A rare event:
Understanding, inhibiting, and optimising nucleation in colloidal systems

On the cover: On the front, the breaking of the fivefold symmetry for the benefit of the periodic sixfold symmetry is represented by the breaking of a pentagon by a stick figure. On the back, the same stick figure looks satisfied at the hexagon resulting from the breaking of the pentagon.

PhD thesis, Utrecht University, the Netherlands, September 2021.

ISBN: 978 94 6423 499 2

Printed by Proefschrift Maken.

The work described in this thesis was performed in the Soft Condensed Matter group, part of the Debye Institute for Nanomaterials Science, at Utrecht University.

A digital version of this thesis is available at http://www.colloid.nl

A rare event:
Understanding, inhibiting, and optimising nucleation in colloidal systems

---

Een zeldzame gebeurtenis:
Inzicht, voorkomen en optimaliseren van nucleatie in colloïdale systemen.

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties in het openbaar te verdedigen
op woensdag 10 november 2021 des middags te 12.15 uur

door

Gabriele Maria Coli

geboren op 22 februari 1993 te Rome, Italië

Promotor: Prof. dr. ir. M. Dijkstra

Assessment committee:
Dr. E. Sanz
Prof. dr. C. Dellago
Prof. dr. C. de Morais Smith
Prof. dr. R. A. Duine
Prof. dr. W. K. Kegel

This research was supported by the “Nederlandse Organisatie voor Wetenschappelijk Onderzoek” (NWO) as part of a Dutch-India collaboration ‘Data-driven science – Computational and Data Science – For Smart and Sustainable energy research’.

Contents

1

Introduction

In this introductory Chapter we describe all the key concepts that will be further explored in the course of this thesis. We begin by defining what is meant by colloidal systems, what characterises them, and the reasons behind the feverish scientific efforts by the scientific community in order to study their behaviour. We introduce the concept of nucleation, providing not only examples from common everyday experience, but also the theoretical framework within which this thesis is placed. Finally, we take the opportunity to mention other recurrent concepts in the present work and inherent to nucleation, such as fivefold symmetry and inverse design methods. Through this Chapter, we provide the reader with all the conceptual tools needed to tackle the subsequent chapters, which are more technical in nature.

1.1 Colloidal systems

When one presents a scientific work that refers to some branch of physics, both at a popular and technical level, usually even non-expert readers have at least a vague idea of what they are dealing with. This happens with a very large set of topics, ranging from gravity to semiconductors. On the contrary, when presenting a thesis on colloids, it is strictly necessary to start this long journey from the mere definition of the system under investigation. This is by the way a rather peculiar phenomenon, as each of us has to deal with colloids in everyday life. So, what are colloids?

One possibility is to call a colloid any particle whose size varies between about 1 nm and 1 μm [1]. This term comes from the Greek word $\kappa o \lambda \lambda \alpha$, which means glue, as it was coined to describe sticky particles on a semipermeable membrane in 1860, by Thomas Graham [2]. Other more familiar examples of colloids include the toothpaste, paints, milk, and various creams.

All of the above-mentioned examples involve not only colloids, but more precisely insoluble colloids dispersed in a medium composed of particles that are much smaller than the colloids themselves. This is the most interesting kind of colloidal system (as well as the one on which this thesis focuses), because it gives rise to the other fundamental characteristic of colloids, which also serves as an alternative definition for them. What distinguishes colloids from other systems in fact, more than their spatial dimensions, is their dynamics. Colloids perform trajectories which are random, but still possess very precise characteristics. This dynamics is called Brownian motion, as a tribute to Robert Brown, the botanist who first described the movement of grains of pollen in water in 1827 [3]. At the dawn of the twentieth century, Einstein and Sutherland [4, 5] gave an explanation to the random movement of colloidal particles suspended in a medium. They understood that the numerous molecules of the solvent are in constant motion due to their thermal energy, and collide continuously onto the surfaces of the colloids. The result of these interactions is a stochastic force that acts on the colloids themselves. Returning to the first attempt to give a definition of a colloid, based on spatial dimensions, we realise that both lower and upper limits (respectively 1 nm and 1 μm) are imposed by the dynamical characteristics just described. In other words, colloids are particles that are large enough not to be affected by quantum effects, but small enough to experience Brownian motion.

In a colloidal system, colloids interact not only with the solvent particles, but also with each other. The resulting collective behaviour is surprisingly rich and diverse. Indeed, thanks to the Brownian motion, a colloidal system effectively explores phase space in search of the minimum free energy. This exploration brings different systems into a multitude of different phases – ranging from the fluid phase to crystalline states, to liquid crystals and even to quasicrystals – through the process of self-assembly [6]. The phase behaviour of colloidal systems is therefore strikingly analogous to that of atoms and molecules, but on larger scales.

This similarity of the collective behaviour of colloidal systems to atomic systems suggests to us one of the main motivations behind the study of their self-assembly processes. Some phase transitions such as nucleation in fact are extremely complicated to study for atomic and molecular systems, because the constituent particles are extremely small and these events happen rather quickly. Conversely, colloids are large enough and slow enough to be observed by conventional optical techniques such as light microscopes, or to be modelled in numerical simulations avoiding the inclusion of quantum effects, which typically slows down simulations by order of magnitudes.

Another aspect that makes the study of colloidal self-assembly relevant, is the possibility of creating new functional materials with optical properties out of the ordinary [7, 8]. Among the best known examples are photonic crystals, ordered and periodic colloidal structures in which the lattice spacing is comparable to the wavelengths of visible light. These materials can create photonic band gaps, in complete analogy with the electronic band gaps formed by semiconductor crystals, which prevent the propagation of light with a specific wavelength in certain directions [9,10]. Nature offers us numerous examples of similar structures, such as gem opals with their wonderful colours, caused by their regular arrangements of silica spheres on a colloidal scale [11, 12]. Other examples come from the plant and animal world, where we find examples of photonic crystals in marble berries or in the wings of *Morpho* butterflies.

Before moving on to the next sections, it is interesting – or simply *beautiful* – to note that the study of colloidal systems is an effort that involves researchers and scientists from very different backgrounds. It is a field where collaborations between physicists, chemists, biologists, mathematicians and engineers are frequent, bringing theoretical analysis, numerical simulations, and experimental techniques together, in order to achieve an ever increasing understanding of these systems.

1.2 Nucleation

Among the many self-assembly processes displayed by colloidal systems, one of the most common and certainly one of paramount importance is nucleation. Nucleation is the process through which a first-order phase transition happens, namely with the birth of a microscopic *nucleus* of the embryo phase inside the parent phase, which in some conditions is able to grow out and become macroscopic [13]. When we talk about first order, we refer to the Ehrenfest classification, for which a first-order phase transition is one where the first derivative of the free energy with respect to a variable is discontinuous. In the case of the vapour-liquid phase transition, as well as the vapour-solid or liquid-solid transitions and many other, the density, which is the inverse of the derivative of the Gibbs free energy per particle with respect to pressure and performed at constant number of particles and temperature, shows a discontinuity at the transition.

Temporarily setting aside technical definitions, it is important to note that all of us deal with nucleation phenomena in our everyday life. It is widely known, for instance, that water boils at 100°C and freezes at 0°C. What is less known though, is that, with the appropriate instrumentation, it is possible to keep distilled water in the liquid state even at negative temperatures like -10°C or even -20°C. Eventually, water freezes, and it does so sooner when the temperature is lower. This behaviour is common to all liquids and not only for water and, in general, it also holds for other examples of first-order phase transitions. From this phenomenon we can understand that nucleation is an *activated process*, where a free-energy barrier must be overcome in order to complete the transition. An obvious consequence is that the waiting time for a nucleation event to happen can be orders of magnitude longer than the event itself. This is why nucleation is referred to as a *rare event*. The fact that nucleation is such a common event in nature makes understanding its mechanism of crucial importance in very diverse areas of science. For instance, ice crystals in the clouds are considered to be key in building a better climate model with more accurate radiative effects. Understanding the multiple pathways through which ice nucleates would therefore considerably help us fighting the current climate crisis. Taking an example from the medical field, many diseases for which we have no cure are initiated by a nucleation phenomenon. One example is sickle cell anemia, which is caused by nucleation of a hemoglobine mutant in the blood. Finally, concluding with an example in the pharmaceutical field, it is a phase transition between a metastable phase and another stable

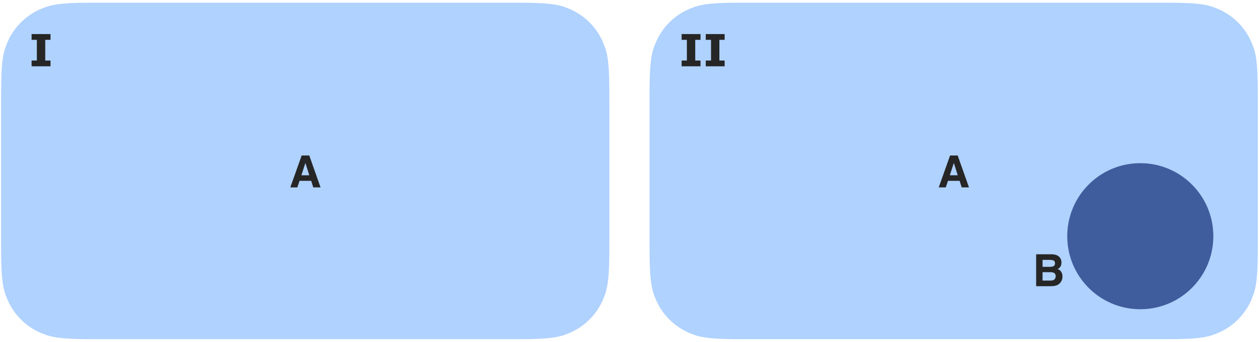


Figure 1.1: On the left, system I is shown, where only the metastable phase A (light blue) is present. Conversely, on the right, system II is shown, where a cluster of the stable child phase B (dark blue) has formed within the metastable parent phase A.

phase that can cause a change in the lattice structure of a specific drug, making it ineffective. This is what happened in 1998 with a AIDS-treating drug called Ritonavir, which was subsequently removed from the market.

The history of nucleation is a few centuries old, and starts with the first experiments of Fahrenheit on crystallisation of water in 1714 [14] – just like the above example – and with analogous experiments on different liquids performed by Gay Lussac [15]. In 1878, Josiah Willard Gibbs made a fundamental contribution to the understanding we still have today of the nucleation process [16]. In fact he was the first to interpret the experiments mentioned above in terms of new concepts such as stability, metastability and instability of phases. He understood, for example, that the system, in the metastable phase, had to do a positive work to form a droplet of the child phase within the parent phase. It is standing on the shoulders of giants then, that Volmer and Weber in 1926 formulated the theoretical framework within which we still interpret the phenomena of nucleation: Classical Nucleation Theory (CNT) [17].

In the following subsections we proceed with a detailed mathematical derivation of all the relevant equations of CNT, as well as the basic assumptions it involves.

1.2.1 Classical Nucleation Theory - Thermodynamics

Let us start by considering a system in the metastable phase A, and a second system where a small cluster of the stable phase B has formed inside the parent phase A. As an example, we can think of phase A as liquid water and phase B as ice, provided that the ice is the stable phase at these thermodynamic conditions (for instance, with an external pressure of 1 atm and a temperature below 0°C). A sketch of these two systems is represented in Fig. 1.1. The goal of this derivation is to find the reversible work performed by system I in order to form the cluster of the stable phase B in system II.

First of all, we compute the internal energy of system I which is given by

$$U^I_A = T^I S^I - P^I_A V + \mu^I_A N \tag{1.1}$$

where T^I, S^I, P^I_A, V, μ^I_A, and N are, respectively, the temperature, entropy, pressure, volume, chemical potential, and number of particles of system I. Conversely, for system II we find

$$U^{II}_{A+B} = T^{II} S^{II}_A + T^{II} S_B - P^{II}_A V^{II}_A - P_B V_B + \mu^{II}_A N^{II}_A + \mu_B N_B + \mathbf{A}\gamma \tag{1.2}$$

where $\mathbf{A}$ is the area of the interface between phase A and phase B, and γ is the free-energy cost per unit of area associated with the formation of this interface. We now assume that the temperature and the pressure are constant during the transformation, namely that $T^I = T^{II} = T$, and that $P_A^I = P_A^{II} = P$. Additionally, we also assume that $N = N_A^{II} + N_B$ and $V = V_A^{II} + V_B$. Note that these two conditions on N and V are only valid if we assume that the interface between phases A and B is sharp, contains no particles and has no volume. The latter statement represents a crucial assumption of CNT and is known as Gibbs *capillarity approximation*. With this further assumptions we can then write

$$U_{A+B}^{II} = TS_A^{II} + TS_B - PV + (P - P_B)V_B + \mu_A^{II}N + (\mu_B - \mu_A^{II})N_B + \mathbf{A}\gamma \tag{1.3}$$

Considering that $N_A \gg N_B$, we also have that $\mu_A^{II} = \mu_A^I = \mu_A$. Operating in the NPT ensemble, the reversible work to form a nucleus of phase B within phase A is expressed in terms of the Gibbs free-energy difference between the two systems, namely

$$\Delta G = G_{A+B}^{II} - G_A^I = (P - P_B)V_B + (\mu_B - \mu_A)N_B + \mathbf{A}\gamma, \tag{1.4}$$

where $\mu_B = \mu_B(P_B)$ and $\mu_A = \mu_A(P)$.

In order to derive the final expression for ΔG, there are a few crucial assumptions to make, which we list here:

- the interfacial free energy γ does not depend on the curvature of the cluster and is instead equal to its value for a flat interface, $\gamma = \gamma_\infty$;
- the properties of the cluster are the same as the bulk phase B;
- the cluster is incompressible, so its density does not change with pressure.

In particular, as a consequence of the incompressibility of the growing cluster, we can re-write the Gibbs-Duhem equation and obtain

$$\mu_B(P_B) = \mu_B(P) + \frac{P_B - P}{\rho_B}. \tag{1.5}$$

If we insert Eq. 1.5 in Eq. 1.4, we finally obtain the reversible work needed to form a cluster of N_B particles of the new phase

$$\Delta G(N_B) = \mathbf{A}(N_B)\gamma - |\Delta\mu|N_B \tag{1.6}$$

where $|\Delta\mu(P)| = |\mu_B(P) - \mu_A(P)|$ is the driving force for nucleation. This equation can be re-written, assuming a spherical shape of the growing cluster. With this assumption, ΔG becomes a function of the radius R of the cluster itself

$$\Delta G(R) = 4\pi R^2\gamma - |\Delta\mu|\frac{4}{3}\pi\rho_B R^3 \tag{1.7}$$

Eq. 1.7 is likely to be the most common and known equation from CNT, and expresses the change in the free energy as a competition between two terms:

1. the first term is the *surface term*, which is the free-energy cost to pay due to the formation of an interface between the two phases. Being always positive, this term acts against nucleation;

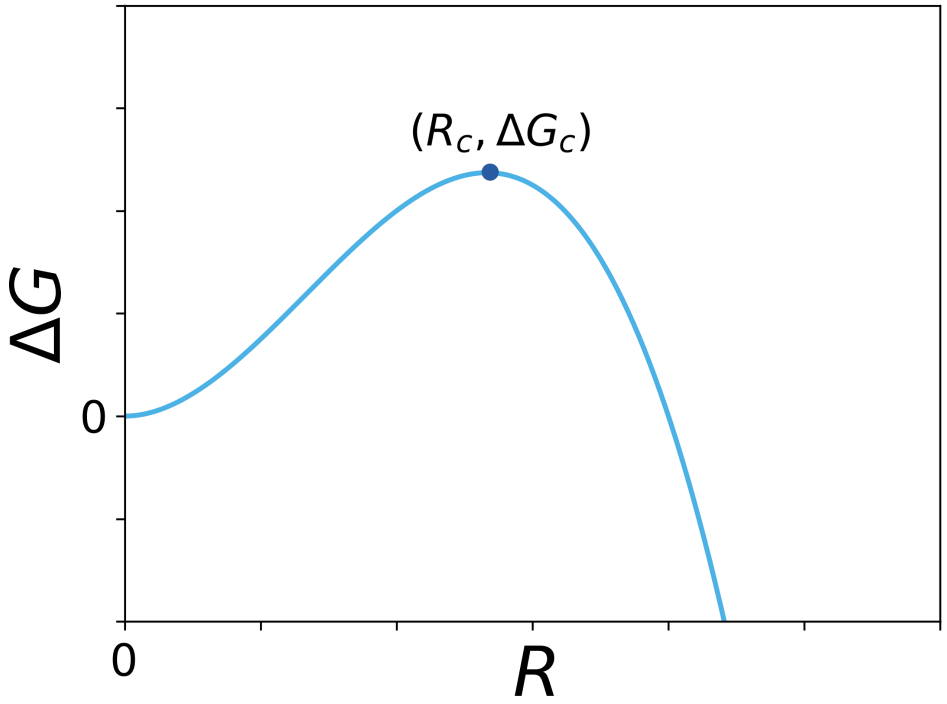


Figure 1.2: Typical profile of the Gibbs free-energy barrier ΔG as a function of the radius R of the nucleus. The dark blue point denotes the critical size of the nucleus (R_c) and the height of the barrier (ΔG_c)

2. the second term is the *bulk term*, which, under the conditions where the parent phase is metastable with respect to the child phase, as we assumed at the beginning of the derivation, represents the free-energy gain due to the formation of a cluster of the new phase. This term is always negative, and therefore is also referred to as the driving force for nucleation.

The competition between these two terms gives birth to a free-energy barrier, with a typical profile as sketched in Fig. 1.2. In order for a nucleation event to happen, a cluster of at least the critical size has to form by means of spontaneous thermal fluctuations. These fluctuations temporarily bring the system in a state of higher free energy, and are therefore rare. However, once a nucleus of at least the critical size has formed, the system is able to lower its free energy by making the cluster grow indefinitely. It is important to note that, when the supersaturation is low, the height of the barrier is high, and nucleation becomes extremely difficult to witness within reasonable time scales. Differently, for higher supersaturation levels, the height becomes lower and lower, until it reaches approximately zero at the *spinodal* point.

We close this subsection noting that from Eq. 1.7 it is straightforward to derive the critical radius of the nucleus R_c which reads

$$R_c = \frac{2\gamma}{\rho_B |\Delta\mu|} \tag{1.8}$$

and the height of the Gibbs free energy ΔG_c reads

$$\Delta G_c = \frac{16\pi}{3} \frac{\gamma^3}{(\rho_B |\Delta\mu|)^2} \tag{1.9}$$

The height of the Gibbs free energy plays a crucial role, but is not the only variable to take into account when making predictions on how frequent a nucleation event should be. In the next subsection, we will see why.

1.2.2 Classical Nucleation Theory - Kinetics

In addition to providing an expression for the free-energy barrier for nucleation, CNT also offers a description of the kinetics of the process. In fact, in this framework, we assume that the cluster of the new phase B grows (shrinks) through the attachment (detachment) of one particle at a time. We can then write

$$\begin{aligned} B_{n-1} + B_1 &\underset{k_{-,n}}{\overset{k_{+,n-1}}{\rightleftharpoons}} B_n \\ B_n + B_1 &\underset{k_{-,n+1}}{\overset{k_{+,n}}{\rightleftharpoons}} B_{n+1} \end{aligned} \tag{1.10}$$

where B_n is a cluster with n particles (and therefore B_1 is a monomer), while $k_{+,n}$ ($k_{-,n}$) is the attachment (detachment) rate of a monomer to (from) a cluster with n particles.

In general, it is possible to find the time-dependent distribution for a cluster with n particles $N_n(t)$ by solving the Master equation [18] of the underlying Markov process

$$\frac{dN_n(t)}{dt} = N_{n-1}(t)k_{+,n-1} - N_n(t)k_{-,n} - N_n(t)k_{+,n} + N_{n+1}(t)k_{-,n+1} \tag{1.11}$$

On the other hand, the net nucleation rate for a cluster of n particles is given by

$$J_{n,t} = N_n(t)k_{+,n} - N_{n+1}(t)k_{-,n+1} \tag{1.12}$$

which corresponds to the net flux of clusters with the same size n. If we assume that the system is in a steady state [19] with constant cluster size distributions and nucleation rate, we can write

$$J = N_n^s k_{+,n} - N_{n+1}^s k_{-,n+1} \tag{1.13}$$

This equation can be solved by recurrence, as in Ref. 20

$$J = N_1^s \left[\sum_{n=1}^{\infty} \frac{1}{k_{+,n}\xi_n} \right]^{-1} \tag{1.14}$$

where

$$\xi_n = \prod_{i=1}^{n-1} \frac{k_{+,i}}{k_{-,i+1}} \tag{1.15}$$

which only holds for $n > 1$. Here, we make a crucial assumption, namely that for clusters which are smaller than the critical size, because the small steady state flux, the cluster distribution is the *equilibrium* one, assuming therefore a kind of quasi-equilibrium in the system. We further assume that clusters of size n are in equilibrium with respect to monomers, and since the interaction between n-sized clusters and monomers is the only one we are considering, we find

$$nN_1 \overset{K}{\rightleftharpoons} N_n \tag{1.16}$$

where N_1 and N_n are respectively the equilibrium distributions of clusters of size 1 and n. Conversely, K is the equilibrium constant. With this assumption, the term on the right of Eq.

1.15 is exactly K, which is also the equilibrium probability of forming a cluster of size n. In mathematical terms, we have

$$\xi_n = K = e^{-\beta \Delta G(n)} \tag{1.17}$$

We can then re-write Eq. 1.14 as

$$J = N_1^s \left[\sum_{n=1}^{\infty} \frac{1}{k_{+,n} e^{-\beta \Delta G(n)}} \right]^{-1} \tag{1.18}$$

To finally calculate J, there are a few more approximations to do:

1. the sum is dominated by the terms corresponding to the top of the barrier;
2. we can expand $\Delta G(n)$ on top of the barrier up to the quadratic term of its Taylor expansion, *i.e.* $\Delta G(n) = \Delta G(n_c) + \frac{1}{2}\Delta G''(n_c)(n - n_c)^2$;
3. we can replace $k_{+,n}$ with k_{+,n_c};
4. the sum is replaced by an integral with respect to a new variable $n^* = n - n_c$ which goes from $-\infty$ to ∞;

We are now able to write the final expression for the nucleation rate, which reads

$$J = N_1 k_{+,n_c} Z e^{-\beta \Delta G(n_c)} \tag{1.19}$$

where Z is known as the Zeldovitch factor [21] and is equal to

$$Z = \left(\frac{|\Delta G''(n_c)|}{2\pi k_B T} \right)^{1/2} = \left(\frac{|\Delta \mu|}{6\pi k_B T n_c} \right)^{1/2} \tag{1.20}$$

Note that the last equality follows from the assumption that the nuclei are spherical, and therefore does not hold in all cases.

1.3 Inhibiting and promoting nucleation

In the previous sections we have described the phenomenon of nucleation and especially outlined how important it is to have a full understanding of it. We have also already mentioned a recurring concept, in the context of nucleation, namely that it is a so-called rare event. It is not surprising, therefore, that over the past decades much attention has been paid to what the inhibitory factors of nucleation are, and at the same time to how it is possible to find optimal conditions for it to occur. In the current Section, we will elucidate the main aspects of both topics, whose histories date back to very different periods of scientific research, but which certainly find common ground in the nucleation of colloidal systems.

1.3.1 Fivefold symmetry

In 1952, Sir Charles Frank published a seminal paper that contributed significantly to the understanding of the behaviour of many-body systems [22]. In this paper, the author starts considering that if one had the task of arranging 12 spheres around a thirteenth sphere, the most efficient packing would be represented by the scenario in which these 12 spheres are positioned at the vertices of a regular icosahedron whose centre coincides with the central particle. This reasoning, despite being purely geometric and apparently harmless, calls for two key considerations. The first is that the regular icosahedron has nothing to do with the typical local environment of a hard sphere in a face-centred-cubic (fcc) crystal, which is the thermodynamically stable structure when dealing with a system of hard spheres at sufficiently high pressures (or densities). The second consideration, possibly even more disruptive, is that the regular icosahedron does not get along with the concept of periodicity, and hence of a crystal lattice [23].

A crystalline structure is in fact by definition periodic, and therefore must necessarily be characterised by precise symmetries. The characteristic of these *selected* symmetries is that the corresponding spatial patterns have the property of being able to repeat themselves in space without creating gaps. As well as, in two dimensions, it is possible to realise a tessellation of a flat surface using squares or hexagons, but not with pentagons, in the same way in nature there are crystals which display fourfold or sixfold symmetries, but not fivefold symmetry. In other words, fivefold symmetry is incommensurate with Euclidean space. For this reason, the regular icosahedron, which is the most iconic representative of the fivefold symmetry clusters (from the vertices of a regular icosahedron it is possible to obtain a total of 12 pentagons!), cannot give rise to periodic, and therefore crystalline, structures.

