

Entropy as a tool for crystal discovery

Thèse N° 9148

Présentée le 25 janvier 2019

à la Faculté des sciences et techniques de l'ingénieur
Laboratoire de théorie et simulation des matériaux
Programme doctoral en science et génie des matériaux

pour l'obtention du grade de Docteur ès Sciences

par

PABLO MIGUEL PIAGGI

Acceptée sur proposition du jury

Prof. P. Muralt, président du jury
Prof. N. Marzari, Prof. M. Parrinello, directeurs de thèse
Prof. M. Mazzotti, rapporteur
Prof. M. Salvalaglio, rapporteur
Prof. M. Ceriotti, rapporteur

2019

Desvarío laborioso y empobrecedor el de componer vastos libros;
el de explayar en quinientas páginas una idea cuya perfecta
exposición oral cabe en pocos minutos.
— Jorge Luis Borges, *Ficciones* (1956)

To Belén and my parents

Acknowledgments

First and foremost I would like to thank Michele Parrinello. Not only has he taught me to perform research and write scientific articles, but he has also transmitted to me his passion for science. His relentless activity as a scientist is an inspiring example that will probably have long lasting effects in my scientific life. One of the reasons why I chose to make this thesis as a compilation of articles is the tremendous effort that was put by Michele and me into writing these four articles. I thought that anything that I would rewrite about them would not do justice to the work done before. Also, as wisely pointed out by Borges in his *Ficciones* preface, sometimes it is best to be concise.

I am also deeply indebted to Nicola Marzari and the NCCR MARVEL. Nicola was crucial to the success of my PhD and supported me at every step of it. The MARVEL project provided the funding for the four years of my PhD and it also allowed me to interact with the simulation community of Switzerland. By virtue of the project I had the opportunity to meet remarkable professors and young scientists that hopefully one day will be my colleagues.

I would also like to thank all the people I had the pleasure to work with during my PhD. An incomplete list is: Omar Valsson, Paramvir Ahlawat, Haiyang Niu, Michele Invernizzi, Tarak Karmakar, Sergio Pérez-Conesa, Luigi Bonati, and Zoran Bjelobrk.

Daniela Wirz has also been instrumental in the development of my thesis work, helping me out from the very beginning with a myriad of administrative procedures.

Lastly, I would like to thank the people that have helped me to arrive here. These are Jorge Kohanoff, Mario Del Popolo, Roberto Pasianot, and Ana María Llois.

Lugano, October 22, 2018

P. M. P.

Abstract

The computational prediction of crystal structures has emerged as an useful alternative to expensive and often cumbersome experiments. We propose an approach to the prediction of crystal structures and polymorphism based on reproducing the crystallization process on the computer. The main hurdle faced by such an approach is that crystallization usually takes place in timescales much longer than those that can be afforded with standard molecular simulations. In order to circumvent this difficulty we construct a bias potential which is a function of one or more collective variables and whose role is to promote crystallization. This approach can only have true predictive power if the collective variable is crystal structure agnostic, that is to say, it does not include information about the geometry of any particular crystal structure. In order to achieve this goal, we take inspiration from thermodynamics and propose to use an entropy surrogate as collective variable. We use an approximation for the entropy based on the radial distribution function $g(r)$. Using this collective variable we are able to explore polymorphism in simple metals and molecular crystals. We study the case of urea and find a new polymorph stabilized by entropic effects. We also propose a projection of the collective variable onto each atom that is useful to characterize atomic environments. Lastly, we introduce a generalized Kullback-Leibler divergence that measures the distance between two radial distribution functions. We apply this divergence to the automatic classification of the polymorphs that crystallize during the simulations.

Riassunto

L'uso di simulazioni al computer per scoprire strutture cristalline offre un'alternativa agli esperimenti che risultano spesso costosi e difficili. In questa tesi, proponiamo un metodo computazionale che permette, tramite la simulazione della cristallizzazione, di scoprire diverse strutture cristalline. Lo studio della cristallizzazione è complesso in quanto è un processo che avviene su scale di tempo molto più lunghe di quelle che si possono raggiungere con simulazioni molecolari standard. Per superare questa difficoltà costruiamo un potenziale di bias in funzione di una o più variabili collettive, il cui scopo è promuovere la cristallizzazione. Affinché questo approccio possa identificare la struttura cristallina del sistema, è necessario che queste variabili non includano informazioni sulla geometria di nessuna struttura in particolare. A tal fine, prendendo ispirazione dalla termodinamica, proponiamo di usare come variabile collettiva un'espressione approssimata dell'entropia, basata sulla funzione di distribuzione radiale $g(r)$. Con questo approccio siamo in grado di trovare diversi polimorfi sia nel caso di metalli semplici che in quello di cristalli molecolari. In particolare, nel caso dell'urea abbiamo identificato un nuovo polimorfo che risulta stabilizzato dall'entropia. Oltre che per promuovere la cristallizzazione, questa approssimazione dell'entropia si rivela utile per altri scopi. Mostriamo che proiettando la variabile collettiva su ogni atomo, siamo in grado di caratterizzare l'ordine atomico locale. Inoltre, se guardiamo all'approssimazione dell'entropia come una distanza tra la $g(r)$ del sistema e quella del gas perfetto, possiamo generalizzarla per misurare la distanza tra due funzioni di distribuzione radiale. Un'applicazione di questa distanza è la classificazione dei polimorfi che emergono durante le simulazioni.

Contents

Acknowledgments	v
Abstract (English/Italiano)	vii
1 Introduction	1
2 Articles	5
2.1 A variational approach to nucleation simulation	5
2.2 Enhancing entropy and enthalpy fluctuations to drive crystallization in atomistic simulations	18
2.3 Entropy fingerprint for local crystalline order	25
2.4 Predicting polymorphism using orientational entropy	35
3 Conclusions	47
4 Future work	49
4.1 Two-body excess entropy collective variable	49
4.2 Generalized Kullback-Leibler divergence	52
Other articles	55
Bibliography	67
Curriculum Vitae	69

1 Introduction

It is hard to think of a material that does not undergo some form of crystallization during its manufacturing process. Recall for instance the crystallization of metal oxides using the sol-gel process and the recrystallization of steel after substantial plastic deformation. One of the main characteristics of all crystallization processes is the existence of a kinetic bottleneck for the transformation. Lets consider an everyday life example to illustrate this point. It is well-known that water freezes at 0°C. However, liquid water can exist well below this temperature and be trapped in the metastable liquid state for relatively long times. In this state, an external perturbation or simply a natural fluctuation can help the system squeeze through the kinetic bottleneck and trigger crystallization. A natural conclusion is that when studying this phenomenon most of the time will be wasted waiting for the crystallization event to occur. This example illustrates that crystallization is an infrequent or *rare event*.

Since the dawn of atomistic computer simulations in the 1960s it was recognized that interesting insights into crystallization could be obtained from atomistic simulations [1–3]. Initially, two approaches were employed to tackle the rare event problem. One was the brute force approach in which one or more independent simulations are performed for very long times waiting for the relevant event to take place [4]. Another approach was to employ deep quenches in order to bring the process within the times accessible to atomistic simulations [5]. The latter procedure has the disadvantage that the supersaturation might be too large to be compared with experiments and moreover one risks entering in the spinodal decomposition regime.

Later on, many methods were developed to address rare events in molecular simulations. Some examples are transition path sampling [6], umbrella sampling [7], adaptive biasing force [8] and metadynamics [9]. The main objective of these methods is sampling all the relevant configurations including metastable states and transitions regions during a comparatively short simulation. Many of these methods are based on introducing a bias potential which is a function of a small number of collective

Chapter 1. Introduction

coordinates or collective variables. These collective variables are typically non linear functions of the atomic coordinates and provide a low dimensional description of the process under consideration. Once that an appropriate collective variable has been found the problem can be considered essentially solved. The main approach to the design of collective coordinates has been using physical or chemical intuition although recently machine learning techniques have also been employed [10–12].

Collective variables employed to study crystallization typically contain information about the geometry of the target crystal structure. Thus, one needs to choose a different collective variable for each crystal structure. Some examples of these variables are the Steinhardt parameters [13, 14], the variables for molecular crystals proposed by Santiso and Trout [15, 16] and the variable for the fcc structure proposed by Angioletti-Uberti et al [17]. It would be useful to develop a collective variable that is not based on any particular feature of the crystal structure. Some advantages of this variable would be that it could be applied effortlessly to different systems and it could even be used for predicting the crystal structure of a system.

However, finding such a collective variable is not trivial. The path that I shall take here is to take inspiration from a general physical principle. Recall that in all first order phase transitions there is an interplay between enthalpy and entropy. In crystallization in particular, the high energy and high entropy phase, the liquid, transforms into a low energy and low entropy phase, the solid. I propose to use enthalpy and entropy as collective variables. Enthalpy or energy have already been used as collective variables [18, 19] for instance in the context of the well tempered ensemble [20]. On the other hand entropy had not been employed before the work described in this thesis¹. This is not surprising since computing entropy can be a daunting task. Here I take a simple approach in which the entropy of the system is expressed in an approximated fashion. Although the idea of using approximated expressions for the entropy is quite general, in this thesis I shall only discuss the use of the so called two-body excess entropy [22, 23]. This approximation is based on a function of paramount importance in the theory of liquids, namely the radial distribution function (see for instance ref. [24]).

This thesis is made as a compilation of articles and is organized as follows. The first article *A variational approach to nucleation simulation* discusses a new method to enhance the sampling of nucleation events based on a variational principle introduced in ref. [25]. The second article *Enhancing entropy and enthalpy fluctuations to drive crystallization in atomistic simulations* is the landmark paper in which we introduced the idea of employing entropy surrogates as collective variables. The third article *Entropy based fingerprint for local crystalline order* describes the use of the two-body

¹Entropy was also used as a collective variable in ref. [21] although in that work the grand canonical ensemble is used and the entropy is a simple function of the energy of the system, namely $S = \frac{U-\mu}{T}$ where S is the entropy, U is the potential energy, μ is the chemical potential, and T is the temperature. Since in the grand canonical ensemble μ and T are constant, their collective variable is a function of the atomic coordinates \mathbf{R} only through the implicit dependence of the energy $U(\mathbf{R})$.

excess entropy to distinguish between liquid-like and solid-like environments. The fourth article *Predicting polymorphism using orientational entropy* presents a variant of the entropic collective variable tailored to study polymorphism in molecular crystals. Finally I present some conclusions and several ideas for the future development of the work done in this thesis.

Since this thesis does not include an introduction to the theoretical underpinnings of molecular simulation, I refer the reader to the book by Frenkel and Smit [26] for information about basic simulation methods and to a review by Valsson, Tiwary and Parrinello [27] for a discussion on enhanced sampling methods.

2 Articles

2.1 A variational approach to nucleation simulation

This article deals with a first order phase transition, but at variance with the other articles in this thesis, here we focus on the condensation of a Lennard-Jones fluid. Furthermore, we choose as collective variable a function of the coordination numbers instead of an entropy surrogate. The main feature of this work is the use of a variational principle to construct the bias potential. The standard approach to solve the variational problem is to expand the bias potential in a basis set. Here we take a different approach and we use as bias potential the recipe for the free energy provided by classical nucleation theory [28]. This work is part of an effort to replace the description of bias potentials using basis sets with expressions that reflect the physics of the phenomena under study. Other work in this direction is using the Ginzburg-Landau theory for the description of second order phase transitions [29].

I present the postprint version of the article published in Faraday Discussions. The supplementary information of this article is not included in this thesis but a link has been provided in the electronic version. My contribution to this article has been implementing the algorithms, performing the simulations, and writing the paper jointly with Prof. Parrinello.

Full bibliographic reference: Pablo M. Piaggi, Omar Valsson, and Michele Parrinello. A variational approach to nucleation simulation. *Faraday discussions*, 195: 557-568, 2016. doi: 10.1039/C6FD00127K. URL <http://dx.doi.org/10.1039/C6FD00127K>.

Copyright © 2016 Royal Society of Chemistry.

A variational approach to nucleation simulation

Pablo M. Piaggi^{1,2}, Omar Valsson^{2,3}, and Michele Parrinello^{2,3}

¹Theory and Simulation of Materials, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

²Facoltà di Informatica, Istituto di Scienze Computazionali, and National Center for Computational Design and Discovery of Novel Materials (MARVEL), Università della Svizzera italiana (USI), Via Giuseppe Buffi 13, CH-6900, Lugano, Switzerland

³Department of Chemistry and Applied Biosciences, ETH Zurich, c/o USI Campus, Via Giuseppe Buffi 13, CH-6900, Lugano, Switzerland

Abstract

We study by computer simulation the nucleation of a supersaturated Lennard-Jones vapor into the liquid phase. The large free energy barriers to transition make the time scale of this process impossible to study by ordinary molecular dynamics simulations. Therefore we use a recently developed enhanced sampling method [Valsson and Parrinello, *Phys. Rev. Lett.* 113, 090601 (2014)] based on the variational determination of a bias potential. We differ from previous applications of this method in that the bias is constructed on the basis of the physical model provided by the classical theory of nucleation. We examine the technical problems associated with this approach. Our results are very satisfactory and will pave the way for calculating the nucleation rates in many systems.

2.1.1 Introduction

Nucleation is a process that plays a prominent role in chemistry, engineering and materials science. Among many others, it finds applications in the pharmaceutical industry where crystal shape and structure dramatically affect the drugs potency [30]. Moreover, nucleation is not only technologically relevant but also scientifically interesting since it is a paradigmatic example of a self-assembly process [31]. There is thus a great interest in understanding, and eventually controlling, the way in which a new phase emerges from the parent one [32, 33]. In spite of the importance of nucleation, the small scales involved represent a formidable hurdle to experimental studies and thus the details of the process are not easy to unveil. In this regard, molecular simulation and theory could pave the way to a better understanding of the early stages of this process.

One of the simplest examples of homogeneous nucleation is the condensation of a supersaturated vapor. Being a fairly well understood phenomenon, it provides a useful

scenario to test new simulation methods. A vapor at a constant temperature and a pressure higher than the vapor pressure is in metastable equilibrium with respect to the liquid phase [28] and thus it will make a transition to the liquid phase in order to minimize its free energy. However, one of the characteristics of first order phase transitions is the ability to remain in the metastable state due to the existence of a free energy barrier. In this metastable state the system experiences density fluctuations that can be described as the fleeting appearance of small clusters of the new phase. Occasionally the fluctuations are so large that a critical cluster is formed and the whole system condensates. In order for this to occur, the system has to overcome a large energy barrier. This makes nucleation a rare event that cannot be sampled in ordinary simulations. Many solutions to this problem have been proposed [6, 7, 9] and applications to liquid-vapor nucleation have also been reported [34–36]. The closest in spirit to our approach is the work of ten Wolde and Frenkel [37]. In this work the free energy landscape of a Lennard-Jones fluid was explored via Monte Carlo (MC) simulations that used umbrella sampling [7] to overcome kinetic barriers.

Our paper has a strong methodological connotation since we aim at applying to this well-studied problem the newly developed variationally enhanced sampling (VES) method [25] in which the bias is determined via a variational procedure based on a suitably defined functional. A few applications of the method have already been presented in the literature [25, 38–41]. The default way of solving the variational problem is to expand the bias potential in a basis set and use the expansion coefficients as variational parameters. We differ from this approach in that, rather than following this procedure, we use a physically motivated expression derived from classical nucleation theory (CNT). This expression contains two empirical parameters, the supersaturation and the effective surface energy. We shall optimize the functional with respect to these two parameters. This will allow us to understand better the properties of the functional and lay the foundations for future work in which we plan to use a variant of the variational method [38, 42] designed to calculate nucleation rates.

2.1.2 Classical nucleation theory

The textbook way of describing nucleation phenomena is CNT. In CNT the cost of forming a cluster of the new phase (in our case the liquid one) can be expressed as:

$$\Delta F^{CNT}(n) = -\Delta\mu n + \sigma n^{2/3}, \quad (2.1.1)$$

where n is the number of atoms in the cluster, $\Delta\mu$ is the difference in chemical potential between the two phases (supersaturation), and σ is an effective interfacial energy. The first term represents the energy gain in going into the new more stable phase whereas the second term expresses the energetic cost of forming an interface between the liquid and the vapor. Since on average liquid clusters are spherical, σ can be related

to the surface free energy γ by $\sigma = (36\pi)^{1/3}\rho^{-2/3}\gamma$ where ρ is the density of the new phase. The supersaturation $\Delta\mu$ is often expressed as a dimensionless quantity called supersaturation ratio (S) and defined through $\Delta\mu = k_B T \log S$. The expression in equation (2.1.1) differs from the one normally found in textbooks in that the latter uses the drop radius R as a measure of its size. If one makes the assumption that the drops have a spherical shape, the relation between n and R is $n = \frac{4}{3}\pi R^3 \rho$. Substituting this relation in equation (2.1.1), the standard expression $\Delta F^{CNT}(R) = -\frac{4}{3}\pi R^3 \rho \Delta\mu + 4\pi R^2 \gamma$ is recovered.

2.1.3 Methods

The presence of high barriers often hinders exhaustive sampling in molecular simulations. Enhanced sampling methods aim at solving this problem. Many rely on the introduction of a bias potential V which is a function of a small number of collective coordinates \mathbf{s} . Among these, it is worth noting the historically important umbrella sampling [7], also used by ten Wolde and Frenkel [37], and metadynamics which has proved to be successful in a variety of fields [9, 27]. In the next subsection we describe the theoretical underpinnings of the recently introduced VES [25].

Variationally enhanced sampling

As in umbrella sampling and many other enhanced sampling methods, we project the high-dimensional \mathbf{R} space of the N particle system into a much smaller and smoother d -dimensional space by introducing the set of collective variables $\mathbf{s}(\mathbf{R}) = (s_1(\mathbf{R}), s_2(\mathbf{R}), \dots, s_d(\mathbf{R}))$ that give a coarse-grained description of the system. The free energy surface (FES) associated to the CV set \mathbf{s} is defined as:

$$F(\mathbf{s}) = -\frac{1}{\beta} \log \int d\mathbf{R} \delta(\mathbf{s} - \mathbf{s}(\mathbf{R})) e^{-\beta U(\mathbf{R})}, \quad (2.1.2)$$

where we have dropped an immaterial constant as we shall also do in the following.

In ref. [25] it was shown how to construct a bias potential $V(\mathbf{s})$ that acts on the CVs via the optimization of the following functional:

$$\Omega[V] = \frac{1}{\beta} \log \frac{\int d\mathbf{s} e^{-\beta[F(\mathbf{s})+V(\mathbf{s})]}}{\int d\mathbf{s} e^{-\beta F(\mathbf{s})}} + \int d\mathbf{s} p(\mathbf{s})V(\mathbf{s}), \quad (2.1.3)$$

where $p(\mathbf{s})$ is a chosen target probability distribution. This functional is convex and it is made stationary by the bias potential:

$$V(\mathbf{s}) = -F(\mathbf{s}) - \frac{1}{\beta} \log p(\mathbf{s}). \quad (2.1.4)$$

It follows that once the functional is minimized, the probability distribution of \mathbf{s} in the biased ensemble $p_V(\mathbf{s})$ is equal to the target distribution $p(\mathbf{s})$:

$$p_V(\mathbf{s}) = \frac{e^{-\beta[F(\mathbf{s})+V(\mathbf{s})]}}{\int d\mathbf{s} e^{-\beta[F(\mathbf{s})+V(\mathbf{s})]}} = p(\mathbf{s}). \quad (2.1.5)$$

The simplest choice for $p(\mathbf{s})$ is to consider the uniform target distribution $p(\mathbf{s}) = 1/\Omega_s$, where $\Omega_s = \int d\mathbf{s}$ is the volume of CV space. In this case $F(\mathbf{s}) = -V(\mathbf{s})$ as in standard metadynamics. If instead one takes $p(\mathbf{s}) \propto p_0(\mathbf{s})^{1/\gamma}$ where γ is greater than one and $p_0(\mathbf{s})$ is the equilibrium probability distribution of \mathbf{s} in the unbiased ensemble [39], the well-tempered metadynamics distribution is recovered. Other choices have been suggested [38, 40] and here we shall also make use of the added flexibility offered by the freedom of choosing $p(\mathbf{s})$. Once the bias has been determined, a standard reweighting procedure can be used to calculate statistical averages in the unbiased ensemble [7]. Details of the reweighting procedure are provided in the Supplementary Information (SI).

