.
- [174] Avik Mahata and Mohsen Asle Zaeem. Size effect in molecular dynamics simulation of nucleation process during solidification of pure metals: investigating modified embedded atom method interatomic potentials. *Modelling and Simulation in Materials Science and Engineering*, 27:85015, 6 2019.
- [175] J R Espinosa, J M Young, H Jiang, D Gupta, C Vega, E Sanz, P G Debenedetti, and A Z Panagiotopoulos. On the calculation of solubilities via direct coexistence simulations: Investigation of NaCl aqueous solutions and Lennard-Jones binary mixtures. *The Journal of Chemical Physics*, 145:154111, 2016.
- [176] Tarak Karmakar, Pablo M Piaggi, and Michele Parrinello. Molecular dynamics simulations of crystal nucleation from solution at constant chemical potential. *Journal of Chemical Theory and Computation*, 15:6923–6930, 2019.
- [177] Angelo Gavezzotti. Can a computer crystallize a liquid? molecular simulation of continuous trajectories from liquid to crystalline n-hexane. *CrystEngComm*, 13:3573–3579, 2011.
- [178] Peng Yi, C Rebecca Locker, and Gregory C Rutledge. Molecular dynamics simulation of homogeneous crystal nucleation in polyethylene. *Macromolecules*, 46:4723–4733, 2013.

- [179] Frank Noé, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi. Machine learning for molecular simulation. *Annual Review of Physical Chemistry*, 71:361–390, 2020.
- [180] Ming Chen. Collective variable-based enhanced sampling and machine learning. *The European Physical Journal B*, 94:211, 10 2021.
- [181] Sapna Sarupria, Steven W. Hall, and Jutta Rogal. Machine learning for molecular simulations of crystal nucleation and growth. *MRS Bulletin*, 47(9):949–957, Sep 2022.
- [182] Eric R Beyerle, Ziyue Zou, and Pratyush Tiwary. Recent advances in describing and driving crystal nucleation using machine learning and artificial intelligence. *Current Opinion in Solid State and Materials Science*, 27:101093, 2023.
- [183] Shams Mehdi, Zachary Smith, Lukas Herron, Ziyue Zou, and Pratyush Tiwary. Enhanced sampling with machine learning. *Annual Review of Physical Chemistry*, 75:347–370, 2024.
- [184] Hendrik Jung, Roberto Covino, A Arjun, Christian Leitold, Christoph Dellago, Peter G Bolhuis, and Gerhard Hummer. Machine-guided path sampling to discover mechanisms of molecular self-organization. *Nature Computational Science*, pages 1–12, 2023.
- [185] Florian M Dietrich, Xavier R Advincula, Gianpaolo Gobbo, Michael A Bellucci, and Matteo Salvalaglio. Machine learning nucleation collective variables with graph neural networks. *Journal of Chemical Theory and Computation*, 20:1600–1611, 2024.
- [186] Nicholas S M Herringer, Siva Dasetty, Diya Gandhi, Junhee Lee, and Andrew L Ferguson. Permutationally invariant networks for enhanced sampling (PINES): Discovery of multimolecular and solvent-inclusive collective variables. *Journal of Chemical Theory and Computation*, 20:178–198, 2024.
- [187] Oren Elishav, Roy Podgaetsky, Olga Meikler, and Barak Hirshberg. Collective variables for conformational polymorphism in molecular crystals. *The Journal of Physical Chemistry Letters*, 14:971–976, 2023.
- [188] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.
- [189] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Saucedo Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems*, 30, 2017.
- [190] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120:145301, 2018.