In a supersaturated fluid, due to the thermal noise that allows the system to explore phase space, regular or defective icosahedra form continuously in the system. However, in order to initiate the nucleation process and thus form a periodic structure, it is necessary for these local clusters to break up and rearrange into a new spatial pattern. All of this obviously comes at a cost, which is paid in terms of free energy. From this perspective, colloidal nucleation can be seen as a competition between different symmetries [24–26].

The consequences of Frank's studies have been enormous. The propensity of some systems to form fivefold symmetry clusters in the disordered phase has been studied closely, as have the relationships between this behaviour and the nucleation of the same systems [27–29].

At the same time, local environment classification algorithms have been developed that are increasingly attentive to these structures. An example is provided by the Common Neighbour Analysis (CNA) which is able to recognise an icosahedral arrangement around a particle [30]. However, the most decisive contribution in this class of algorithms is certainly represented by the Topological Cluster Classification (TCC) [31]. Unlike other algorithms (such as CNA) TCC does not aim to classify each particle in a system based on its local environment, instead it detects a whole zoo of topological clusters without relying on a single reference particle. Among the clusters recognised by TCC, many have five-membered rings, and for this reason they can be called fivefold symmetry clusters (see Fig. 1.3). The possibility of following the behaviour of these clusters during simulations or experiments where nucleation phenomena are present, has contributed significantly to our understanding of the nucleation mechanism itself, in the light of that competition between the above-mentioned symmetries.

An exemplary case of how the classification of fivefold symmetry clusters has allowed to attribute to these clusters the role of nucleation inhibitors is represented by Ref. 29. In this work, Taffs and Royall simulated hard-sphere systems with a bias that favoured or penalised the formation of a specific fivefold symmetry cluster, namely the pentagonal bipyramid. It was observed that by favouring the formation of this cluster, the nucleation rate of the system was significantly lowered, making nucleation an increasingly rarer event. On the contrary, penalising the formation of the same cluster and therefore lowering its concentration, the nucleation was found to be more frequent compared to the unbiased system.

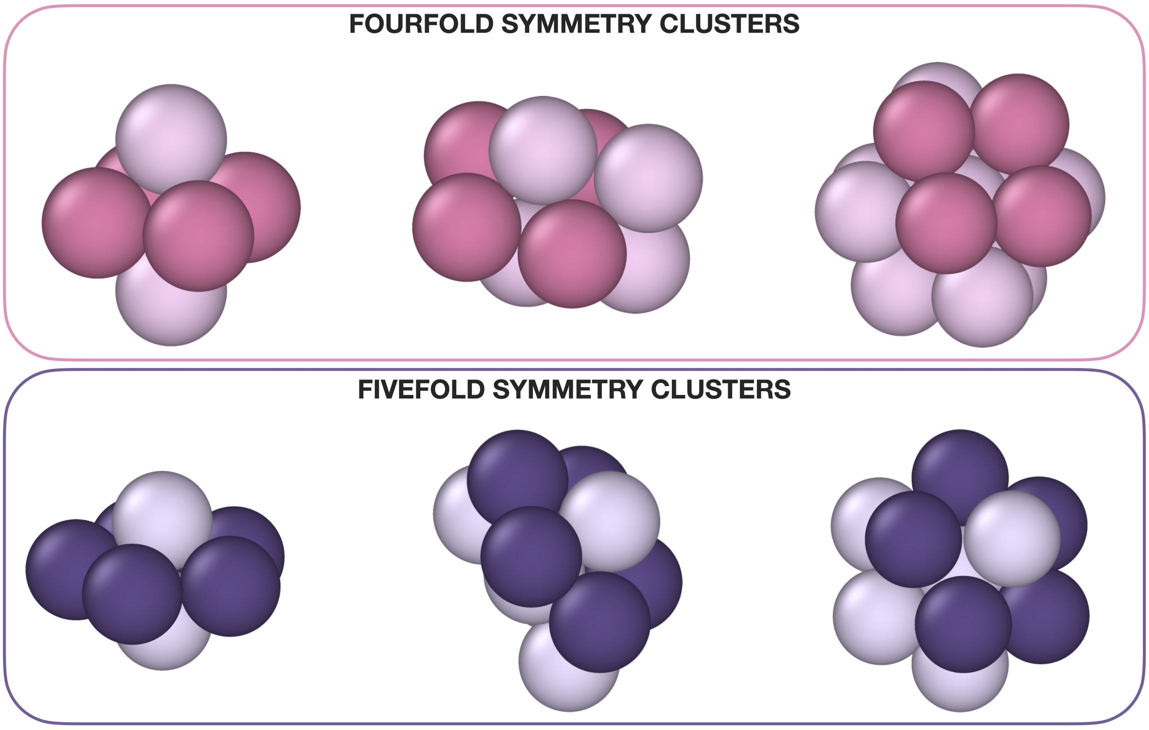


Figure 1.3: Examples of clusters detected by the TCC. In the upper panel, we show three fourfold symmetry clusters, where we highlight particles arranged in a four-membered ring pattern. In the lower panel, we show examples of fivefold symmetry clusters, where we highlight particles arranged in a five-membered ring pattern.

1.3.2 Inverse Design

Unlike what we have just discussed, where the focus was on the nucleation process, nucleation is often not interesting in itself as a mechanism, but on the contrary one is more interested in obtaining its products. As we have already discussed in Section 1.1, colloidal self-assembly – the spontaneous organisation of matter into ordered arrangements – has been considered key to producing the next generation of materials. It is in this context of research that *inverse design* methods (IDMs), in which the optimal thermodynamic conditions and interaction potentials are determined for a structure to form, were born.

Since its formulation, statistical mechanics has gained enormous success due to its ability to predict the properties of a material from the knowledge of the interactions between the constituent particles. In general, the assumption that the characteristics of a material are intimately related to the properties of its elementary constituents has always guided materials discovery in soft matter, through the process we call *forward design*: shape and interaction potential of the system are pre-determined (or in other words, the starting building blocks are characterised), and then the properties which result from this choice are investigated [32, 33].

Nowadays, thanks to a continuous and exponential improvement of synthesis techniques, the range of available building blocks to build new functional materials is so wide to be difficult to process [34–38]. In other words, with the enlargement of the pool of possible building blocks, also the time and resources required for a systematic investigation of the parameter space have grown enormously. An approach that is therefore more natural and effective is based on a different and opposite logic to that just described. IDMs have their starting point in the

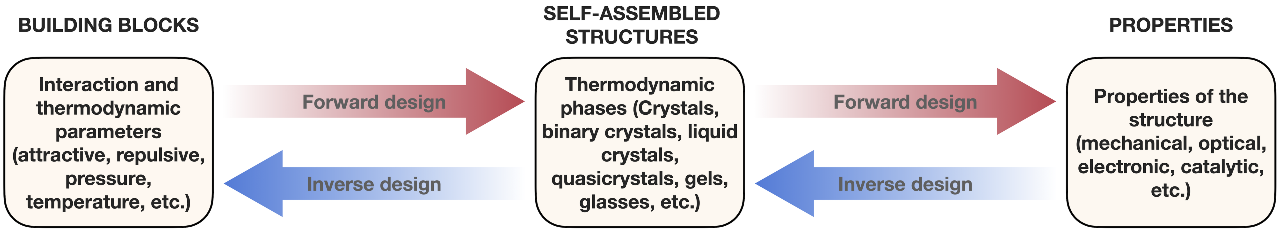

BUILDING BLOCKS
Interaction and thermodynamic parameters (attractive, repulsive, pressure, temperature, etc.)
Forward design
Inverse design
SELF-ASSEMBLED STRUCTURES
Thermodynamic phases (Crystals, binary crystals, liquid crystals, quasicrystals, gels, glasses, etc.)
Forward design
Inverse design
PROPERTIES
Properties of the structure (mechanical, optical, electronic, catalytic, etc.)

Figure 1.4: Sketch of the opposite logics that characterise the forward design methods from the inverse design methods. In the first case, we start from the building blocks, use free-energy calculations to find the stable structures that are obtained from the selected building blocks, and derive the physical properties that derive from these structures. On the other hand, within the inverse design framework, the starting point is the target properties that you want to obtain from your system. The optimal interaction and thermodynamic parameters of the system in order to reach the target properties are obtained through parameter optimisation techniques.

desired properties of the final material and, through a process of parameters optimisation, the characteristics of the constituent building blocks which permit the self-assembly of the target phase are obtained.

In the past few years, many different IDMs have emerged with the goal of finding the optimal interaction and thermodynamic parameters, so that the building blocks can spontaneously self-assemble into the target structure [39–42]. Among the cases we can find in literature, IDMs have been successfully employed for finding regions of stability of crystal structures with exceptional photonic properties [43], for predicting crystal and protein structures [44–46], and for tuning the mechanical and transport properties of materials [47].

In order to set up an IDM to reverse engineer different phases, from crystals to liquid crystals, and even quasicrystals, two separate ingredients are needed. First of all, it is crucial to define an order parameter which is able to distinguish the phase we want to target from the competing ones. Specifically, this order parameter will be translated into a fitness function which indicates how “close" our system is with respect to the desired phase. Secondly, one has to devise a mathematical scheme to update the design parameters based on the chosen fitness function.

The latter requirement can be easily satisfied by choosing among several techniques, either borrowed from classical optimisation algorithms (like Particle Swarm Optimisation or Covariance Matrix Adaption Evolutionary Strategy) or inspired by statistical physics [48–51].

Conversely, the choice of an effective fitness function represents the real bottleneck for any IDM to succeed. In the last decade, a plethora of order parameters has been used to define fitness functions for all kinds of phases. For instance, free-energy or chemical-potential differences with respect to the competing structures have been employed to reverse engineer 3D crystal lattices starting from (non)spherical colloids [52,53]. Often, full knowledge of the target crystal has been translated into a fitness function by computing the mean square displacements of the particles with respect to their target lattice points [39], or through the radial distribution function [54–56]. The sometimes unrealistic resulting potentials have been explicitly filtered by Adorf *et al.* in order to obtain smooth and short-range interactions [57]. All these fitness functions have different advantages and disadvantages and are usually tailored on specific classes of materials. The search of a general fitness function which makes the IDM more robust and versatile is still on the go.

We stress here that we have just discussed a general overview of IDMs, touching only those points which are relevant for the understanding of this thesis. The topic is much wider though, and we invite the interested reader to consult Ref. 42 and 58 for a more comprehensive perspective of the role of inverse methods for materials design.

1.4 Scope of the thesis

Nucleation in soft matter systems is an extremely broad topic, with many open, challenging, and fascinating questions. With the current thesis, we aim to address some of these questions, trying to understand its underlying mechanisms for several systems, and particularly paying attention to the inhibitory factors, as well as the optimal conditions for nucleation to happen.

In Chapter 2, we study the nucleation of the Laves Phases (LPs) from a binary mixture of nearly hard spheres. We focus on the role of the softness in the interparticle potential, and its influence on the high-order spatial correlations of the particles, analysing several fivefold symmetry clusters. We then study the effect these clusters have on the dynamics and, ultimately, on the interplay between crystallisation and the glass transition.

In Chapter 3 we focus on another binary crystal, and certainly the most bizarre one, namely the AB_{13} crystal. Due to the icosahedral arrangement of the small spheres, a satisfying distinction between these particles and the fluid phase is extremely challenging. We resort to an artificial neural network, which reaches unprecedented classification accuracy, and allows for an extraordinary spatial resolution.

In Chapter 4 we turn our attention on hard-sphere nucleation, and on the polymorph selection mechanism which leads to a predominance of face-centred-cubic-like particles, with respect to hexagonal-closed-packed-like ones. We use two different classification schemes, one that is sensitive to symmetries shown by the local environment of each particle, and another which detects topological clusters. The combination of the two techniques reveals the geometric mechanism which shows how fivefold symmetry clusters break up and attach to the crystal phase during nucleation. The specific rearrangement of the particles ultimately sheds light on the polymorph selection.

In Chapter 5 we keep our attention on the same system, and define an unsupervised learning procedure to better classify the different polymorphs appearing during hard-sphere nucleation. This is done including a huge amount of information about each particle's local environment, and using this information to find new interpretable order parameters.

In Chapter 6 we turn to IDMs, in order to study the optimal conditions for a body-centred-crystal to form, using repulsive Yukawa particles. We achieve the goal in two separate ways, testing two algorithms of different nature – the first based on the statistical fluctuations of the system and the second derived from classical optimisation techniques.

Finally, in Chapter 7, building on the knowledge we acquire in Chapter 6, we develop a new IDM, based on the use of a convolutional neural network (CNN) as an order parameter. This IDM not only manages to reverse-engineer a large number of phases, including phases notoriously difficult to classify, as for instance quasicrystals, but is equally applicable to two-dimensional and three-dimensional cases, and enables us to find a quasicrystal not yet reported in the literature for a given model.

2

Tuning the glass transition: Enhanced crystallisation of Laves Phases in nearly hard spheres

In this Chapter we study the nucleation process of the colloidal Laves phase by means of numerical simulations. Although Laves phases have been proven to be stable in a binary hard-sphere system, they have never been observed to spontaneously crystallise in a binary fluid in simulations nor in experiments of micron-sized hard spheres due to slow dynamics. Here we demonstrate, using computer simulations, that softness in the interparticle potential suppresses the degree of fivefold symmetry in the binary fluid phase and enhances crystallisation of Laves phases in nearly hard spheres.

Based on: T. Dasgupta / G. M. Coli, and M. Dijkstra, *Tuning the glass transition: Enhanced crystallisation of Laves Phases in nearly hard spheres*, ACS Nano, **14**, 3957-3968 (2020)

2.1 Introduction

Photonic crystals (PCs) are periodic dielectric structures that possess a photonic bandgap that forbids the propagation of light at certain frequency ranges. The ability to control the flow of light is attractive for numerous applications, ranging from lossless dielectric mirrors, bending of light around sharp corners in optical waveguides, telecommunications, to optical transistors in optical computers.

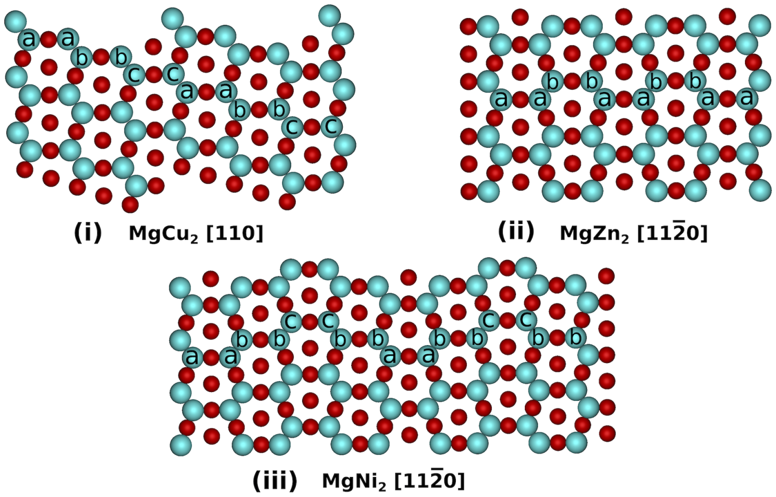
A highly promising route to fabricate photonic crystals is *via* self-assembly of optical wavelength sized colloidal building blocks. PCs that display a wide omnidirectional photonic bandgap at low refractive index contrasts are related to the family of either the diamond or the pyrochlore structure. However, these low-coordinated crystals are notoriously difficult to self-assemble from colloids with simple isotropic pair interactions.

One strategy to form open lattices is by employing long-range Coulomb interactions with a range that exceeds multiple times the particle size [59, 60]. The range of the screened Coulomb interaction is set by the Debye screening length of the solvent, like water or other polar solvents, which is why this approach will fail for particle sizes that are required for opening up a photonic bandgap in the visible region.

To circumvent these problems associated with the self-assembly of low-coordinated crystal structures, one can also employ a different route in which both the diamond and pyrochlore structure are self-assembled in a single close-packed $MgCu_2$ crystal structure from a binary colloidal dispersion. By selectively removing one of the species, one can obtain either the diamond (Mg, large spheres) or the pyrochlore (Cu, small spheres) structure. $MgCu_2$ is one of the three binary LS_2 crystal structures (L = large species, S = small species), also known as Laves phases (LPs), as first found in intermetallic compounds. The three structural prototypes of the LPs are the hexagonal $MgZn_2$, cubic $MgCu_2$ and hexagonal $MgNi_2$ structures, which can be distinguished by the stacking of the large-sphere dimers in the crystal structures (Fig. 2.1). Experimentally, LPs have been observed in binary nanoparticle suspensions [61, 62], and in submicron-sized spheres interacting *via* soft repulsive potentials [63–68].

Although free-energy calculations in Monte Carlo (MC) simulations have demonstrated that the LPs are thermodynamically stable for a binary hard-sphere (BHS) mixture with a diameter ratio of $0.76 \leq q = \sigma_S/\sigma_L \leq 0.84$ [9], LPs have never been observed to spontaneously crystallise *via* nucleation in such a binary fluid mixture in computer simulations, but only through spinodal decomposition, in conditions where the fluid phase is unstable with respect to the competing crystal structure [69]. There are numerous possible reasons. First of all, the freezing transition of the LPs in a BHS fluid is located at very high densities. Nucleation can thus only occur when the system is sufficiently dense. At these high concentrations, nucleation is severely hampered by slow dynamics. Binary mixtures with a diameter ratio of $q \sim 0.8$, identical to the range where the LPs are stable, are known to be excellent glass formers [70]. This, in conjunction with the above factor makes the self-assembly of LPs in BHS mixtures an extremely rare event. Furthermore, due to small free-energy differences, the three LPs are strongly competing during the crystallisation process, leading to numerous stacking faults in the final crystal structure [71–74].

The suppression of crystallisation due to glassy behaviour is often rationalised by the prevalence of icosahedral clusters of spheres whose short-range fivefold symmetry is incompatible with the long-range translational order as exhibited by crystals [22]. The icosahedral order arises when one maximises the density of a packing of 12 identical spheres in contact with a central sphere of the same size. The densest packing is obtained by arranging the outer spheres


(i) $MgCu_2$ [110]

(ii) $MgZn_2$ [$11\bar{2}0$]

(iii) $MgNi_2$ [$11\bar{2}0$]

Figure 2.1: Structure of the three types of Laves phases, showing the different stacking sequences of the large-sphere dimers, marked as "aa", "bb" and "cc", when viewed along specific projection planes. The stacking of the large-sphere dimers is (i) "...aa-bb-cc..." for $MgCu_2$, (ii) "...aa-bb..." for $MgZn_2$, and (iii) "...aa-bb-cc-bb..." for $MgNi_2$.

on the vertices of an icosahedron, rather than by using 13-sphere subunits of face-centred cubic and hexagonal close-packed bulk crystals.

Here we demonstrate that spontaneous crystallisation of the LPs is strongly suppressed by the presence of fivefold symmetry structures in a binary fluid of hard spheres. Interestingly, we show that softness of the interaction potential reduces the degree of fivefold symmetry in the binary fluid phase. We systematically study the role of softness in the interaction potential on the structure, phase behaviour, and nucleation of the LPs. By carefully tuning the particle softness, we observe for the first time spontaneous nucleation of the LPs in a nearly hard-sphere system in computer simulations, thereby providing evidence that the LPs are stable in a binary hard-sphere system. The key result of this study is that soft repulsive spheres can be mapped onto a hard-sphere system in such a way that the structure and thermodynamics are invariant, but that the dynamics and therefore the kinetic glass transition are strongly affected by higher-body correlations, *i.e.* fivefold symmetry clusters, which can be tuned both in simulations and in experiments by the softness of the particle interactions. In this way, softness suppresses fivefold symmetry and enhances crystallisation of the LPs.

2.2 Freezing transition and fivefold symmetry

2.2.1 The model

We first study the effect of particle softness on the freezing transition of the LPs in a binary fluid of soft repulsive spheres. To vary the softness of the pair interaction, we consider a binary mixture of N_L large (L) and N_S small (S) spheres in a volume V interacting with a Weeks-Chandler-Andersen (WCA) potential $u_{\alpha\beta}(r_{ij})$ between species $\alpha = L, S$ and $\beta = L, S$ [75]

$$u_{\alpha\beta}\left(r_{ij}\right) = \begin{cases} 4\epsilon\left[\left(\frac{\sigma_{\alpha\beta}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{\alpha\beta}}{r_{ij}}\right)^{6} + \frac{1}{4}\right], & r_{ij} < 2^{\frac{1}{6}}\sigma_{\alpha\beta} \\ 0, & r_{ij} \geq 2^{\frac{1}{6}}\sigma_{\alpha\beta} \end{cases} \tag{2.1}$$

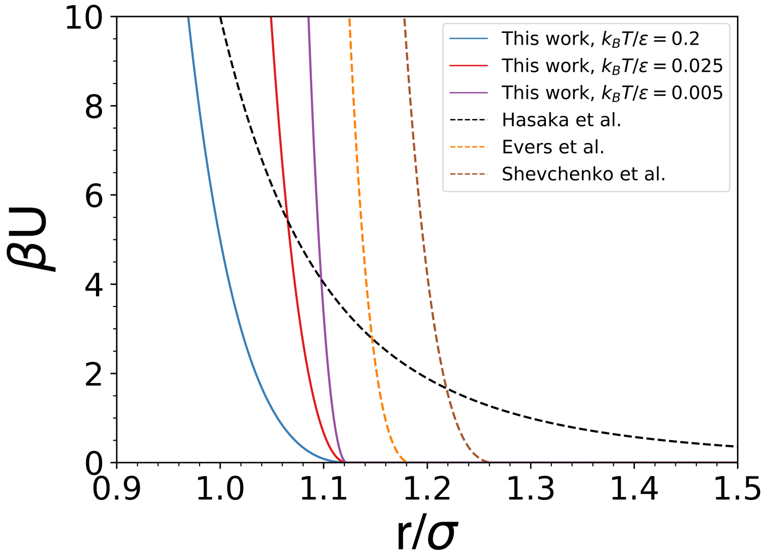


Figure 2.2: The WCA potential at $T^* = k_\mathrm{B}T/\epsilon = 0.2$, 0.025 and 0.005 along with the pair potentials of the experimental systems (dashed lines) for which the Laves phase has been reported in literature, which are binary nanoparticle suspensions of Evers *et al.* [62] and Shevchenko *et al.* [61], and polystyrene spheres of Hasaka *et al.* [64].

where $r_{ij} = |r_i - r_j|$ denotes the centre-of-mass distance between particle i and j, r_i the position of particle i, and ϵ the interaction strength. For a BHS mixture, LPs are thermodynamically stable for a diameter ratio $q \in [0.76, 0.84]$ [9]. In this Chapter, we set the diameter ratio $q = \sigma_S/\sigma_L = 0.78$, which is close to the value of a recent experimental study [76]. Here σ_α denotes the diameter of species α and $\sigma_{\alpha\beta} = (\sigma_\alpha + \sigma_\beta)/2$. The *softness* of the potential can be tuned by changing the reduced temperature $T^* = k_BT/\epsilon$ with k_B Boltzmann's constant and T the temperature. The WCA potential has been previously used to mimic the interactions between hard spheres as it reduces to the hard-sphere potential in the limit of $T^* \to 0$ [77–80]. In Fig. 2.2, we plot the WCA potential for $T^* = 0.005$, 0.025 and 0.2 along with the pair potentials of some notable experimental systems for which the Laves phases have been reported in literature, which are polystyrene latex spheres [64], and binary nanoparticle suspensions [61, 62]. Fig. 2.2 shows that the WCA potential at $T^* = 0.025$ agrees well with that of the nanoparticle systems of Evers *et al.* [62] and Shevchenko *et al.* [61], whereas the WCA potential at $T^* = 0.2$ is slightly softer. The pair potential of the polystyrene latex spheres of Hasaka *et al.* [64] is considerably softer and more long ranged than the WCA pair interactions used in the present study.