Collective variable

In order to use the CNT free energy expression to construct the bias, we need to define properly a CV that expresses in analytical and differentiable form the variable n in equation (2.1.1) as a function of \mathbf{R} . This requires first defining what is meant by liquid cluster. For this, we follow the procedure suggested by ten Wolde and Frenkel [37]. A pair of atoms is considered to belong to the same cluster if their distance is below an assigned radius r_c and each of them has at least n_c neighbors within r_c . Once the clusters have been defined, in order to write a CV that corresponds to the one in CNT one would have had to sort the clusters by their size and consider also their multiplicity, a procedure that would have been expensive and cumbersome.

In their work based on MC, ten Wolde and Frenkel [37] decided to use as CV the largest cluster. The idea behind this choice is that, as the system climbs the barrier, only the largest cluster survives. Eventually the free energy is reweighted to obtain the cluster size distribution (n) as needed in the CNT expression. In the present work, that is based on MD, the use of the largest cluster as CV would have caused problems in the calculation of the forces. In fact, the flag of the largest cluster can change abruptly from one set of atoms to another. Although this could have been remedied, we preferred to use as CV the total number of liquid-like atoms (n_l), a quantity that is easy to calculate. In the SI we describe in detail the calculation of n_l . As we shall see, for small systems where the probability of observing several clusters is negligible, this is a good choice and the resulting free energy as a function of this variable is similar to the CNT expression. For larger systems this is no longer the case but still using a reweighting procedure the cluster size distribution can be obtained (see the SI for details).

Using the CNT model for the bias potential

Having decided how to represent n , albeit in an approximate form, we first take $p(s)$ to be uniform and write for the bias the functional form:

$$\begin{aligned} V(s; \Delta\mu, \sigma) &= -\Delta F^{CNT}(s; \Delta\mu, \sigma) \\ &= -\left(-\Delta\mu s + \sigma s^{2/3}\right), \end{aligned} \quad (2.1.6)$$

where $s = n_l$. In equation (2.1.6) we have used the fact that if $p(s)$ is uniform, $V(s) = -F(s)$. Expression (2.1.6) is then inserted into $\Omega[V]$ and the functional is minimized relative to $\Delta\mu$ and σ . As already discussed in the introduction, this differs from the usual approach in which the bias $V(s)$ is expanded in an orthogonal basis set and the expansion coefficients are used as variational parameters.

In principle at this stage we could have moved to describe the calculation. However, before doing so, a practical issue needs to be addressed. It is in fact convenient to restrict the accessible CV space such that the region in which the system is totally converted into liquid is not explored. This region is not of interest since here we focus on the nucleation barrier and restricting the CV space accelerates the convergence. We shall use $p(s)$ to limit the exploration of the CV space. In particular we shall choose a $p(s)$ that is uniform until a value s_0 and vanishes smoothly beyond it, i.e.:

$$p(s) = \begin{cases} \frac{1}{C} & \text{if } s < s_0 \\ \frac{1}{C} e^{-\frac{1}{2}\beta\kappa(s-s_0)^2} & \text{if } s > s_0 \end{cases}, \quad (2.1.7)$$

where s_0 should lie beyond the barrier region, κ is a constant that determines how fast $p(s)$ goes to zero, and C is a normalization constant. The bias potential would be able to produce this $p(s)$ provided that it had sufficient variational flexibility. Since we use instead a bias potential with minimal flexibility, $V(s)$ must be constructed in such a way that it is capable of satisfying equation (2.1.4) in all the CV space. A bias potential that is able to do so is:

$$V(s) = \begin{cases} -\Delta F^{CNT}(s; \Delta\mu, \sigma) & \text{if } s < s_0 \\ -\Delta F^{CNT}(s; \Delta\mu, \sigma) + \frac{1}{2}\kappa(s - s_0)^2 & \text{if } s > s_0 \end{cases}. \quad (2.1.8)$$

Appropriate values for κ and s_0 can be easily chosen based on a very approximate knowledge of the free energy landscape. In Figure 2.1.1 the functional forms of $F(s)$, $V(s)$ and $p(s)$ are depicted.

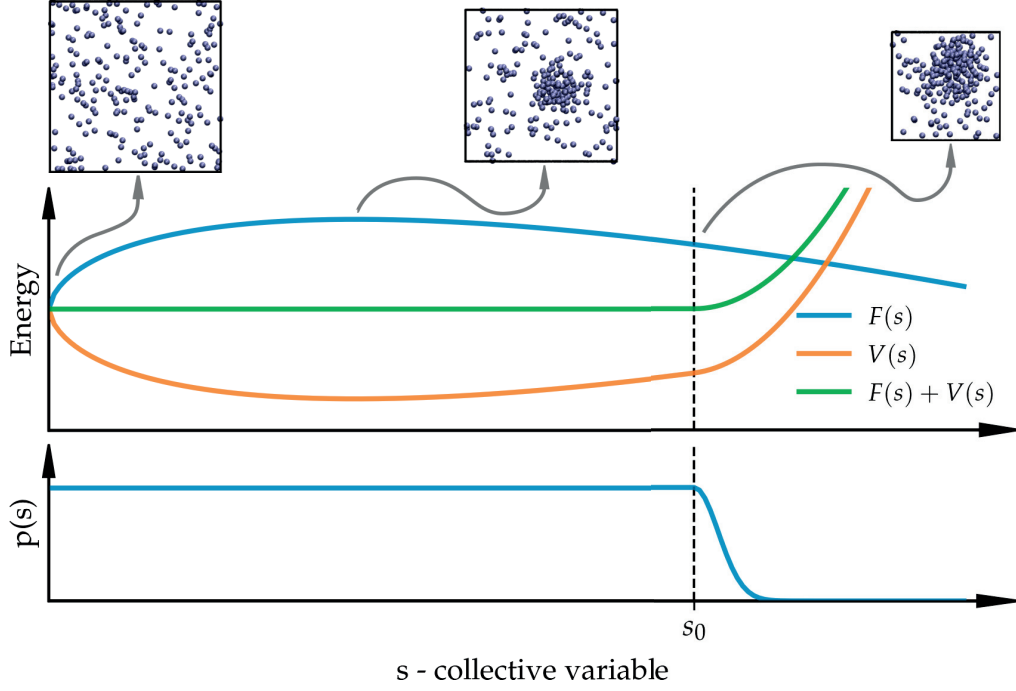


Figure 2.1.1: Top plot) One dimensional nucleation free energy surface $F(s)$ (blue line), bias potential $V(s)$ (orange line), and effective free energy surface $F(s) + V(s)$ (green line). Bottom plot) Target distribution function $p(s)$. The plots are here for illustrative purposes only and do not reflect the properties of specific physical systems.

Optimization algorithm

Inserting equations (2.1.8) into $\Omega[V]$, the functional becomes a function of $\Delta\mu$ and σ , $\Omega(\Delta\mu, \sigma)$. In order to optimize the functional we need to evaluate the gradient:

$$\begin{aligned} \frac{\partial \Omega(\Delta\mu, \sigma)}{\partial \Delta\mu} &= \langle s \rangle_V - \langle s \rangle_p \\ \frac{\partial \Omega(\Delta\mu, \sigma)}{\partial \sigma} &= -\langle s^{2/3} \rangle_V + \langle s^{2/3} \rangle_p \end{aligned} \quad (2.1.9)$$

and the Hessian matrix:

$$H(\Delta\mu, \sigma) = \begin{bmatrix} H_{\Delta\mu \Delta\mu} & H_{\Delta\mu \sigma} \\ H_{\sigma \Delta\mu} & H_{\sigma \sigma} \end{bmatrix} = \beta \cdot \begin{bmatrix} \langle s^2 \rangle_V - \langle s \rangle_V^2 & -\langle s^{5/3} \rangle_V + \langle s \rangle_V \langle s^{2/3} \rangle_V \\ -\langle s^{5/3} \rangle_V + \langle s \rangle_V \langle s^{2/3} \rangle_V & \langle s^{4/3} \rangle_V - \langle s^{2/3} \rangle_V^2 \end{bmatrix}. \quad (2.1.10)$$

These expressions involve only expectation values either in the biased ensemble $\langle \cdot \rangle_V$ or over the target distribution $\langle \cdot \rangle_p$.

Crucial to a successful optimization is the use of the averaged stochastic gradient descent algorithm [43]. In the present case this algorithm can be written as an iterative

procedure with a fixed step size μ :

$$\begin{aligned}\Delta\mu^{(k+1)} &= \Delta\mu^{(k)} - \mu \left[\frac{\partial\Omega}{\partial\Delta\mu} + H_{\Delta\mu\Delta\mu} \cdot (\Delta\mu^{(k)} - \overline{\Delta\mu}^{(k)}) + H_{\Delta\mu\sigma} \cdot (\sigma^{(k)} - \overline{\sigma}^{(k)}) \right] \\ \sigma^{(k+1)} &= \sigma^{(k)} - \mu \left[\frac{\partial\Omega}{\partial\sigma} + H_{\sigma\Delta\mu} \cdot (\Delta\mu^{(k)} - \overline{\Delta\mu}^{(k)}) + H_{\sigma\sigma} \cdot (\sigma^{(k)} - \overline{\sigma}^{(k)}) \right],\end{aligned}\tag{2.1.11}$$

where $\Delta\mu$ and σ are the instantaneous parameters whereas $\overline{\Delta\mu}^{(k)} = k^{-1} \sum_{i=1}^k \Delta\mu^{(i)}$ and $\overline{\sigma}^{(k)} = k^{-1} \sum_{i=1}^k \sigma^{(i)}$ are their averaged counterparts. At each iteration k , both the gradient and the Hessian matrix are estimated in the bias ensemble with a bias potential given by the averaged parameters $\overline{\Delta\mu}^{(k)}$ and $\overline{\sigma}^{(k)}$.

In previous calculations we have used only the diagonal part of the Hessian. Here the use of the full Hessian is essential for a successful optimization. In fact the minimization problem is ill-conditioned with a condition number $\sim 10^4$.

Although using the full Hessian allowed us to reach the minimum, we found that a faster convergence could be achieved by making the following change of variables,

$$s(x) = N_0 \left(\frac{x+1}{2} \right)^3, \tag{2.1.12}$$

where N_0 is a number slightly larger than the point s_0 where $p(s)$ starts decaying towards zero. This change of variables is akin to the one described in subsection 2.1.2 that transforms the number of atoms n in the droplet into its radius R . Therefore x is related to a characteristic length of the droplet. With this change of variables the new CV is defined in the interval $[-1, 1]$ and, for $s < s_0$, $V(x; \Delta\mu, \sigma)$ can be written as a polynomial:

$$V(x; \Delta\mu, \sigma) = N_0 \Delta\mu \left(\frac{x+1}{2} \right)^3 - N_0^{2/3} \sigma \left(\frac{x+1}{2} \right)^2. \tag{2.1.13}$$

It is therefore natural to express $V(x; \Delta\mu, \sigma)$ in terms of Chebyshev polynomials,

$$V(x; \Delta\mu, \sigma) = \sum_{i=0}^3 \alpha_i \cdot T_i(x) \tag{2.1.14}$$

where $T_i(x)$ is the Chebyshev polynomial of degree i . Comparison with the expression

in equation (2.1.13) gives a set of relations between the α_i , and $\Delta\mu$ and σ :

$$\begin{aligned}
 \alpha_0 &= \frac{5\Delta\mu N_0}{16} - \frac{3\sigma N_0^{2/3}}{8} \\
 \alpha_1 &= \frac{15\Delta\mu N_0}{32} - \frac{\sigma N_0^{2/3}}{2} \\
 \alpha_2 &= \frac{3\Delta\mu N_0}{16} - \frac{\sigma N_0^{2/3}}{8} \\
 \alpha_3 &= \frac{\Delta\mu N_0}{32}.
 \end{aligned} \tag{2.1.15}$$

We can now use, say, α_2 and α_3 as variational parameters, constraining the other two via equations (2.1.15). In these new variables, the problem is better behaved and the condition number of the Hessian is reduced to the manageable value of ~ 5 . Although $V(s; \alpha_2, \alpha_3)$ depends only on two variables, the fact that it can be formally expressed as a linear expansion in orthogonal polynomials allows us to use the optimization machinery previously developed to handle this case [25].

2.1.4 Computational details

We have studied condensation from the vapor phase in a Lennard-Jones system in which the interaction potential was truncated and shifted at a cutoff radius $r = 2.5\sigma_{LJ}$, with σ_{LJ} the particle diameter. All MD simulations were performed using LAMMPS [44] patched with a private development version of the PLUMED 2 enhanced sampling plugin [45]. In the following we shall measure all quantities in Lennard-Jones units [26], such that the Lennard-Jones well depth ϵ is the unit of energy and the Lennard-Jones diameter σ_{LJ} is the unit of length. In the definition of the liquid-like atoms we used the values $r_c = 1.5$ and $n_c = 5$ as suggested in ref. [37].

Periodic boundary conditions and a time-step of 0.001 were used in the simulations [26]. In all cases the stochastic velocity rescaling thermostat [46] and the isotropic version of the Parrinello-Rahman barostat [47] were employed. The relaxation time for the thermostat and the barostat were 0.05 and 50, respectively. We employed cubic boxes with different number of particles and a temperature of 0.741 ($T_c = 1.085$). The target pressure of the barostat was set to 0.016. This system setup is similar to that of ref. [37], although the supersaturation is higher in our case.

Each iteration in the optimization of Ω corresponded to 500 MD steps and the step size μ in the optimization was chosen to be 0.001. In all cases 4 multiple walkers were employed, starting half of them in the vapor basin and the rest beyond the nucleation barrier. Due to the highly non-local nature of the basis sets employed, the use of multiple walkers proved to be instrumental in accelerating the optimization. The initial variational parameters were taken as $\Delta\mu = 0$ and $\sigma = 0$ such that initially the

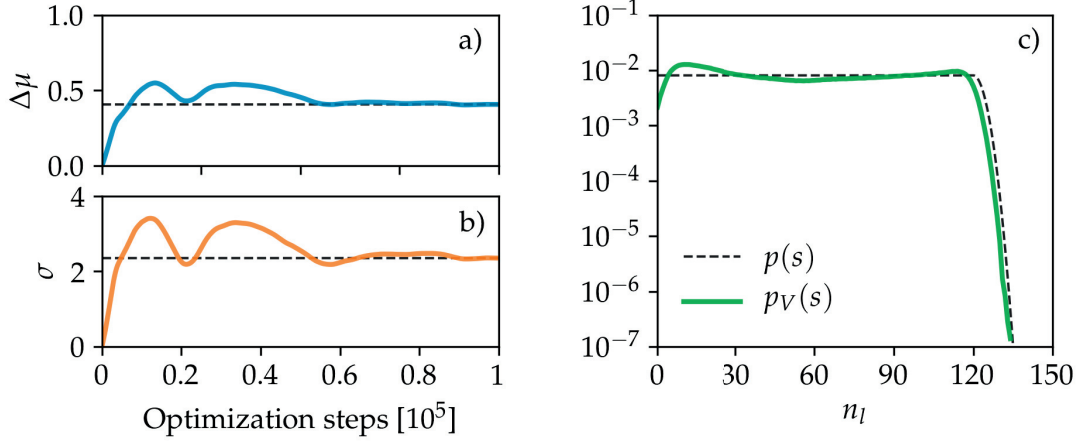


Figure 2.1.2: (a)-(b) Evolution of $\Delta\mu$ and σ during the optimization process for a system of 216 particles. (c) Comparison between the sampled probability distribution of the CV $n_l p_V(s)$ and the target probability distribution $p(s)$.

bias was $V(s) = -\frac{1}{\beta} \log p(s)$. The parameters of $p(s)$ were $s_0 = 120$ and $\kappa = 0.1$.

2.1.5 Results

As an example of a typical behavior of the optimization process, we show in Figure 2.1.2 (a)-(b) the convergence of the variational parameters $\Delta\mu$ and σ as a function of the number of optimization steps in a system with 216 atoms. If we use Lennard-Jones parameters appropriate to argon, the total length of the optimization corresponds to a ~ 100 ns long simulation. A measure of the quality of the variational ansatz in equation (2.1.6) is how much the biased distribution $p_V(s)$ differs from the target distribution $p(s)$ (see Figure 2.1.2(c)). According to equation (2.1.5) these should be identical. Indeed, they are very close but not totally identical. The small discrepancies are a result of the limited variational flexibility of $V(s)$ (equation (2.1.6)).

We now turn our attention to the behavior of the free energy as a function of the collective variable n_l for three different system sizes (see Figure 2.1.3(a)). By and large, the curves are rather similar, however, some differences are observed. This is not surprising since the probability of observing several liquid clusters at one time depends on the system size. This is particularly evident for the low n_l region on which the largest size dependence is observed. This behavior has already been reported for the free energy associated with the size of the largest cluster [48]. For larger n_l the probability of finding more than one cluster is small, at least for the system sizes studied here, and the size effects are negligible. These finite size effects are quantified in Table 2.1.1 where we compare the estimates of $\Delta\mu$ and σ obtained from minimizing Ω to a fit to $F(n_l)$ of the CNT expression. It is seen that the smaller the system, the smaller the

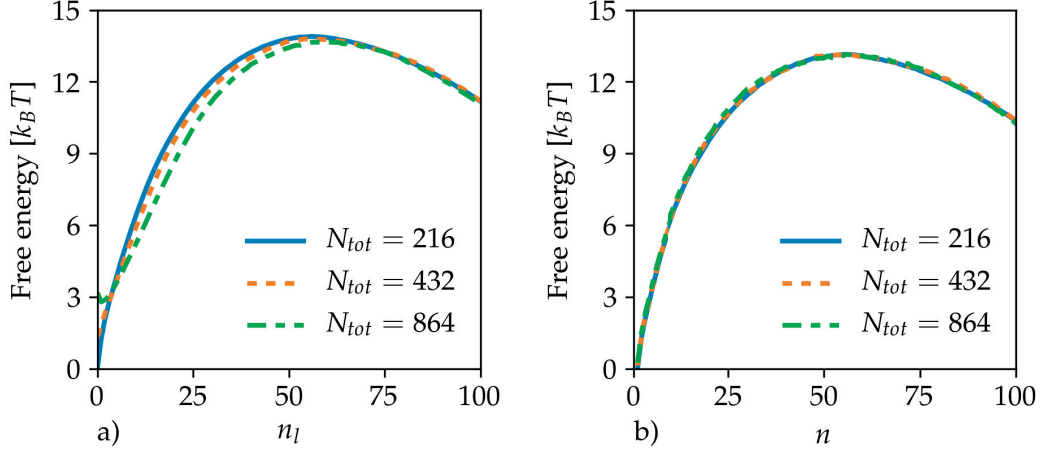


Figure 2.1.3: (a) Reweighted free energy as a function of n_l for system sizes $N_{tot} = 216, 432$ and 864 . For small n_l the curves show discrepancies. (b) Reweighted free energy associated to n for the same system sizes as above. All free energies are equal within the statistical error.

deviation of $F(n_l)$ from a CNT-like behavior. However if we reweight our data so as to obtain $F(n)$ (see Figure 2.1.3(b)), all the finite size effects disappear. Details of the reweighting procedure and of the calculation of $F(n)$ can be found in the SI. The results can be all fitted to the CNT expression and give what is possibly our best estimate for these parameters (see Table 2.1.1), assuming that our data can be described by CNT. The values obtained are consistent with the estimates of $\Delta\mu = 0.530$ and $\sigma = 2.85$ that can be calculated using a coexistence pressure $P_0 = 0.00783$ obtained from Gibbs ensemble simulation [37] and a surface free energy $\gamma = 0.494$ [37, 49].