- [191] Peter Bjørn Jørgensen, Karsten Wedel Jacobsen, and Mikkel N Schmidt. Neural message passing with edge updates for predicting properties of molecules and materials. *arXiv preprint arXiv:1806.03146*, 2018.
- [192] Alireza Moradzadeh, Hananeh Oliaei, and Narayana R Aluru. Topology-based phase identification of bulk, interface, and confined water using an edge-conditioned convolutional graph neural network. *The Journal of Physical Chemistry C*, 127:2612–2621, 2023.
- [193] QHwan Kim, Joon-Hyuk Ko, Sunghoon Kim, and Wonho Jhe. GCIceNet: a graph convolutional network for accurate classification of water phases. *Physical Chemistry Chemical Physics*, 22:26340–26350, 2020.
- [194] Fabio Pietrucci and Wanda Andreoni. Graph theory meets ab initio molecular dynamics: atomic structures and transformations at the nanoscale. *Physical Review Letters*, 107:85504, 2011.
- [195] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87:184115, 2013.
- [196] Martin Sipka, Andreas Erlebach, and Lukas Grajciar. Constructing collective variables using invariant learned representations. *Journal of Chemical Theory and Computation*, 19:887–901, 2023.
- [197] Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57:238–247, 1989.
- [198] Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3693–3702, 2017.
- [199] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016.
- [200] David Eppstein, Michael S Paterson, and F Frances Yao. On nearest-neighbor graphs. *Discrete Computational Geometry*, 17:263–282, 1997.
- [201] Saulius Gražulis, Daniel Chateigner, Robert T Downs, A F T Yokochi, Miguel Quirós, Luca Lutterotti, Elena Manakova, Justas Butkus, Peter Moeck, and Armel Le Bail. Crystallography open database – an open-access collection of crystal structures. *Journal of Applied Crystallography*, 42:726–729, 6 2009.
- [202] Alice Dawson, David R Allan, Scott A Belmonte, Stewart J Clark, William I F David, Pamela A McGregor, Simon Parsons, Colin R Pulham, and Lindsay Sawyer. Effect of high pressure on the crystal structures of polymorphs of glycine. *Crystal Growth & Design*, 5:1415–1427, 2005.
- [203] Yoichi Iitaka. The crystal structure of β -glycine. *Acta Crystallographica*, 13:35–45, 1960.

- [204] Clare F Macrae, Ioana Sovago, Simon J Cottrell, Peter T A Galek, Patrick McCabe, Elna Pidcock, Michael Platings, Greg P Shields, Joanna S Stevens, Matthew Towler, et al. Mercury 4.0: From visualization to analysis, design and prediction. *Journal of Applied Crystallography*, 53:226–235, 2020.
- [205] Aidan P Thompson, H Metin Aktulga, Richard Berger, Dan S Bolintineanu, W Michael Brown, Paul S Crozier, Pieter J in't Veld, Axel Kohlmeyer, Stan G Moore, Trung Dac Nguyen, et al. Lammmps-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271:108171, 2022.
- [206] Herman J C Berendsen, David van der Spoel, and Rudi van Drunen. GROMACS: a message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91:43–56, 1995.
- [207] Joseph W Lynch. Molecular structure and function of the glycine receptor chloride channel. *Physiological Reviews*, 84:1051–1095, 2004.
- [208] E V Boldyreva, V A Drebuschak, T N Drebuschak, I E Paukov, Yu A Kovalevskaya, and E S Shutova. Polymorphism of glycine, part I. *Journal of Thermal Analysis and Calorimetry*, 73:409–418, 2003.
- [209] Alexander Stukowski. Structure identification methods for atomistic simulations of crystalline materials. *Modelling and Simulation in Materials Science and Engineering*, 20:45021, 2012.
- [210] Tongsik Lee, Michael I Baskes, Steven M Valone, and J D Doll. Atomistic modeling of thermodynamic equilibrium and polymorphism of iron. *Journal of Physics: Condensed Matter*, 24:225404, 2012.
- [211] X Ou. Molecular dynamics simulations of fcc-to-bcc transformation in pure iron: A review. *Materials Science and Technology*, 33:822–835, 2017.
- [212] A Fernández Guillermet and Per Gustafson. An assessment of the thermodynamic properties and the(p, T) phase diagram of iron. *High Temperatures-High Pressures*, 16:591–610, 1984.
- [213] S Alireza Etesami and Ebrahim Asadi. Molecular dynamics for near melting temperatures simulations of metals using modified embedded-atom method. *Journal of Physics and Chemistry of Solids*, 112:61–72, 2018.
- [214] Byeong-Joo Lee and Michael I Baskes. Second nearest-neighbor modified embedded-atom-method potential. *Physical Review B*, 62:8564, 2000.
- [215] J Dana. Honeycutt and Hans C Andersen. Molecular dynamics study of melting and freezing of small Lennard-Jones clusters. *The Journal of Physical Chemistry*, 91:4950–4963, 1987.