2.2.2 Free-energy calculations

The equilibrium phase diagram of the WCA mixture is calculated by determining the free energies of the binary fluid at composition $x_L = 1/3$ and the three LPs - MgCu$_2$, MgZn$_2$, and MgNi$_2$. The Helmholtz free energy per particle $f = F/N$ as a function of density ρ for all these phases is calculated using thermodynamic integration of the equations of state

$$\beta f(\rho) = \beta f(\rho_0) + \int_{\rho_0}^{\rho} d\rho' \frac{\beta P(\rho')}{\rho'^{\,2}}, \tag{2.2}$$

where $\rho = N/V$ is the density with N the number of particles and V the volume of the system, $f(\rho_0)$ denotes the Helmholtz free energy per particle for the reference density ρ_0, $\beta = 1/k_BT$ is the inverse temperature, and P is the pressure. Generally, the Helmholtz free energy can be split in two contributions, the one of the ideal gas (βf_{id}), and the excess part (βf_{ex}). The equations of state for the binary fluid and the binary LPs are calculated using MC simulations in the NPT ensemble. Isotropic volume change moves are used for the fluid phase and the cubic $\mathrm{MgCu_2}$ phase while anisotropic volume change moves are used for the hexagonal $\mathrm{MgZn_2}$ and $\mathrm{MgNi_2}$ phases. We use the ideal gas as a reference state for the binary fluid phase. For the LPs, we employ the Frenkel-Ladd method to calculate the Helmholtz free energy at a reference density ρ_0 using MC simulations in the NVT ensemble. In the Frenkel-Ladd method, we start from an Einstein crystal, where the particles are coupled *via* harmonic springs with a dimensionless spring constant λ to the ideal positions of the crystal structure under consideration. We then construct a reversible path from the crystal of interest to the Einstein crystal using the auxiliary potential energy function

$$\beta U_{Ein}(\mathbf{r}^N;\lambda) = \beta U(\mathbf{r}_0^N) + (1-\frac{\lambda}{\lambda_{max}})[\beta U(\mathbf{r}^N) - \beta U(\mathbf{r}_0^N)] + \lambda\sum_{i=1}^{N}\frac{(\mathbf{r}_i-\mathbf{r}_{0,i})^2}{\sigma_L^2}, \tag{2.3}$$

where $U(\mathbf{r}^N) = \sum_{i<j}^{N} u(r_{ij})$ is the potential energy of the system due to the interparticle interactions, $\mathbf{r}_{0,i}$ represents the ideal lattice position of particle i, and λ is the dimensionless spring constant, which ranges from 0 to a value λ_{max}. At λ_{max}, the particles are so strongly tied to their respective lattice sites, that the system reduces to an Einstein crystal of non-interacting particles, whereas $\lambda = 0$ corresponds to the interacting system of interest for which we want to compute the free energy [81, 82]. The Helmholtz free energy of the crystal $\beta f(\rho)$ can be approximated to that of the Einstein crystal using [83]:

$$\beta f(\rho) = \beta f_{Ein}(\lambda_{max}) - \frac{1}{N}\int\limits_{\lambda=0}^{\lambda_{max}} d\lambda \left\langle \frac{\partial \beta U_{Ein}(\mathbf{r}^N;\lambda)}{\partial\lambda}\right\rangle_{\lambda}^{CM} \tag{2.4}$$

where $\langle\cdots\rangle_\lambda^{CM}$ denotes that the ensemble average is sampled for a solid with a fixed centre of mass using the Boltzmann factor $\exp[-\beta U_{Ein}(\mathbf{r}^N;\lambda)]$, and $f_{Ein}(\lambda_{max})$ denotes the free energy per particle of an ideal Einstein crystal given by

$$\beta f_{Ein}(\lambda_{max}) = \frac{\beta U(\mathbf{r}_0^N)}{N} + \frac{3(N-1)}{2N}\ln\left(\frac{\lambda_{max}}{\pi}\right) + \frac{1}{N}\ln\left(\frac{N}{V}\Lambda^3\right) - \frac{3}{2N}\ln(N), \tag{2.5}$$

where Λ is the thermal wavelength. We note that it is convenient to rewrite the integral in Eq. 2.4 as

$$\frac{1}{N}\int\limits_{\ln c}^{\ln(\lambda_{max}+c)} (\lambda+c)\left\langle \sum_{i=1}^{N}\frac{(\mathbf{r}_i-\mathbf{r}_{0,i})^2}{\sigma_L^2} - \frac{1}{\lambda_{max}}[\beta U(\mathbf{r}^N) - \beta U(\mathbf{r}_0^N)]\right\rangle_{\lambda}^{CM} d\left[\ln(\lambda+c)\right], \tag{2.6}$$

with

$$c = \frac{1}{\left\langle \sum_{i=1}^{N}(\mathbf{r}_i-\mathbf{r}_{0,i})^2/\sigma_L^2 - \frac{1}{\lambda_{max}}[\beta U(\mathbf{r}^N) - \beta U(\mathbf{r}_0^N)]\right\rangle_{\lambda=0}^{CM}}. \tag{2.7}$$

The integral in Eq. 2.6 is calculated numerically using a 40 or 60 point Gauss-Legendre quadrature, yielding the LP free energy at the reference density. We calculate the free energy for the three LPs for varying system sizes. In Fig. 2.3 we plot the excess Helmoltz free energy $\beta F_{ex}/N + \ln(N)/N$ as a function of $1/N$ in order to investigate the finite-size scaling. We find that the $MgCu_2$ and $MgNi_2$ LPs are metastable with respect to the $MgZn_2$ LP in the thermodynamic limit, *i.e.* $1/N \to 0$ (see Fig. 2.3 and Table 2.1).

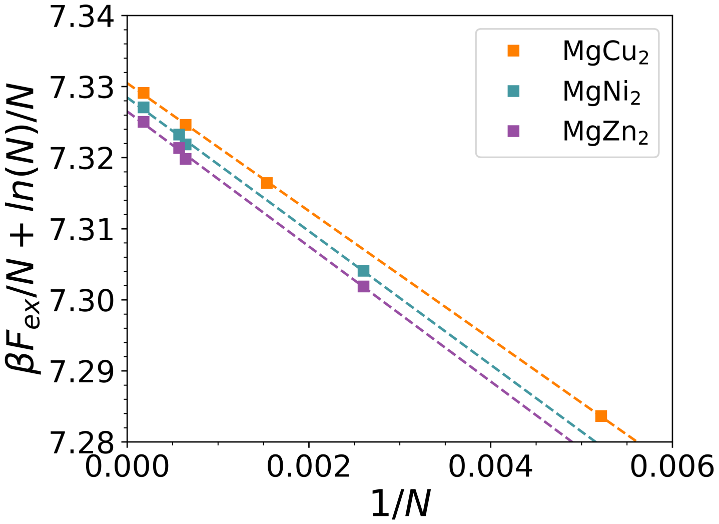


Figure 2.3: Finite-size scaling of the excess Helmholtz free energy $F_{ex}/Nk_BT + \ln(N)/N$ *versus* $1/N$ of the three Laves phases $MgCu_2$, $MgZn_2$, and $MgNi_2$ of a binary mixture of WCA spheres with a size ratio $q = 0.78$, at the melting density $\rho\sigma_L^3 = 1.4668$ and temperature $k_BT/\epsilon = 0.2$. $MgZn_2$ is the Laves phase structure with the lowest free energy. The lines are linear fits to the data points.

Subsequently, by employing a common-tangent construction on the free-energy curves in the $\beta f\rho - \rho$ plane, we obtain the fluid-LP coexistence for different temperatures T^*, which is plotted in Fig. 2.4 in the temperature k_BT/ϵ - reduced density $\rho\sigma_L^3$ plane, where $\rho = (N_L + N_S)/V$ denotes the density. We find that the freezing transition shifts to higher $\rho\sigma_L^3$ with increasing temperature or softness of the particle interaction.

The bulk densities of the fluid-LP coexistence for a BHS mixture with a diameter ratio $q = 0.78$ correspond to packing fractions $\eta_{\mathrm{BHS}}^{(f)} = 0.5356$ and $\eta_{\mathrm{BHS}}^{(LP)} = 0.5943$ for the fluid and LP, respectively. As the freezing transition of the LPs is located at relatively high densities, crystallisation is likely suppressed by slow dynamics.

2.2.3 High-coordination clusters

In the case of monodisperse spheres, glassy dynamics and suppression of crystallisation are often linked to the presence of icosahedral clusters with fivefold symmetry in the supersaturated fluid, which is incompatible with the long-range periodic order of a crystal. To investigate whether or not fivefold symmetry structures suppress crystallisation of the LPs in a binary fluid mixture at composition $x_L = N_L/N = 1/3$, we perform MC simulations with $N = N_L + N_S = 1200$ particles (WCA spheres and BHS) in the NVT ensemble.

We then measure the number fraction of three significant representatives of the fivefold symmetry structures, *i.e.* the pentagonal bipyramids, defective icosahedra, and regular icosahedral clusters as depicted in Fig. 2.5, using the topological cluster classification (TCC) [31] for varying softness of the interparticle potential. The algorithm is used regardless of the species of the

Phase	N	$n_{\text{eq}} \times 10^3$	$n_{\text{prod}} \times 10^3$	$\beta F_{ex}/N$	±
$MgCu_2$	192	100	1000	7.25715	0.00051
$MgCu_2$	648	100	1000	7.30665	0.00022
$MgCu_2$	1536	100	1000	7.31999	0.00006
$MgCu_2$	5184	100	1000	7.32760	0.00006
$MgZn_2$	384	100	1000	7.28680	0.00027
$MgZn_2$	1536	100	1000	7.31516	0.00002
$MgZn_2$	1728	100	1000	7.31722	0.00006
$MgZn_2$	5184	100	1000	7.32348	0.00004
$MgNi_2$	384	100	1000	7.28900	0.00027
$MgNi_2$	1536	100	1000	7.31713	0.00011
$MgNi_2$	1728	100	1000	7.31902	0.00003
$MgNi_2$	5184	100	1000	7.32545	0.00006

Table 2.1: The excess Helmholtz free energies per particle $\beta F_{ex}/N$ for different system sizes for a binary mixture of WCA particles with size ratio $\sigma_S/\sigma_L = 0.78$ at density $\rho\sigma_{\text{av}}^3 = 0.953$, and dimensionless temperature $k_{\text{B}}T/\epsilon = 0.2$. n_{eq} and n_{prod} are the number of Monte Carlo cycles in the equilibration and production (sampling) runs, respectively, and the error estimate is given by the standard deviation of three independent simulations.

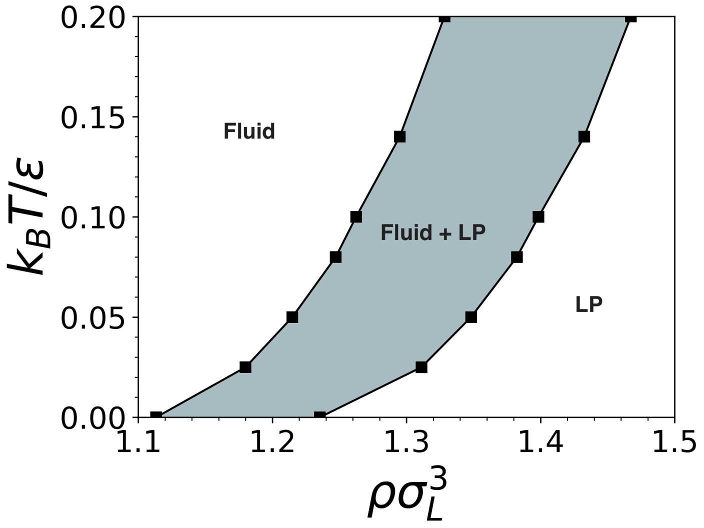


Figure 2.4: Fluid-Laves Phase (LP) coexistence as denoted by the grey region of a binary mixture of WCA spheres with a diameter ratio $q = 0.78$ at a fixed composition $x_L = N_L/(N_L + N_S) = 1/3$ in the reduced temperature k_BT/ϵ - reduced density $\rho\sigma_L^3$ plane. In the limit of $k_BT/\epsilon \to 0$, the system reduces to a binary mixture of hard spheres.

particles forming the clusters and the bonds between particles are detected using a modified Voronoi construction method. The free parameter f_c, controlling the amount of asymmetry that a four-membered ring can show before being identified as two three-membered rings, is set to 0.82. The cluster concentrations are averaged over 100 independent snapshots in each simulation.

The effect of the presence of these clusters on the kinetics and nucleation of monodisperse

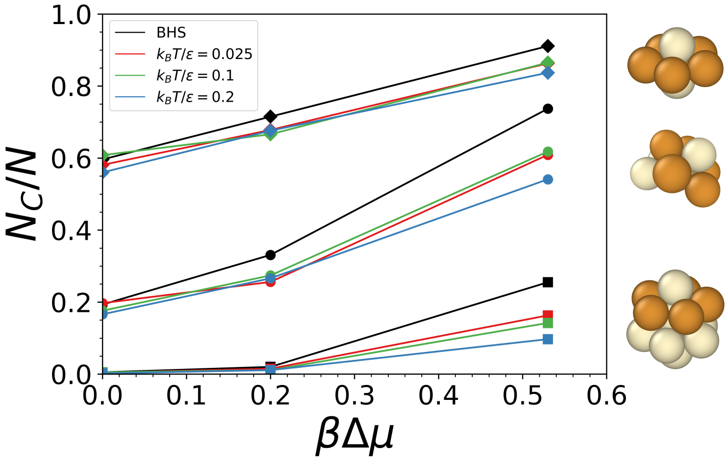


Figure 2.5: Number fraction of particles N_{CL}/N belonging to three different fivefold symmetry clusters as a function of the supersaturation $\beta\Delta\mu$ of the fluid phase of a binary mixture of WCA spheres at varying temperatures corresponding to different degrees of particle softness, and for a binary hard-sphere mixture. The three data sets correspond to pentagonal bipyramids (diamonds), defective icosahedra (bullets) and icosahedra (squares). The error bars determined from 5 independent simulation runs are smaller than the symbols. Sketches of these clusters are shown on the right. We highlight (one of the) pentagons in the respective clusters.

hard-sphere systems has already been investigated [29, 84]. In order to investigate the effect of particle softness, we compare the number fraction of these clusters at fixed supersaturation $\beta\Delta\mu$ for varying temperatures T^*. The supersaturation $\beta\Delta\mu = \beta\mu_{\text{fluid}}(P) - \beta\mu_{\text{LP}}(P)$ is defined as the chemical potential difference between the supersaturated fluid and the stable LP at pressure P, and is determined by employing the Gibbs-Duhem relation $\int_{\mu(P_{coex})}^{\mu(P)} d\mu' = \int_{P_{coex}}^{P} \frac{1}{\rho(P')} dP'$ with P_{coex} and μ_{coex} the bulk pressure and chemical potential at the fluid-LP coexistence.

In Fig. 2.5, we plot the number fraction N_{CL}/N of the three investigated clusters in a binary fluid mixture at composition $x_L = 1/3$ *versus* $\beta\Delta\mu$ for varying T^* corresponding to different particle softness. We clearly see that the number fraction of fivefold symmetry clusters increases with $\beta\Delta\mu$, but more significantly, it decreases substantially with a small increase in particle softness. Notably, the five-membered rings as observed in the clusters highlighted in Fig. 2.5 are also present in the three LP crystal structures. It is thus not immediately clear whether these pentagons in the supersaturated fluid act as nucleation precursors or are responsible for the slow dynamics.

We therefore analyse the five-membered rings further in the BHS fluid phase as well as in the three ideal LPs, and classify them according to their large/small sphere composition and topology. In particular, to ensure that the systems at high supersaturation ($\beta\Delta\mu = 0.53$) are well-equilibrated, we run the simulations for a very long time ($> 10^7$ MC cycles). We distinguish 8 topologies (indexed $\mathcal{N}$) in Fig. 2.6b, and measure the probability to observe a specific topology $P(\mathcal{N})$ in the ideal LPs and the metastable BHS fluid at a high supersaturation $\beta\Delta\mu \simeq 0.53$. We reason that if these five-membered rings are formed randomly in a fluid mixture, $P(\mathcal{N})$ should follow a binomial distribution where the probability to observe a large sphere in a pentagonal cluster is determined by the composition x_L. We present the probability distributions $P(\mathcal{N})$ for the LPs, the metastable fluid, and binomial distribution all at a composition $x_L = 1/3$ in Fig. 2.6a. We find that the probability distribution $P(\mathcal{N})$ of the pentagons in

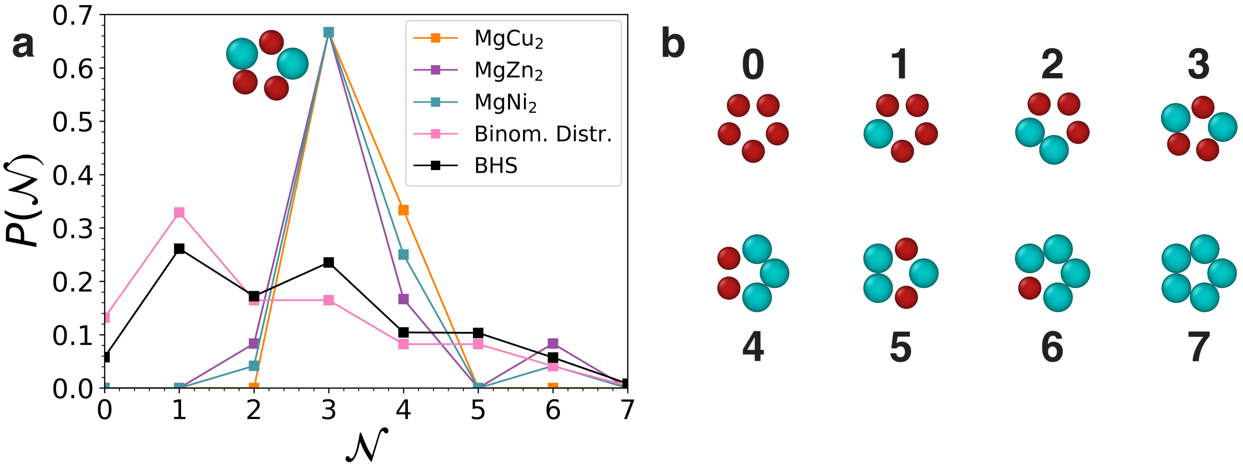


Figure 2.6: a) Probability distribution to observe a specific cluster topology $P(\mathcal{N})$ for the five-membered rings for the three ideal LPs, a supersaturated binary hard-sphere (BHS) fluid ($\beta\Delta\mu \simeq 0.533$), and a binomial distribution, all at composition $x_L = 1/3$. The 8 distinct cluster topologies are shown in b) with their index label $\mathcal{N}$. The error bars determined from 5 independent simulation runs are smaller than the symbols.

the BHS fluid mixture (black line in Fig. 2.6a) and in the WCA mixtures follows reasonably well the binomial distribution (pink line in Fig. 2.6a) for all topologies, demonstrating that the pentagons are formed randomly in the fluid. Furthermore, the pentagons with a topology $\mathcal{N} = 1$ are predominant in the supersaturated BHS fluid phase, whereas pentagons with a topology $\mathcal{N} = 3$ and 4 are prevalent in the ideal LPs. We therefore conclude that the fivefold symmetry clusters in the supersaturated fluid do not act as precursors for crystallisation, but are responsible for the slowing down of the dynamics and the kinetic arrest.

For future studies, it may be interesting to investigate whether a slightly different composition of the mixture - namely one that increases the probability of finding pentagonal clusters with a topology that is found in LPs - may enhance LP crystallisation [85]. More importantly, we find that the presence of these fivefold symmetry clusters can be reduced significantly by particle softness. This unexpected finding raises the immediate question whether or not crystal nucleation of the LPs can be enhanced or suppressed by tuning the softness of the interparticle potential.

2.3 Identification of Laves phases

In order to study the nucleation of the Laves phases, we require a criterion that distinguishes crystalline clusters with the Laves phase (LP) symmetry from the binary fluid phase. To this end, we first make a distinction between particles that have a solid-like environment with an LP-like symmetry and a fluid-like environment. We use the local bond orientational order parameters $q^{\alpha}_{l,m}(i)$ to determine the symmetry of the local environment of particle i with identity $\alpha(i) = L$ or S [86]

$$q^{\alpha}_{l,m}(i) = \frac{1}{N_b(i)} \sum_{j=1}^{N_b(i)} Y_{l,m}\left(\theta_{i,j}, \phi_{i,j}\right), \tag{2.8}$$

where $N_b(i)$ represents the number of all neighbours of particle i irrespective of its identity $\beta(j)$ (with $\beta(j) = L$ or S), $Y_{l,m}(\theta, \phi)$ are the spherical harmonics for m ranging from $[-l, l]$, and $\theta_{i,j}$ and $\phi_{i,j}$ are the polar and azimuthal angles of the centre-of-mass distance vector $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ with $\mathbf{r}_i$ the position of particle i. The neighbours of particle i with identity $\alpha(i)$ include all particles j that lie within a radial distance r_c of particle i, and we set r_c equal to the distance corresponding to the first minimum of the respective radial distribution function. Subsequently, we calculate the dot product $d_l^{\alpha\beta}(i,j)$ for each particle pair i with identity $\alpha(i)$ and j with identity $\beta(j)$

$$d_{l,\alpha\beta}(i,j) = \frac{\sum\limits_{m=-l}^{l} q_{l,m}^{\alpha}(i)^* q_{l,m}^{\beta}(j)}{\left(\sum_{m=-l}^{l} |q_{l,m}^{\alpha}(i)|^2\right)^{1/2} \left(\sum_{m=-l}^{l} |q_{l,m}^{\beta}(j)|^2\right)^{1/2}}. \tag{2.9}$$

We note that the dot product is symmetric in α and β, *i.e.*, $d_{l,\alpha\beta}(i,j) = d_{l,\beta\alpha}(i,j)$. We choose a symmetry index $l = 6$. Only in the case $\alpha = \beta = L$, we employ the *average* bond order parameter $\overline{q}_{l,m}^{\alpha}(i)$ as introduced in Ref. 87

$$\overline{q}_{l,m}^{L}(i) = \frac{1}{N_b(i)+1} \sum_{k=0}^{N_b(i)} q_{l,m}(k), \tag{2.10}$$

where $N_b(i)$ denotes the number of neighbours of particle i (with identity $\alpha(i) = L$) and $k = 0$ represents the particle i itself. We use this average bond order parameter in the dot-product expression of Eq. 2.9. The bond between particle i and j is classified as a *solid* bond if the dot product $d_{l,ij}$ lies in between a lower and a upper threshold value denoted by $d_{\alpha\beta}^{\downarrow}$ and $d_{\alpha\beta}^{\uparrow}$. In order to discriminate crystalline clusters with the Laves phase symmetry from the fluid phase, we use the following cut-off values for the three different species correlations: For the (i) large-large correlation, we employ $\overline{d}_{6,LL} > d_{LL}^{\downarrow} = 0.88$, (ii) small-small correlation $d_{SS}^{\downarrow} = -0.25 < d_{6,SS} < d_{SS}^{\uparrow} = -0.02$ and (iii) small-large $d_{LS}^{\downarrow} = -0.49 < d_{6,\mathrm{LS}} < d_{LS}^{\uparrow} = -0.1$ (see Fig. 2.7). Using these threshold values, we define a particle i with identity α as *crystalline* if the number of solid bonds satisfies $\xi^{\alpha}(i) > \xi_c^{\alpha}$ where

$$\xi^{\alpha}(i) = \sum_{j=1}^{N_b(i)} H(d_{l,\alpha\beta}(i,j) - d_{\alpha\beta}^{\downarrow}) - H(d_{l,\alpha\beta}(i,j) - d_{\alpha\beta}^{\uparrow}) \tag{2.11}$$

and H is the Heaviside step function. We employ the threshold values $\xi_c^L = 10$ and $\xi_c^S = 9$ for the large and small species, respectively. The above criteria are sufficient to distinguish the crystalline particles with LP-like symmetry from the surrounding fluid.