From $F(n)$ we can estimate the nucleus size and the barrier height. The nucleus size n^* at the supersaturation condition studied here corresponds to 56 atoms. We estimated the barrier height defined as $F(n^*) - F(1)$ to be $\sim 13 k_B T$.

In order to test the correctness of the approach described in this work, we have per-

System size	Optimization		Fit $F(n_l)$			Fit $F(n)$		
	$\Delta\mu$	σ	$\Delta\mu$	σ	rms error	$\Delta\mu$	σ	rms error
216	0.42(1)	2.4(1)	0.409(6)	2.38(3)	0.39	0.439(2)	2.51(1)	0.12
432	0.41(2)	2.4(3)	0.408(8)	2.37(4)	0.54	0.432(2)	2.48(1)	0.12
864	0.37(4)	2.2(5)	0.38(1)	2.22(7)	0.83	0.436(2)	2.49(1)	0.11

Table 2.1.1: We compare estimations of $\Delta\mu$ and σ for different system sizes. From left to right we show the direct result of the optimization at $5 \cdot 10^5$ steps, a fit of the CNT expression to $F(n_l)$ in the interval $[0, N]$, and a similar fit to $F(n)$ in the interval $[1, N]$. In the last two cases, we show also the error associated to the fit.

formed a benchmark simulation employing well-tempered metadynamics [50]. The details of this simulation can be found in the SI. The free energies obtained with the methodology described in this work are equal to those calculated employing well-tempered metadynamics within the statistical error (see Figure S1 in the SI).

2.1.6 Discussion and conclusions

We have developed an enhanced sampling method based on the introduction of a bias potential with a physically motivated functional form. In particular, in the context of nucleation, we have employed the functional form of classical nucleation theory for the free energy of formation of a cluster. This idea is put into practice employing a variational principle that allows the estimation of the parameters of the model. In this way the bias potential compensates the underlying free energy of the system.

Our results are encouraging, however they underline the fact that attention must be paid to the choice of the collective variables. In particular for the system of interest here, the choice of n_l as CV has clear computational advantages but it is best applied to small systems where $F(n_l)$ is very close to the free energy associated to the cluster size distribution $F(n)$, that is system size independent. It must also be noted that for small systems the values for $\Delta\mu$ and σ obtained via the optimization, are very close to those obtained by reweighting the trajectory to get the cluster size distribution and fitting the CNT expression to the results. This might provide an expedite way to estimate the vapor pressure and the surface energy, two quantities of great practical interest.

The lesson learned here will be applied to nucleation rates calculation. In fact, given the results obtained, one could use n_l as the collective variable and employ the approaches suggested either in ref. [42] or ref. [38]. Both methodologies rely on introducing a bias potential that leaves the transition region between metastable states untouched. Under this assumption the physical transition time (τ) can be related to the one calculated in a biased simulation (τ_V) by [42, 51, 52],

$$\tau = \tau_V \langle e^{\beta V(s,t)} \rangle_V. \quad (2.1.16)$$

However, the two methodologies differ from each other in the manner in which the bias potential is constructed. On the one hand, ref. [42] describes a metadynamics based methodology with infrequent deposition of kernels. If the transition is rare but fast then the procedure leads to bias free transition regions therefore fulfilling the assumption that leads to equation (2.1.16). On the other hand, the approach described in ref. [38] is based on the construction of a bias potential by means of the variational principle [25] also used in the present work. This bias potential floods the free energy surface up to a predefined energy level. By construction this approach guarantees bias free transition states and accurate times can be extracted using equation (2.1.16). We point out that the independence of $F(n)$ from the system size provides a strong

encouragement to limit ourselves to the study of small systems.

Finally, our results provide yet another confirmation of the validity of CNT for liquid-vapor nucleation.

Acknowledgments

P.M.P would like to thank Matteo Salvalaglio and Federico Giberti for useful discussions. The authors acknowledge funding from the National Center for Computational Design and Discovery of Novel Materials MARVEL and the European Union Grant No. ERC-2014-AdG-670227 / VARMET. The computational time for this work was provided by the Swiss National Supercomputing Center (CSCS).

2.2 Enhancing entropy and enthalpy fluctuations to drive crystallization in atomistic simulations

This is the seminal article in which we introduce the idea of using a proxy for entropy as a collective variable. In this case we combine entropy with enthalpy to accelerate the frequency of reversible transitions between the liquid and the solid. The most remarkable result of this paper is the fact that the collective variable does not contain any information of the crystal structures and yet it is able to crystallize different substances into their equilibrium crystal structure. All subsequent articles described in this thesis will be extensions of the ideas presented here. The collective variable has been implemented in the PLUMED 2 enhanced sampling plugin.

I present the postprint version of the article published in Physical Review Letters. The supplementary information of this article is not included in this thesis but a link has been provided in the electronic version. My contribution to this article has been implementing the algorithms, performing the simulations, and writing the paper jointly with Prof. Parrinello.

Full bibliographic reference: Pablo M. Piaggi, Omar Valsson, and Michele Parrinello. Enhancing entropy and enthalpy fluctuations to drive crystallization in atomistic simulations. *Physical Review Letters*, 119 (1): 015701, 2017. doi: 10.1103/PhysRevLett.119.015701. URL <https://link.aps.org/doi/10.1103/PhysRevLett.119.015701>.

Copyright © 2017 American Physical Society.

Enhancing entropy and enthalpy fluctuations to drive crystallization in atomistic simulations

Pablo M. Piaggi^{1,2}, Omar Valsson^{2,3}, and Michele Parrinello^{2,3}

¹Theory and Simulation of Materials, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

²Facoltà di Informatica, Istituto di Scienze Computazionali, and National Center for Computational Design and Discovery of Novel Materials (MARVEL), Università della Svizzera italiana (USI), Via Giuseppe Buffi 13, CH-6900, Lugano, Switzerland

³Department of Chemistry and Applied Biosciences, ETH Zurich, c/o USI Campus, Via Giuseppe Buffi 13, CH-6900, Lugano, Switzerland

Abstract

Crystallization is a process of great practical relevance in which rare but crucial fluctuations lead to the formation of a solid phase starting from the liquid. Like in all first order first transitions there is an interplay between enthalpy and entropy. Based on this idea, in order to drive crystallization in molecular simulations, we introduce two collective variables, one enthalpic and the other entropic. Defined in this way, these collective variables do not prejudge the structure the system is going to crystallize into. We show the usefulness of this approach by studying the case of sodium and aluminum that crystallize in the bcc and fcc crystalline structure, respectively. Using these two generic collective variables, we perform variationally enhanced sampling and well tempered metadynamics simulations, and find that the systems transform spontaneously and reversibly between the liquid and the solid phases.

Crystallization is a remarkable physical process in which the disordered atoms of a liquid spontaneously form beautifully ordered periodic patterns. It is also a phenomenon of great importance in many areas of application, from metallurgy to material science, pharmacology and even biology. Not surprisingly this problem has received considerable attention. Understanding the way in which crystallization proceeds holds also the key to improve many scientific and technological processes. It suffices to recall here the difficult art of protein crystallization. Since experiments can only provide a limited insight into this phenomenon, already in the very early days of computer simulation much effort has been devoted to the study of crystal nucleation [1–3].

Unfortunately in most cases the time scale of nucleation is much longer than what can be reached in an atomistic simulation. Early on this problem was tackled by forcing nucleation using unphysically deep temperature quenches so as to bring the

nucleation time scale within the reach of simulation [3]. This procedure is not without problems since it can even change the nature of the nucleation process [5]. For this reason enhanced sampling methods have been extensively used [14, 53, 54]. Most of them are based on the definition of appropriate collective variables (CVs) able to distinguish one type of local order from another. Typical examples of CVs used in this context are the Steinhardt order parameters [13, 14] or the ones introduced by Santiso and Trout [15, 16, 54].

However, the use of these kind of CVs can prejudge the structure the system is going to crystallize into. Thus, it would be extremely useful to have an enhanced sampling method that does not assume from the start the final structure. Such a method could illuminate important details of nucleation and complement structure prediction methods that are based on finding the crystal structure of lowest energy [55, 56]. This would be extremely valuable in all those cases in which the crystal structure is stabilized by strong entropic effects as in superionic or plastic crystals.

In order to achieve this result the CV should not be related to this or that feature of the geometry of the crystal. It thus comes natural to recall that in crystallization, like in all first order transformations, there is a trade off between enthalpy and entropy, and introduce two CVs able to describe these changes. Of course enthalpy $H = E + PV$ where E is the total energy, P the pressure and V the volume, is easy to estimate while for the entropy there is no exact expression. However in order to bias the system we do not need to compute the exact entropy and even an approximate expression will do. The liquid state theory provides an expression in which the excess entropy per atom is expanded in an infinite series of terms involving multiparticle correlation functions [23]. The first such term S_2 , used in ref. [57–59] to perform insightful analysis, includes only two body correlations and is given by:

$$S_2 = -2\pi\rho k_B \int_0^\infty [g(r) \ln g(r) - g(r) + 1] r^2 dr. \quad (2.2.1)$$

In this equation $g(r)$ is the radial distribution function and ρ is the density of the system. We shall use S_2 as CV. In a different context the use of S_2 has already been suggested to address rare events problems in which the entropy plays a dominant role [60]. Both H and S_2 have a proper thermodynamic meaning only when averaged. However, for the purpose of this paper we use their instantaneous values to define useful CVs.

In this spirit we introduce two CVs, one enthalpic s_H and the other entropic s_S . The former is defined as:

$$s_H = \frac{U(\mathbf{R}) + PV}{N}, \quad (2.2.2)$$

where $U(\mathbf{R})$ is the potential energy and N the number of atoms in the system [20]. We

thus do not include the kinetic energy contribution to the total energy in the definition of s_H . The latter is slightly more complex. We first define a mollified version of the radial distribution function:

$$g_m(r) = \frac{1}{4\pi N \rho r^2} \sum_{i \neq j} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(r-r_{ij})^2/(2\sigma^2)}, \quad (2.2.3)$$

where r_{ij} is the distance between particles i and j , and σ is a broadening parameter. The mollification is necessary to ensure that the derivatives of $g_m(r)$ relative to the atomic positions are continuous. The resulting $g_m(r)$ is inserted into equation (2.2.1) and the integral calculated numerically using the trapezoid rule up to a cut off distance r_{\max} . In the spirit of the work of Tiwary and Berne [61] r_{\max} is chosen so as to optimize the frequency of transitions between solid and liquid.

Before describing the details of our simulation we would like to pause a little to discuss a small but crucial technical issue. It has long been recognized that a given crystalline order can conflict with the periodic boundary conditions of molecular dynamics (MD) simulations and that allowing the MD cell shape to vary is essential to study processes that involve crystals. This is at the heart of the Parrinello-Rahman method that allows the MD cell to change under conditions of constant stress [47]. The MD cell shape is expressed by an upper diagonal matrix $\mathbf{h} = [h_{11}, h_{22}, h_{33}, h_{23}, h_{13}, h_{12}]$. Changes in \mathbf{h} are driven by the unbalance between the internal and the external stress. A straightforward application of the Parrinello-Rahman method to the liquid state is however problematic since a liquid offers no resistance to shear and left to its own devices the MD box shape would fluctuate randomly assuming even inconvenient cigar like shapes. In order to remedy this we take $h_{11} = h_{22} = h_{33}$ and this forces the volume to be always close to a cube but still allows the off diagonal elements to vary, and thus accommodate different structures.

We now exemplify the usefulness of this approach in the cases of sodium and aluminum that crystallize in the bcc and fcc structures, respectively. We simulated Na and Al using embedded atom models [62, 63] whose melting temperatures have been determined. Biased MD simulations were performed using LAMMPS [44] patched with PLUMED 2 [45] and the VES code [64]. The integration of the equations of motion was carried out with a timestep of 2 fs. We employed the stochastic velocity rescaling thermostat [46] with a relaxation time of 0.1 ps. The target pressure of the barostat was set to its standard atmospheric value and a relaxation time of 10 ps was used. Systems composed of 250 and 256 atoms were employed for Na and Al, respectively. In order to determine the Gibbs free energy surface $G(\mathbf{s})$ as a function of s_H and s_S we used well tempered metadynamics [50] and the variationally enhanced sampling (VES) [25] method in its well tempered variant [39]. Details of the calculations can be found in the Supplemental Material. For the calculation of s_S we used a cutoff of $r_{\max} = 0.65$ nm for Na and $r_{\max} = 0.70$ for Al. The broadening parameter σ in equation (2.2.3) and the step size in

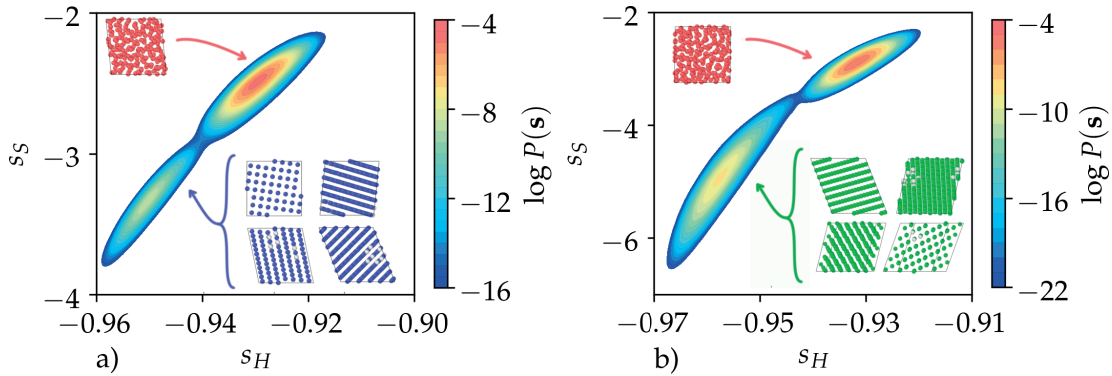


Figure 2.2.1: Marginal probability distribution with respect to s_H and s_S . a) Na at 350 K. b) Al at 800 K. A liquid and a solid basin are observed in the plots and some characteristic configurations in each basin are depicted. The solid structures were identified using CNA. Bcc-like atoms are colored in blue and fcc-like atoms are colored in green. As expected, the solid configurations in Na and Al, have a bcc and fcc crystalline structure, respectively. A liquid configuration is also shown with liquid-like atoms colored in red. s_H is expressed in units of the cohesive energy and s_S is in k_B .

the numerical integration of equation (2.2.1) was 0.0125 nm both for Na and Al.

As shown in Figure 2.2.1, employing these CVs we were able to describe accurately the phase transition between the liquid and crystalline phases in Na and Al without having had to feed any prior information on the systems. In both cases two basins were observed, one of high enthalpy and high entropy, and another of low enthalpy and low entropy, corresponding to the liquid and solid basins, respectively. This reflects the trade off discussed in the introduction between enthalpy and entropy in first order phase transitions. As hoped for, Na crystallized into the bcc structure whereas Al crystallized in the fcc structure. In all cases the transition is reversible (see Supplemental Material) allowing the free energy surface to be estimated.

We performed a detailed analysis of the crystalline structures in the trajectories using common neighbor analysis (CNA) [65, 66]. For Na the only solid phase that can be identified in the simulation is bcc. In the case of Na either s_S or s_H can act as order parameters, since any of them by itself could be able to distinguish between the liquid and bcc phases. However they alone do not fully capture the nature of the transition. We have performed simulations biasing only s_S or s_H and we found that their efficiency in metadynamics is very low. Instead if they are biased together the sampling efficiency is greatly improved. Furthermore, the FES is essentially one dimensional thus no extra computational burden is posed by the use of two CVs rather than one. For Al the situation is slightly more complex since there is a bcc phase at high free energies. In this case the use of both s_S and s_H is necessary to distinguish between the liquid, fcc, and bcc phases. For compact structures such as fcc, the formation of stacking faults

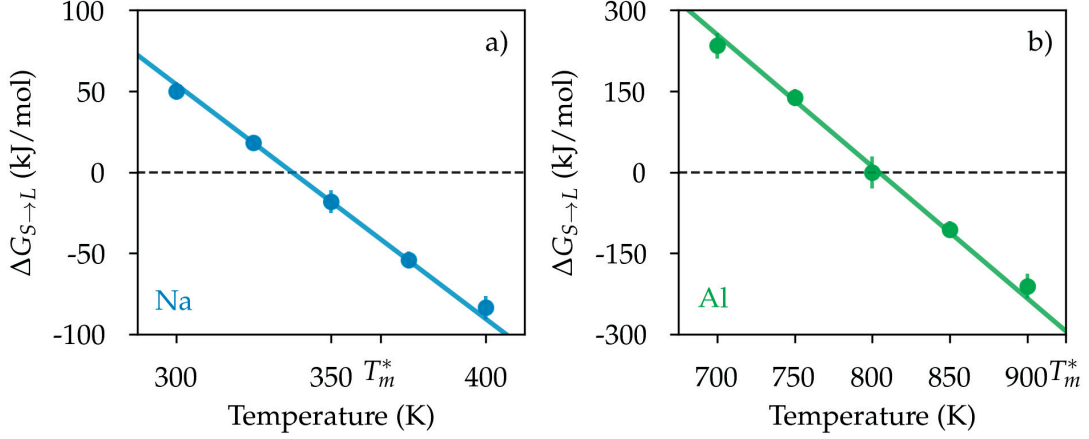


Figure 2.2.2: Difference in free energy between the liquid and solid phases as a function of temperature as calculated from equation (2.2.4). Subplots a) and b) correspond to Na and Al, respectively. The melting temperatures reported in ref. [62] and [63] are marked in the abscissa axis with T_m^* . The straight lines have only been fitted to the three middle points.

must be considered. In our simulations these defects rarely form since the stacking fault energy of the Al potential [63] reproduces well the high experimental value. These results are discussed to a greater detail in the Supplemental Material.

From our calculation we can also get the difference in free energy between the two phases as:

$$\Delta G_{S \rightarrow L} = -\frac{1}{\beta} \log \left(\frac{\int_L ds e^{-\beta G(s)}}{\int_S ds e^{-\beta G(s)}} \right) \quad (2.2.4)$$

where s is the set of CVs s_H and s_S and the integrals are restricted to the liquid (L) and solid (S) basins, respectively. In Figure 2.2.2a we plot $\Delta G_{S \rightarrow L}$ for Na in the temperature range 300-400 K. From this calculations, we estimate the melting temperature to be around 340 K in agreement with results obtained using a different technique (366 K) [62], if one takes into account that the melting temperature is a strong function of system size [67]. For details of these calculations we refer the reader to the Supplemental Material. In Figure 2.2.2b we plot $\Delta G_{S \rightarrow L}$ for Al in the temperature range 700-900 K. The melting temperature is around 800 K, somewhat below the melting temperature calculated from coexistence simulations (926 K). Again this is due to strong finite size effects.