- [216] Daniel Faken and Hannes Jónsson. Systematic analysis of local atomic structure combined with 3D computer graphics. *Computational Materials Science*, 2:279–286, 1994.
- [217] Craig L Bull, Giles Flowitt-Hill, Stefano de Gironcoli, Emine Küçükbenli, Simon Parsons, Cong Huy Pham, Helen Y Playford, and Matthew G Tucker. ζ -glycine: insight into the mechanism of a polymorphic phase transition. *IUCrJ*, 4:569–574, 6 2017.
- [218] Elena Boldyreva. Glycine: The gift that keeps on giving. *Israel Journal of Chemistry*, 61:828–850, 2021.
- [219] S V Goryainov, E V Boldyreva, and E N Kolesnik. Raman observation of a new (ζ) polymorph of glycine? *Chemical Physics Letters*, 419:496–500, 2006.
- [220] Bojun Liu, Mingyi Xue, Yunrui Qiu, Kirill A Konovalov, Michael S O’Connor, and Xuhui Huang. Graphvampnets for uncovering slow collective variables of self-assembly dynamics. *The Journal of Chemical Physics*, 159:94901, 2023.
- [221] Tarak Karmakar, Aaron R Finney, Matteo Salvalaglio, A Ozgur Yazaydin, and Claudio Perego. Non-equilibrium modeling of concentration-driven processes with constant chemical potential molecular dynamics simulations. *Accounts of Chemical Research*, 56:1156–1167, 2023.
- [222] Daan Frenkel and Anthony JC Ladd. New Monte Carlo method to compute the free energy of arbitrary solids. application to the fcc and hcp phases of hard spheres. *The Journal of Chemical Physics*, 81:3188–3193, 1984.
- [223] Evan Pretti and Jeetain Mittal. Extension of the Einstein molecule method for solid free energy calculation to non-periodic and semi-periodic systems. *The Journal of Chemical Physics*, 151:54105, 2019.
- [224] Yihang Wang, Lukas Herron, and Pratyush Tiwary. From data to noise to data for mixing physics across temperatures with generative artificial intelligence. *Proceedings of the National Academy of Sciences*, 119:e2203656119, 2022.
- [225] Peter Schwerdtfeger and David J Wales. 100 years of the Lennard-Jones potential. *Journal of Chemical Theory and Computation*, 20:3379–3405, 2024.
- [226] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.
- [227] Peter G Bolhuis, Christoph Dellago, and David Chandler. Reaction coordinates of biomolecular isomerization. *Proceedings of the National Academy of Sciences*, 97:5877–5882, 2000.
- [228] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

- [229] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks?, 2022.
- [230] Suemin Lee, Dedi Wang, Markus Seeliger, and Pratyush Tiwary. Calculating protein-ligand residence times through state predictive information bottleneck based enhanced sampling. *bioRxiv*, 2024.
- [231] Bodhi P Vani, Akashnathan Aranganathan, Dedi Wang, and Pratyush Tiwary. AlphaFold2-RAVE: From sequence to Boltzmann ranking. *Journal of Chemical Theory and Computation*, 19:4351–4354, 2023.
- [232] Bodhi P Vani, Akashnathan Aranganathan, and Pratyush Tiwary. Exploring kinase Asp-Phe-Gly (DFG) loop conformational stability with AlphaFold2-RAVE. *Journal of Chemical Information and Modeling*, 64:2789–2797, 2024.
- [233] Xinyu Gu, Akashnathan Aranganathan, and Pratyush Tiwary. Empowering alphafold2 for protein conformation selective drug discovery with alphafold2-rave. 2024.
- [234] Ziyue Zou and Pratyush Tiwary. Enhanced sampling of crystal nucleation with graph representation learnt variables. *The Journal of Physical Chemistry B*, 128:3037–3045, 2024.
- [235] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.
- [236] David Wales. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge Molecular Science. Cambridge University Press, 2004.
- [237] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.
- [238] Glenn J Martyna, Douglas J Tobias, and Michael L Klein. Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics*, 101:4177–4189, 1994.
- [239] Yuriy G Bushuev, Svetlana V Davletbaeva, and Oscar I Koifman. Molecular dynamics simulations of aqueous glycine solutions. *CrystEngComm*, 19:7197–7206, 2017.