2.4 Nucleation behaviour

To investigate the effect of particle softness on the nucleation of the LPs, we determine the nucleation barrier height, critical nucleus size, and nucleation rate for a binary mixture of WCA spheres at composition $x_L = 1/3$, size ratio $q = 0.78$, and varying temperature T^*, using the seeding approach [88, 89]. This technique involves inserting a crystalline seed of a pre-determined shape and size into a metastable fluid. The configuration is subjected to a two-step equilibration process in the NPT ensemble where (i) the interface between the crystalline cluster and surrounding fluid is equilibrated by keeping the cluster fixed and then (ii) the

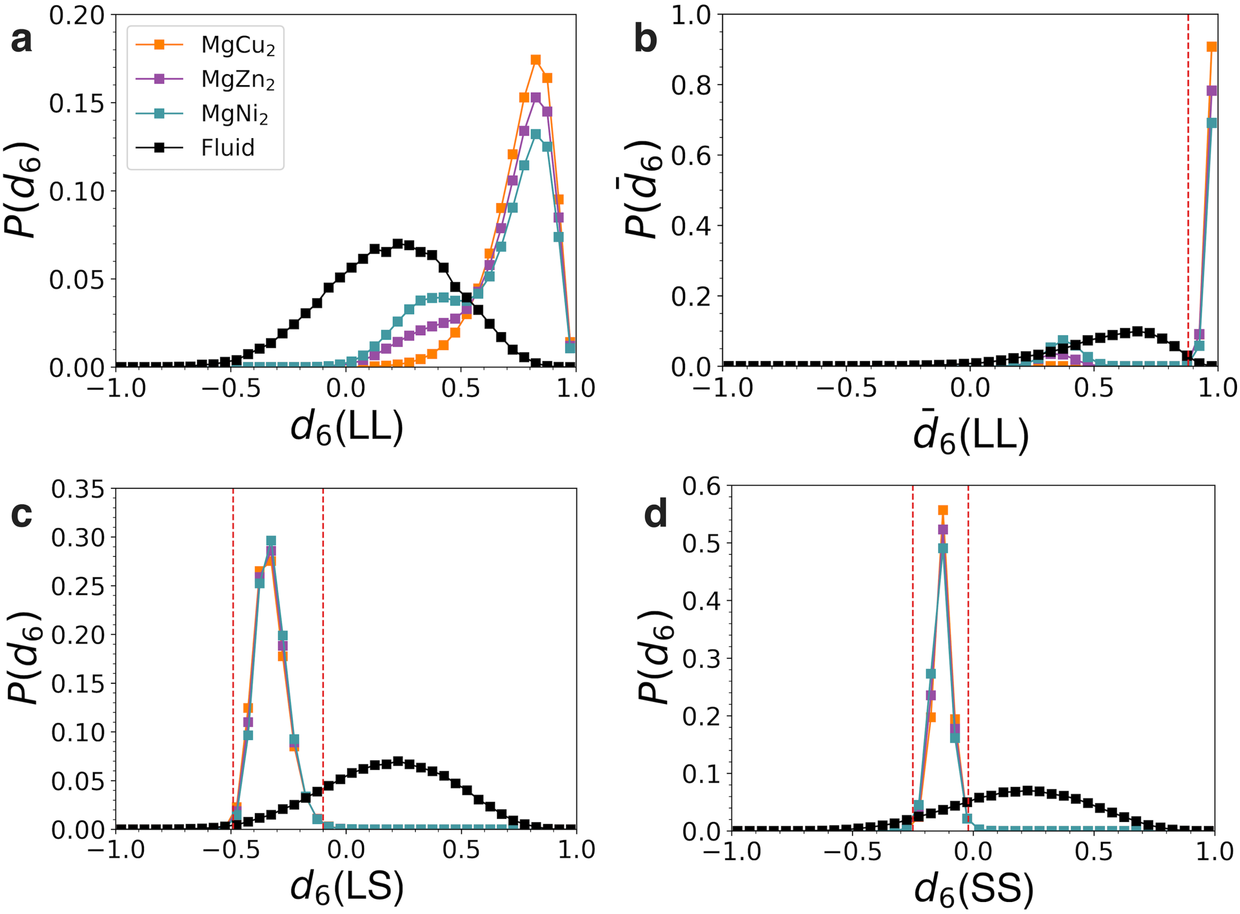


Figure 2.7: Probability distribution functions of the dot-product values $d_{6,LL}$, $\bar{d}_{6,LL}$, $d_{6,SS}$, and $d_{6,LS}$ for the large-large, averaged large-large, small-small and large-small correlations, respectively, of the fluid phase and the three Laves phases of a binary mixture of WCA spheres with a size ratio $q = 0.78$, temperature $T^* = 0.2$, and at bulk coexistence. The threshold values that we use in our LP crystal criterion are denoted by the vertical dashed lines. We specify that the hump in the probability distributions in a) and b) represents a very little number of bonds which are not considered to be solid-like, even if they are. However, this does not mean that particles will be wrongly classified, as a particle needs a certain number of solid bonds ξ in order to be classified as crystalline.

constraint on the cluster is relaxed and the system is equilibrated further. Subsequently, the equilibrated configuration is simulated for a range of pressures in order to determine the critical pressure $\beta P_c \sigma_L^3$ at which the critical cluster size of N_{CL} particles stabilises. We note that seeding simulations in the NVT ensemble have a strong dependence on system size due to a progressive depletion of particles in the supersaturated fluid phase when a crystal nucleus starts to grow and due to a drop in pressure when the fluid starts to crystallize. In order to avoid finite-size effects, we perform simulations in the NPT ensemble, where the system is subject to a fixed pressure [79]. An illustration is shown in Fig. 2.8, where a MgZn$_2$ LP seed melts, stabilises and grows out, as can be observed from the evolution of the size of the largest cluster N_{CL} as a function of time t/τ_{MD}, at $\beta P \sigma_L^3 = 22.6$ (red curve), 23 (green) and 25 (orange), respectively. The two equilibration phases (i) and (ii) were performed using MC simulations in the NPT ensemble involving isotropic volume scale moves in addition to the particle translation moves. The final step of the seeding is carried out with MD simulations performed using HOOMD-blue

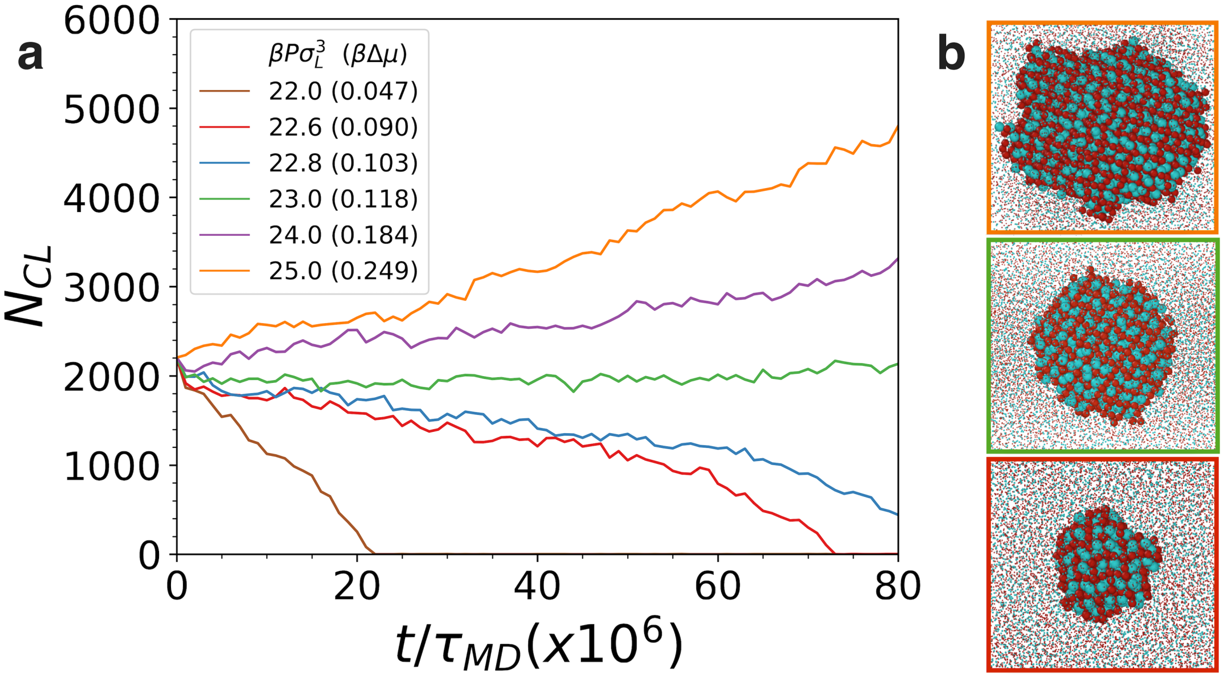


Figure 2.8: (a) The largest cluster size N_{CL} with LP symmetry as a function of time t/τ_{MD} using the seeding approach in MD simulations of a binary mixture of WCA spheres in the NPT ensemble at temperature $T^* = 0.2$, composition $x_L = 1/3$ and a diameter ratio $q = 0.78$ for varying pressures $\beta P\sigma_L^3$ with corresponding supersaturations $\beta\Delta\mu$ between brackets in order to estimate the critical pressure $\beta P_c\sigma_L^3$. The initial seed size is 2205 particles of the MgZn$_2$ Laves phase. The snapshots in (b) show the melting of the seed at $\beta P\sigma_L^3 = 22.6$ (red box), growth of the seed at $\beta P\sigma_L^3 = 25$ (orange box), and a stable seed at the critical pressure $\beta P_c\sigma_L^3 = 23$ (green box). The large (small) spheres are coloured blue (red). Fluid particles (particles with a disordered neighbourhood, see S.I.) are reduced in size for visual clarity.

(Highly Optimised Object-oriented Many-particle Dynamics) [90, 91] in the NPT ensemble. The temperature T and isotropic pressure P are kept constant *via* the Martyna-Tobias-Klein (MTK) [92] integrator, with the thermostat and barostat coupling constants $\tau_T = 1.0\ \tau_{MD}$ and $\tau_P = 1.0\ \tau_{MD}$ respectively, where $\tau_{MD} = \sigma_L\sqrt{m/\epsilon}$ is the MD time unit. The time step is set to $\Delta t = 0.004\tau_{MD}$ and the simulations are run for $10^9\tau_{MD}$ time steps. The simulation box is cubic and periodic boundary conditions are applied in all directions.

In order to discriminate crystalline clusters with LP-like symmetry from the fluid phase, we use the local bond-orientational order parameter described in Section 4.3.1.

The height of the Gibbs free-energy barrier ΔG_c for a critical nucleus size N_{CL} can subsequently be obtained from Classical Nucleation Theory

$$\beta\Delta G_c(N_{CL}) = N_{CL}\ \beta\Delta\mu/2. \tag{2.12}$$

By using different critical cluster sizes N_{CL} in the seeding approach, we obtain ΔG_c for varying critical pressures, corresponding to different supersaturations $\beta\Delta\mu$. We repeat these calculations for the three distinct LPs – MgZn$_2$, MgCu$_2$ and MgNi$_2$ – as crystalline seeds in the seeding approach. In Fig. 2.9(a), we present ΔG_c as a function of $\beta\Delta\mu$ for the three LPs and for temperatures $k_BT/\epsilon = 0.005$, 0.025, 0.1 and 0.2, corresponding to varying particle softness. We observe that for all temperatures and the three LP types, ΔG_c goes to infinity upon approaching bulk coexistence at $\beta\Delta\mu = 0$, decreases with increasing supersaturation $\beta\Delta\mu$, and approaches zero at sufficiently high $\beta\Delta\mu$. We find that all our ΔG_c data coincides within statistical error bars for all the three LPs, which is to be expected as the free-energy differences between the three bulk LPs are extremely small. More interestingly, we observe that the ΔG_c data collapses onto a master curve for all four temperatures, yielding an intriguing thermodynamic invariance for the different degrees of softness in the WCA interaction potential.

In addition, we calculate the nucleation rate J, which is determined by a thermodynamic term related to the Gibbs free-energy barrier $\beta\Delta G_c$ and a kinetic pre-factor

$$\frac{J\sigma_L^5}{D_L} = \sqrt{\frac{\beta\Delta\mu}{6\pi N_{CL}}}\frac{f^+\sigma_L^2}{D_L}\rho_f\sigma_L^3\exp(-\beta\Delta G_c), \tag{2.13}$$

where $f^+ = \langle (N(t) - N_{CL})^2 \rangle / t$ is the attachment rate of particles to the critical cluster, t the time, $\rho_f(\beta P_c\sigma_L^3)$ is the critical density of the fluid at the critical pressure, and D_L is the long-time diffusion coefficient at the same ρ_f. The attachment rate f^+ is measured from 10 independent simulation trajectories at the critical density ρ_f. We present the nucleation rates as a function of $\beta\Delta\mu$ in Fig. 2.9b, and find that they collapse for all four temperatures and three LPs onto a master curve - in a similar way as we observed for the nucleation barriers in Fig. 2.9a. This finding can be rationalised by the fact that the nucleation rate is predominantly determined by the thermodynamic term and simply echoes the thermodynamic invariance as observed for $\beta\Delta G_c$ for the three LPs and the four temperatures.

The data points on the free-energy barrier height and nucleation rate profiles are obtained from simulations with five different seed sizes (N_{seed}): (i) $N_{seed} = 96$ (all three LPs) and total system size $N = 4140$, (ii) $N_{seed} = 192$ ($MgCu_2$), 384 ($MgZn_2$, $MgNi_2$) and $N = 4160$, (iii) $N_{seed} = 648$ ($MgCu_2$), 1080 ($MgZn_2$), 720 ($MgNi_2$) and $N = 8100$, (iv) $N_{seed} = 1536$ ($MgCu_2$), 1728 ($MgZn_2$), 1440 ($MgNi_2$) and $N = 12500$, and (v) $N_{seed} = 3000$ ($MgCu_2$), 2688 ($MgZn_2$, $MgNi_2$) and $N = 17000$ particles, respectively. The critical pressures and attachment rates for the different seed sizes were obtained from MD simulations in the NPT ensemble, where the simulations were initialised with the MC-equilibrated configurations. The details of the MD simulations are described above.

To summarise, we find by employing the seeding approach that both the nucleation barriers and nucleation rates collapse onto a master curve for the different degrees of particle softness and for the three different types of LPs. This finding is in contrast to a previous simulation study that showed that the enhanced crystal nucleation rate for softer spheres is caused by a lower nucleation barrier [93]. Moreover, we find not only that the nucleation barrier decreases with supersaturation $\beta\Delta\mu$, but more importantly this method also allows us to pinpoint the supersaturation range, where the nucleation barrier of the LP becomes so low (less than several k_BT) that spontaneous nucleation should occur.

2.5 Spontaneous nucleation

Guided by the seeding approach results, we perform MD simulations in the NPT ensemble (see Section 3.4 for the simulations settings) using 1536 WCA spheres, and search for spontaneous nucleation of the LP in a highly supersaturated binary fluid phase of soft repulsive spheres, which has hitherto never been observed in previous simulation studies. In Fig. 2.10, we determine the size of the largest cluster N_{CL} with LP symmetry as a function of time t/τ_{MD} for a

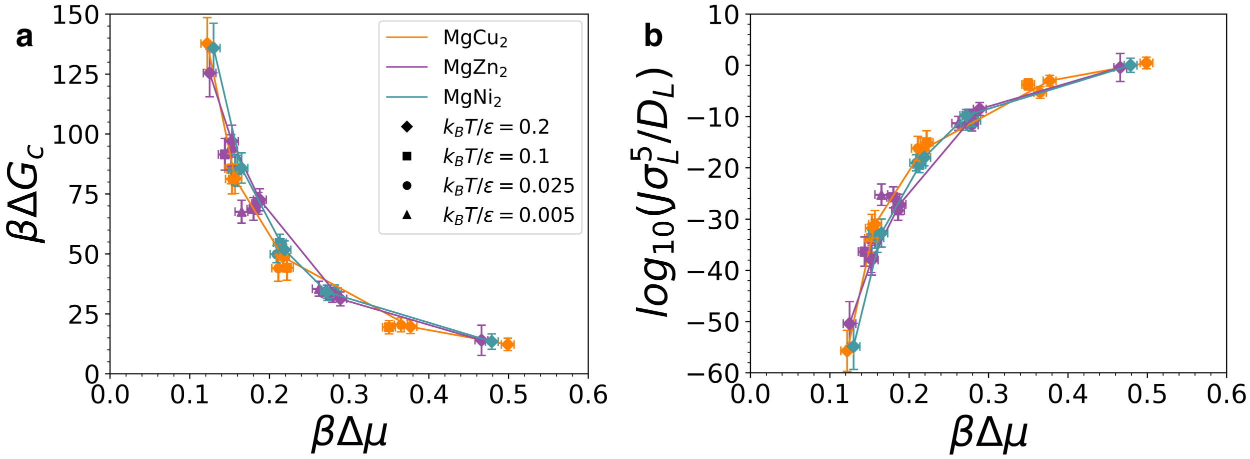


Figure 2.9: a) The height of the Gibbs free-energy barrier $\beta\Delta G_c$ and b) the nucleation rate $J\sigma_L^5/D_L$ as a function of the chemical potential difference $\beta\Delta\mu$ between the fluid and the LP for a binary WCA mixture for the three different LPs and for temperatures $k_BT/\epsilon = 0.005, 0.025, 0.1$, and 0.2.

range of pressures beyond coexistence, for $k_BT/\epsilon = 0.2$, where the coexistence pressure $\beta P\sigma_L^3$ $= 21.32$. The results yield some interesting observations. At pressure $\beta P\sigma_L^3 = 29$, no crystallisation is observed within our long simulation times. At a slightly higher pressure, $\beta P\sigma_L^3 =$ 29.5 ($\beta\Delta\mu = 0.524$), we observe that the system stays in a metastable fluid phase for a certain induction time until a nucleation event occurs, *i.e.* a crystalline nucleus of the MgZn$_2$ phase forms that subsequently grows out and transforms into the MgCu$_2$ phase as soon as the cluster spans the whole simulation box. Upon increasing the pressure further $\beta P\sigma_L^3 \geq 30$, the crystallisation exhibits features of spinodal-like behaviour as the supersaturated fluid is unstable with respect to the crystal phase and small crystalline nuclei appear immediately throughout the system. We later refer to this spinodal-like behaviour as the instability line, which we define as the lowest pressure $\beta P\sigma_L^3$ (or supersaturation $\beta\Delta\mu$) where crystallisation sets in immediately as soon as the simulation is started. For still higher pressures, we again see immediate crystallisation, but the clusters grow less, which we attribute to glassy behaviour.

More importantly, we also investigate whether or not the nucleation of the LP proceeds *via* a classical pathway by analysing different particle configurations of a spontaneous nucleation event in time. We observe in the initial stage of the simulation the formation and dissolution of small crystalline nuclei in the binary fluid phase until a crystal nucleus of the LP exceeds its critical size at an intermediate time and subsequently grows out. We thus show that crystal nucleation of the LP follows a classical pathway. This finding is important as our approach for estimating the Gibbs free-energy barrier heights and nucleation rates from seeding simulations and classical nucleation theory is only valid in the case that nucleation proceeds *via* a classical nucleation pathway. We hereby validate our method used in the previous section for measuring the nucleation barriers and nucleation rates.

To study the effect of temperature or particle softness on the spontaneous nucleation of the LP in a binary mixture of WCA spheres, we perform MD simulations in the NPT ensemble for $T^* = 0.1$ and 0.025 and pressures higher than the coexistence pressures $\beta P\sigma_L^3 = 20.03$ for $T^* = 0.1$, and $\beta P\sigma_L^3 = 18.35$ for $T^* = 0.025$, respectively. We find a similar pressure-dependence (not shown) as described above for $T^* = 0.2$: absence of crystallisation at low pressures, nucleation in a tiny intermediate pressure regime, and immediate spinodal-like crystallisation

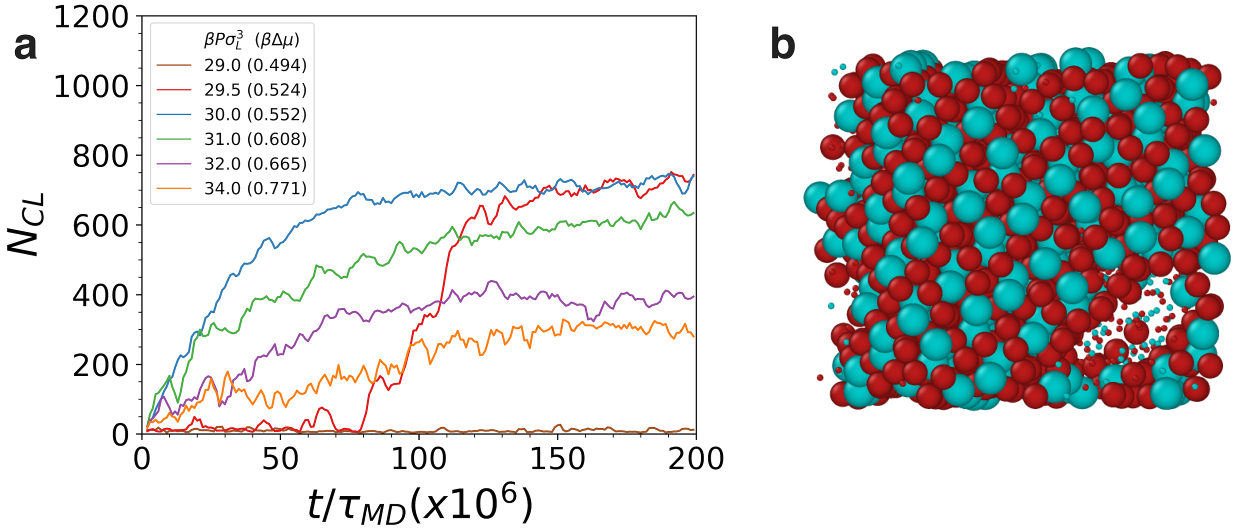


Figure 2.10: a) Size of the largest crystalline cluster N_{CL} for a binary mixture of WCA spheres with a diameter ratio $q = 0.78$ and temperature $T^* = 0.2$ as a function of time t/τ_{MD} for varying pressures $\beta P\sigma_L^3$ with corresponding supersaturations $\beta\Delta\mu$ between brackets using MD simulations in the NPT ensemble. b) Final configuration of a spontaneous nucleation event at $\beta P\sigma_L^3 = 29.5$ ($\beta\Delta\mu = 0.524$), initiated by the formation of a crystalline nucleus of the $MgZn_2$ phase that subsequently grows out and transforms into the $MgCu_2$ phase as the cluster spans the simulation box.

at $\beta P\sigma_L^3 \simeq 27.5$ at $T^* = 0.1$, and $\beta P\sigma_L^3 \simeq 25.5$ for $T^* = 0.025$. Surprisingly, the values of the thermodynamic driving force $\beta\Delta\mu$ corresponding to the pressures at which the fluid is unstable are given by $\beta\Delta\mu = 0.53 \pm 0.02$ for all three temperatures, which yields again an intriguing “universality" for the onset of spinodal-like behaviour.

2.6 Invariance with hard spheres

Perturbation and integral equation theories of simple liquids are based on the premise that the structure of monatomic fluids at high densities resembles that of hard spheres [75,94,95]. Hence, a system of hard spheres serves as a natural reference system for determining the properties of more realistic systems. On this basis one expects invariance of the structure along the melting and freezing line of simple fluids, and of thermodynamic properties such as the relative density change upon freezing and melting [96, 97].

Here, we observe an *invariance* of the Gibbs free-energy barriers, nucleation rates, and the onset of spinodal-like behaviour as a function of $\beta\Delta\mu$ for binary WCA mixtures at different temperatures. Inspired by this significant observation, we investigate whether other thermodynamic quantities, *and* structural properties are invariant along the freezing line of our WCA systems. Such an invariance is very interesting as it allows us to map the WCA mixture onto a simple binary hard-sphere system. A BHS mixture with a fixed composition depends on only one thermodynamic variable, the overall packing fraction, thereby yielding a simple one-dimensional phase diagram with a singular freezing and melting transition. In addition, the invariance may enable us to make predictions on the nucleation of the LP in a binary hard-sphere mixture, and may shed light on why LP nucleation is observed in a binary mixture of soft repulsive spheres and not in a system of hard spheres.

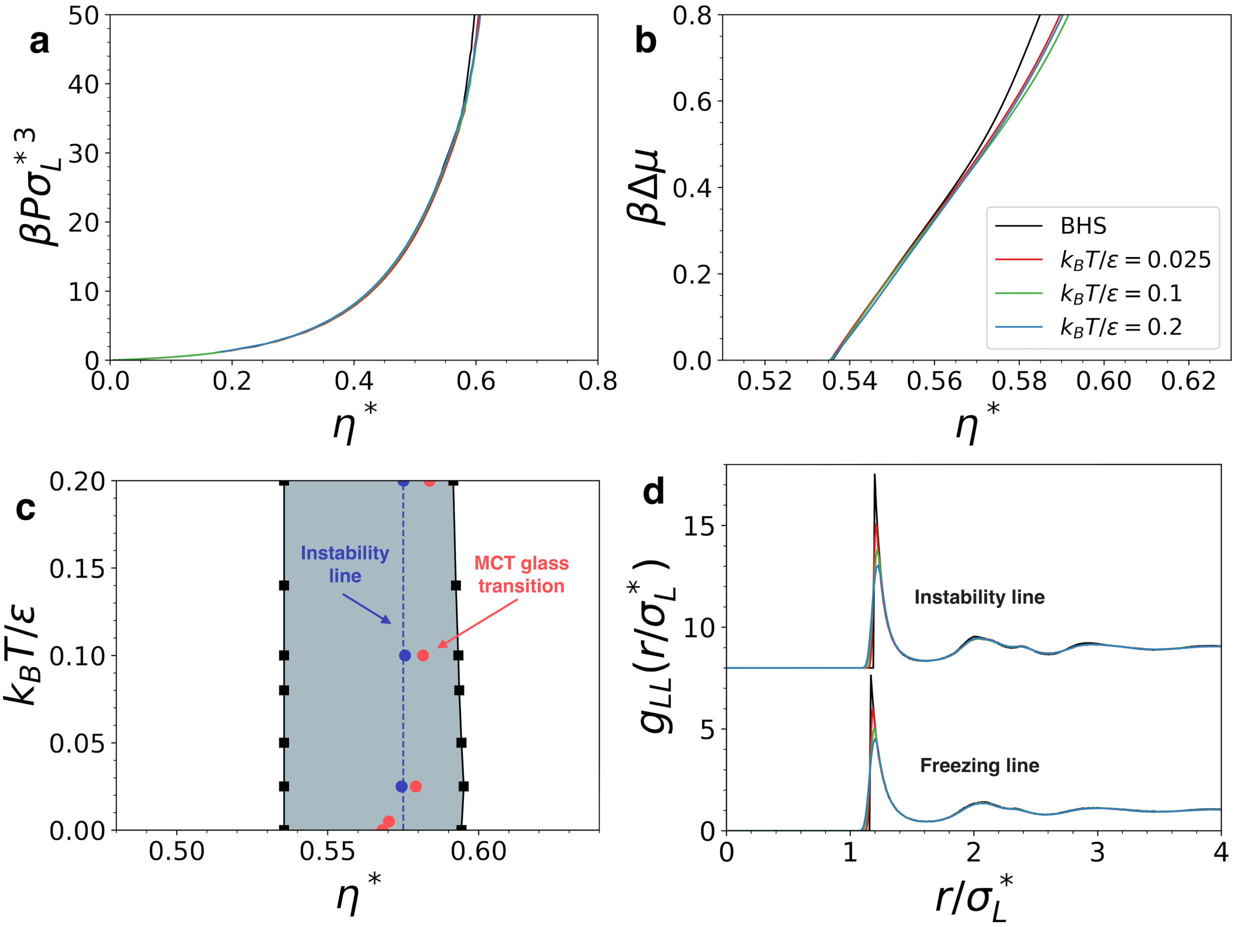


Figure 2.11: a) The reduced pressure $\beta P\sigma_L^{*3}$ and b) the supersaturation $\beta\Delta\mu$ *versus* effective packing fraction η^* for a binary hard-sphere mixture and a binary WCA mixture for varying temperatures with a diameter ratio $q = 0.78$ and composition $x_L = 1/3$. The error bars in (a) and (b) are $\leq 0.1\%$ and invisible with respect to the line thickness. c) The phase diagram of this binary WCA mixture in the reduced temperature k_BT/ϵ - η^* plane. The blue circles connected by a vertical dashed line denote the instability line where the fluid is unstable with respect to freezing. The kinetic MCT glass transition points are denoted by the red circles. For $k_BT/\epsilon = 0.005$, we find that kinetic glass transition precedes the spinodal-like instability, similar to hard spheres. Therefore, the instability point is not marked for this temperature as we do not observe crystallisation within reasonable timescales. d) The pair correlation function $g_{LL}(r)$ for the large spheres as a function of the scaled radial distance r/σ_L^* for a binary mixture of WCA spheres at three different temperatures and for a BHS mixture using MC simulations along the freezing line ($\beta\Delta\mu = 0$) and along the instability line ($\beta\Delta\mu \simeq 0.53$).