Once that $\Delta G_{S \rightarrow L}$ is determined as a function of temperature, the difference in entropy between the liquid, $\Delta S_{S \rightarrow L}$, can be calculated from the thermodynamic identity

$\Delta S_{S \rightarrow L} = -\left. \frac{\partial \Delta G_{S \rightarrow L}}{\partial T} \right|_{N,P}$. Since also $\Delta H_{S \rightarrow L}$ can be calculated, the definition of Gibbs free energy $\Delta G_{S \rightarrow L} = \Delta H_{S \rightarrow L} - T \Delta S_{S \rightarrow L}$ provides an independent estimate of $\Delta S_{S \rightarrow L}$. In Table 2.2.1 we compare these two estimates of the entropy and calculations from experiments, both for Na and Al. The estimates are comfortably similar and in line with the experimental values [68].

In conclusion, the use of a collective variable that couples directly to entropy has proven to be very promising. We have illustrated the power of the approach by crystallizing Na and Al in their minimum free energy structures. The success of our calculation suggests a general strategy for tackling those problems in which entropy alone or in combination with enthalpy plays a role. This in the practice means that one only needs to find approximate ways of expressing the entropy and the enthalpy. Once that the variables have been chosen, the use of metadynamics [9] or VES can amplify the CV fluctuations and accelerate the observation of the desired transition. We stress that a rigorous definition of entropy is not necessary and metadynamics or VES are very accommodating in this respect. We can anticipate here that the strategy of biasing an entropic CV is being adapted to the folding of small proteins.

Acknowledgments

This research was supported by the NCCR MARVEL funded by the Swiss National Science Foundation. The authors also acknowledge funding from the European Union Grant No. ERC-2014-AdG-670227 / VARMET. The computational time for this work was provided by the Swiss National Supercomputing Center (CSCS). Calculations were performed in CSCS clusters Piz Daint and Mönch. We are grateful to Paolo V. Giaquinta and David Chandler for useful discussions.

System	$-\left. \frac{\partial \Delta G_{S \rightarrow L}}{\partial T} \right _{N,P} \left(\frac{\text{J}}{\text{K mol}} \right)$	$\frac{\Delta H_{S \rightarrow L} - \Delta G_{S \rightarrow L}}{T} \left(\frac{\text{J}}{\text{K mol}} \right)$	Experimental [68] $\left(\frac{\text{J}}{\text{K mol}} \right)$
Na	5.8	6.6	7.017
Al	9.5	10.7	11.475

Table 2.2.1: Difference in entropy between the solid and liquid phases $\Delta S_{S \rightarrow L}$ at the melting temperature. $\Delta S_{S \rightarrow L}$ was calculated using two different approaches and the results are compared with the experimental value.

2.3 Entropy fingerprint for local crystalline order

Another area that can profit from structure agnostic collective variables is the analysis of atomic environments in molecular simulations. For this purpose in the next article we define a local, atomic version of our entropy inspired collective variable that we dub pair entropy fingerprint. We have applied our method to distinguish between solid and liquid-like atomic environments, including the challenging and spectacular case of nanocrystals. In order to make the pair entropy fingerprint available to the community I have implemented it in the widely used softwares LAMMPS and PLUMED 2.

I present the postprint version of the article published in the *Journal of Chemical Physics*. My contribution to this article has been implementing the algorithms, performing the simulations, and writing the paper jointly with Prof. Parrinello.

Full bibliographic reference: Pablo M. Piaggi and Michele Parrinello. Entropy fingerprint for local crystalline order. *Journal of Chemical Physics*, 147 (11): 114112, 2017. doi: 10.1063/1.4998408. URL <https://doi.org/10.1063/1.4998408>.

Copyright © 2017 American Institute of Physics.

Entropy based fingerprint for local crystalline order

Pablo M. Piaggi^{1,2} and Michele Parrinello^{2,3,*}

¹Theory and Simulation of Materials, École Polytechnique Fédérale de Lausanne,
CH-1015 Lausanne, Switzerland

²Facoltà di Informatica, Istituto di Scienze Computazionali, and National Center for
Computational Design and Discovery of Novel Materials (MARVEL), Università della
Svizzera italiana (USI), Via Giuseppe Buffi 13, CH-6900, Lugano, Switzerland

³Department of Chemistry and Applied Biosciences, ETH Zurich, c/o USI Campus, Via
Giuseppe Buffi 13, CH-6900, Lugano, Switzerland

Abstract

We introduce a new fingerprint that allows distinguishing between liquid-like and solid-like atomic environments. This fingerprint is based on an approximate expression for the entropy projected on individual atoms. When combined with a local enthalpy, this fingerprint acquires an even finer resolution and it is capable of discriminating between different crystal structures.

2.3.1 Introduction

Atomistic computer simulation is an important technique used in the study of a broad range of phenomena in materials science, chemistry, and condensed matter physics. In these fields, very often one is faced with the problem of identifying different local arrangements. A paradigmatic case is that of the nucleation of a crystal from the liquid where one is required to distinguish between solid-like and liquid-like atomic environments. The situation is even more complicated in systems exhibiting polymorphism since in these cases it is desirable to classify the atoms as belonging to one of the different polymorphic structures. This is a common occurrence in nucleation studies where Ostwald's step rule is observed [16, 69] or where clusters exhibit a core-shell structure [70, 71]. Another area where the ability to distinguish between different local arrangements plays a role is in the identification of crystallites in nanocrystalline materials [72].

Several methods have been proposed to distinguish between liquid-like and solid-like atoms and to identify local crystalline structures. One such method is the common neighbor analysis (CNA) [66, 73] which is an efficient algorithm able to distinguish between liquid, bcc, fcc, and hcp phases. However, it lacks robustness with respect to particle displacements such as those arising from thermal motion or stresses. Another popular method is based on the local Steinhardt parameters [74] which are

local, averaged versions of the original Steinhardt parameters [13]. Although they are more general than other choices, Steinhardt parameters require choosing the angular component l that defines them in order to distinguish between different crystal structures.

This work is inspired by a recent progress in the study of nucleation using metadynamics [9, 50] to enhance the probability of inducing the crystal formation in an accessible computer time. Metadynamics relies on the identification of appropriate collective variables (CVs). In Ref. [75] we found that enthalpy and an approximate expression for entropy based on the two body correlation function, were useful CVs in this context. One of the features of this work was that the CVs did not contain any information on the geometry of the crystal structure. This suggested that perhaps from these two quantities one could extract fingerprints able to distinguish between different local atomic arrangements.

Enthalpy and entropy are global properties and in order to be able to use them as local parameters we have to project them onto each atom. We propose a method that is able to do so. We find that the local entropy thus defined is able to distinguish extremely well between solid-like and liquid-like atoms. Furthermore, in conjunction with local enthalpy it can distinguish well between different polymorphs, even in the subtle case of the difference between fcc-like and hcp-like arrangements.

2.3.2 Entropy approximation based on the two body correlation function

Ref. [75] was based on the consideration that in the liquid to solid transition there is a trade-off between entropy and enthalpy. The role of metadynamics was there to enhance the fluctuations of these two quantities so as to accelerate crystallization. This required designing CVs able to describe these two quantities. Enthalpy is easy to compute but entropy is extremely costly to evaluate. However, an expression that gives an approximate evaluation of the entropy is sufficient for the purpose of driving crystallization. Such an expression was derived from an expansion of the configurational entropy in terms of multibody correlation functions [22, 23, 76]. In simple liquids the second term of the expansion, often called two-body excess entropy, involves only the pair correlation function and accounts for about 90% of the configurational entropy [76–79]. This term is given by,

$$S_2 = -2\pi\rho k_B \int_0^\infty [g(r) \ln g(r) - g(r) + 1] r^2 dr, \quad (2.3.1)$$

where ρ is the system's density, and $g(r)$ is the radial distribution function. Extensions of the expansion to multicomponent [80, 81] and inhomogeneous [82] systems are also available. We also recall that entropy series expansions have been used to study

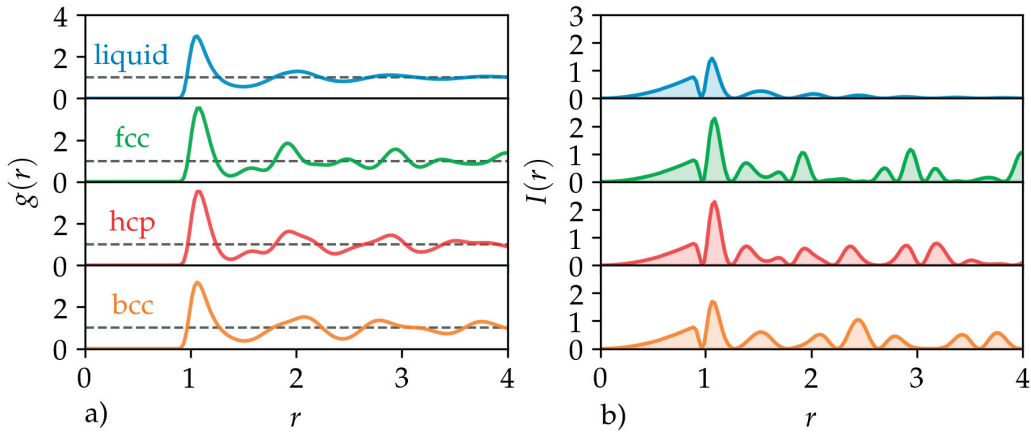


Figure 2.3.1: Analysis of functions related to the entropy approximation S_2 . a) Radial distribution function $g(r)$, and b) integrand in equation (2.3.1) $I(r) = [g(r) \ln g(r) - g(r) + 1]r^2$. These functions are compared for the liquid, fcc, hcp, and bcc phases of a Lennard-Jones fluid at the melting temperature. We use Lennard-Jones units, i.e. $\sigma = 1$.

order-disorder phenomena starting with the landmark work of Kikuchi [83].

In order to come to grasp with S_2 and understand better why it works, we first contrast in Figure 2.3.1 the different behaviors of $g(r)$ and the integral in equation 2.3.1 $I(r) = [g(r) \ln g(r) - g(r) + 1]r^2$. The data were taken from a system with Lennard-Jones interactions at temperature $T = 1.15$ and pressure $P = 5.68$, that corresponds to the fcc-liquid coexistence point [84]. The Lennard-Jones potential was truncated at 2.5 and tail corrections were included. We refer the reader to the Appendix for further computational details. As usual we use Lennard-Jones units [26], i.e. $\sigma = 1$ and $\epsilon = 1$. We have chosen these thermodynamic conditions because at this temperature and pressure the fcc, hcp, bcc, and liquid phases are all metastable allowing a fair comparison. The first observation is that while $g(r)$ has some difficulty at distinguishing between solid and liquid, it strikes the eye that $I(r)$ in the liquid phase is much more short ranged than in the solid phases. Furthermore, the $g(r)$ for the solid phases can hardly distinguish between the different polymorphs. In contrast, the bcc $I(r)$ appears clearly different from that of the closed packed structures. More subtle is the difference between fcc and hcp, that is revealed only if one goes as far out as the third neighbor shell.

2.3.3 Entropy fingerprint for solid-like and liquid-like environments

The analysis of $I(r)$ suggests that, if properly projected onto the different atoms, S_2 could be used as a fingerprint to identify local structures. The projection on atom i can

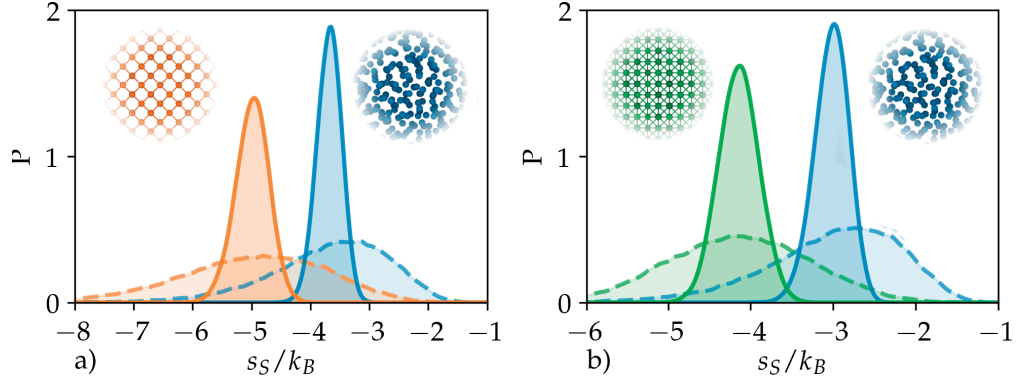


Figure 2.3.2: Distributions of s_S and \bar{s}_S for a) bcc Na [62] at 350 K, b) fcc Al [67] at 900 K. Orange, green, and blue lines refer to the bcc, fcc, and liquid phases, respectively. Dotted and full lines refer to the non-averaged s_S and averaged parameters \bar{s}_S , respectively. The probability distributions of each phase are normalized to one. The solid atomic configurations correspond to $\{100\}$ planes of bcc and fcc crystals at 0 K. The parameters r_m , r_a , and σ that were used are summarized in Table 2.3.1.

be achieved using the expression:

$$s_S^i = -2\pi\rho k_B \int_0^{r_m} [g_m^i(r) \ln g_m^i(r) - g_m^i(r) + 1] r^2 dr, \quad (2.3.2)$$

where r_m is an upper integration limit that in principle should be taken to infinity, and g_m^i is the radial distribution function centered at the i -th atom. To obtain a continuous and differentiable order parameter, we define a mollified version of the radial distribution function [75],

$$g_m^i(r) = \frac{1}{4\pi\rho r^2} \sum_j \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(r-r_{ij})^2/(2\sigma^2)}, \quad (2.3.3)$$

where j are the neighbors of atom i , r_{ij} is the distance between atoms i and j , and σ is a broadening parameter. We shall choose σ so small that $g_m(r) \sim g(r)$ yet large enough for the derivatives relative to the atomic positions to be manageable [75]. A similar projection of S_2 has been used in Ref. [85].

If we use s_S^i as defined in equation (2.3.2) it can be seen in Figure 2.3.2 that, in the cases of Na [62] at 350 K and Al [67] at 900 K (see Appendix for technical details), the distribution of s_S^i in the liquid and solid phases are peaked at two different positions but exhibit a large overlap. In order to calculate local order parameters whose distributions are more clearly distinct, we take cue from Lechner and Dellago [74] and define an

Structure	Model	T (K)	r_m (a)	r_a (a)
bcc	Na	350	1.8 (5NS)	1.2 (2NS)
fcc	Al	900	1.4 (3NS)	0.9 (1NS)

Table 2.3.1: Parameters in the definition of s_S and \bar{s}_S for different structures. The columns represent the crystal structure, the model system, the temperature (T) at which the distributions of solid and liquid phases are compared, and the parameters r_m and r_a defined in equation (2.3.2) and (2.3.3). r_m and r_a are in units of the lattice constant, $a = 4.23 \text{ \AA}$ for Na and $a = 4.05 \text{ \AA}$ for Al. We report the number of neighbor shells (NS) corresponding to r_m and r_a . For both cases $\sigma = 0.02 \text{ nm}$.

average local entropy:

$$\bar{s}_S^i = \frac{\sum_j s_S^j f(r_{ij}) + s_S^i}{\sum_j f(r_{ij}) + 1} \quad (2.3.4)$$

where j runs over the neighbors of atom i and $f(r_{ij})$ is a switching function with cutoff r_a . Switching functions have a value of 1 for $r_{ij} \ll r_a$, 0 for $r_{ij} \gg r_a$, and decay smoothly from 1 to 0 for $r_{ij} \approx r_a$. We have used a switching function with the functional form:

$$f(r_{ij}) = \frac{1 - (r_{ij}/r_a)^N}{1 - (r_{ij}/r_a)^M} \quad (2.3.5)$$

with $N = 6$ and $M = 12$. Such a form has proven useful in many other contexts [45]. At variance with s_S^i , the distributions of \bar{s}_S^i of the liquid and solid phases now have a negligible overlap (see Figure 2.3.2). Henceforth, we shall drop the index i when referring to distributions and we shall refer to \bar{s}_S as entropy fingerprint.

The ability to distinguish sharply between solid-like and liquid-like molecules depend on a wise choice of the parameters r_m and r_a . As r_m is increased, more of the long range part of the integrand is included making the difference between liquid and solid more and more evident. On the other hand by increasing r_a , more neighbors are included in the summation in equation (2.3.4) and eventually the locality of \bar{s}_S is lost. In the practice we have chosen for r_m and r_a the smallest values that still ensure sharp distinction between solid-like and liquid-like atoms. The parameters r_m , r_a , and σ that were used are summarized in Table 2.3.1.

It is interesting to investigate whether the entropy fingerprint can identify ordered structures in a complex situation, in a context different from nucleation. To this effect we generated a nanocrystalline structure (see Figure 2.3.3) using a procedure described in the Appendix. The system is Al, as described by the potential in Ref. [63]. It can be seen that the entropy fingerprint clearly brings out the nanostructure of the system

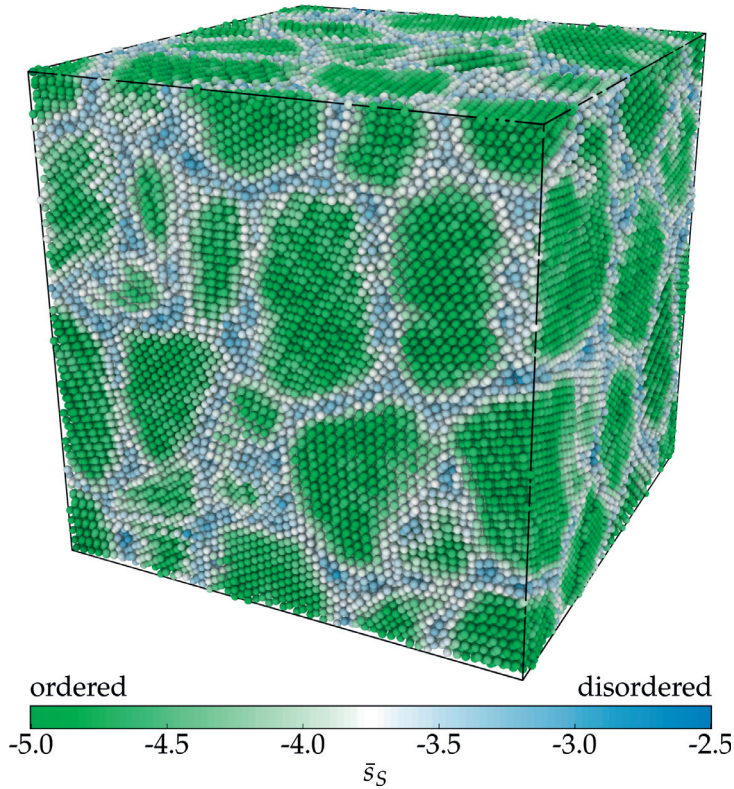


Figure 2.3.3: Nanocrystalline Al with mean grain size 5 nm at 300 K. Atoms are colored according to \bar{s}_S (see text for details). The colorscale is such that green and blue atoms have ordered and disordered environments, respectively. Image obtained with OVITO [65].

and the network of grain boundaries. This indicates that the entropy fingerprint can also work in inhomogeneous situations where different atomic environments coexist.

2.3.4 Identification of crystal structures

In the previous subsection we have shown that \bar{s}_S is able to distinguish liquid-like from solid-like atomic environments. We will now explore the possibility of distinguishing between fcc, hcp, bcc and liquid-like atomic environments. As we shall see, this is best achieved if we accompany our definition of local entropy with a measure of local enthalpy.

The local enthalpy is easily defined if we consider an interatomic potential $U(\mathbf{R})$ that can be decomposed into energies $U_i(\mathbf{R})$ associated to individual atoms. Here \mathbf{R} denotes the atomic coordinates of an N atom system. The expression that we shall use is then,

$$s_H^i = U_i(\mathbf{R}) + PV/N \quad (2.3.6)$$

where P and V are the system's pressure and volume, respectively and, for simplicity, we have partitioned the volume of the system into N equal parts. A more complex partition criterion is also possible, such as using the Voronoi volume of each atom. As

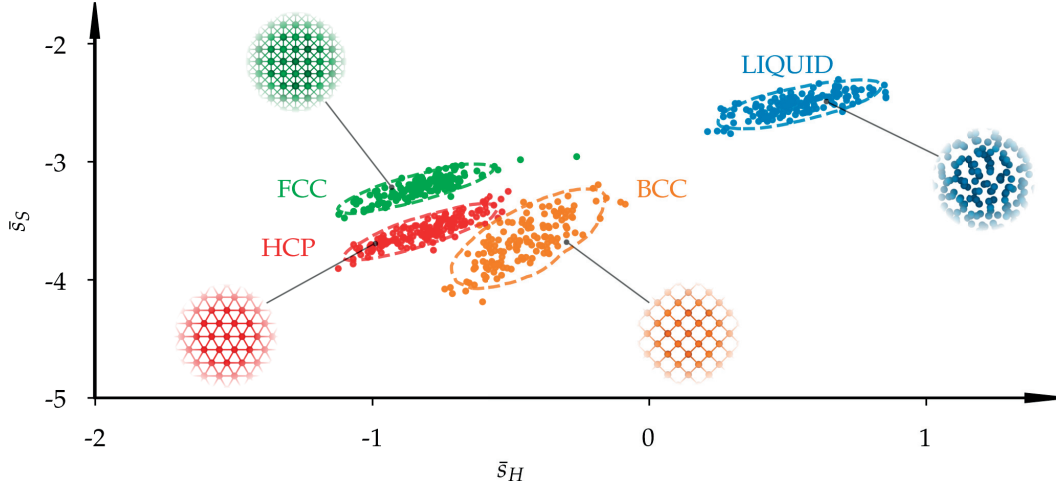


Figure 2.3.4: Joint probability distributions of \bar{s}_H and \bar{s}_S ($P(\bar{s}_H, \bar{s}_S)$) of the fcc, hcp, bcc, and liquid phases of the Lennard-Jones system (see text for simulation details). The dashed lines are the iso-probability lines for a probability equal to $1/10 \max\{P(\bar{s}_H, \bar{s}_S)\}$. The scattered points are 150 random samples of \bar{s}_H and \bar{s}_S over the trajectory in each phase. The solid atomic configurations correspond to $\{100\}$ planes of bcc and fcc crystals, and the basal plane of an hcp crystal at 0 K.

done for the local entropy, we define an average local enthalpy,

$$\bar{s}_H^i = \frac{\sum_j s_H^j f(r_{ij}) + s_H^i}{\sum_j f(r_{ij}) + 1} \quad (2.3.7)$$

where the symbols have the same meaning as in equation (2.3.4).

We calculated the joint probability distributions of \bar{s}_H and \bar{s}_S ($P(\bar{s}_H, \bar{s}_S)$) of the fcc, hcp, bcc, and liquid phases of the Lennard-Jones system described in Section 2.3.2. For this purpose we simulated systems in each of those phases for 200 ps. The thermodynamic conditions were the same as described in Section 2.3.2. We used the following parameters to define \bar{s}_H and \bar{s}_S : $r_m = 2.5$, and $\sigma = 0.1$. We studied two choices for the averaging cutoff, $r_a = 1.4$ and $r_a = 2.5$. The $P(\bar{s}_H, \bar{s}_S)$ of each phase are shown in Figure 2.3.4 for $r_a = 2.5$. Each $P(\bar{s}_H, \bar{s}_S)$ was normalized to one.

We now discuss the results in Figure 2.3.4. We first notice that the distributions of the different phases in Figure 2.3.4 have minimal overlap and therefore \bar{s}_H and \bar{s}_S are useful fingerprints. As in the case of Na and Al, the distributions of liquid and solid phases are very far apart and therefore the fingerprints distinguish very well between liquid-like and solid-like environments. The distributions in the solid phases are clustered together in the region of low enthalpy and entropy, and it is easy to distinguish between the structures using \bar{s}_H and \bar{s}_S . We analyze in detail the challenging case of fcc and hcp. Both fcc and hcp structures are formed by stacking of close-packed planes. However,

phases	$r_a = 1.4$	$r_a = 2.5$
fcc-hcp	0.041177	0.000022
fcc-bcc	0.003090	0.000000
bcc-hcp	0.154870	0.002675
liquid-fcc	0.000086	0.000000
liquid-bcc	0.000749	0.000000
liquid-hcp	0.000010	0.000000

Table 2.3.2: Overlap integrals of each pair of phases in the Lennard-Jones system (see text for details). Calculations have been performed both for $r_a = 1.4$ and $r_a = 2.5$.

they differ in the way the close-packed planes are stacked. For this reason, these structures are usually not easy to discriminate. As seen in Figure 2.3.4, the fingerprints introduced in this work discriminate well between fcc and hcp configurations.

In order to make our results quantitative we have calculated the integrals of the product of the probability distributions of each pair of phases. We employed the definition of the overlap integral used in Ref. [74]. The results are shown in Table 2.3.2 both for $r_a = 1.4$ and $r_a = 2.5$. For both choices of r_a there is negligible or non measurable overlap between the liquid and solid phases. For the small cutoff, $r_a = 1.4$, there is a small overlap between the solid phases, that however is somewhat larger for the bcc-hcp case. On the other hand, for the large cutoff, $r_a = 2.5$, the fingerprint is much sharper and identifies each solid phase with negligible probability of misclassifying them.

2.3.5 Conclusions

To conclude, the degree of success of the entropy based fingerprint is at first sight surprising. However, the root of this success must lie on the point of view taken here that does not directly focus on the local geometry but on properties of deeper thermodynamic significance, like local entropy and enthalpy. It also points to the usefulness of looking at old problems from a different standpoint.

Appendix: Computational details

We performed molecular dynamics (MD) simulations using LAMMPS [44]. We employed an anisotropic Parrinello-Rahman barostat [47] and the stochastic velocity rescaling thermostat [46]. The fingerprints were programmed in a development version of PLUMED 2 [45].

The Lennard-Jones simulations were performed at temperature $T = 1.15$ and pressure $P = 5.68$ (fcc-liquid coexistence [84]). As usual, we use Lennard-Jones units [26], i.e. $\sigma = 1$ and $\epsilon = 1$. The Lennard-Jones potential was truncated at 2.5 and tail corrections were included. The time step for the integration of the equations of motion was 2 fs.

The relaxation times of the barostat and thermostat were 5 and 0.05, respectively.

Na and Al were simulated using embedded atom models (EAM) [62, 67]. For Na we set the temperature at 350 K, close to the melting temperature (366 K) of the model. For Al the temperature was set to 900 K, near the melting temperature 931 K. In both cases the pressure was set to its standard atmospheric value. The relaxation times of the barostat and thermostat were 10 ps and 0.1 ps, respectively. The results presented in Figure 2.3.2 were obtained by performing independent simulations in the liquid and solid phases of Na and Al at the above cited temperatures. Each simulation had a length of 200 ps and the distributions of s_S and \bar{s}_S were calculated taking samples every 1 ps.

The configuration of the nanocrystalline Al was constructed using Voronoi tessellation [72, 86]. The mean grain size was 5 nm and the system contained 255064 atoms. We performed an annealing at 600 K for 0.2 ns, then the temperature was ramped to 300 K in 0.2 ns, and finally the temperature was kept constant at 300 K for 0.2 ns. For these simulations we employed a different EAM potential [63]. The configuration in Figure 2.3.3 corresponds to the last in this trajectory. The simulation details were the same as those used for Al above.

EAM potentials [87, 88] have a natural way to partition the energy between the atoms as needed in equation (2.3.6), i.e.

$$U_i(\mathbf{R}) = \sum_{j \neq i} \phi(r_{ij}) + F \left(\sum_{j \neq i} \rho_{\text{atom}}(r_{ij}) \right) \quad (2.3.8)$$

where ϕ is a pairwise potential, F is the embedding energy function, and ρ_{atom} is the electron charge density function. We have used this partition criterion.

Acknowledgments

This research was supported by the NCCR MARVEL funded by the Swiss National Science Foundation. The authors also acknowledge funding from the European Union Grant No. ERC-2014-AdG-670227 / VARMET. The computational time for this work was provided by the Swiss National Supercomputing Center (CSCS) under project ID mr3. Calculations were performed in CSCS cluster Piz Daint.

2.4 Predicting polymorphism using orientational entropy

In the previous articles I have discussed simple atomic systems. However, the idea of using approximate expressions for entropy is quite general and is not circumscribed to these simple systems. Therefore, it would be interesting to assess whether the method can also be applied in more complex scenarios. In the following article I introduce an extension of the entropy collective variable that can be applied to small and relatively rigid molecules. The collective variable is still based on the two body excess entropy but in this case the molecules are described using a center and an orientation in space. We successfully apply the method to the cases of urea and naphthalene and discover an interesting polymorph of urea.

I present the preprint version of the article published in Proceeding of the National Academy of Sciences of the USA. The supplementary information of this article is not included in this thesis but a link has been provided in the electronic version. My contribution to this article has been implementing the algorithms, performing the simulations, and writing the paper jointly with Prof. Parrinello.

Full bibliographic reference: Pablo M. Piaggi and Michele Parrinello. Predicting polymorphism using orientational entropy. *Proceedings of the National Academy of Sciences*, 115(41):10251, 2018. URL <http://www.pnas.org/content/115/41/10251>. doi :10.1073/pnas.1811056115.

Copyright © 2018 National Academy of Sciences.

Predicting polymorphism in molecular crystals using orientational entropy

Pablo M. Piaggi^{1,2} and Michele Parrinello^{2,3}

¹Theory and Simulation of Materials, École Polytechnique Fédérale de Lausanne,
CH-1015 Lausanne, Switzerland

²Facoltà di Informatica, Istituto di Scienze Computazionali, and National Center for
Computational Design and Discovery of Novel Materials (MARVEL), Università della
Svizzera italiana (USI), Via Giuseppe Buffi 13, CH-6900, Lugano, Switzerland

³Department of Chemistry and Applied Biosciences, ETH Zurich, c/o USI Campus, Via
Giuseppe Buffi 13, CH-6900, Lugano, Switzerland

Abstract

We introduce a computational method to discover polymorphs in molecular crystals at finite temperature. The method is based on reproducing the crystallization process starting from the liquid and letting the system discover the relevant polymorphs. This idea, however, conflicts with the fact that crystallization has a time scale much longer than that of molecular simulations. In order to bring the process within affordable simulation time, we enhance the fluctuations of a collective variable by constructing a bias potential with well tempered metadynamics. We use as collective variable an entropy surrogate based on an extended pair correlation function that includes the correlation between the orientation of pairs of molecules. We also propose a similarity metric between configurations based on the extended pair correlation function and a generalized Kullback-Leibler divergence. In this way, we automatically classify the configurations as belonging to a given polymorph using our metric and a hierarchical clustering algorithm. We apply our method to urea and naphthalene. We find different polymorphs for both substances and we predict new polymorphs. One of them is stabilized at finite temperature by entropic effects.

Polymorphism is the ability that substances have to crystallize into different structures. A paradigmatic example is carbon that in its two main polymorphs, graphite and diamond, exhibits amazingly different properties. Polymorphism is also important from a practical point of view since controlling which crystal structure forms is of the utmost importance in many manufacturing processes. The pharmaceutical industry suffers in particular the consequences of polymorphism [89,90]. Active pharmaceutical ingredients are usually small, organic molecules that frequently exist in a plethora of crystalline forms. Different polymorphs can be patented separately and usually lead to

different drug performances. Therefore a comprehensive screening of polymorphs is crucial to avoid a rival company from releasing to the market the same molecule in a different polymorph [91], and to anticipate the transformation of one polymorph into another during the manufacturing process or the shelf life [92].

The screening of polymorphs was traditionally performed experimentally in spite of the large costs involved [90]. In the last 15 years the increase in computer power and the development of algorithms able to screen a large number of polymorphs has led to very significant successes in polymorph prediction [56, 93–96]. Such methods are based on the search of local minima on the potential energy surface. The minima are ordered by energy and typically corrected for thermal effects using the harmonic approximation. Typically these methods find hundreds of structures most of which are thermodynamically unstable. In addition, entropic effects beyond the harmonic approximation can be significant. Not only they can alter the delicate energetic balance between the different polymorphs but even stabilize structures that are not local minima of the potential energy surface, for instance in the case of superionic conductors such as *AgI* [97]. Another issue that is often overlooked is the kinetic side of crystallization, for instance a given polymorph can be favored relative to energetically lower ones by the fact that is kinetically more accessible. For all these reasons we take here a different approach and we try to reproduce on the computer the crystallization process starting from the liquid state and letting the system discover different polymorphs.

The above ambition conflicts with the fact that crystallization is a process that occurs on a time scale that is much longer than that of computer simulations. This requires the use of enhanced sampling methods that bring the time scale of crystallization within affordable simulation time [27]. Some enhanced sampling methods require the definition of order parameters or collective variables (CVs). These methods channel and enhance the fluctuations so as to favor the reversible observation of multiple freezing and melting events. Thus far such order parameters have been based on some structural geometrical information on the phase the system is going to crystallize into. If one is interested in discovering new polymorphs this approach defeats the purpose. Recently, however, we have shown that in simple systems this can be circumvented by using as CVs surrogates of enthalpy and entropy [75]. The idea is to mimic what happens in a real system in which there is a trade off between entropy and enthalpy. We dealt with simple one component [75] or two component [97] atomic systems. In this case the following expression is used as a surrogate for entropy:

$$S_2 = -2\pi\rho k_B \int_0^\infty [g(r) \ln g(r) - g(r) + 1] r^2 dr, \quad (2.4.1)$$

where r is a distance, $g(r)$ is the radial distribution function and ρ is the density of the system. Together with enthalpy, S_2 proved successful in predicting the lattice into

which the system was going to crystallize. For a discussion of S_2 we refer the reader to Ref. [75]. This has been a simple proof of principle to show that crystal structures, even the ones that are stabilized by entropy [97], can be predicted.

Molecular systems that are of interest to pharmaceutical industry present a complexity much larger than the relatively simple systems so far handled in which most of the times only one polymorph was stable. Here we enlarge considerably the scope of these calculations and move to study molecular systems that, as we shall see, present a large number of polymorphs.

We shall consider a system of molecules and, for the purpose of developing a CV, we shall represent each molecule by the position of its center of mass and a vector that characterizes its orientation in space. We can define a correlation function $g(r, \theta)$ akin to $g(r)$ but including the relative orientation between two molecules. θ is defined as $\theta = \arccos\left(\frac{\mathbf{v}_i \cdot \mathbf{v}_j}{|\mathbf{v}_i||\mathbf{v}_j|}\right)$ where \mathbf{v}_i and \mathbf{v}_j are the vectors characterizing the orientation of molecules i and j . Statistical mechanics provides us with an expression for the entropy of such a system equivalent to the one in equation (2.4.1), this is [81],

$$S_\theta = -\pi\rho k_B \int_0^\infty \int_0^\pi [g(r, \theta) \ln g(r, \theta) - g(r, \theta) + 1] r^2 \sin \theta dr d\theta. \quad (2.4.2)$$

We shall use S_θ as CV to drive simulations. A similar CV was introduced in ref. [98] although in that case the probability as a function of the angle of the molecules with respect to a fix reference frame was used to define the entropy.

An important part in the definition of our CV is the choice of angles to characterize the relative orientation between neighboring molecules. In principle, three angles are needed to specify completely the relative orientation between two rigid molecules, for instance the three Euler angles ϕ, θ, ψ . This would imply the construction of a function $g(r, \phi, \theta, \psi)$ whose calculation would be cumbersome. Here we take a different approach and we use several CVs each involving one angle.

We shall choose two systems to test the ability of S_θ to explore polymorphism, namely urea and naphthalene. We have chosen two CVs and therefore two angles for each system. In the case of urea we use the angles θ_1 and θ_2 to define the CVs s_{θ_1} and s_{θ_2} . The first one is defined using the direction of the dipole moment and the second with the direction of the vector joining the two nitrogens. In the case of naphthalene we use the direction of the longest axis of the molecule and the direction perpendicular to the aromatic rings to define the CVs s_{θ_2} and s_{θ_1} . Other choices are possible, for instance the directions of the eigenvectors of the moment of inertia tensor. At variance with our previous work [75], here we do not bias enthalpy. This would add a third CV making the calculation more cumbersome. Also, the scope here is different since our objective is to discover new structures, and observing multiple and reversible transitions is less

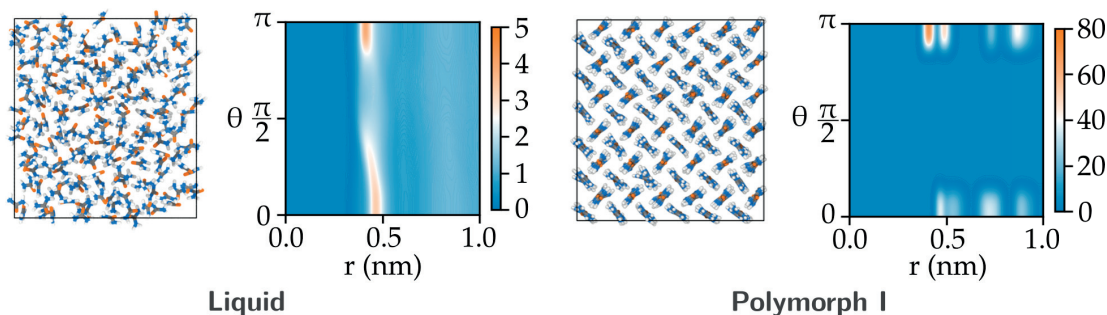


Figure 2.4.1: $g(r, \theta)$ for the liquid and polymorph I of urea at 450 K. Snapshots of the system in each of the phases are shown. Polymorph I is viewed down the c axis. C, O, N and H atoms are shown in orange, grey, blue and white, respectively.

of an issue.

Since the use of $g(r, \theta)$ is not so widespread, we thought it useful to help the reader get a feeling of its behavior by plotting $g(r, \theta_1)$ for the liquid and polymorph I of urea at 450 K (see Figure 2.4.1). The liquid $g(r, \theta_1)$ exhibits some structure at very short distances and almost no correlations at distances larger than 0.8 nm. On the other hand the $g(r, \theta_1)$ of polymorph I shows a well defined structure that persists at long distances as expected from a solid phase. As can be observed in the figure, one of the main characteristics of polymorph I is that molecules have parallel or antiparallel dipole moments. Thus, $g(r, \theta)$ contains important orientational information that can help to distinguish between phases.

We briefly describe the polymorphs found experimentally so far for each system. Urea shows a rich polymorphism and up to five polymorphs have been reported [99–101]. The most stable form at ambient conditions is form I and it has been extensively studied. Two other forms exist at higher pressures, namely forms III and IV. Another high pressure polymorph, form V, has been found although to our knowledge the details of the structure have not been reported. There has also been theoretical work that found other polymorphs [16, 102]. In particular, for urea as described by the Amber force field, the so called form A [16, 102] is highly relevant having an energy very close to that of the ground state. At variance with urea, naphthalene has only one solid form and in spite of several investigations at high pressure [103, 104] no new forms have yet been found.

We have used well-tempered metadynamics (WTMetaD) [50] to enhance the fluctuations of s_{θ_1} and s_{θ_2} . In WTMetaD a time-dependent potential is constructed as a sum of kernels, typically chosen to be Gaussians. The potential discourages frequently visited configurations thus boosting the exploration of configuration space. We simulated urea at 450 K and naphthalene at 300 K. Both temperatures were chosen close to the

melting point of each substance. Further details can be found in the Materials and Methods section. In the 200 ns biased simulations both urea and naphthalene explore thoroughly the space spanned by the CVs, although understanding the nature of the configurations explored requires further analysis. A visual inspection of the trajectories shows many transitions to different crystal forms. The crystalline configurations have different orientations in space and some of them contain small crystalline defects. The wealth of information that these simulations contain, however, cannot be analyzed with the naked eye. It would therefore be useful to have an automatic method to identify and classify the polymorphs that crystallize in the course of the simulation. In the following paragraphs we propose one such automatic method.