2.6.1 Thermodynamic invariance

To investigate the thermodynamic invariance of our WCA systems, we relate the thermodynamic properties of the WCA systems to those of a reference hard-sphere system. For this purpose, we scale the freezing number density of the binary WCA system for $q = 0.78$ at temperature T^* to the binary hard-sphere freezing packing fraction $\eta_{\mathrm{BHS}}^{(f)}$, which allows us to determine an *effective diameter* σ_L^* as well as an effective packing fraction η^* at each temperature, in a similar way as in Refs. 78 and 98. For a BHS mixture with a diameter ratio $q = 0.78$, the freezing packing fraction is $\eta_{\mathrm{BHS}}^{(f)} = 0.5356$. In Fig. 2.11c, we present the phase diagram in the temperature k_BT/ϵ - effective packing fraction η^* plane. As the freezing density for all temperatures of the WCA system is scaled to the freezing density of hard spheres, the freezing line becomes a vertical line in this representation. In addition, we find that the softness of the interactions has only a minor effect on the melting line and the width of the coexistence region. We find that the melting line shifts slightly to lower packing fractions and that the width of the coexistence region decreases marginally upon increasing the softness of the potential, *i.e.* increasing the reduced temperature T^*. In addition, we also plot the effective packing fractions corresponding to the state points where the fluid becomes unstable with respect to freezing, *i.e.* $\beta\Delta\mu \simeq 0.53$. This instability line lies well inside the two-phase coexistence region.

Subsequently, we use the same effective diameter σ_L^* to scale the equations of state, $\beta P \sigma_L^{*3}$ *versus* η^* for our WCA systems at different temperatures, and compare them with the equation of state for a BHS mixture with a diameter ratio $q = 0.78$. As seen in Fig. 2.11a, we find a perfect collapse of the equations of state, demonstrating a thermodynamic invariance for the equations of state for the WCA systems with temperature.

Finally, we plot the chemical potential difference $\beta\Delta\mu = \beta\mu_{\text{fluid}}(P) - \beta\mu_{\text{LP}}(P)$ between the fluid and the Laves phase as a function of the effective packing fraction η^* in Fig. 2.11b for both the WCA systems at varying temperatures and the BHS mixture. The deviation of the BHS system from the WCA systems at very high packing fractions $\eta^* \simeq 0.575$ is due to equilibration issues beyond the kinetic glass transition.
The collapse of the chemical potential difference for the WCA systems with different temperatures and the BHS system yields a fascinating “universality" in the thermodynamic driving force for nucleation of the LPs, explaining our observation of the thermodynamic invariance of the Gibbs free-energy barriers, nucleation rates, and the onset of spinodal-like behaviour as a function of $\beta\Delta\mu$ for different temperatures as described above.

We note that relating the properties of a system with realistic pair interactions to those of a hard-sphere system forms the basis of perturbation theories. The method that we use here to determine the effective diameter of an equivalent hard-sphere system is the simplest one and resembles the expression given by Barker and Henderson [99]. A more accurate prescription is given by the Weeks-Chandler-Andersen (WCA) theory, which equates the free energy of the reference system to that of a hard-sphere system at the same density and temperature, yielding a density- and temperature-dependent effective diameter [100]. In lowest order, WCA theory reduces to the Barker-Henderson expression, which is accurate for the WCA systems in the temperature and density range considered here, *i.e.* $T^* \leq 0.2$ and $\eta^* \leq 0.575$.

2.6.2 Structural invariance

In order to investigate whether the structure is also invariant along lines in the phase diagrams identified by equal $\beta\Delta\mu$ values, we measure the pair correlation function $g_{LL}(r)$ for the large spheres as a function of the radial distance expressed in terms of the effective diameter of the large spheres for the WCA systems at the three different temperatures and for the BHS mixture along the (i) freezing line $\beta\Delta\mu = 0$ and (ii) instability line $\beta\Delta\mu \simeq 0.53$, where we made sure that the system remained in the fluid state during our sampling. The results are presented in Fig. 2.11d. We find a good collapse of both sets of $g_{LL}(r)$'s as the peak positions coincide, showing a structural invariance of the two-body correlation functions along the freezing and instability lines. Additionally, one observes that the height of the first peak of the $g_{LL}(r)$ increases and the peak becomes narrower, and thus the $g_{LL}(r)$ becomes more hard-sphere-like upon lowering the temperature.

The collapse of the phase diagram, equations of state, and pair correlation functions demonstrate an invariance of the binary WCA mixtures for varying temperatures along lines of equal $\beta\Delta\mu$, *i.e.* thermodynamic driving force, in the phase diagram. We thus find that a binary mixture of soft repulsive spheres can be mapped onto a hard-sphere system in such a way that the two-body structure and thermodynamics are invariant. However, this invariance does not yet explain why nucleation of LPs is observed in the case of WCA systems and not for a binary hard-sphere mixture.

2.7 Kinetic glass transition

To shed light on this counterintuitive result, we investigate the kinetics of the WCA systems as a function of temperature. We already demonstrated above that the degree of fivefold symmetry clusters can be tuned by the softness of the interaction potential. To investigate the effect of fivefold clusters on the kinetics of the system further, we determine the kinetic glass transition of the WCA systems at varying temperatures. To this end, we calculate the self-intermediate scattering function $F_s(q,t) = 1/N_L \sum_{j=1}^{N_L} \langle \exp\{i\mathbf{q}.[\mathbf{r}_j(0) - \mathbf{r}_j(t)]\}\rangle$ at the wave vector $q = |\mathbf{q}|$ $= 2\pi/\sigma_L$ as a function of time tD_0/σ_L^2 using MC simulations. The summation runs over all large particles N_L and the short-time diffusion coefficient of the large spheres D_0 is computed in separate MC simulations. We repeat this calculation for a binary mixture of WCA spheres at temperatures $T^* = 0.2$, 0.1 and 0.025, and for a binary mixture of hard spheres, all at a diameter ratio of $q = 0.78$, using MC simulations at varying effective packing fractions η^*. Exemplarily, we plot $F_\mathrm{S}(q,t)$ for a WCA mixture at $T^* = 0.2$ in Fig. 2.12a for varying η^*. The dynamics slows down dramatically with increasing η^*. We plot the structural relaxation time τ_α in reduced units, defined by $F_\mathrm{S}(q,\tau_\alpha) = e^{-1}$, as a function of $|\eta^* - \eta_c^*|/\eta_c^*$ in Fig. 2.12b for our WCA systems at $T^* = 0.025, 0.1$, and 0.2 and the BHS mixture. At sufficiently high densities, the structural relaxation time τ_α diverges algebraically, and we find a perfect collapse of all the data. We employ the prediction from mode coupling theory (MCT) [101], $\tau_\alpha \sim |\eta^* - \eta_c^*|^{-\gamma}$, to fit the structural relaxation times τ_α as a function of η^* using η_c^* and γ as fit parameters. Here, η_c^* denotes the critical packing fraction corresponding to the kinetic glass transition as described by MCT. We note that η_c^* serves solely as a proxy for a qualitative change in the dynamics and mention that for $\eta > \eta_c^*$ the relaxation time diverges exponentially as shown for both active and passive hard-sphere systems [102, 103]. The structural relaxation τ_α is well described by an exponential divergence at a packing fraction corresponding to the ideal glass transition as described by the Vogel-Fulcher-Tammann law. We also stress that the structural relaxation times are determined here from simulations, thereby incorporating all many-body correlations in the binary fluid phase in contrast to theoretical predictions from MCT that uses the structure factor (two-body correlation) as input.

We list the critical MCT packing fractions η_c^* with the corresponding supersaturation $\beta\Delta\mu$, critical exponents γ, and the effective diameters σ_L^* in Table 2.2 for the WCA systems at varying temperatures, *i.e.* softness of the interaction potential and the BHS mixture. The statistical error on γ ($< 3\%$) and η_c^* ($< 0.2\%$) are determined by the fit, and on τ_α ($< 0.1\%$) by the real part of the $F_\mathrm{S}(q,t)$). We also plot the critical MCT packing fractions η_c^* in Fig. 2.11c. We clearly observe that the critical MCT effective packing fraction η_c^* and corresponding supersaturation $\beta\Delta\mu$ decreases with decreasing softness of the interaction potential. In fact, in the BHS case, the kinetic glass transition, as predicted by MCT, precedes the state point where the fluid becomes unstable with respect to freezing.

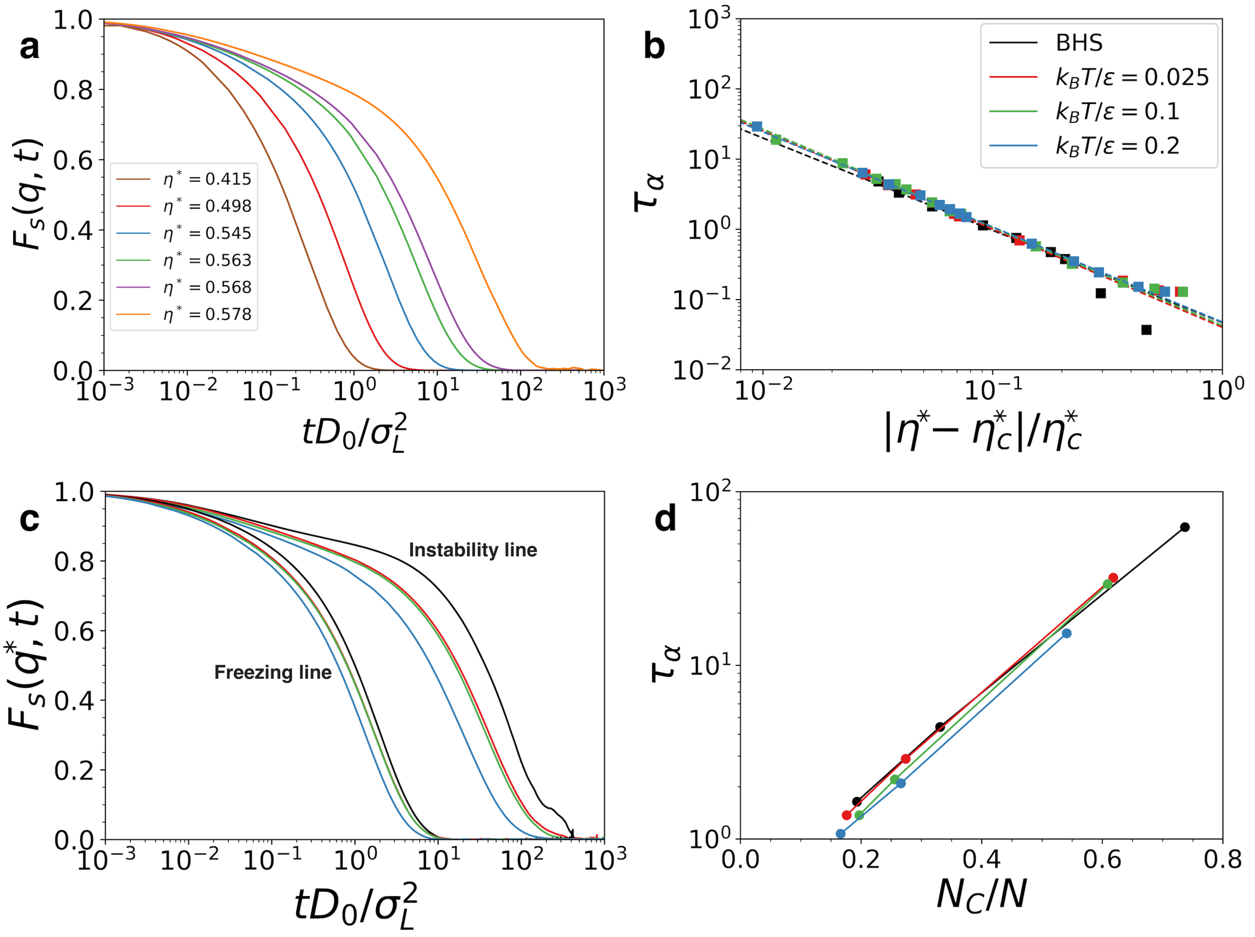


Figure 2.12: a) The self-intermediate scattering function $F_S(q,t)$ for the large spheres as a function of time tD_0/σ_L^2 for a binary WCA mixture with a diameter ratio $q = 0.78$ at $T^* = 0.2$ for varying effective packing fractions η^* as obtained from MC simulations. b) The structural relaxation time τ_α as a function of $| \eta^* - \eta_c^* | /\eta_c^*$ for a binary mixture of WCA spheres at $T^* = 0.025, 0.1$, and 0.2 and a binary hard-sphere mixture. Note that in c) and d) the same colour coding is used as in b). c) The self-intermediate scattering function $F_S(q^*,t)$ for the large spheres for a binary mixture of WCA spheres at three different temperatures and for a BHS mixture using MC simulations along the freezing line and along the instability line. The four curves on the left correspond to $\beta\Delta\mu = 0$, and the four curves on the right correspond to $\beta\Delta\mu = 0.533$. d) Exponential relation between the relaxation time τ_α at three different levels of supercooling ($\beta\Delta\mu = 0$, $\beta\Delta\mu = 0.2$ and $\beta\Delta\mu \simeq 0.53$) and the number fraction of particles N_{CL}/N belonging to a defective icosahedron.

In order to make a further comparison between the investigated systems, we compare $F_s(q^*,t)$ at the wave vector $q^* = |\mathbf{q}^*| = 2\pi/\sigma_L^*$, for state points along the freezing line $\beta\Delta\mu = 0$ and along the instability line $\beta\Delta\mu \simeq 0.53$ for a BHS system and WCA systems at $T^* = 0.025$, 0.1 and 0.2, in Fig. 2.12c. Not only do we observe a different dynamical behaviour for varying softness, but we also note a strong correlation between the relaxation times τ_α and the number fraction of fivefold symmetry defective icosahedron clusters. We display this correlation in Fig. 2.12d, where we find an exponential relation between the structural relaxation times in the fluid phase for varying particle softness and different supersaturations $\beta\Delta\mu$, and the fraction of particles belonging to a defective icosahedron. Such a correlation between the defective icosahedron and the structural relaxation times was also found for weakly polydisperse hard

System	η_c^*	$\beta\Delta\mu$	γ	σ_L^*
$k_BT/\epsilon = 0.2$	0.5837	0.673	1.3545	1.0583
$k_BT/\epsilon = 0.1$	0.5816	0.620	1.3984	1.0764
$k_BT/\epsilon = 0.025$	0.5792	0.604	1.3993	1.1009
BHS	0.5681	0.452	1.3140	1.0000

Table 2.2: The critical MCT effective packing fraction η_c^* corresponding to the kinetic glass transition for a binary mixture of WCA spheres at varying temperatures and for a BHS mixture, all at a diameter ratio $q = 0.78$, the corresponding supersaturation level $\beta\Delta\mu$, the critical exponents γ of the MCT fits, and the effective large-sphere diameters σ_L^*.

spheres [104].

To summarise, we find that a BHS mixture gets kinetically arrested at a lower packing fraction than the packing fraction where we expect to find spontaneous nucleation of the LP. However, for a slightly softer interaction potential, a binary WCA mixture at $T^* = 0.025$, we find the reverse situation, and hence spontaneous nucleation is observed at a packing fraction that is lower than that of the kinetic glass transition. This finding may explain why LP nucleation is never observed in a binary mixture of hard spheres, and is observed here for a binary WCA system.

2.8 Conclusions

In 2007, a novel self-assembly route towards a photonic bandgap material was proposed in which the diamond and pyrochlore structure are self-assembled from a binary mixture of colloidal hard spheres into a closely packed $MgCu_2$ Laves phase [9]. Despite numerous efforts, spontaneous nucleation of the LPs has never been observed in simulations of BHS mixtures or in experiments on micron-sized colloidal hard spheres, casting doubts on the thermodynamic stability of these crystal structures in binary hard spheres. Recent MC simulations have shown, however, that by introducing size polydispersity, either in a static or dynamic way, and by using unphysical particle swap moves, LPs may be nucleated from a dense hard-sphere fluid [105, 106].

Alternatively, to alleviate problems with the degeneracy of the three competing LPs and with the metastability of the $MgCu_2$ with respect to the $MgZn_2$ phase, one may resort to another self-assembly route in which the $MgCu_2$ phase, stable in the present system, is formed from a binary mixture of colloidal spheres and preassembled tetrahedral clusters of spheres as shown both in simulations [107] and experiments using DNA-mediated interactions [108].

To better understand why the nucleation of LP is severely hampered in a binary fluid of hard spheres, we investigated the degree of fivefold symmetry in the binary fluid phase as the presence of fivefold symmetry structures may suppress nucleation. In order to study the effect of softness of the interaction potential, we measured the number fraction of three significant representatives of the fivefold symmetry structures in a binary fluid of WCA spheres at varying temperatures k_BT/ϵ thereby altering the softness of the interaction potential. In the limit of $k_BT/\epsilon \to 0$, this system reduces to the binary hard-sphere system. Surprisingly, we found that particle softness significantly reduces the degree of fivefold symmetry in the binary fluid phase.

To investigate the repercussions of this finding on LP nucleation, we subsequently performed simulations with a crystalline seed to measure the nucleation barrier and nucleation rate for the three LP types and for varying temperatures, *i.e.* degrees of particle softness. These results enabled us to study, for the first time, spontaneous nucleation of the LPs in simulations of nearly hard spheres. We thus find that the seeding approach is versatile and robust [88, 89] – it not only enables one to determine the nucleation barrier and nucleation rate, but also locate the regime in the phase diagram where spontaneous nucleation may occur and provides information on how a crystal nucleus grows and melts.

Our observation of spontaneous nucleation of the LP in a system of soft spheres is important and intriguing for two reasons. On the one hand, our simulations provide evidence that the LP is stable in the phase diagram of such a binary mixture, as predicted theoretically more than a decade ago [9]. On the other hand, it immediately begs the question why LP nucleation has never been seen in simulations of BHS mixtures or in experiments on micron-sized colloidal hard spheres despite numerous attempts by many research groups, whereas it nucleates spontaneously with a tiny degree of particle softness.

To address this question, we studied the role of softness in the interaction potential on the structure, phase behaviour, and dynamics of the LPs, and found that a system of soft repulsive spheres can be mapped onto a binary hard-sphere system in such a way that the structure and thermodynamics are invariant in reduced units for varying softness of the interaction potential. However, the invariance of the nucleation barrier and nucleation rate as a function of supersaturation for varying softness of the potential seems to be at odds with the observation of LP nucleation in WCA systems for $T^* \geq 0.025$ and the absence of it in binary hard spheres.

In order to shed light on this counterintuitive result, we determined the kinetic glass transition by fitting the structural relaxation times as obtained from the self-intermediate scattering functions with an MCT fit for the various WCA systems. Surprisingly, we found that the packing fraction corresponding to the kinetic glass transition strongly depends on the softness of the particle interactions, which in turn affects the presence of fivefold symmetry clusters in the supersaturated fluid phase. It will be interesting to investigate in future work whether or not there is a connection with previous studies that show that the fragility of a glass can be tuned by particle softness [109–112]. We thus find that crystallisation can be enhanced by tuning the softness of the particle interactions, either by charge, ligands, or a stabiliser, in simulations or experiments. This finding is indeed consistent with the experimental observations of the LPs as they all seem to involve particles interacting with (slightly) soft repulsive interactions [61–68]. Moreover, introducing a small degree of softness in the particle interactions can be exploited in a wealth of other crystallisation studies. For instance, there are still many open questions on how and why binary crystal phases nucleate. A systematic study of binary nucleation has been hampered so far by either slow dynamics or by finding the right regime in the phase diagram where nucleation may occur.

Finally, we note that the structure as characterised by the two-body correlation functions as well as the thermodynamics which is predominantly determined by also two-body correlations is invariant in reduced units for varying softness of the pair potential. However, the structural relaxation time and the kinetics depend strongly on the presence of fivefold structures, and thus on higher-body correlations. We hope that this finding will inspire the development of new theories for predicting the kinetic glass transition that take into account higher-body correlations.

Acknowledgements

I would like to thank Tonnishtha Dasgupta, who equally contributed to this work. Also, I would like to extend my gratitude to Dr. Da Wang, Dr. Ernest van der Wee for useful discussions regarding the Laves phases.

3

An artificial neural network reveals the nucleation mechanism of a binary colloidal AB$_{13}$ crystal

In this Chapter we investigate binary nucleation of the AB$_{13}$ crystal from a binary fluid phase of nearly-hard spheres. We calculate the nucleation barrier and nucleation rate as a function of supersaturation and draw a comparison with nucleation of single component and other binary crystals. To follow the nucleation process, we employ a neural network to identify the AB$_{13}$ phase from the binary fluid phase and the competing fcc crystal with single-particle resolution and significant accuracy in the case of bulk phases. We show that AB$_{13}$ crystal nucleation proceeds *via* a co-assembly process where large spheres and icosahedral small-sphere clusters simultaneously attach to the nucleus. Our results lend strong support for a classical pathway that is well-described by classical nucleation theory, even though the binary fluid phase is highly structured and exhibit local regions of high bond orientational order.

Based on: G. M. Coli and M. Dijkstra, *An artificial neural network reveals the nucleation mechanism of a binary colloidal AB$_{13}$ crystal*, ACS Nano, **15 (3)**, 4335-4346 (2021)

3.1 Introduction

Understanding crystallisation is important in many research fields such as protein crystallisation for resolving the molecular structure, drugs design in the pharmaceutical industry, ice crystal formation in clouds for weather forecasts, and crystallisation of colloidal and nanoparticle suspensions with application perspectives in catalysis, opto-electronics, and plasmonics. Hence, it is not surprising that over the past decades many experimental and simulation studies have been devoted to studying crystal nucleation in a fluid of hard spheres, which is indisputably one of the simplest possible model systems to describe colloidal and nanoparticle systems, and serves as a reference for systems with more complicated interactions, *e.g.* depletion interactions, or electrostatic interactions.

Nucleation describes the process, in which a crystal nucleus spontaneously forms due to a statistical fluctuation in the metastable fluid phase. Despite the significant amount of work spent on understanding crystal nucleation in hard spheres, the mechanism by which a hard-sphere fluid transforms into a crystal phase remains to be settled. Several scenarios such as a classical one-step crystallisation process, a non-classical two-step crystallisation mechanism with precursors consisting of local regions with either high density, high bond-orientational order or competing orders, or a spinodal-like process have been proposed, but all of these crystallisation mechanisms are still heavily debated [77, 113–121].