A key ingredient for an automatic method to identify and classify polymorphs is a metric for the similarity between two given configurations. Several structural similarity metrics exist in the literature [105] but in this work we shall propose a new one. In the present context, it is natural to use for this purpose the very function $g(r, \theta)$ that defines the CVs to characterize the configuration of the system. However, we still need a measure of distance between two $g(r, \theta)$. We can define a distance taking inspiration in the pair entropy expression. We first note that equation (2.4.2) is a measure of the distance between the $g(r, \theta)$ of the present configuration and the $g(r, \theta)$ of the ideal gas, i.e. $g(r, \theta) = 1 \forall r, \theta$. Inspired by this observation we introduce a divergence of $g_1(r, \theta)$ with respect to $g_2(r, \theta)$,

$$D(g_1||g_2) = \int_0^\infty \int_0^\pi \left[g_1(r, \theta) \ln \frac{g_1(r, \theta)}{g_2(r, \theta)} - g_1(r, \theta) + g_2(r, \theta) \right] r^2 \sin \theta \, dr \, d\theta. \quad (2.4.3)$$

This is a generalization of the Kullback-Leibler divergence for non-normalized functions. This divergence is a special case of Bregman divergence and has some interesting properties such as that of being convex and having a minimum at $g_1 = g_2$ [106]. Strictly speaking $D(g_1||g_2)$ is not a distance since it is not symmetric. For applications in which a well defined distance is needed we shall use a symmetrized version of equation (2.4.3), namely,

$$d(g_1, g_2) = \frac{D(g_1||g_2) + D(g_2||g_1)}{2}. \quad (2.4.4)$$

Equipped with this metric, we can compare configurations and analyze the rich and complex trajectories resulting from the biased simulations. We will exemplify our approach by analyzing the trajectory of urea. The configurations in the trajectory were clustered using a hierarchical clustering approach [107, 108] based on the distance defined in equation (2.4.4). We used the average distance between points in two clusters as linkage criterion. As a result of the clustering, we obtain a tree diagram (see

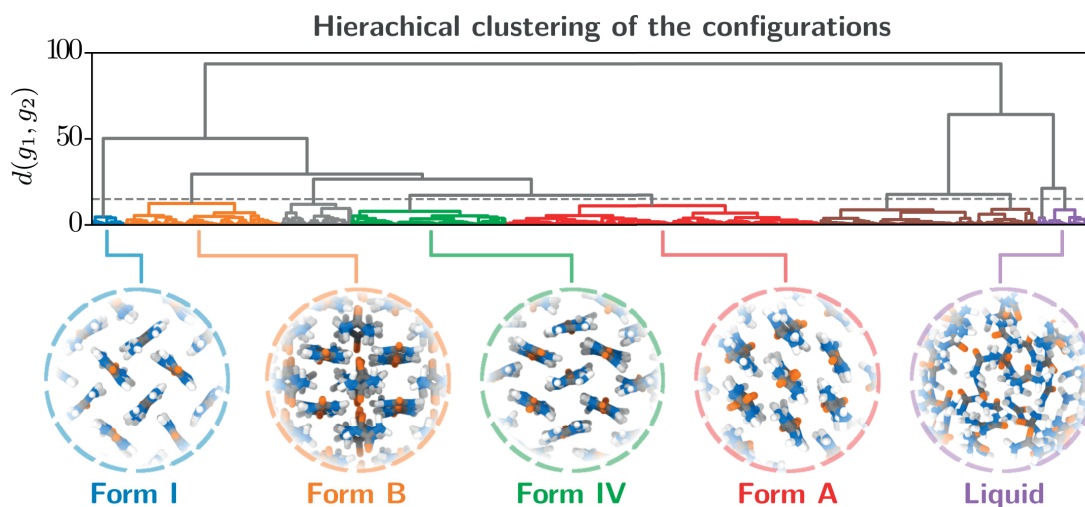


Figure 2.4.2: Tree diagram resulting from the clustering according to the distance in equation (2.4.4) of the trajectory of urea at 450 K. The threshold distance used to join clusters is shown with a grey dashed line. Configurations at 450 K for selected clusters are shown.

Figure 2.4.2) that shows the similarity between different configurations in the trajectory. We can now choose a threshold distance d_c and join together all configurations that belong to a branch with maximum distance d_c between configurations. The choice of d_c allows us to focus on the dominant structures that appear in the simulation. In Figure 2.4.2 d_c is shown with a dashed line and the resulting clusters of structures are shown with different colors.

We still have to determine the phases that each cluster represents. A possible way to do so is by choosing the minimum energy configuration within each cluster. This configuration will be the one with the least number of defects and less affected by the thermal motion of molecules. In some cases this approach is not appropriate, for instance when structures are stabilized by large entropic effects. In these cases one can choose the configuration that has an energy close to the average energy of the cluster. We have chosen with this criteria the configurations that are used to determine the nature of each cluster. Some of these configurations are shown in Figure 2.4.2.

We now describe the phases that were found. The tree diagram has two main branches. The right branch contains liquid-like configurations (violet cluster in Figure 2.4.2) and interesting partially ordered configurations (brown cluster in Figure 2.4.2) in which the dipole moments are oriented in the same direction but do not exhibit long range translational order. The left branch contains solid-like configurations and it can be further subdivided into five relevant clusters. One of these clusters contains an unstable structure and we shall disregard it (grey cluster in Figure 2.4.2). The other four clusters correspond to form I, to a new polymorph that we shall name form B, to form

IV and form A. To the best of our knowledge it is the first time that form B has been reported. The other structures were expected based on previous studies [16, 100, 102]. All polymorphs are metastable at 450 K and they do not transform during a 1 ns unbiased simulation. We include the configurations of all relevant structures in the SI. We have also performed a similar analysis for naphthalene. The clustering identifies the experimentally known form I, the liquid, and a new structure that we shall name form A. The results can be found in the SI.

We have estimated the free energy difference between the polymorphs and the liquid using:

$$\Delta G = -\frac{1}{\beta} \log \left(\frac{p_i}{p_l} \right) \quad (2.4.5)$$

where p_i and p_l the probabilities to observe polymorph i and the liquid, respectively. The states that we consider in these calculations are either purely solid or purely liquid. In an unbiased MD simulation one could calculate the probabilities p_i and p_l directly from the simulation. However, since we have introduced the WTMetaD potential that alters the probability of observing a given configuration, the p_i 's must be calculated with the reweighting procedure described in ref. [109]. We have employed the clustering described above to identify the phase of each configuration. The resulting free energy differences are shown in Figure 2.4.3. The error bars in this figure are relatively large since free energy differences calculated in this way are not easy to converge and the simulation contains transitions between many different structures. Figure 2.4.3 shows also the enthalpy ΔH of the polymorphs with respect to the liquid phase. Using ΔG and ΔH the entropy ΔS can be calculated from the definition of free energy $\Delta G = \Delta H - T \Delta S$. The results show that form I of urea is close to equilibrium with the liquid at 450 K, in line with ref. [110] and [16] where the melting temperature was found to be around 420 K. Similarly, form I of naphthalene is close to equilibrium with the liquid at 300 K, as expected from the estimated melting temperature (330 K).

We shall now consider in detail the newly discovered polymorphs. We first discuss form B of urea that has a $P4_2/mbc$ space group and is shown in Figure 2.4.4. This polymorph is particularly interesting because it has a relatively high enthalpy, roughly $k_B T$ above form I (see Figure 2.4.3). Based only on energy arguments one would conclude that this structure cannot compete with form I. However, strong entropic effects stabilize it. The entropies shown in Figure 2.4.3 indeed show a greater contribution to the stability in form B than in form I. We suggest that an important factor that contributes to the entropy is the fast rotation about the C-O axis. We have calculated the characteristic rotation time using the time autocorrelation function of the N-N unit vector and fitting an exponential function to it. We show the results in Figure 2 of the SI and we compare them with those of form I. The characteristic time of rotation in form I is ~ 800 ps while in form B it is ~ 7 ps. We have also computed the probability $p(\theta)$ as

Predicting polymorphism using orientational entropy

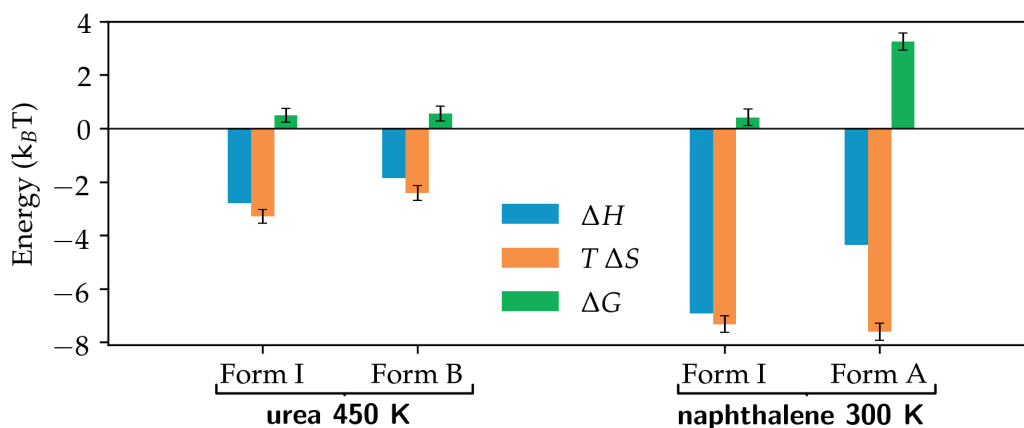


Figure 2.4.3: Enthalpy, entropy and free energy for selected polymorphs of urea at 450 K and naphthalene at 300 K. All quantities have the liquid as reference state.

a function of the rotation angle θ about the C-O axis. From $p(\theta)$ the free energy can be calculated as $G(\theta) = -k_B T \log p(\theta) \sin \theta$. We show the results in Figure 2 of the SI. Both in form I and B $G(\theta)$ exhibits a barrier separating two molecular configurations in which the N are exchanged. The barrier height is ~ 18 kJ/mol in form B while it is ~ 34 kJ/mol in form I. The entropy contribution from this rotation can be calculated from $k_B T \int p(\theta) \log p(\theta) \sin \theta d\theta$. The difference in entropy between form I and B accounts for about 1.5 kJ/mol ($0.4 k_B T$). As the temperature is lowered, the structure undergoes a phase transition at around 200 K. Therefore, methods that search structures at zero temperature would only find the low temperature form instead of the high temperature one. The change in structure cannot be accounted for using harmonic corrections.

We now turn to discuss polymorph A of naphthalene. Form A has a layered structure and its space group is Pnmm [112]. The structure is shown in Figure 2.4.4. During an unbiased simulation at 300 K, form A decays to the liquid. This is consistent with the calculated free energy (see Figure 2.4.3) that shows that form A has a free energy around $3 k_B T$ higher than form I and the liquid. In spite of the relatively high free energy, it is possible that this polymorph could be kinetically trapped.

We have presented a method to explore polymorphism in molecular crystals in finite temperature molecular dynamics simulations. An important feature of our method is that not only does it discover polymorphs but also pinpoints which are the relevant ones at finite temperature. In fact, the new polymorph of urea, form B, could have not been predicted from a zero temperature search with harmonic corrections. A key ingredient of our approach is the structure similarity metric defined using $g(r, \theta)$ and the new distance in equation (2.4.3). This metric allows us to automatically assign configurations to a given polymorph, thus reducing the burden of the analysis

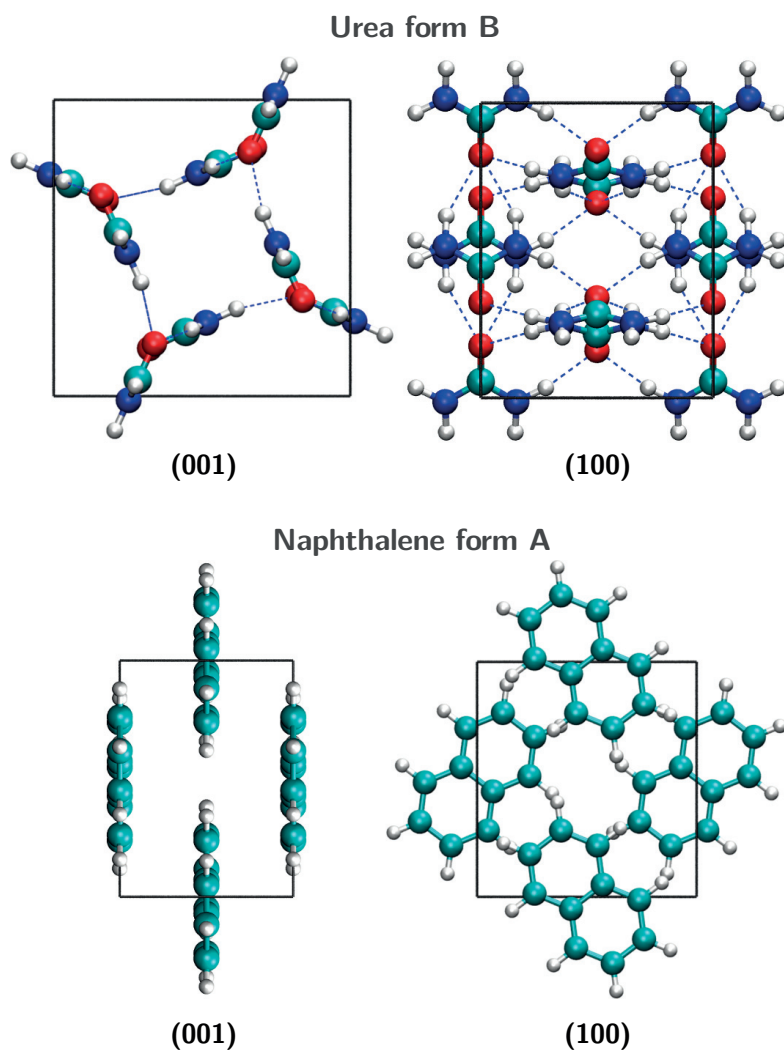


Figure 2.4.4: Crystal structures of the new forms of urea and naphthalene. C, O, N and H atoms are shown in cyan, red, blue and white, respectively. Images obtained with VMD [111].

of the simulations. We are also able to calculate free energies and entropies from the simulation using a reweighting procedure [109]. In the future, we plan to generalize our approach to crystals with more complex hydrogen bond networks and to molecules with internal degrees of freedom.

Materials and methods

Urea and naphthalene were described using the generalized amber force field (GAFF) [113]. For naphthalene, the electrostatic potential was calculated at the B3LYP/6-31+G(d,p) level using Gaussian 09 [114] and the partial charges of the atoms were fitted using the restrained electrostatic potential (RESP) method [115]. The partial charges of urea were those provided with the Amber 03 database [116]. Biased MD simulations

were performed using Gromacs 5.1.4 [117] patched with a development version of PLUMED 2 [45]. Van der Waals interactions and the electrostatic interaction in real space were calculated with cutoffs 0.9 nm and 0.75 nm for urea and naphthalene, respectively. The electrostatic interaction in reciprocal space was calculated using the particle mesh ewald (PME) method [118]. The atomic bonds involving hydrogen were constrained using the LINCS algorithm [119] and the equations of motion were integrated with a 2 fs timestep. The temperature was controlled using the stochastic velocity rescaling thermostat [46] with a relaxation time of 0.1 ps. The target temperature of the thermostat was 450 K and 300 K for urea and naphthalene, respectively. We maintained the pressure at its atmospheric value employing the isotropic version of the Parrinello-Rahman [47] barostat with a 10 ps relaxation time. We employed systems of 108 and 36 molecules for urea and naphthalene, respectively.

It must be added here that we do not expect that the observed nucleation process is realistic. In fact small size effects and periodic boundary conditions will artificially promote crystallization. In the present context this is a desirable feature. If we were to study the nucleation process much larger systems should be considered.

We now provide the parameters used for the WTMetaD simulations [50]. The Gaussians had a width of $0.1 k_B$ and $0.2 k_B$ for urea and naphthalene, respectively. In all cases the Gaussians had an initial height of $5 k_B T$ and were deposited every 1 ps. The bias factor was 200 for all simulations. The maximum free energy explored by WTMetaD is roughly the bias factor in $k_B T$. In our case we use a relatively large bias factor that ensures the exploration of high free energy regions.

We now discuss some practical aspects of the use of S_θ as a CV. In order to calculate the forces arising from the WTMetaD bias, S_θ should be continuous and differentiable. This can be achieved by constructing the function $g(r, \theta)$ using Gaussian kernels of width σ_r and σ_θ , as done in previous work. Furthermore, the integration in equation (2.4.2) cannot have an infinite upper limit, and in practice a finite cutoff r_m is taken. The integration is performed numerically using the trapezoid rule with steps of size σ_r and σ_θ in the r and $\cos \theta$ dimension, respectively. We report in Table 2.4.1 the chosen parameters. σ_θ is reported in units of $\cos \theta$. A subtlety in the calculation of S_θ is the periodicity of $g(r, \theta)$ in its θ argument. For a general molecule $g(r, \theta)$ is periodic in θ with period π . However, for a molecule that has a mirror symmetry with respect to the plane perpendicular to the vector v defining the orientation of the molecule, $g(r, \theta)$ has a period $\pi/2$. We report in Table 2.4.1 whether a given CV is defined based on a direction of the molecule with mirror symmetry.

	r_m (nm)	σ_r (nm)	σ_θ	mirror symmetry
urea				
S_{θ_1}	0.6	0.05	0.25	no
S_{θ_2}	0.6	0.05	0.125	yes
naphthalene				
S_{θ_1}	0.7	0.05	0.125	yes
S_{θ_2}	0.7	0.05	0.125	yes

Table 2.4.1: Parameters used in the definition of CVs S_{θ_1} and S_{θ_2} for urea and naphthalene. See text for details.

Acknowledgments

We are grateful to Zoran Bjelobrk for providing the force field for naphthalene. We would also like to thank Haiyang Niu for his valuable assistance in the analysis using the structure factor. This research was supported by the NCCR MARVEL, funded by the Swiss National Science Foundation. The authors also acknowledge funding from European Union Grant No. ERC-2014-AdG-670227/VARMET. The computational time for this work was provided by the Swiss National Supercomputing Center (CSCS) under Project ID mr3. Calculations were performed in CSCS cluster Piz Daint.

3 Conclusions

In this thesis I have presented different methods to study crystallization using computer simulations. The main method is based on employing approximate expressions for the entropy as collective variables. In particular we make use of the so called two-body excess entropy approximation [22, 23] that is based on the radial distribution function $g(r)$. I have shown that these collective variables can drive crystallization both in simple metals [75] and in molecular crystals [120]. Moreover, since they are crystal-structure agnostic, they can be used for the prediction of crystal structures. At variance with traditional crystal structure prediction methods that work at 0 K, our approach probes the stability of crystal structures at finite temperature thus automatically including the effects of entropy. In order to illustrate this point, I have shown an application of the method to the discovery of polymorphs in urea and naphthalene [120]. The simulations are able to find the already known polymorphs, and also a new interesting polymorph of urea that is stabilized by entropic effects. This finding illustrates that the effects of temperature must be treated carefully when performing crystal structure prediction.