To enhance the structural diversity and functional composition of the self-assembled structures, one may resort to binary mixtures of large and small colloidal hard spheres with diameters σ_L and σ_S, respectively. Although the number of distinct binary crystal structures is relatively small, we wish to remark here that the structural diversity can be enhanced significantly by taking into account varying interaction potentials, *e.g.* suspensions consisting of two types of particles with opposite charges can form a dazzling variety of binary superlattice structures [122, 123]. In this Chapter, we focus on binary mixtures of particles interacting with hard-sphere like potentials as they serve as a reference for a wider class of soft repulsive interaction potentials, mimicking the interactions of many nanoparticle systems.

The phase behaviour of binary hard-sphere mixtures is well studied by now, and display a wide variety of behaviours ranging from a spindle-type to azeotropic and eutectic phase diagram, wide coexistence regions between phases with different compositions, pure single-component crystals, substitutionally disordered crystalline phases, interstitial solid solutions, and various binary crystal structures with different stoichiometries $x_L = N_L/(N_L + N_S)$, with N_L (N_S) denoting the number of large (small) particles [124]. Depending on the diameter ratio $q = \sigma_S/\sigma_L$, binary hard-sphere systems exhibit entropically-stabilised binary superlattice structures analogous to their atomic counterparts NaCl ($0.2 \leq q \leq 0.42$), AlB_2 ($0.42 \leq q \leq 0.59$), $NaZn_{13}$ ($0.48 \leq q \leq 0.62$), and the Laves phases ($0.74 \leq q \leq 0.84$) [124].

The most intriguing structure of the above-mentioned binary crystals is without any doubt the $NaZn_{13}$, also termed the icosahedral AB_{13} structure in order to distinguish it from the cuboctahedral AB_{13} structure [125], which has been found to be metastable due to a less efficient packing of the small spheres in the case of binary hard-sphere mixtures [126, 127]. The stability of the AB_{13} structure has gained much attention because of its bizarre lattice. The large spheres are remarkably distant from each other as shown in Fig. 3.1, and are arranged on a simple cubic lattice, which is an unusual crystal structure in the case of plain hard spheres. Furthermore, each unit cell of this simple cubic lattice of large spheres contains an icosahedral cluster of 13 small spheres, which are all rotated by 90° with respect to their neighbouring icosahedral clusters. Hence, the full unit cell of an icosahedral AB_{13} structure consists of 8 unit cells of this simple cubic lattice of large spheres with 8 icosahedral clusters of 13 small spheres in their centres, resulting in 112 particles in total.

The colloidal analog of the $NaZn_{13}$ was for the first time observed by Sanders *et al.* in natural gem opals consisting of two sizes of silica spheres in 1987 [12,128]. The same AB_{13} structure was later observed in systems of charged-stabilised colloids or PMMA particles [63,129–132], and in various nanoparticle systems, *e.g.* mixtures of semiconductor, metaloxide, magnetic, silica, and polymer-grafted nanoparticles, as well as polyoxometalate clusters [61,62,123,125,133–142].

In contrast to the considerable amount of work that has been devoted to studying crystal nucleation in single-component hard-sphere fluids, only a few studies have been focused on crystal nucleation in binary mixtures. Crystallisation in fluid mixtures is generally much harder than in single-component systems. Spindle-, azeotropic-, eutectic-like phase transitions in binary systems usually involve fractionation as the composition of the solid phase deviates from that of the supersaturated phase. Fractionation is known to slow down the rate of crystallisation [70,143]. Additionally, the surface tension of the solid-fluid interface will increase when the compositions of the fluid and solid phase deviate substantially, leading to higher nucleation barriers and lower nucleation rates [144,145]. Moreover, nucleation of a binary (compound) crystal is believed to be orders of magnitude slower than that of pure crystals or substitutionally disordered crystalline phases due to a loss of mixing entropy, making binary crystal nucleation an extremely rare event [143]. This is one of the reasons that the number of simulation and experimental nucleation studies on binary colloidal crystals is very limited.

Yet, the few simulation studies on binary crystal nucleation revealed several interesting observations. For instance, simulations on homogeneous nucleation of a binary AB_2 crystal in a mixture of hard spheres revealed that in the case of multiple competing crystal structures, the phase that nucleates is the one whose composition is closest to that of the fluid phase even when it is metastable [143,145]. In addition, it was found by simulations that kinetic barriers also play an important role in determining which crystal phase nucleates. In the case of oppositely charged colloids it was found that the disordered face-centred-cubic (fcc) phase that nucleates is metastable and has a higher free-energy barrier for nucleation than the thermodynamically stable binary CsCl crystal [146]. In this case the disordered fcc phase was favoured by non-equilibrium nucleation. These results greatly challenge the commonly held assumption that subcritical clusters are always in quasi-equilibrium with the fluid phase [147]. Another simulation study showed that homogeneous nucleation of an interstitial solid solution in a binary mixture of hard spheres is driven by the nucleation of large spheres into an fcc crystal while maintaining chemical equilibrium of the small spheres throughout the system. Additionally, as shown in Chapter 2, nucleation of Laves phases is severely suppressed by the presence of icosahedral clusters in a binary hard-sphere mixture, but softness of the interaction potential reduces the degree of fivefold symmetry in the binary fluid and enhances crystallisation [148]. Finally, we also mention for completeness that spontaneous spinodal-like crystallisation of structures isostructural to AlB_2, $NaZn_{13}$, and the Laves phases has been observed in simulations on highly supersaturated binary hard-sphere fluids with and without unphysical moves that swap the identities of large and small spheres [149].

All nucleation studies share a common challenge: being able to recognise different phases starting from the raw particle coordinates of the system. In particular, one requires a criterion that is able to distinguish on a single-particle level particles belonging to the growing phase from those of the metastable parent phase. Most crystal nucleation studies are based on describing the local environment of each particle in terms of the so-called bond orientational order parameters, *i.e.* rotational invariant combinations of the spherical harmonics of degree l, as

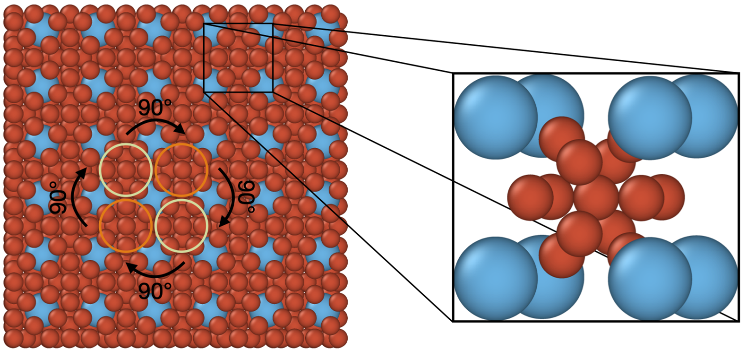


Figure 3.1: The icosahedral-AB_{13} structure consists of a simple cubic lattice of large (A) spheres and icosahedral clusters of 13 small (B) spheres denoted in blue and red, respectively. The icosahedral clusters of 13 small spheres are rotated by 90° with respect to their neighbouring clusters as indicated by the differently coloured circles. The zoom-in displays the icosahedral cluster formed by the small spheres inside a simple cubic subunit cell of large spheres.

introduced by Steinhardt *et al.* in 1983 [86]. Specifically, the fourfold and sixfold bond order parameters, q_4 and q_6, suffice to distinguish the crystalline particles from the fluid-like particles as most crystals exhibit either cubic and/or hexagonal symmetry. In the case of binary crystals, the local environment of each particle of each species can deviate substantially from cubic and hexagonal symmetry, and other symmetries should be taken into account in identifying the different crystal phases. Here, we describe the local environment of a particle by using a full expansion in spherical harmonics, and we train an artificial neural network to identify the different phases on a single-particle level using a set of bond order parameters up to degree $l = 12$ as input. We demonstrate the effectiveness of this method by studying nucleation of the AB_{13} crystal structure in binary mixtures of nearly hard spheres using simulations. We show that standard techniques fail in identifying the different phases and that machine learning is useful in achieving this goal.

Employing the trained neural network as an order parameter, we investigate how an AB_{13} crystal nucleates and grows and we shed light on the formation mechanism during the early stages. In addition, we study how icosahedral clusters of small spheres arrange themselves inside this simple cubic lattice and whether the growth of the binary nucleus proceeds *via* the attachment of individual small spheres, small clusters, or *via* perfect or defective icosahedral clusters of small spheres. Specifically, the role of the icosahedral clusters on the AB_{13} nucleation is intriguing as the presence of fivefold clusters is often attributed to glassy dynamics and suppression of crystallisation [22]. However, in Ref. 126, it was conjectured that the abundance of icosahedral clusters in both the fluid and AB_{13} crystal and thus the structural similarity of these two phases may result in an ultra-low surface tension and hence a low nucleation barrier and high nucleation rate. To investigate this, we determine the nucleation barrier height and the nucleation rate using the seeding approach and we compare our results with crystal nucleation in pure hard spheres and with nucleation of the Laves phases. Finally, we analyse the kinetic pathways of the spontaneous crystallisation events of the AB_{13} phase in binary mixtures of hard spheres using brute force Molecular Dynamics (MD) simulations.

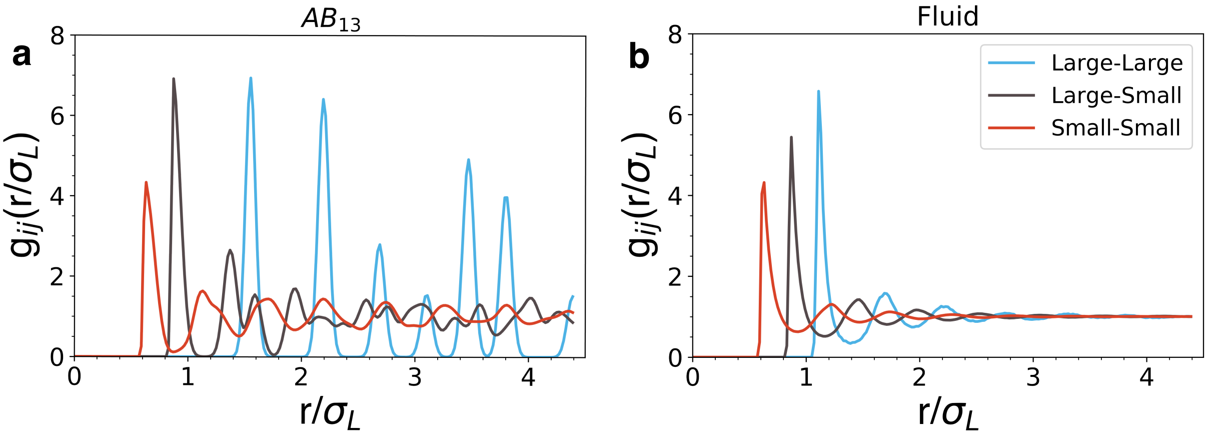


Figure 3.2: Pair correlation functions $g_{ij}(r/\sigma_L)$ of a) the AB$_{13}$ crystal phase and b) the binary fluid phase at composition $x_L = 1/14$ with $i, j \in \{L, S\}$ denoting the large (L) and small (S) species, for a mixture of WCA spheres at $k_BT/\epsilon = 0.025$ to mimic hard spheres and at coexistence pressure $\beta P_{coex}\sigma_L^3 = 45.35$. The large-large pair correlation function $g_{LL}(r)$ of the AB$_{13}$ shows that the large spheres are unusually distant from each other in comparison with the binary fluid phase. The small-small pair correlation function $g_{SS}(r)$ of the AB$_{13}$ phase demonstrates that the small spheres exhibit fluid-like behaviour, making it difficult to distinguish the binary fluid phase and the AB$_{13}$ crystal on the basis of the small spheres.

3.2 The Model

We consider a binary mixture of N_L large (L) hard spheres with a diameter σ_L and N_S small (S) hard spheres with a diameter σ_S. For a binary hard-sphere (BHS) mixture, the AB$_{13}$ phase is thermodynamically stable for a diameter ratio $q = \sigma_S/\sigma_L \in [0.54, 0.61]$ [124]. In this work, we set $q = 0.55$. The particles interact *via* a Weeks-Chandler-Andersen (WCA) pair potential, which can straightforwardly be employed in Molecular Dynamics (MD) simulations and which reduces to the hard-sphere potential in the limit that the temperature $T \to 0$. The WCA potential $u_{\alpha\beta}(r_{ij})$ between species $\alpha, \beta \in \{L, S\}$ reads [75]

$$u_{\alpha\beta}\left(r_{ij}\right) = \begin{cases} 4\epsilon\left[\left(\frac{\sigma_{\alpha\beta}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{\alpha\beta}}{r_{ij}}\right)^{6} + \frac{1}{4}\right], & r_{ij} < 2^{\frac{1}{6}}\sigma_{\alpha\beta} \\ 0, & r_{ij} \geq 2^{\frac{1}{6}}\sigma_{\alpha\beta} \end{cases} \tag{3.1}$$

with $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ the centre-of-mass distance between particle i and j, $\mathbf{r}_i$ the position of particle i, ϵ the interaction strength, and $\sigma_{\alpha\beta} = (\sigma_\alpha + \sigma_\beta)/2$. The steepness of the repulsion between the particles can be controlled by tuning the temperature k_BT/ϵ. We set $k_BT/\epsilon = 0.025$, which has been used extensively in previous simulation studies to mimic hard spheres [77–80].

3.2.1 Free-energy calculations

Using free-energy calculations in Monte Carlo (MC) simulations, we determine phase coexistence between the AB$_{13}$ crystal and the binary fluid phase with the same composition as that of the AB$_{13}$ crystal, by computing the Helmholtz free energy per particle $f = F/N$ as a function of density $\rho = N/V$ for both phases with N the number of particles and V the volume of the system. We calculate f using thermodynamic integration of the equations of state

$$\beta f\left(\rho\right)=\beta f\left(\rho_{0}\right)+\int_{\rho_{0}}^{\rho}d\rho'\frac{\beta P(\rho')}{\rho'^{\ 2}}, \tag{3.2}$$

where $f\left(\rho_0\right)$ denotes the Helmholtz free energy per particle for a reference density ρ_0, $\beta=1/k_BT$ is the inverse temperature, and P is the pressure. We use the ideal gas as a reference state for the binary fluid phase, and we employ the Frenkel-Ladd method to calculate the Helmholtz free energy at a reference density ρ_0 using MC simulations in the NVT ensemble [83].

Subsequently, we calculate for both the AB_{13} crystal and the binary fluid phase, the chemical potential $\beta\mu$ at pressure P:

$$\beta\mu=\frac{\beta G}{N}=\beta f+\frac{\beta P}{\rho}, \tag{3.3}$$

with βG the dimensionless Gibbs free energy. The chemical potential difference or supersaturation is obtained *via* $\beta\Delta\mu=\beta\mu_{fluid}-\beta\mu_{AB_{13}}$, whereas two-phase coexistence between the AB_{13} and fluid phase is determined by imposing $\beta\Delta\mu=0$, resulting in

$$\frac{\beta\Delta f}{\Delta(1/\rho)}=-\beta P, \tag{3.4}$$

which is equivalent to the common tangent construction on the free-energy curves in the $\beta f-1/\rho$ plane with $\Delta f=f_{fluid}-f_{AB_{13}}$ and $\Delta(1/\rho)=(1/\rho_{fluid})-(1/\rho_{AB_{13}})$. The pressure at which the crystal and fluid are at coexistence reads $\beta P_{coex}\sigma_L^3=45.35$. To study nucleation of the AB_{13} crystal, we perform simulations at pressures $P>P_{coex}$, which determine the regime where the fluid phase is metastable with respect to the crystal phase.

3.3 Local structure detection

3.3.1 Bond Order Parameters

In order to follow the nucleation process of the AB_{13} phase, we need to find a way to detect an embryo of the stable AB_{13} crystal structure in the supersaturated binary fluid phase with single-particle resolution. In many simulation studies, local bond orientational order parameters have been used to study crystal nucleation [86, 150, 151]. To calculate these local bond order parameters, we first have to define the local environment of particle i by determining a list of neighbours using, for instance, a distance criterion based on the first minimum of the pair correlation function or by employing a Voronoi construction. We then define the complex vector $q_{lm}(i)$ for each particle i

$$q_{lm}(i)=\frac{1}{N_b(i)}\sum_{j=1}^{N_b(i)}Y_{lm}(\theta(\mathbf{r}_{ij}),\phi(\mathbf{r}_{ij})), \tag{3.5}$$

where $N_b(i)$ is the number of neighbours of particle i, $Y_{lm}(\theta(\mathbf{r}_{ij}),\phi(\mathbf{r}_{ij}))$ denotes the spherical harmonics, $m\in[-l,l]$, $\theta(\mathbf{r}_{ij})$ and $\phi(\mathbf{r}_{ij})$ are the polar and azimuthal angles of the distance vector $\mathbf{r}_{ij}=\mathbf{r}_j-\mathbf{r}_i$, and $\mathbf{r}_i$ denotes the position of particle i. Subsequently, we define rotationally invariant quadratic and cubic order parameters as

$$q_l(i) \quad = \quad \sqrt{\frac{4\pi}{2l+1}\sum_{m=-l}^{l}|q_{lm}(i)|^2} \tag{3.6}$$

and

$$w_l(i) = \frac{\sum\limits_{m_1+m_2+m_3} \begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} q_{lm_1}(i) q_{lm_2}(i) q_{lm_3}(i)}{(\sum\limits_{m=-l}^{l} |q_{lm}(i)|)^{3/2}}. \tag{3.7}$$

Additionally, we also use the averaged bond-orientational order parameters. The averaged $\bar{q}_{lm}(i)$ is defined as

$$\bar{q}_{lm}(i) = \frac{1}{\tilde{N}_b(i)} \sum_{j=1}^{\tilde{N}_b(i)} q_{lm}(j), \tag{3.8}$$

where $\tilde{N}_b(i)$ is the number of neighbours including particle i itself. The rotationally invariant quadratic and cubic averaged bond order parameters are defined as

$$\bar{q}_l(i) \quad = \quad \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |\bar{q}_{lm}(i)|^2} \tag{3.9}$$

and

$$\bar{w}_l(i) = \frac{\sum\limits_{m_1+m_2+m_3} \begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} \bar{q}_{lm_1}(i) \bar{q}_{lm_2}(i) \bar{q}_{lm_3}(i)}{(\sum\limits_{m=-l}^{l} |\bar{q}_{lm}(i)|)^{3/2}}. \tag{3.10}$$

Due to the cubic or hexagonal symmetry of most crystals, the fourfold and sixfold bond order parameters, q_4 and q_6, have been extensively employed in literature to study crystal nucleation. We first study whether the fourfold and sixfold bond order parameters can be used to distinguish the AB$_{13}$ crystal from the binary fluid phase with a composition $x_L = N_L/(N_L + N_S) = 1/14$ and from the pure fcc phase. For this purpose, we use the averaged bond order parameters $\bar{q}_l$, thereby taking into account also the second shell of neighbours of a particle [87]. We perform MC simulations in the isobaric-isothermal ensemble, *i.e.* we fix the pressure P, the temperature T, and the number of particles $N = N_L + N_S$. We carry out bulk simulations of the AB$_{13}$ crystal and the binary fluid at coexistence pressure $\beta P_{coex}\sigma_L^3 = 45.35$, and of the pure fcc crystal at $\beta P_{coex}\sigma^3 = 8.87$ corresponding to the pressure at bulk coexistence with the single-component fluid phase.

In Fig. 3.2, we plot the pair correlation functions $g_{ij}(r)$ of the AB$_{13}$ crystal and the binary fluid phase with $i, j \in \{L, S\}$ denoting the large (L) and small (S) species. We first observe from the small-small pair correlation function $g_{SS}(r)$ of the AB$_{13}$ phase that the small spheres exhibit fluid-like behaviour even though they are in a solid state. The structural similarity of the small spheres in the binary fluid and the AB$_{13}$ phase makes it difficult to distinguish the two phases on the basis of the local symmetry of the small spheres. Additionally, we observe that the main peak of the large-large pair correlation function $g_{LL}(r)$ of the AB$_{13}$ crystal is at an unusually large distance in comparison with that of the binary fluid phase. In order to nucleate the AB$_{13}$ phase in the binary fluid phase, the large spheres have to be pushed away from each other to much larger distances to make room for the icosahedral clusters of small spheres. In addition, the huge difference in the position of the main peak of the $g_{LL}(r)$ of the binary fluid and the AB$_{13}$ phase complicates the identification of neighbouring particles on the basis of a simple cut-off distance.

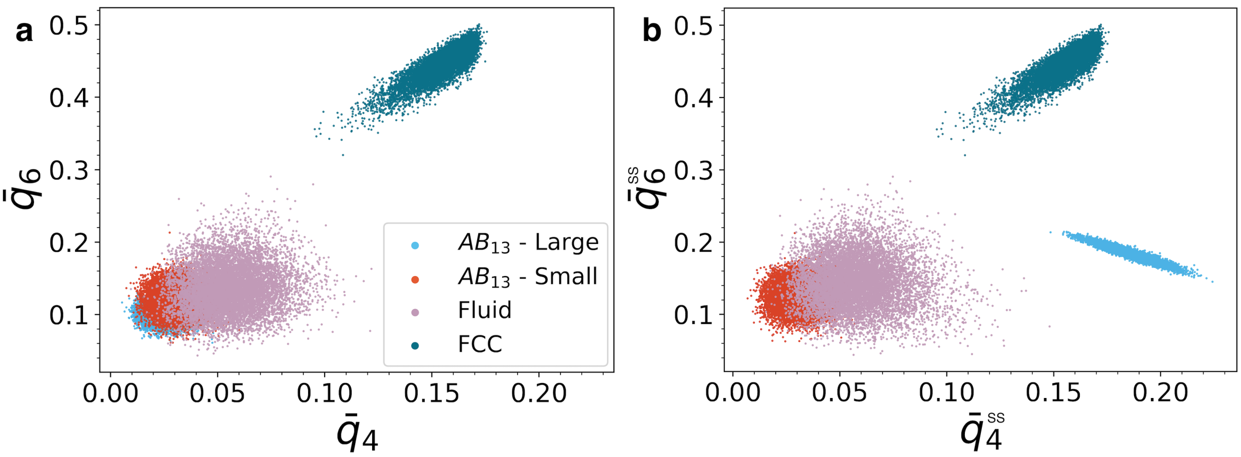


Figure 3.3: Scatter plot in the averaged bond order parameter $\bar{q}_4$ - $\bar{q}_6$ plane for the four local particle environments we wish to distinguish: large species (light blue) and small species (red) of the AB_{13} phase, binary fluid phase (light purple) and the pure fcc phase (dark blue). (a) Averaged bond order parameters calculated using the solid-angle based nearest-neighbour criterion irrespective of particle species, showing significant overlap of 3 of the 4 local structures thereby making the classification impossible. (b) Averaged bond order parameters calculated by taking into account only the neighbours belonging to the same species. The distribution of the small species of the AB_{13} phase still overlaps with that of the binary fluid phase.

To circumvent this problem, we use the parameter-free solid-angle based nearest-neighbour (SANN) algorithm to identify the neighbours of each particle [152]. Using the SANN algorithm, we measure the fourfold and sixfold averaged bond order parameters, $\bar{q}_4$ and $\bar{q}_6$, and we show scatter plots in the $\bar{q}_4 - \bar{q}_6$ plane for the large and small species of the AB_{13} phase, the binary fluid phase, and the fcc phase in Fig. 3.3. We observe from Fig. 3.3a that the distributions for the large and small species of the AB_{13} phase and the binary fluid phase overlap, making the distinction between the different phases very hard. To improve the separation of the bond order parameter distributions, we calculate $\bar{q}_4^{ss}$ and $\bar{q}_6^{ss}$, where the superscript ss means that bond order parameters are calculated by considering only particles of the same species as particle i. We plot the results in Fig. 3.3b, and observe that the distribution of the large species in the AB_{13} is well separated from the other structures. However, the distribution of the small species of the AB_{13} phase still overlaps with that of the binary fluid phase. We thus find that it remains a major challenge to correctly classify the small species of the AB_{13} phase which can be misidentified as fluid particles due to their icosahedral arrangement.

3.3.2 Feed forward neural network

In order to overcome this problem, we describe the local environment of a particle using a full expansion in spherical harmonics, and we train an artificial neural network (ANN) to identify distinct structures on a single-particle level, thereby extending the approach of Ref. 153. The main goal of the ANN here is to detect the birth of a crystal nucleus of the AB_{13} structure in a binary fluid phase, which presents additional difficulties with respect to the identification of bulk phases due to the solid-fluid interfaces of the crystal nuclei [153]. Moreover, the identification of crystalline particles and estimating the number of crystalline particles are crucial for determining the barrier height and the nucleation rate using the seeding approach

[88,154]. To minimise the effect of interfaces, we employ the standard, instead of the averaged coarse-grained, bond order parameters thereby increasing the spatial resolution at the expense of the accuracy in the local structure detection. The idea behind the choice of non-averaged bond order parameters is as follows: given the great predictive capacity of neural networks, due to the extremely effective way of combining information from many features and using non-linear functions, we can use *less precise* but *more local* descriptors. A high accuracy can be reached thanks to the versatility of the neural network, and in this way it is possible to obtain, for each particle, an estimate of the class based only on the *first shell* of neighbours.

We employ the bulk simulations of the AB$_{13}$ phase, the pure fcc phase, and the binary fluid as described in Section 4.3.1, and build a training set of 10^5 training samples for each of the local particle environments we wish to distinguish: large particles of the AB$_{13}$ phase, small species in the AB$_{13}$ phase, particles in a pure fcc phase, and particles irrespective of species in a binary fluid. In order to build such a data set, we perform MC simulations in the NPT ensemble of the AB$_{13}$ crystal and the binary fluid phase with a composition $x_L = 1/14$, both at coexistence pressure $\beta P_{coex}\sigma_L^3 = 45.35$, and of the pure fcc phase at bulk coexistence with the fluid phase at pressure $\beta P_{coex}\sigma^3 = 8.87$. We describe the local environment of each particle i by a $36-$*dimensional* input vector of bond order parameters

$$I(i) = (\{q_l(i)\}, \{w_{l'}(i)\}, \{q_l^{ss}(i)\}, \{w_{l'}^{ss}(i)\}), \tag{3.11}$$

where $l \in [1, 12]$ and l' varies in the same range but only assumes even values.

In Ref. 153, a single-layer ANN, *i.e.* only an input and output layer and no hidden layers, was employed to successfully classify the AB$_{13}$ phase from a binary fluid phase with a composition $x_L = 1/3$ using the *averaged* bond order parameters as input. However, this network architecture with the *averaged* bond order parameters as input vector is not accurate enough to distinguish the AB$_{13}$ phase from the binary fluid with a composition $x_L = 1/14$ equal to the stoichiometry of the AB$_{13}$ phase, which is mostly due to the structural similarity of the small spheres in the AB$_{13}$ and the fluid phase, both exhibiting an abundance of (defective) icosahedral clusters. In addition, the standard *non-averaged* bond order parameters that we employ offer a poorer characterisation of the bulk phases with respect to their averaged counterparts. In order to improve the accuracy of the classification, we add hidden layers to our neural network.

To be more specific, we employ a fully connected neural network with two hidden layers consisting of 72 neurons. Each neuron uses a rectified linear unit (ReLU) activation function to guarantee fast convergence and good generalisation. The output layer has 4 neurons, corresponding to the four distinct local particle environments (classes) we wish to distinguish, and is activated with a Softmax function (Fig. 3.4). The training of the network was done using the Keras package, enabled by Tensorflow backend. Specifically, we trained the network minimising the categorical cross-entropy loss function with the addition of a L_2 regularisation term using a weight decay pre-factor of 10^{-4}. The minimisation was carried out using minibatch stochastic gradient descent with momentum [155–157], and we set the learning rate to 10^{-2}. We employ 20% of the samples as validation data to predict the accuracies for each output node corresponding to the 4 different particle environments. The accuracies related to each specific class are shown in Table 8.1.

3.4 Seeding approach

Numerical simulations have helped elucidating nucleation for a plethora of model systems, but despite the possibility of following each single particle during the crystallisation process, they

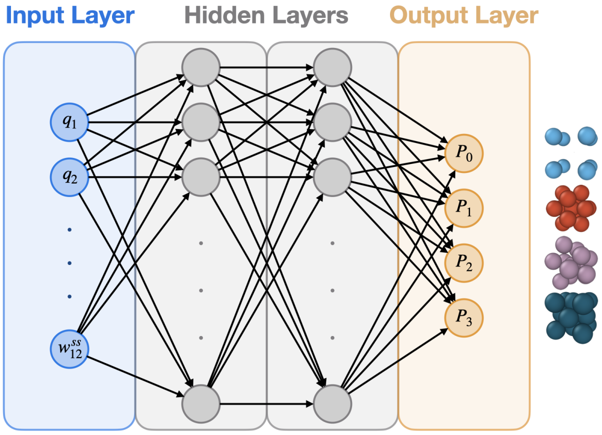


Figure 3.4: Architecture of our fully connected artificial neural network (ANN). The input layer has 36 units, as described by Eq. 5.5, while both hidden layers contain 72 neurons. The output layer consist of four neurons, yielding the probability that a particle corresponds to a certain class.

Class	Accuracy
AB_{13} - Large	100.0%
AB_{13} - Small	98.1%
Fluid	97.8%
fcc	99.4%

Table 3.1: Accuracies of the ANNs on the validation set calculated for all four classes.

suffer from an important drawback. In fact, the accessible time scales in MC or MD simulations are typically much shorter than in experiments. This is particularly disadvantageous for nucleation studies as the birth of a crystalline nucleus in a metastable fluid is a rare event.

For this reason, it is often necessary to use special sampling schemes in simulations like umbrella sampling (US) [158–161], forward flux sampling (FFS) [160, 162], metadynamics [163, 164], or transition path sampling [165–167]. These techniques are mainly employed to enhance or bias the sampling of the system in order to observe rare events like nucleation. However, these simulation techniques are extremely expensive from a computational point of view, restricting nucleation studies to highly metastable conditions.

Recently, another technique has been proposed to study nucleation, known as the seeding approach [88, 154]. The great merit of this technique is that it enables the determination of all relevant physical quantities to describe nucleation, *e.g.* barrier height and nucleation rate, and that it divides the simulation study into short simulations with a standard and unbiased sampling of phase space. Moreover, the computational cost of these simulations is moderate, which allows studying nucleation under weakly metastable conditions where the critical nucleus consists of several thousands of particles. The seeding technique involves the following steps: 1) inserting a seed of the crystal structure of interest in a metastable fluid, and running simulations to equilibrate the interface while keeping the crystalline particles fixed, 2) releasing this constraint and equilibrating the full system at a sufficiently high pressure that the seed does not melt, and finally 3) running simulations of this carefully equilibrated system for a wide range of pressures in order to determine the critical pressure P_c at which the probability that the seed will grow or melt will be equal, while for $P < P_c$ the seeds will predominately melt, or grow for $P > P_c$.

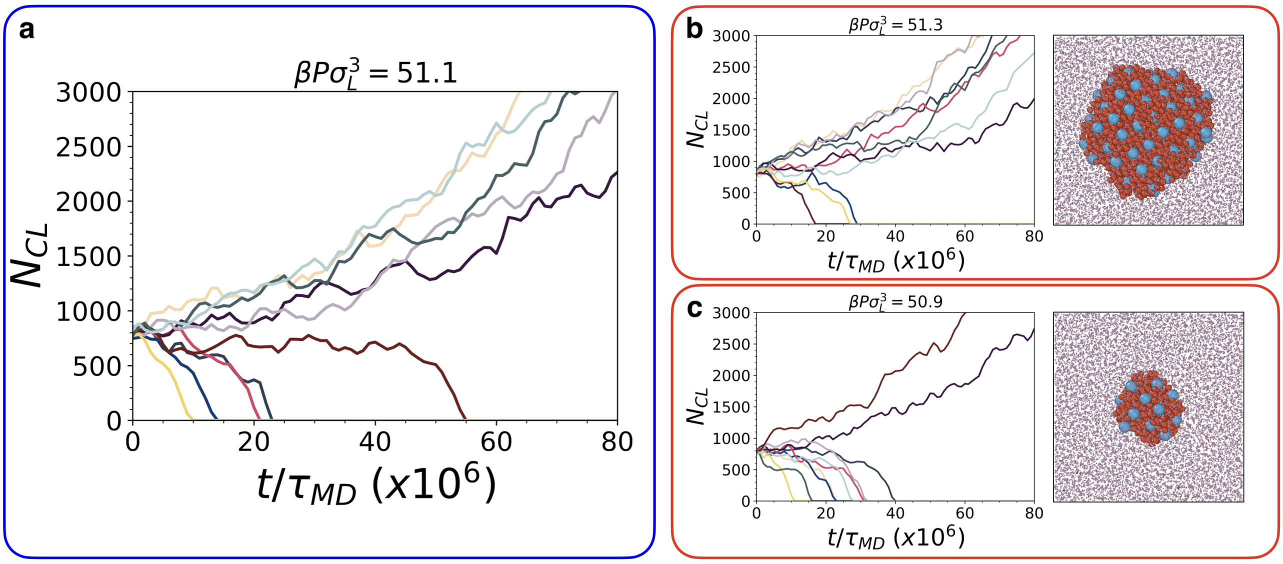


Figure 3.5: a) Largest cluster size N_{CL} of the AB$_{13}$ phase as recognised by the ANN as a function of time t/τ_{MD} using the seeding approach in 10 independent MD simulations of a binary mixture of WCA spheres at temperature $k_BT/\epsilon = 0.025$ with a diameter ratio $q = 0.55$ in the NPT ensemble, composition $x_L = 1/14$, and at a pressure a) $\beta P\sigma_L^3 = 51.1$, where the cluster melts and grows with equal probability, indicating that this pressure value corresponds to the critical pressure for this cluster size, b) $\beta P\sigma_L^3 = 51.3$, where the cluster grows in the majority of the simulations, and c) $\beta P\sigma_L^3 = 50.9$, at which the cluster melts in most of the simulations. Typical configurations of the growth and melting of the cluster are shown on the right in b) and c), respectively. Note that the size of fluid-like and fcc-like particles are reduced in size for visual clarity.

In order to avoid finite-size effects, we perform simulations in the NPT ensemble. Specifically, both equilibration parts of the seeding approach have been carried out using MC simulations in the NPT ensemble, using pressure $\beta P\sigma_L^3 = 56.0$ and a total number of MC cycles equal to 10^3. Differently, for our investigations on the seeded growth, we perform MD simulations using HOOMD-blue (Highly Optimised Object-oriented Many-particle Dynamics) [168, 169] in the NPT ensemble. The temperature T and pressure P are kept constant *via* the Martyna-Tobias-Klein (MTK) integrator [92], with the thermostat and barostat coupling constants $\tau_T = 1.0\ \tau_{MD}$ and $\tau_P = 1.0\ \tau_{MD}$, respectively, and $\tau_{MD} = \sigma_L\sqrt{m/\epsilon}$ is the MD time unit. The time step is set to $\Delta t = 0.004\tau_{MD}$, which is small enough to ensure stability of the simulations. The simulation box is cubic and periodic boundary conditions are applied in all directions.

An illustration of the last step of the seeding approach is shown in Fig. 3.5, where we plot the size of the largest cluster N_{CL} as a function of time t/τ_{MD}, for 10 independent simulations, at pressure $\beta P\sigma_L^3 = 50.9, 51.1,$ and 51.3. Here, $\tau_{MD} = \sigma_L\sqrt{m/k_BT}$ denotes the MD time unit and m the mass of the particles. In Fig. 3.5b (3.5d) the majority of the simulations show a growing (melting) cluster, which means that the pressure $\beta P\sigma_L^3 = 51.3$ (50.9) is higher (lower) than the critical one. In Fig. 3.5a, we observe that at $\beta P\sigma_L^3 = 51.1$ the cluster melts and grows with equal probability, indicating that this value corresponds to the critical pressure for a critical cluster size $N_{CL} = 770$. The number of particles belonging to the main cluster is determined using the neural network-based order parameter as described in Section 3.3.2 together with a clustering algorithm to identify clusters of mutually bonded solid particles.

Subsequently, several physical quantities can be calculated using classical nucleation theory

(CNT), such as the height of the Gibbs free-energy barrier ΔG_c using

$$\Delta G_c(N_{CL}) = N_{CL}\Delta\mu/2, \tag{3.12}$$

and the nucleation rate J

$$\frac{J\sigma_L^5}{D_L} = \sqrt{\frac{\beta\Delta\mu}{6\pi N_{CL}}\frac{f^+\sigma_L^2}{D_L}}\rho_f\sigma_L^3 \exp(-\beta\Delta G_c), \tag{3.13}$$

where N_{CL} denotes the critical nucleus size, $\beta\Delta\mu$ the supersaturation, *i.e.* the difference in chemical potential between the supersaturated fluid and the stable crystal phase, $f^+ = \langle (N(t) - N_{CL})^2 \rangle / t$ is the attachment rate of particles to the critical cluster, t the time, $\rho_f(\beta P_c\sigma_L^3)$ is the critical density of the fluid at the critical pressure P_c, and D_L is the long-time diffusion coefficient at the same ρ_f [88, 154]. The attachment rate f^+ is measured from 10 independent simulation trajectories at density ρ_f. Assuming an on average spherical cluster shape, the crystal-fluid interfacial free energy γ can be calculated from

$$N_{CL} = \frac{32\pi\gamma^3}{3\rho_s^2\Delta\mu^3} \tag{3.14}$$

with ρ_s the density of the solid phase. We note that these equations rely on the validity of CNT, which will be checked and proven in Section 3.5. We emphasise that all variables computed through the seeding approach using Eqs. 3.12, 3.13, and 3.14, are sensitive to the numerical value of N_{CL}, and thus to an estimate of the nucleus interface. This particular estimate is problematic for all classification algorithms and can, in principle, lead to systematic errors when evaluating the aforementioned variables. In Section 3.4.1 we show a detailed analysis of the performance of the ANN with respect to the interface detection.

Using different seed sizes in the seeding approach, we determine ΔG_c, J, γ and ρ_s for different critical pressures corresponding to different supersaturations $\Delta\mu$. We plot our results as a function of $\Delta\mu$ in Fig. 3.6, and present the numerical data in Table 3.2. For comparison, we also plot the results from Chapter 2 on the nucleation of the Laves phase in a binary hard-sphere mixture and of the fcc phase in a fluid of pure hard spheres [79, 80].

Comparing the nucleation barrier ΔG_c for the three different phases at the same thermodynamic driving force $\Delta\mu$, we clearly observe from Fig. 3.6a that ΔG_c is consistently higher for the fcc phase than for the AB$_{13}$ crystal and the Laves phases. This is in contradiction with the assumption that the nucleation barrier for binary nucleation should be higher due to a loss of mixing entropy. As the nucleation rate J is predominantly determined by ΔG_c, we find a similar behaviour for J, where J of the fcc phase seems to be smaller compared to that of the binary crystals [80]. In addition, we plot the interfacial free energy γ for the three phases in Fig. 3.6c. We wish to remark here that we express the surface tensions in units of k_BT/σ_L^2, which is an arbitrary choice, and hence a direct comparison of the three systems cannot be made as the dimensions of the spheres and the compositions are very different for the fcc, Laves and AB$_{13}$ phase. Hence, the conjecture of Ref. 126 that the surface tension of the AB$_{13}$-fluid interface may be low due to the structural similarity of these phases is difficult to verify. Moreover, one might expect that the much higher dimensionless interfacial tension $\beta\gamma\sigma_L^2$ of the AB$_{13}$ phase may give rise to a much higher ΔG_c, but this is counterbalanced by a higher reduced crystal density $\rho_s\sigma_L^3$ in Eq. 3.14. On the other hand, by comparing nucleation of the fcc phase with that of the Laves phase, we find that although the dimensionless interfacial free energies $\beta\gamma\sigma_L^2$ are comparable, the difference in crystal density $\rho_s\sigma_L^3$ can yield a difference in ΔG_c. Thus,

N_{CL}	N	$\beta\Delta\mu$	$\beta P_c\sigma_L^3$	$\rho_f\sigma_L^3$	$\rho_s\sigma_L^3$	$\beta\Delta G_c$	$\beta\gamma\sigma_L^2$	$\log_{10}(J\sigma_L^5/D_L)$
2706	40334	0.095	48.40	3.383	3.778	128.3	0.994	-55.66
1977	29204	0.110	48.90	3.390	3.786	108.9	1.042	-47.47
1290	19376	0.141	49.90	3.405	3.802	91.01	1.161	-39.72
770	13692	0.178	51.10	3.423	3.820	68.38	1.234	-29.37
488	8988	0.214	52.30	3.440	3.838	52.22	1.281	-22.10
176	4746	0.344	56.70	3.499	3.897	30.29	1.482	-12.08

Table 3.2: Values of the most significant variables involved in the seeding approach calculations. Each row corresponds to a different critical nucleus size N_{CL}. See the main text for the meaning of each variable.

in order to compare the effect of interfacial tension and crystal density on the nucleation of different crystal structures, one should compare the ratio γ^3/ρ_s^2 for the various systems as this ratio is directly related to the barrier height and critical nucleus size *via* Eq. 3.12 and Eq. 3.14, and is independent of an arbitrary choice of length scale.

Finally, we observe not only much higher reduced surface tensions $\beta\gamma\sigma_L^2$ for the AB$_{13}$ phase with respect to the other examined crystals, but also a much stronger increase of γ with supersaturation. This steep rise in surface tension with $\Delta\mu$ is responsible for the flattening of the nucleation barrier and nucleation rate at high supersaturation in Fig. 3.6a and Fig. 3.6b, indicating that spontaneous nucleation, *i.e.* where the nucleation barrier is sufficiently low, may be at surprisingly high driving forces $\Delta\mu$, with respect to what we found in Chapter 2.

3.4.1 Interface detection

One of the main difficulties common to all nucleation studies is to correctly identify the location of the interface between the solid nucleus and the surrounding fluid phase. In this section we show the performance of the ANN in carrying out this task.

A common approach to classify particles according to the different thermodynamic phases is to determine the typical BOP values of all the reference phases and to select *by hand* a threshold value, called a *decision boundary*. Subsequently, the BOP values can be measured for each particle in the system, and classified according to these thresholds. In the case of a two-phase coexistence, the particles at the interface of the two coexisting phases will have BOP values which are close to these decision boundaries. Hence, the choice of these thresholds is fundamental from a quantitative point of view.

In the case of an ANN, the classification is again performed through a decision boundary. The main difference is that in this case the decision boundary is not selected manually but by the machine learning algorithm through the minimisation of a loss function evaluated for a training set.

To verify the performance of the ANN in identifying the interface, we consider a system with an approximately spherical nucleus of an AB$_{13}$ crystal surrounded by a metastable fluid as obtained from a seeding simulation at a pressure where the crystalline seed has grown out and almost doubled its original size. For each particle, we determine the probability of belonging to the AB$_{13}$ crystal phase or the probability of belonging to the fluid phase. These probabilities are calculated through the trained ANN. In Fig. 3.7b, we show a cut-through image of the grown crystal nucleus along with the ANN output for 5 exemplary particles, *i.e.* the probability that a particle belongs to the AB$_{13}$-Large, AB$_{13}$-Small, fluid, and fcc phase. From the image,

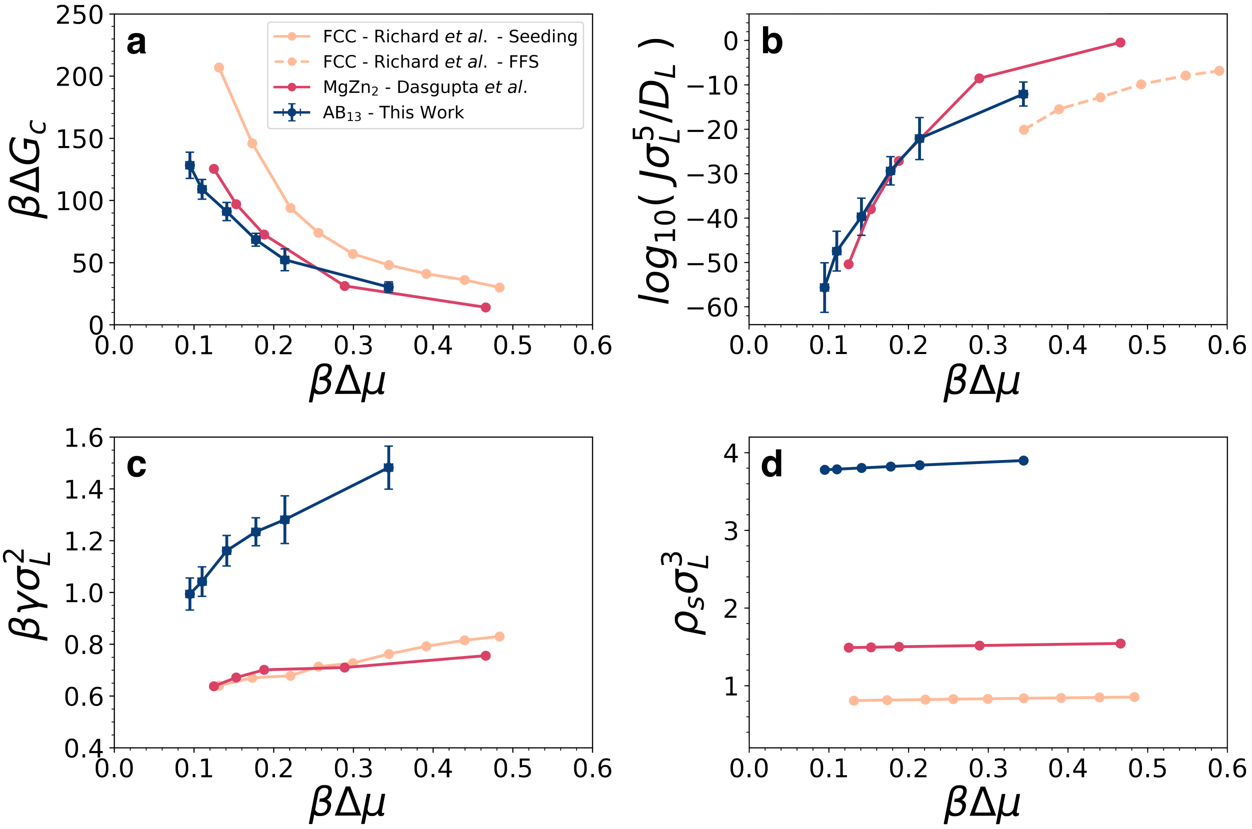


Figure 3.6: (a) Height of the Gibbs free-energy barrier $\beta\Delta G_c$, (b) nucleation rate $J\sigma_L^5/D_L$, (c) interfacial free energy $\beta\gamma\sigma_L^2$, and (d) the crystal number density $\rho_s\sigma_L^3$ as a function of the chemical potential difference $\beta\Delta\mu$ between the fluid and the AB$_{13}$ phase of a binary WCA mixture at $k_BT/\epsilon = 0.025$, with a diameter ratio $q = 0.55$ and composition $x_L = 1/14$ as obtained from the seeding approach. For comparison, we also plot the results on nucleation of the Laves phase in a binary WCA mixture from Chapter 2 and of the fcc phase in a fluid of WCA spheres from Refs. 79 and 80.

we observe that particles well-inside the crystalline nucleus are correctly classified with great certainty (probabilities around 99.9%) as belonging to the AB$_{13}$ crystal, while particles at the interface, *i.e.* where the crystal transforms into the disordered fluid phase, the ANN gives similar scores to the probabilities of being crystalline or fluid-like. In the fluid phase, the ANN correctly identifies the fluid particles with again probabilities around 99.9%.

Subsequently, we calculate the radial distance of each particle i with respect to the centre-of-mass of the nucleus, $d_{CM} = |\mathbf{r}_i - \mathbf{r}_{CM}|/\sigma_L$, where $\mathbf{r}_i$ is the position of particle i, and $\mathbf{r}_{CM}$ is the centre-of-mass position of the nucleus. In Figure 3.7a, we plot the radial averaged probability that a particle belongs to one of the two classes of the AB$_{13}$ crystal (dark blue curve) or to the fluid phase (light purple curve) as a function of the radial distance $d_{CM} = |\mathbf{r}_i - \mathbf{r}_{CM}|/\sigma_L$ from the centre-of-mass position of the nucleus. Figure 3.7a shows that particles with a radial distance $d_{CM} < 3$, which corresponds to the solid bulk of the nucleus, are indeed classified by the ANN as particles belonging to one of the two classes of the AB$_{13}$ phase with a probability close to 1. The probability that these particles belong to the fluid phase is nearly zero. Upon approaching the interface at $d_{CM} \simeq 4.3$, the probability that a particle belongs to the AB$_{13}$ phase gradually decreases, while the probability that a particle belongs to the fluid phase increases. The probability that a particle belongs to the fluid phase is about equal to the

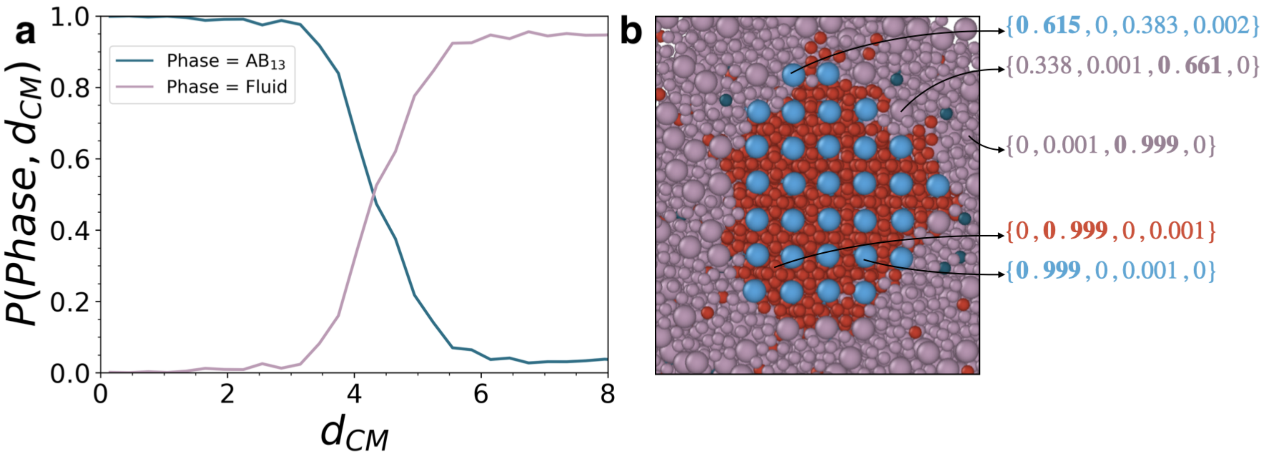


Figure 3.7: a) Probability that a particle belongs to one of the two classes of the AB_{13} crystal (dark blue curve) or to the fluid phase (light purple curve) as a function of the radial distance $d_{CM} = |\mathbf{r}_i - \mathbf{r}_{CM}|/\sigma_L$ from the centre-of-mass position of the nucleus as predicted by the ANN. The analysis has been carried out on a single snapshot of a seeding simulation performed at $\beta P \sigma_L^3 = 51.3$. In b), we show a cut-through image of the grown nucleus along with the ANN output, *i.e.* the probability to belong to the AB_{13}-Large, AB_{13}-Small, fluid, and fcc phase, for 5 exemplary particles.

probability that a particle belongs to the AB_{13} phase at $d_{CM} \simeq 4.3$, and exceeds the probability of the AB_{13} phase for $d_{CM} > 4.3$. Almost all particles are identified as fluid-like for $d_{CM} > 6$, which corresponds to the surrounding fluid phase.