I have also presented a novel fingerprint based on the two-body excess entropy [121] that allows to characterize atomic environments. This fingerprint can distinguish well between liquid-like and solid-like local environments, even if it does not contain any information about the particular crystal structure under study. Recently it has also been employed to characterize atomic environments in amorphous solids [122]. Furthermore, when this entropic fingerprint is combined with an enthalpic counterpart, they are able to distinguish between atomic environments compatible with different polymorphs.

In this thesis I have also introduced a generalized Kullback-Leibler divergence between two radial distribution functions. This divergence is inspired by the interpretation of the two-body excess entropy as a distance between the radial distribution function of the system and that of the ideal gas. I have generalized the two-body excess entropy

Chapter 3. Conclusions

expression to a divergence that measures the distance between any two radial distribution functions and employed it as a metric to compare atomistic configurations. The metric is very sensitive to changes in the structure and I have applied it to the unsupervised classification of the polymorphs that appear during a simulation [120].

Finally, I have presented a variational formalism to construct a bias potential based on the free energy expression of classical nucleation theory [123]. In this method, efficient sampling of the free energy surface is achieved simultaneously with the determination of the parameters of classical nucleation theory, namely the supersaturation and the surface free energy.

4 Future work

In this chapter I discuss ideas for the future development of the methods described in this thesis and I also propose new applications.

4.1 Two-body excess entropy collective variable

Using approximate expressions for the entropy in order to construct collective variables is a relatively general idea. However, in order to construct good approximations one must include information about the physics and chemistry of the system being studied. Consider the case of small organic molecules that were tackled using Eq. (2.4.2). This equation encodes one of the main characteristics of these systems, namely that correlations in the orientations between molecules are key to describe molecular crystals. Therefore the application to other type of systems will also require to include relevant physical information. In this section I discuss the cases of ionic crystals and molecules with conformational flexibility. I analyze the characteristics of each type of system and I propose possible collective variables.

4.1.1 Ionic crystals

A large number of materials fall in the category of ionic crystals, from traditional ceramics to state-of-the-art materials for electronic applications. Here I shall focus only on binary AB compounds. The most frequent crystal structures in this class are that of NaCl and CsCl [124]. In order to construct a collective variable for these systems we first note that in this case correlations between different pairs of species will be relevant. These correlations can be described through the radial distribution functions $g_{AA}(r)$, $g_{BB}(r)$, and $g_{AB}(r)$ that involve AA, BB, and AB pairs, respectively. An entropy expansion in terms of multibody correlation functions for a multicomponent system has been discussed in refs. [80, 81]. In particular for a binary AB compound the

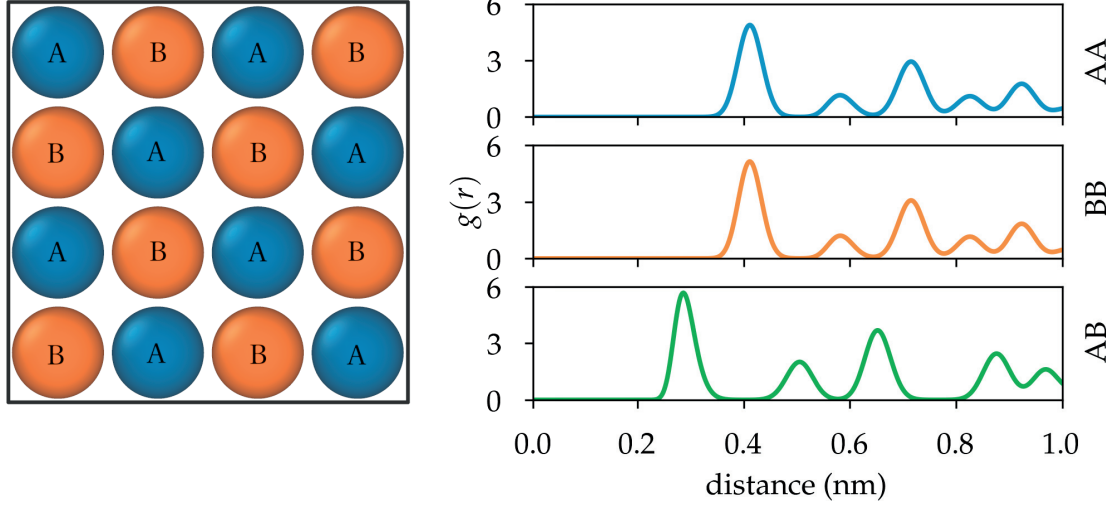


Figure 4.1.1: Left: Representation of the {100} plane of a rocksalt structure. Atoms have been labeled as A and B to match the discussion in the text. Right: Radial distribution function of NaCl rocksalt structure at 600 K. $g_{AA}(r)$, $g_{BB}(r)$, and $g_{AB}(r)$ are shown in the top, middle, and bottom plot respectively.

two-body term is:

$$\begin{aligned}
 \frac{S_2}{k_B(N_A + N_B)} &= -2\pi \frac{\rho_A^2}{\rho} \int_0^\infty [g_{AA}(r) \ln g_{AA}(r) - g_{AA}(r) + 1] r^2 dr \\
 &\quad -4\pi \frac{\rho_A \rho_B}{\rho} \int_0^\infty [g_{AB}(r) \ln g_{AB}(r) - g_{AB}(r) + 1] r^2 dr \\
 &\quad -2\pi \frac{\rho_B^2}{\rho} \int_0^\infty [g_{BB}(r) \ln g_{BB}(r) - g_{BB}(r) + 1] r^2 dr \quad (4.1.1)
 \end{aligned}$$

where ρ_A and ρ_B are the number densities of species A and B, respectively, and ρ is the global density. One could employ this expression as a collective variable in order to explore polymorphism.

Figure 4.1.1 shows a representation of the rocksalt structure for species A and B (left), and an example of the radial distribution function for NaCl (right). Although NaCl at ambient conditions is the paradigmatic example of the rocksalt structure, it also has a high pressure polymorph with the structure of CsCl [125]. It is therefore an interesting candidate to test a collective variable based on Eq. (4.1.1). Another possible candidate is ZnO that can crystallize both in the zinc blende and in the rocksalt structure [126].

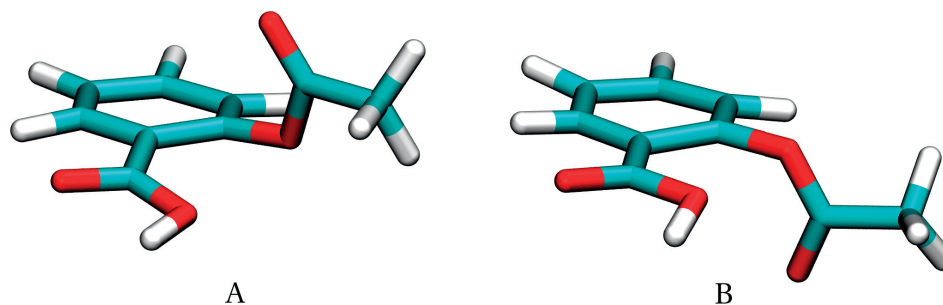


Figure 4.1.2: Main conformational isomers or conformers of aspirin.

4.1.2 Molecular crystals with conformational flexibility

Most molecules employed as active pharmaceutical ingredients exhibit conformational isomerism. It is often the case that the barriers for the interconversion of different conformers are high enough to hinder good sampling. In the case of crystallization from solution or the melt, the formation of the crystal requires that each lattice position is occupied by the right conformer. This requirement increases the complexity of crystallizing these molecules compared to molecules without conformational flexibility.

Consider the case of aspirin that has two main conformers that I shall name A and B. The two conformers differ in the rotation of the ester functional group. Other conformers exist [127], although the probability of finding them is relatively low compared with that of A and B. The most common polymorph of aspirin, form I, hosts both conformers A and B simultaneously in its crystal structure. The conformers are shown in Figure 4.1.2.

If the conformers A and B are regarded as two different components, then the two body entropy for a binary mixture discussed in the previous section can be employed. However, the inclusion of the correlations in the orientations between molecules is still crucial to describe the system. Thus, combining Eq. (4.1.1) with the extended radial distribution function $g(r, \theta)$ in Eq. (2.4.2) we obtain,

$$\begin{aligned}
 \frac{S_2}{k_B(N_A + N_B)} = & -\pi \frac{\rho_A^2}{\rho} \int_0^\infty \int_0^\pi [g_{AA}(r, \theta) \ln g_{AA}(r, \theta) - g_{AA}(r, \theta) + 1] r^2 \sin \theta \, dr \, d\theta \\
 & -2\pi \frac{\rho_A \rho_B}{\rho} \int_0^\infty \int_0^\pi [g_{AB}(r, \theta) \ln g_{AB}(r, \theta) - g_{AB}(r, \theta) + 1] r^2 \sin \theta \, dr \, d\theta \\
 & -\pi \frac{\rho_B^2}{\rho} \int_0^\infty \int_0^\pi [g_{BB}(r, \theta) \ln g_{BB}(r, \theta) - g_{BB}(r, \theta) + 1] r^2 \sin \theta \, dr \, d\theta.
 \end{aligned}
 \tag{4.1.2}$$

This expression encodes information both on the relative distance and orientation between molecules and on the spatial distribution of conformers. Although aspirin has been used and studied for over a century, it has recently been shown to exhibit polymorphism [128]. Therefore, it would be interesting to study whether Eq. (4.1.2) can be used as a collective variable to study polymorphism in aspirin.

4.2 Generalized Kullback-Leibler divergence

The generalized Kullback-Leibler divergence described in Eq. (2.4.3) is an important outcome of this thesis. The divergence can be used in all scenarios in which a distance between radial distribution functions is required. Here, I propose one application that could profit from this divergence.

Consider the crystallization of a one component system. We assume that the only phases that exist in some thermodynamic condition are the liquid and the solid, and we characterize each of them by their radial distribution functions $g_l(r)$ and $g_s(r)$, respectively. We can construct two divergences $D(g||g_l)$ and $D(g||g_s)$ that measure the distance between the current $g(r)$ of the system and a target $g(r)$, i.e. $g_l(r)$ for the liquid and $g_s(r)$ for the solid. These divergences are:

$$D(g||g_l) = \int_0^{\infty} \left[g(r) \ln \frac{g(r)}{g_l(r)} - g(r) + g_l(r) \right] r^2 dr,$$

$$D(g||g_s) = \int_0^{\infty} \left[g(r) \ln \frac{g(r)}{g_s(r)} - g(r) + g_s(r) \right] r^2 dr. \quad (4.2.3)$$

We consider the case of elemental sodium and analyze a biased simulation in which the liquid and solid phases are visited reversibly (see ref. [75]). In Figure 4.2.3a we plot $D(g||g_l)$ vs $D(g||g_s)$ for this trajectory. The biased probability of finding the system in a given point is shown with contour lines. Some points of the trajectory are shown and colored with the common neighbor analysis [65, 66] technique that quantifies the fraction of bcc atoms. We note that the line $D(g||g_l) = D(g||g_s)$ divides the space in two important regions. The region $D(g||g_l) < D(g||g_s)$ corresponds to the configurations that are closer to the liquid than to the solid, and for the region $D(g||g_l) > D(g||g_s)$ the opposite is true.

It is clear that this result can be used for classification purposes but it can also be employed to define new collective variables. A very simple collective variable is:

$$s = D(g||g_l) - D(g||g_s). \quad (4.2.4)$$

This collective variable has the property that for positive s the system is closer to

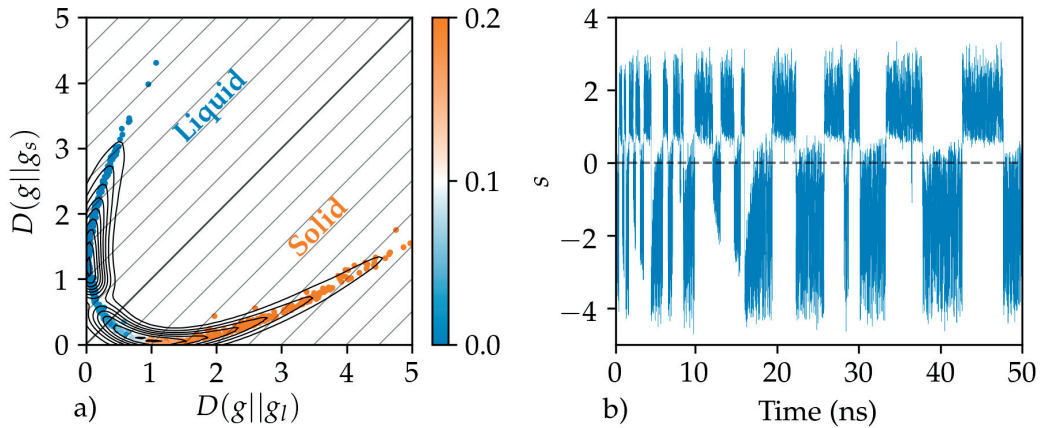


Figure 4.2.3: Analysis of a biased trajectory using the divergences defined in equation (4.2.3). a) Scatter plot and contour lines of the joint probability density of $D(g||g_l)$ and $D(g||g_s)$. The regions closer to the solid than to the liquid are separated with a thick black line. Lines of constant s are shown in thin grey lines. The points in the scatter plot are colored according to the fraction of bcc atoms as obtained from common neighbor analysis [65, 66]. Note that the solid is not fully bcc due to the thermal noise. b) Collective variable s vs time for the biased trajectory.

the solid than to the liquid, and for negative s the system is closer to the liquid than to the solid. In Figure 4.2.3b) s is plotted against time for the biased trajectory of sodium. It is clear that s is able to distinguish very well between the liquid and solid states. A remarkable characteristic of this collective variable is that all the information contained in the $g(r)$ is leveraged in an elegant way to distinguish between the liquid and the solid.

The preliminary results presented here are very encouraging and I envisage that these ideas could be used to distinguish multiple polymorphs and the liquid.

Other articles

In this chapter I present a list of articles which I co-authored during my PhD with topics unrelated to this thesis work. These are:

- 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(21):5348–5352, 2018. doi: 10.1073/pnas.1803919-115. URL <http://www.pnas.org/content/115/21/5348>.
- Tarak Karmakar, Pablo M Piaggi, Claudio Perego, and Michele Parrinello. A cannibalistic approach to grand canonical crystal growth. *Journal of chemical theory and computation*, 14(5):2678–2683, 2018. doi: 10.1021/acs.jctc.8b00191. URL <https://doi.org/10.1021/acs.jctc.8b00191>.
- Dan Mendels, James McCarty, Pablo M Piaggi, and Michele Parrinello. Searching for entropically stabilized phases: The case of silver iodide. *The Journal of Physical Chemistry C*, 122(3):1786–1790, 2018. doi: 10.1021/acs.jpcc.7b11002. URL <https://doi.org/10.1021/acs.jpcc.7b11002>.
- Sergio Perez-Conesa, Pablo M Piaggi, and Michele Parrinello. A local fingerprint for hydrophobicity and hydrophilicity: from methane to peptides. *Submitted to Journal of chemical theory and computation*, arXiv preprint arXiv:1808.09712, 2018. URL <https://arxiv.org/abs/1808.09712>.
- Paramvir Ahlawat, Pablo M Piaggi, Michele Parrinello, Michael Graetzel, and Ursula Röthlisberger. Atomistic mechanism of the nucleation of methylammonium lead iodide from solution. *arXiv preprint arXiv:1810.00759*, 2018. URL <https://arxiv.org/abs/1810.00759>

Bibliography

- [1] Frank H Stillinger and Thomas A Weber. Study of melting and freezing in the gaussian core model by molecular dynamics simulation. *The Journal of Chemical Physics*, 68(8):3837–3844, 1978.
- [2] MJ Mandell, JP McTague, and A Rahman. Crystal nucleation in a three-dimensional lennard-jones system: A molecular dynamics study. *The journal of chemical physics*, 64(9):3699–3702, 1976.
- [3] CS Hsu and Aneesur Rahman. Crystal nucleation and growth in liquid rubidium. *The Journal of Chemical Physics*, 70(11):5234–5240, 1979.
- [4] Masakazu Matsumoto, Shinji Saito, and Iwao Ohmine. Molecular dynamics simulation of the ice nucleation and growth process leading to water freezing. *Nature*, 416(6879):409, 2002.
- [5] Federica Trudu, Davide Donadio, and Michele Parrinello. Freezing of a lennard-jones fluid: From nucleation to spinodal regime. *Physical review letters*, 97(10):105701, 2006.
- [6] Peter G Bolhuis, David Chandler, Christoph Dellago, and Phillip L Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual review of physical chemistry*, 53(1):291–318, 2002.
- [7] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.
- [8] Eric Darve and Andrew Pohorille. Calculating free energies using average force. *The Journal of Chemical Physics*, 115(20):9169–9183, 2001.
- [9] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.
- [10] 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(7):072301, 2018.

Bibliography

- [11] Mohammad M Sultan and Vijay S Pande. Automated design of collective variables using supervised machine learning. *The Journal of chemical physics*, 149(9):094106, 2018.
- [12] Dan Mendels, Giovanni Maria Piccini, and Michele Parrinello. Collective variables from local fluctuations. *The journal of physical chemistry letters*, 9(11):2776–2781, 2018.
- [13] Paul Steinhardt, David Nelson, and Marco Ronchetti. Bond-orientational order in liquids and glasses. *Phys. Rev. B*, 28:784–805, Jul 1983.
- [14] JS Van Duijneveldt and D Frenkel. Computer simulation study of free energy barriers in crystal nucleation. *The Journal of chemical physics*, 96(6):4655–4668, 1992.
- [15] Erik E. Santiso and Bernhardt L. Trout. A general set of order parameters for molecular crystals. *The Journal of Chemical Physics*, 134(6):–, 2011.
- [16] Federico Giberti, Matteo Salvalaglio, Marco Mazzotti, and Michele Parrinello. Insight into the nucleation of urea crystals from the melt. *Chemical Engineering Science*, 121:51–59, 2015.
- [17] Stefano Angioletti-Uberti, Michele Ceriotti, Peter D Lee, and Mike W Finnis. Solid-liquid interface free energy through metadynamics simulations. *Physical Review B*, 81(12):125416, 2010.
- [18] Christian Bartels and Martin Karplus. Probability distributions for complex systems: adaptive umbrella sampling of the potential energy. *The Journal of Physical Chemistry B*, 102(5):865–880, 1998.
- [19] Omar Valsson and Michele Parrinello. Thermodynamical description of a quasi-first-order phase transition from the well-tempered ensemble. *Journal of chemical theory and computation*, 9(12):5267–5276, 2013.
- [20] Massimiliano Bonomi and Michele Parrinello. Enhanced sampling in the well-tempered ensemble. *Physical review letters*, 104(19):190601, 2010.
- [21] Caroline Desgranges and Jerome Delhommelle. Free energy calculations along entropic pathways. i. homogeneous vapor-liquid nucleation for atomic and molecular systems. *The Journal of chemical physics*, 145(20):204112, 2016.
- [22] Herbert S Green. *The molecular theory of fluids*. North-Holland Publishing Company Amsterdam, 1952.
- [23] RE Nettleton and MS Green. Expression in terms of molecular distribution functions for the entropy density in an infinite system. *The Journal of Chemical Physics*, 29(6):1365–1370, 1958.

-
- [24] Mark Tuckerman. *Statistical mechanics: theory and molecular simulation*. Oxford university press, 2010.
- [25] Omar Valsson and Michele Parrinello. Variational approach to enhanced sampling and free energy calculations. *Physical review letters*, 113(9):090601, 2014.
- [26] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*, volume 1. Academic press, 2001.
- [27] Omar Valsson, Pratyush Tiwary, and Michele Parrinello. Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint. *Annual Review of Physical Chemistry*, 67(1), 2016.
- [28] Dimo Kashchiev. *Nucleation: Basic Theory with Applications*. Butterworth-Heinemann, Oxford, 2000.
- [29] Michele Invernizzi, Omar Valsson, and Michele Parrinello. Coarse graining from variationally enhanced sampling applied to the ginzburg–landau model. *Proceedings of the National Academy of Sciences*, page 201618455, 2017.
- [30] Kwok Chow, Henry H.Y. Tong, Susan Lum, and Albert H.L. Chow. Engineering of pharmaceutical materials: An industrial perspective. *Journal of Pharmaceutical Sciences*, 97(8):2855–2877, 2008.
- [31] William M Jacobs and Daan Frenkel. Self-assembly of structures with addressable complexity. *Journal of the American Chemical Society*, 138(8):2457–2467, 2016.
- [32] Peter G Vekilov. Nucleation. *Crystal growth & design*, 10(12):5007–5019, 2010.
- [33] Roger J Davey, Sven LM Schroeder, and Joop H ter Horst. Nucleation of organic crystals—a molecular perspective. *Angewandte Chemie International Edition*, 52(8):2166–2179, 2013.
- [34] Bin Chen, J Ilja Siepmann, Kwang J Oh, and Michael L Klein. Aggregation-volume-bias monte carlo simulations of vapor-liquid nucleation barriers for lennard-jonesium. *The Journal of Chemical Physics*, 115(23):10903–10913, 2001.
- [35] Isamu Kusaka and David W Oxtoby. Identifying physical clusters in vapor phase nucleation. *The Journal of chemical physics*, 110(11):5249–5261, 1999.
- [36] Joonas Merikanto, Hanna Vehkamäki, and Evgeni Zapadinsky. Monte carlo simulations of critical cluster sizes and nucleation rates of water. *The Journal of chemical physics*, 121(2):914–924, 2004.
- [37] 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(22):9901–9918, 1998.

Bibliography

- [38] James McCarty, Omar Valsson, Pratyush Tiwary, and Michele Parrinello. Variationally optimized free-energy flooding for rate calculation. *Physical review letters*, 115(7):070601, 2015.
- [39] Omar Valsson and Michele Parrinello. Well-tempered variational approach to enhanced sampling. *Journal of chemical theory and computation*, 11(5):1996–2002, 2015.
- [40] Patrick Shaffer, Omar Valsson, and Michele Parrinello. Enhanced, targeted sampling of high-dimensional free-energy landscapes using variationally enhanced sampling, with an application to chignolin. *Proceedings of the National Academy of Sciences*, page 201519712, 2016.
- [41] James McCarty, Omar Valsson, and Michele Parrinello. Bespoke bias for obtaining free energy differences within variationally enhanced sampling. *Journal of Chemical Theory and Computation*, pages 2162–2169, 2016.
- [42] Pratyush Tiwary and Michele Parrinello. From metadynamics to dynamics. *Physical review letters*, 111(23):230602, 2013.
- [43] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.
- [44] Steve Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics*, 117(1):1–19, 1995.
- [45] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. Plumed 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, 2014.
- [46] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1):014101, 2007.
- [47] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.
- [48] Lutz Maibaum. Comment on “elucidating the mechanism of nucleation near the gas-liquid spinodal”. *Physical review letters*, 101(1):019601, 2008.
- [49] Cynthia D Holcomb, Paulette Clancy, and John A Zollweg. A critical study of the simulation of the liquid-vapour interface of a lennard-jones fluid. *Molecular Physics*, 78(2):437–459, 1993.
- [50] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Physical review letters*, 100(2):020603, 2008.

-
- [51] Arthur F Voter. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Physical Review Letters*, 78(20):3908, 1997.
- [52] Helmut Grubmüller. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Physical Review E*, 52(3):2893, 1995.
- [53] Daniele Moroni, Pieter Rein ten Wolde, and Peter G. Bolhuis. Interplay between structure and size in a critical crystal nucleus. *Phys. Rev. Lett.*, 94:235703, Jun 2005.
- [54] 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.
- [55] Stefan Goedecker. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *The Journal of chemical physics*, 120(21):9911–9917, 2004.
- [56] Artem R Oganov and Colin W Glass. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *The Journal of chemical physics*, 124(24):244704, 2006.
- [57] Emanuela Giuffré, Santi Prestipino, Franz Saija, A Marco Saitta, and Paolo V Giaquinta. Entropy from correlations in tip4p water. *Journal of chemical theory and computation*, 6(3):625–636, 2010.
- [58] Thomas M Truskett, Salvatore Torquato, and Pablo G Debenedetti. Towards a quantification of disorder in materials: distinguishing equilibrium and glassy sphere packings. *Physical Review E*, 62(1):993, 2000.
- [59] Cui Zhang, Leonardo Spanu, and Giulia Galli. Entropy of liquid water from ab initio molecular dynamics. *The Journal of Physical Chemistry B*, 115(48):14190–14195, 2011.
- [60] Xin Zhou, Yi Jiang, Kurt Kremer, Hans Ziock, and Steen Rasmussen. Hyperdynamics for entropic systems: Time-space compression and pair correlation function approximation. *Physical Review E*, 74(3):035701, 2006.
- [61] Pratyush Tiwary and BJ Berne. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proceedings of the National Academy of Sciences*, page 201600917, 2016.
- [62] SR Wilson, KGSH Gunawardana, and MI Mendeleev. Solid-liquid interface free energies of pure bcc metals and b2 phases. *The Journal of chemical physics*, 142(13):134705, 2015.

Bibliography

- [63] MI Mendeleev, MJ Kramer, Chandler A Becker, and M Asta. Analysis of semi-empirical interatomic potentials appropriate for simulation of crystalline and liquid al and cu. *Philosophical Magazine*, 88(12):1723–1750, 2008.
- [64] *VES Code*, a library that implements enhanced sampling methods based on Variationally Enhanced Sampling written by O. Valsson. For the current version, see <http://www.ves-code.org>.
- [65] 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.
- [66] J Dana Honeycutt and Hans C Andersen. Molecular dynamics study of melting and freezing of small lennard-jones clusters. *Journal of Physical Chemistry*, 91(19):4950–4963, 1987.
- [67] Jess B Sturgeon and Brian B Laird. Adjusting the melting point of a model system via gibbs-duhem integration: Application to a model of aluminum. *Physical Review B*, 62(22):14720, 2000.
- [68] Malcolm W Chase. Janaf thermochemical tables. *JANAF thermochemical tables, by Chase, MW Washington, DC: American Chemical Society; New York: American Institute of Physics for the National Bureau of Standards, c1986.. United States. National Bureau of Standards.*, 1, 1986.
- [69] Pieter Rein ten Wolde and Daan Frenkel. Homogeneous nucleation and the ostwald step rule. *Physical Chemistry Chemical Physics*, 1(9):2191–2196, 1999.
- [70] Pieter Rein ten Wolde, Maria J. Ruiz-Montero, and Daan Frenkel. Numerical evidence for bcc ordering at the surface of a critical fcc nucleus. *Phys. Rev. Lett.*, 75:2714–2717, Oct 1995.
- [71] Wolfgang Lechner, Christoph Dellago, and Peter G Bolhuis. Role of the prestructured surface cloud in crystal nucleation. *Physical review letters*, 106(8):085701, 2011.
- [72] Marc A Meyers, A Mishra, and David J Benson. Mechanical properties of nanocrystalline materials. *Progress in materials science*, 51(4):427–556, 2006.
- [73] Alexander Stukowski. Structure identification methods for atomistic simulations of crystalline materials. *Modelling and Simulation in Materials Science and Engineering*, 20(4):045021, 2012.
- [74] Wolfgang Lechner and Christoph Dellago. Accurate determination of crystal structures based on averaged local bond order parameters. *The Journal of chemical physics*, 129(11):114707, 2008.

-
- [75] Pablo M Piaggi, Omar Valsson, and Michele Parrinello. Enhancing entropy and enthalpy fluctuations to drive crystallization in atomistic simulations. *Physical Review Letters*, 119(1):015701, 2017.
- [76] Andras Baranyai and Denis J Evans. Direct entropy calculation from computer simulation of liquids. *Physical Review A*, 40(7):3817, 1989.
- [77] Duane C Wallace. On the role of density fluctuations in the entropy of a fluid. *The Journal of chemical physics*, 87(4):2282–2284, 1987.
- [78] Duane C Wallace. Statistical mechanical theory of liquid entropy. *International journal of quantum chemistry*, 52(2):425–435, 1994.
- [79] Brian B Laird and ADJ Haymet. Calculation of the entropy from multiparticle correlation functions. *Physical Review A*, 45(8):5680, 1992.
- [80] JA Hernando. Thermodynamic potentials and distribution functions: I. a general expression for the entropy. *Molecular Physics*, 69(2):319–326, 1990.
- [81] Santi Prestipino and Paolo V Giaquinta. The entropy multiparticle-correlation expansion for a mixture of spherical and elongated particles. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(09):P09008, 2004.
- [82] Tohru Morita and Kazuo Hiroike. A new approach to the theory of classical fluids. iii: General treatment of classical systems. *Progress of Theoretical Physics*, 25(4):537–578, 1961.
- [83] Ryoichi Kikuchi. A theory of cooperative phenomena. *Physical review*, 81(6):988, 1951.
- [84] Jean-Pierre Hansen and Loup Verlet. Phase transitions of the lennard-jones system. *physical Review*, 184(1):151, 1969.
- [85] Mathieu Leocmach, John Russo, and Hajime Tanaka. Importance of many-body correlations in glass transition: An example from polydisperse hard spheres. *The Journal of chemical physics*, 138(12):12A536, 2013.
- [86] PM Piaggi, EM Bringa, RC Pasianot, N Gordillo, M Panizo-Laiz, J del Río, C Gómez de Castro, and R Gonzalez-Arrabal. Hydrogen diffusion and trapping in nanocrystalline tungsten. *Journal of Nuclear Materials*, 458:233–239, 2015.
- [87] Murray S Daw and Michael I Baskes. Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals. *Physical Review B*, 29(12):6443, 1984.
- [88] MW Finnis and JE Sinclair. A simple empirical n-body potential for transition metals. *Philosophical Magazine A*, 50(1):45–55, 1984.

Bibliography

- [89] Joel Bernstein. *Polymorphism in molecular crystals*, volume 14. Oxford University Press, 2002.
- [90] Rolf Hilfiker. *Polymorphism: in the pharmaceutical industry*. John Wiley & Sons, 2006.
- [91] Walter Cabri, Paolo Ghetti, Giovanni Pozzi, and Marco Alpegiani. Polymorphisms and patent, market, and legal battles: cefdinir case study. *Organic process research & development*, 11(1):64–72, 2007.
- [92] John Bauer, Stephen Spanton, Rodger Henry, John Quick, Walter Dziki, William Porter, and John Morris. Ritonavir: an extraordinary example of conformational polymorphism. *Pharmaceutical research*, 18(6):859–866, 2001.
- [93] 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(14):5984–5991, 2002.
- [94] Sarah L Price. From crystal structure prediction to polymorph prediction: interpreting the crystal energy landscape. *Physical Chemistry Chemical Physics*, 10(15):1996–2009, 2008.
- [95] Chris J Pickard and RJ Needs. Ab initio random structure searching. *Journal of Physics: Condensed Matter*, 23(5):053201, 2011.
- [96] Tang-Qing Yu and Mark E Tuckerman. Temperature-accelerated method for exploring polymorphism in molecular crystals based on free energy. *Physical review letters*, 107(1):015701, 2011.
- [97] Dan Mendels, James McCarty, Pablo M Piaggi, and Michele Parrinello. Searching for entropically stabilized phases: The case of silver iodide. *The Journal of Physical Chemistry C*, 122(3):1786–1790, 2018.
- [98] Gianpaolo Gobbo, Michael A Bellucci, Gareth A Tribello, Giovanni Ciccotti, and Bernhardt L Trout. Nucleation of molecular crystals driven by relative information entropy. *Journal of chemical theory and computation*, 14(2):959–972, 2018.
- [99] FJ Lamelas, ZA Dreger, and YM Gupta. Raman and x-ray scattering studies of high-pressure phases of urea. *The Journal of Physical Chemistry B*, 109(16):8206–8215, 2005.
- [100] Anna Olejniczak, Kinga Ostrowska, and Andrzej Katrusiak. H-bond breaking in high-pressure urea. *The Journal of Physical Chemistry C*, 113(35):15761–15767, 2009.

-
- [101] 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(4):2380–2387, 2017.
- [102] 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(47):32125–32131, 2017.
- [103] Francesca PA Fabbiani, David R Allan, Simon Parsons, and Colin R Pulham. Exploration of the high-pressure behaviour of polycyclic aromatic hydrocarbons: naphthalene, phenanthrene and pyrene. *Acta Crystallographica Section B: Structural Science*, 62(5):826–842, 2006.
- [104] Anna Y Likhacheva, Sergey V Rashchenko, and Konstantin D Litasov. High-pressure structural properties of naphthalene up to 6 gpa. *Journal of Applied Crystallography*, 47(3):984–991, 2014.
- [105] Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769, 2016.
- [106] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.
- [107] Daniel Müllner et al. fastcluster: Fast hierarchical, agglomerative clustering routines for r and python. *Journal of Statistical Software*, 53(9):1–18, 2013.
- [108] Eric Jones, Travis Oliphant, and Pearu Peterson. Scipy: open source scientific tools for python. 2014.
- [109] Pratyush Tiwary and Michele Parrinello. A time-independent free energy estimator for metadynamics. *The Journal of Physical Chemistry B*, 119(3):736–742, 2014.
- [110] 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(41):17221–17233, 2012.
- [111] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.
- [112] HT Stokes, DM Hatch, and BJ Campbell. Isotropy software suite, iso. byu. edu.
- [113] 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(9):1157–1174, 2004.

Bibliography

- [114] MJE Frisch, GW Trucks, Hs B Schlegel, GE Scuseria, MA Robb, JR Cheeseman, G Scalmani, V Barone, B Mennucci, GA Petersson, et al. Gaussian 09, revision a. 02, gaussian. Inc., Wallingford, CT, 200, 2009.
- [115] Christopher I Bayly, Piotr Cieplak, Wendy Cornell, and Peter A Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model. *The Journal of Physical Chemistry*, 97(40):10269–10280, 1993.
- [116] David A Case, Thomas E Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The amber biomolecular simulation programs. *Journal of computational chemistry*, 26(16):1668–1688, 2005.
- [117] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, 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] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh ewald method. *The Journal of chemical physics*, 103(19):8577–8593, 1995.
- [119] Berk Hess. P-lincs: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation*, 4(1):116–122, 2008.
- [120] Pablo M Piaggi and Michele Parrinello. Predicting polymorphism in molecular crystals using orientational entropy. *arXiv preprint arXiv:1806.06006, accepted to Proceedings of the National Academy of Sciences*, 2018.
- [121] Pablo M Piaggi and Michele Parrinello. Entropy based fingerprint for local crystalline order. *The Journal of chemical physics*, 147(11):114112, 2017.
- [122] James E Hallett, Francesco Turci, and C Patrick Royall. Local structure in deeply supercooled liquids exhibits growing lengthscales and dynamical correlations. *Nature communications*, 9(1):3272, 2018.
- [123] Pablo M Piaggi, Omar Valsson, and Michele Parrinello. A variational approach to nucleation simulation. *Faraday discussions*, 195:557–568, 2017.
- [124] David G Pettifor. *Bonding and structure of molecules and solids*. Oxford University Press, 1995.
- [125] William A Bassett, Taro Takahashi, Ho-Kwang Mao, and J Scott Weaver. Pressure-induced phase transformation in nacl. *Journal of Applied Physics*, 39(1):319–325, 1968.

- [126] Carl H Bates, William B White, and Rustum Roy. New high-pressure polymorph of zinc oxide. *Science*, 137(3534):993–993, 1962.
- [127] Rainer Glaser. Aspirin. an ab initio quantum-mechanical study of conformational preferences and of neighboring group interactions. *The Journal of organic chemistry*, 66(3):771–779, 2001.
- [128] Peddy Vishweshwar, Jennifer A McMahon, Mark Oliveira, Matthew L Peterson, and Michael J Zaworotko. The predictably elusive form ii of aspirin. *Journal of the American Chemical Society*, 127(48):16802–16803, 2005.

Pablo Piaggi

Via Giuseppe Parini 11
6900 Lugano
Switzerland
☎ +41 (0)76 296 96 03
✉ pablo.piaggi@epfl.ch



Education

- 2015-2018 **Doctoral program in Materials Science and Engineering**, *École Polytechnique Fédérale de Lausanne*, Lausanne, Switzerland.
Advisors: Prof. Michele Parrinello and Prof. Nicola Marzari.
Date of oral examination: November 26, 2018
- 2010-2014 **Materials Science and Engineering**, *Instituto Sabato, Comisión Nacional de Energía Atómica*, Buenos Aires, Argentina, Grade point average: 9.58 out of 10.
Materials Engineer
- 2008-2010 **Industrial Engineering**, *Universidad Tecnológica Nacional, Facultad Regional Buenos Aires*, Buenos Aires, Argentina.

Employment history

- November 2014-Present **PhD thesis**, *Advisors: Prof. Michele Parrinello and Prof. Nicola Marzari*, EPFL, Lausanne and Università della Svizzera italiana, Lugano, Switzerland.
- September-October 2014 **Research fellowship**, *Hydrogen embrittlement of nanocrystalline tungsten*, Advisor: Dr. Roberto C. Pasianot, Centro Atómico Constituyentes, Buenos Aires, Argentina.
- January-July 2014 **Engineering thesis**, *Atomistic simulation of hydrogen in nanocrystalline tungsten*, Advisor: Dr. Eduardo M. Bringa, Universidad de Cuyo, Mendoza, Argentina.
- February 2013 **Summer internship**, *Computer generation of substitutional binary alloy crystals with short range order*, Advisor: Dr. Roberto C. Pasianot, Centro Atómico Constituyentes, Buenos Aires, Argentina.
- February 2012 **Summer internship**, *Influence of defects in the electronic and transport properties of graphene ribbons*, Advisor: Dr. Ana María Llois, Centro Atómico Constituyentes, Buenos Aires, Argentina.
- February 2011 **Summer internship**, *MgB₂ superconductors*, Advisor: Dr. Adriana Serquis, Instituto Balseiro, Bariloche, Argentina.

Organization of conferences and schools

- February 2017 Organizer (together with Dr. Omar Valsson) of the Variationally Enhanced Sampling School, USI, Lugano, Switzerland

Prizes, awards and fellowships

- July 2018 **Poster prize "Faraday Discussion: Methods and Applications of Crystal Structure Prediction"**, *Royal Society of Chemistry, Cambridge, United Kingdom.*
- September 2016 **Poster prize "Faraday Discussion: Reaction Rate Theory"**, *Royal Society of Chemistry, Cambridge, United Kingdom.*
- 2015 **Award "Best Graduates of Argentine Universities 2014"**, *National Academy of Engineering, Buenos Aires, Argentina.*
- 2014 **Award "Ing. Federico José Bonavita"**, *Highest grade point average in class 2014, Materials Science and Engineering, Instituto Sabato, Buenos Aires, Argentina.*
- 2014 **Research scholarship**, *Granted by Comisión Nacional de Energía Atómica, Argentina.*
- 2010-2014 **Scholarship**, *Granted by Comisión Nacional de Energía Atómica and YPF, Argentina.*

Memberships in panels, boards, and individual scientific reviewing activities

- 2015-2018 **Reviewer of two articles for Physical Review Letters and one article for Physical Review E .**

Personal skills

Language	Spanish	<i>Native</i>
Language	English	<i>Advanced - Certificate of Advanced English, Cambridge</i>
Language	Italian	<i>Advanced - Level C1</i>
Digital competence	C++, Python, MPI, openMP	
Digital competence	Usage of large computer clusters	