In conclusion, the ANN effectively recognises the solid core of the AB_{13} phase and the surrounding fluid, as well as the interface between the two bulk phases present in the system. The interfacial width is about $2\sigma_L$ as expected. We remark that the *non-averaged* BOPs enable us to locate the interface much more accurately than the averaged BOPS as the non-averaged BOPs only take into account the first shell of neighbouring particles. Hence, we used the non-averaged BOPs in our nucleation study.

3.5 Spontaneous nucleation

Fig. 3.6a shows that the nucleation barrier $\beta \Delta G_c$ decreases with supersaturation $\beta \Delta \mu$. When $\beta \Delta G_c$ is sufficiently low, we expect to observe spontaneous nucleation of the AB_{13} phase using brute force MD simulations, *i.e.* without any non-physical biasing of the sampling of phase space.

In order to observe spontaneous AB_{13} nucleation, we initialise a system of 3024 particles with stoichiometry $x_L = 1/14$ in a highly supersaturated binary fluid phase, and perform MD simulations for a wide range of pressures in the NPT ensemble. All simulations in this section are performed using HOOMD-blue software, retaining the same settings as in Section 3.4. Using our trained ANN to identify the AB_{13} particles, we monitor and plot the size of the largest cluster N_{CL} of the AB_{13} phase as a function of time t/τ_{MD} in Fig. 3.8a. We distinguish three different regimes. At pressure $\beta P \sigma_L^3 = 71.0$, the metastable fluid does not show any sign of crystallisation within our simulation times, and hence, the supersaturation $\beta \Delta \mu \simeq 0.74$ is too low to observe spontaneous nucleation. At a slightly higher pressure $\beta P \sigma^3 = 71.4$ ($\beta \Delta \mu \simeq 0.75$), we find a critical nucleus appearing after some waiting time, which subsequently grows out in time, showing a spontaneous crystallisation event of the AB_{13} structure proceeding

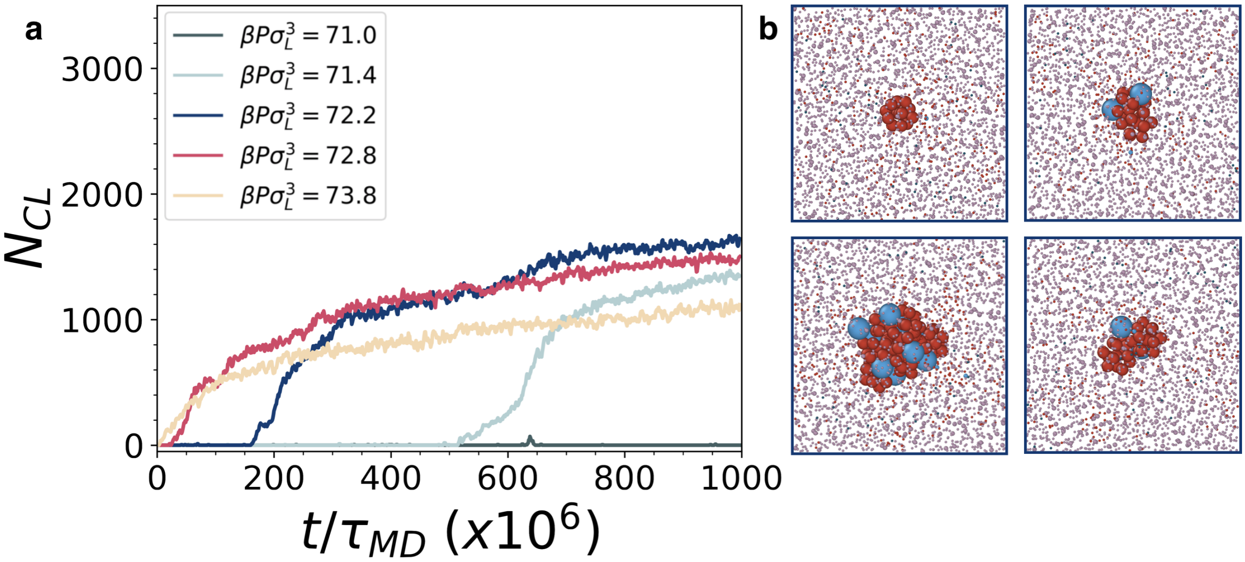


Figure 3.8: (a) Size of the largest AB_{13} cluster N_{CL} as recognised by the ANN for a binary mixture of WCA spheres at $k_BT/\epsilon = 0.025$, with a diameter ratio $q = 0.55$ and composition $x_L = 1/14$ as a function of time t/τ_{MD} for five different pressures $\beta P\sigma_L^3$ using MD simulations in the NPT ensemble. (b) Four configurations of a spontaneous nucleation event at $\beta P\sigma_L^3 = 72.2$, showing a time sequence of the early stages of AB_{13} nucleation with time increasing from the upper-left corner and then proceeding clockwise. Particles that do not belong to the main crystalline cluster have been reduced in size for visual clarity.

via nucleation. Increasing the pressure even further, we enter the third regime, where as soon as the simulation is started, multiple nuclei form immediately throughout the fluid. In this regime, the supersaturated fluid phase is mechanically unstable, and hence, crystallisation sets in immediately and exhibits spinodal-like behaviour. We thus confirm that the instability regime of the binary fluid phase with respect to AB_{13} crystallisation is at much higher driving forces $\beta\Delta\mu > 0.75$ than in the case of the Laves phases for which, as shown in Chapter 2, we found $\beta\Delta\mu > 0.53$. The significant increase of interfacial tension with supersaturation of the AB_{13} phase (Fig. 3.6c) implies a relatively slow decrease (increase) of the nucleation barrier $\beta\Delta G_c$ (nucleation rate J). Hence, spontaneous nucleation of the AB_{13} is found at surprisingly high $\beta\Delta\mu$ with respect to the Laves phases or the fcc phase. Moreover, the validation of our estimate of $\beta\Delta\mu$ where we should expect spontaneous nucleation based on the results from the seeding approach support our assumption that AB_{13} nucleation is well-described by classical nucleation theory.

In Fig. 3.8b we show a time sequence of the early stages of the spontaneous AB_{13} nucleation from the metastable binary fluid phase at $\beta P\sigma_L^3 = 72.2$ with time increasing from the upper-left corner and then proceeding clockwise. We make two remarks here. First, we observe that the ANN classification, combined with the clustering algorithm, is capable of following the nucleation process from the early stages, thereby revealing the kinetic pathways towards the formation of an embryo. Second, we find that the nucleation starts with a (defective) icosahedral cluster of small spheres around which large spheres start to order themselves on a simple cubic lattice. We thus show that the local bond orientational order of small spheres into clusters with icosahedral symmetry plays a crucial role in the kinetic pathway of the fluid-to-solid transition. This immediately begs the question whether nucleation of the AB_{13}

phase proceeds *via* a classical pathway or a non-classical two-step crystallisation scenario where relatively dense or bond orientational ordered structures in the fluid phase act as precursors for nucleation. The nucleation kinetics is of paramount importance, as the seeding approach for estimating the Gibbs free-energy barrier heights and nucleation rates is only valid in the case that nucleation proceeds *via* a classical nucleation pathway. To this end, we analyse particle configurations of spontaneous nucleation events in time. We observe that the system remains in the metastable binary fluid phase in which small crystalline nuclei appear and dissolve until a crystal nucleus of the AB_{13} phase exceeds its critical size at intermediate times and grows out. The induction time and the growth of a crystal nucleus when its size is larger than the critical nucleus size demonstrate that binary nucleation of the AB_{13} phase proceeds *via* a classical nucleation pathway. In Fig. 3.8b, we observe the spontaneous formation of a crystalline nucleus in the metastable fluid phase, which grows further when its size is larger than the critical size. This observation, together with the video that we include in the Appendix, shows that AB_{13} formation occurs *via* classical nucleation.

3.5.1 Analysis of local motifs

In order to shed light on the early stages of the AB_{13} crystal nucleation, we employ a recently developed method called topological cluster classification (TCC) algorithm to detect pre-determined particle arrangements in particle configurations [31]. The algorithm is used separately on both species, *i.e.* we take into account one species at a time. Bonds between particles are detected using a modified Voronoi construction method. The free parameter f_c, controlling the amount of asymmetry that a four-membered ring can show before being identified as two three-membered rings, is set to 0.82.

Using the TCC, we focus on two topological clusters, the square four-membered ring (sp4a) and the regular icosahedral cluster of 13 particles (13A), which are relevant particle clusters of the large and small species, respectively, in the AB_{13} crystal (see Fig. 3.9b). The fraction of particles belonging to these two clusters has on average a non-zero value in the fluid phase, and reaches one as crystallisation proceeds.

In Fig. 3.9a we plot the evolution of the fraction of large and small particles belonging to the square and the icosahedral clusters, respectively, during a spontaneous nucleation event. In order to facilitate the comparison between the two clusters, we subtract the corresponding averaged particle fraction observed in the fluid phase. Hence, the curves fluctuate around zero until nucleation occurs. interestingly, the fraction of square and icosahedral clusters increases both at the same time when nucleation occurs, and the growth behaviour of both particle clusters is similar. Hence, we conclude that the nucleation of the AB_{13} phase proceeds *via* a co-assembly process, in which the large spheres form a simple cubic lattice and the small species form the icosahedral clusters.

3.6 Conclusions

In conclusion, we have investigated homogeneous nucleation of an AB_{13} crystal in a binary fluid of hard spheres with a size ratio of $q = 0.55$ and a composition corresponding to the stoichiometry of the AB_{13} phase. To achieve this, we have trained a neural network using a large set of non-averaged bond order parameters as input to distinguish the AB_{13} phase from all competing phases, *i.e.* the binary fluid and the fcc phase, with significant accuracy in the case

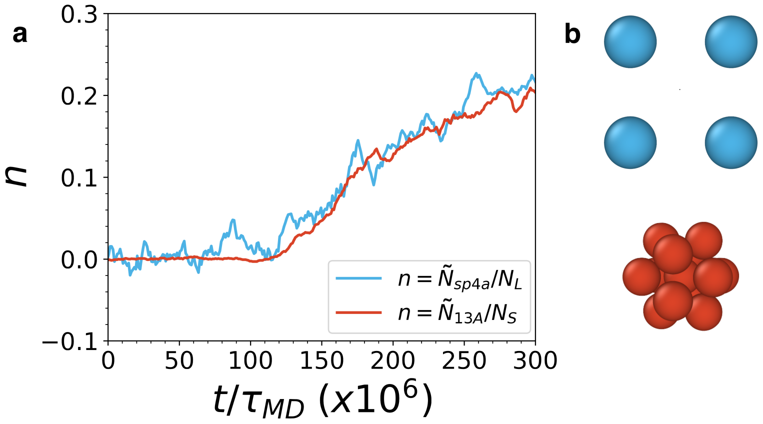


Figure 3.9: a) Normalised number of large $\tilde{N}_{sp4a}/N_L$ (small $\tilde{N}_{13A}/N_S$) particles belonging to a *sp*4*a* (13*A*) cluster as a function of time $t/\tau_{\rm MD}$ in a spontaneous nucleation event, corrected for the averaged number observed in the fluid phase. The large statistical fluctuations in $\tilde{N}_{sp4a}/N_L$ is due to a much lower number of large species in the system. b) A sketch of the square four-membered ring *sp*4*a* (blue) and the 13*A* (red) icosahedral cluster.

of bulk phases. We showed that using two (averaged) bond order parameters is not sufficient to identify the different phases of interest on a single-particle level, while an artificial neural network with two hidden layers provides an elegant and powerful way of combining a high number of bond order parameters and to successfully distinguish the different phases with high accuracy.

Using the trained neural network as an order parameter in our nucleation study, we were able to follow crystal nucleation of the AB$_{13}$ phase in a supersaturated binary fluid phase. We used the seeding approach to calculate Gibbs free-energy barriers and nucleation rates without prior knowledge about the system. Subsequently, we made a comparison of the nucleation of the AB$_{13}$ phase with another binary hard-sphere crystal, the Laves phase, and with a single-component fcc phase of pure hard spheres. Our key findings are that 1) the assumption that the nucleation barrier for binary nucleation is higher due to a loss of mixing entropy is incorrect, *e.g.* the barrier for the pure fcc phase is higher than for the AB$_{13}$ and the Laves phases in the case of hard spheres at the same thermodynamic driving force $\Delta\mu$, and that 2) the assumption that the nucleation barrier is high due to a high interfacial free energy is not always valid as it also depends on the number density of the solid phase, *e.g.* the reduced surface tensions $\beta\gamma\sigma_L^2$ for the fcc and the Laves phase are very similar, but fcc has a higher barrier height due to a lower reduced solid density $\rho_s\sigma_L^3$. Hence, in order to compare the effect of interfacial tension and crystal density on the nucleation of different crystal structures, one should compare the ratio γ^3/ρ_s^2 for the various systems as this ratio is directly related to the barrier height and critical nucleus size *via* Eq. 3.12 and Eq. 3.14, and is independent of an arbitrary choice of length scale.

Subsequently, we used the dependence of the nucleation barrier height $\beta\Delta G_c$ on supersaturation $\Delta\mu$ to obtain an estimate of $\Delta\mu$ where spontaneous nucleation should occur. In this way, we were able to observe spontaneous nucleation of the AB$_{13}$ phase using brute force MD simulations. To shed light on the nucleation mechanism, we analysed the spontaneous nucleation events by measuring the fraction of large particles belonging to a square shortest-path four-membered ring (sp4a) and the fraction of small particles belonging to an icosahedral (13*A*) cluster. We observed a similar growth behaviour of both clusters, demonstrating that the AB$_{13}$

nucleation proceeds *via* a co-assembly process. This finding is corroborated by our analysis of the early stages of nucleation when the first embryo forms using the trained neural network as an order parameter. Fig. 3.8b shows clearly that the embryo grows by the attachment of both large particles and icosahedral clusters of small particles. Our results show that AB_{13} crystal nucleation proceeds *via* a classical pathway that can be well-described by CNT, even though the binary fluid is highly structured. Due to thermal fluctuations, the local regions of high bond orientational order and many-body correlations appear and disappear in the metastable fluid phase. The crystal only forms when also the large species are coordinated in the right way around the icosahedral clusters, thereby making AB_{13} nucleation possible.

Finally, our method for the identification of local structures using a neural network can straightforwardly be extended to other crystal structures, liquid crystal phases, and glasses, and can be employed for future nucleation studies, analysing not only numerical, but also experimental data stacks.

Acknowledgements

I would like to thank Jayden Savage for a preliminary study of the AB_{13} crystal nucleation.

3.7 Appendix

3.7.1 Multiple bond order parameters: the Shape Matching algorithm

The exemplary 2D projection of the (averaged) bond order parameters in Fig. 3.3 reveals that two bond order parameters are not sufficient to achieve a satisfactory classification. An artificial neural network (ANN) is an efficient and optimised way of using multiple descriptors in the classification. In order to provide a more local description of a particle, we employed 36 *non-averaged* bond order parameters (BOP) as descriptors, even though they offer a poorer characterisation of the particles with respect to their *averaged* counterparts. An ANN combines these parameters by giving them different weights in the classification and adding non-linearity if one or more hidden layers are present in the chosen ANN architecture. However, there are also other techniques, outside the realm of machine learning, that are based on more than 2 descriptors in the classification. One of these methods is the shape matching technique (SM) [170].

The SM technique requires a set of *shape descriptors* or order parameters that can identify on a single-particle level the different phases we wish to classify. For each particle i in the system, we build a vector $\vec{\mathbf{S}}_i$ consisting of all the values of the shape descriptors for particle i. Subsequently, we determine the reference vector $\vec{\mathbf{S}}_A$ of all the shape descriptors for the respective phases, *e.g.* phase A, we want to identify. By employing a *similarity metrics*, we find the reference vector that is most similar to the one determined for particle i, and particle i will be classified accordingly.

In order to compare the performance of the SM algorithm with the ANN used in this work, we define the shape descriptor vector to be the same as the input vector of the ANN, which is an array of 36 non-averaged BOPs (Eq. 5.5).

Subsequently, we determine the reference vectors for the four classes that we wish to identify. Instead of using the values of the shape descriptors for the perfect crystalline structures, we take the average values of the BOPs for each class. This approach corresponds to selecting the central point of the cloud in the 36-dimensional BOP space of each phase. We find that this choice leads to better results in the classification of particles than using the perfect crystalline structures. Moreover, it can be straightforwardly be extended to non-crystalline structures—the fluid phase in this case.

Finally, a similarity metrics has to be defined that quantifies the similarity of the shape descriptor vector $\vec{\mathbf{S}}_i$ of particle i and the reference vector $\vec{\mathbf{S}}_A$ of phase A, $M(\vec{\mathbf{S}}_i, \vec{\mathbf{S}}_A)$. The most common metrics is the Euclidean distance

$$M_{dist}(\vec{\mathbf{S}}_i, \vec{\mathbf{S}}_A) = 1 - [(\vec{\mathbf{S}}_i - \vec{\mathbf{S}}_A)/(|\vec{\mathbf{S}}_i| + |\vec{\mathbf{S}}_A|)]. \tag{3.15}$$

Another choice is based on the projection of vector $\vec{\mathbf{S}}_i$ onto the reference vector $\vec{\mathbf{S}}_A$

$$M_{proj}(\vec{\mathbf{S}}_i, \vec{\mathbf{S}}_A) = 1/2[1 + (\vec{\mathbf{S}}_i \cdot \vec{\mathbf{S}}_A)/(|\vec{\mathbf{S}}_i||\vec{\mathbf{S}}_A|)] \tag{3.16}$$

We find that M_{proj} gives the best results. The accuracy of the SM method based on M_{proj} is reported in Table 3.3 along with the accuracy of the ANN used in this work.

Although the accuracy is high for *e.g.* the large species of the AB_{13} crystal, the overall accuracy is not satisfactory enough for the purpose of this work. The reason is that in the SM technique each BOP is equally important as the classification is simply based on a dot-product.

Class	SM	ANN
AB_{13} - Large	100.0%	100.0%
AB_{13} - Small	89.8%	98.1%
Fluid	91.8%	97.8%
fcc	93.2%	99.4%

Table 3.3: Accuracies of the SM technique and the ANN for all four classes.

Class	$N_{hidden} = 0$	$N_{hidden} = 1$	$N_{hidden} = 2$	$N_{hidden} = 3$
AB_{13} - Large	100.0%	100.0%	100.0%	100.0%
AB_{13} - Small	95.9%	97.4%	98.1%	99.2%
Fluid	94.1%	96.3%	97.8%	96.9%
fcc	97.9%	99.1%	99.4%	99.6%

Table 3.4: Accuracies for all four classes of the ANNs with varying number of hidden layers N_{hidden}.

Consequently, if a BOP is very noisy in the training set, it will obfuscate the classification. The remedy for this problem is to employ different weights for different BOPs, based on their importance. This is exactly what the ANN learns during the training phase, when the weights are tuned.

3.7.2 Details on the artificial neural network

Choice of architecture

In this Chapter, we employ an ANN composed of two hidden layers, in addition to the input and output layer. In principle, multiple architectures are possible and the choice of the number of hidden layers requires a trial-and-error tuning [155, 171]. In particular, adding hidden layers to the neural network means that more hyperparameters (namely weights and biases) have to be tuned during the training phase, which allow for a better total score on the training set, but may also result in a higher chance of *overfitting* the data. The latter is signaled by the score on the validation set, which is not as high as the score on the training set.

In order to select the architecture of the ANN, we trained multiple ANNs, differing from each other solely by the number of hidden layers. In Table 3.4 we show the accuracy of each of these ANNs on the test set, while in Fig. 3.10 we display the total accuracy as a function of the number of hidden layers N_{hidden}.

We observe from Fig. 3.10 that, even though the total accuracy of the ANN increases with the addition of up to 3 hidden layers, the third layer does not constitute a significant improvement in the performance of the ANN. We therefore decided to employ an ANN with 2 hidden layers as a compromise between efficiency and accuracy.

Fluid particles

When building the training and validation sets to train the ANN, one of the first questions to ask is how many classes the ANN should identify. This choice depends also on the features that are used in the input layer, as they must be able to distinguish the different classes in order to achieve a successful score at the end of the training phase.

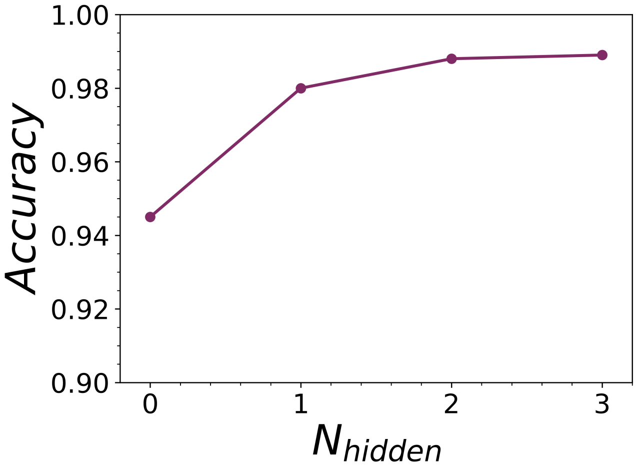


Figure 3.10: Total accuracy of the ANNs as a function of the number of hidden layers N_{hidden}. The total accuracy increases significantly up to 3 hidden layers, but the addition of the third hidden layer does not constitute a significant improvement in the performance of the ANN.

Class	ANN
AB_{13} - Large	100.0%
AB_{13} - Small	98.3%
Fluid - Large	99.9%
Fluid - Small	97.3%
fcc	99.3%

Table 3.5: Accuracies of the ANN trained to distinguish the small-species and large-species binary fluid particles.

In this work, we have selected 4 target classes: large species of the AB_{13} crystal, small species of the AB_{13} crystal, fcc particles and fluid particles. Regarding the fluid particles, one can make a further distinction based on species. This choice would split the fluid particles into large-species and small-species fluid particles, leading to a total of 5 different classes.

This different choice is valid based on the assumption that in the binary fluid with stoichiometry $x_L = 1/14$, large and small species have different local environments. This assumption is indeed correct, as can be seen from Fig. 3.11, where we plot the probability distribution functions of 4 exemplary BOPs, namely q_4, q_6, w_4 and w_6 for the large and small species of the binary fluid phase separately. We clearly see from Fig. 3.11 that almost all probability distribution functions overlap considerably, hampering the distinction of the two classes. Nevertheless, the two q_6 distribution functions are well separated for the large and small species, allowing for the possibility of selecting an output layer with 5 different classes.

To this end, we build a second training set with 10^5 large-species and 10^5 small-species fluid samples. We note that in the previous training set, the ratio between large-species fluid samples and small-species fluid samples was approximately equal to the stoichiometry of the system under investigation, and hence the number of large-species fluid samples was much lower than for the small-species counterpart. In Table 3.5 we show the accuracy of the ANN when the fluid particles are distinguished in large and small species.

We observe from Table 3.5 that the ANN correctly classifies almost all the large species of the fluid phase, whereas the accuracy of the large species of the AB_{13} crystal and of the fcc

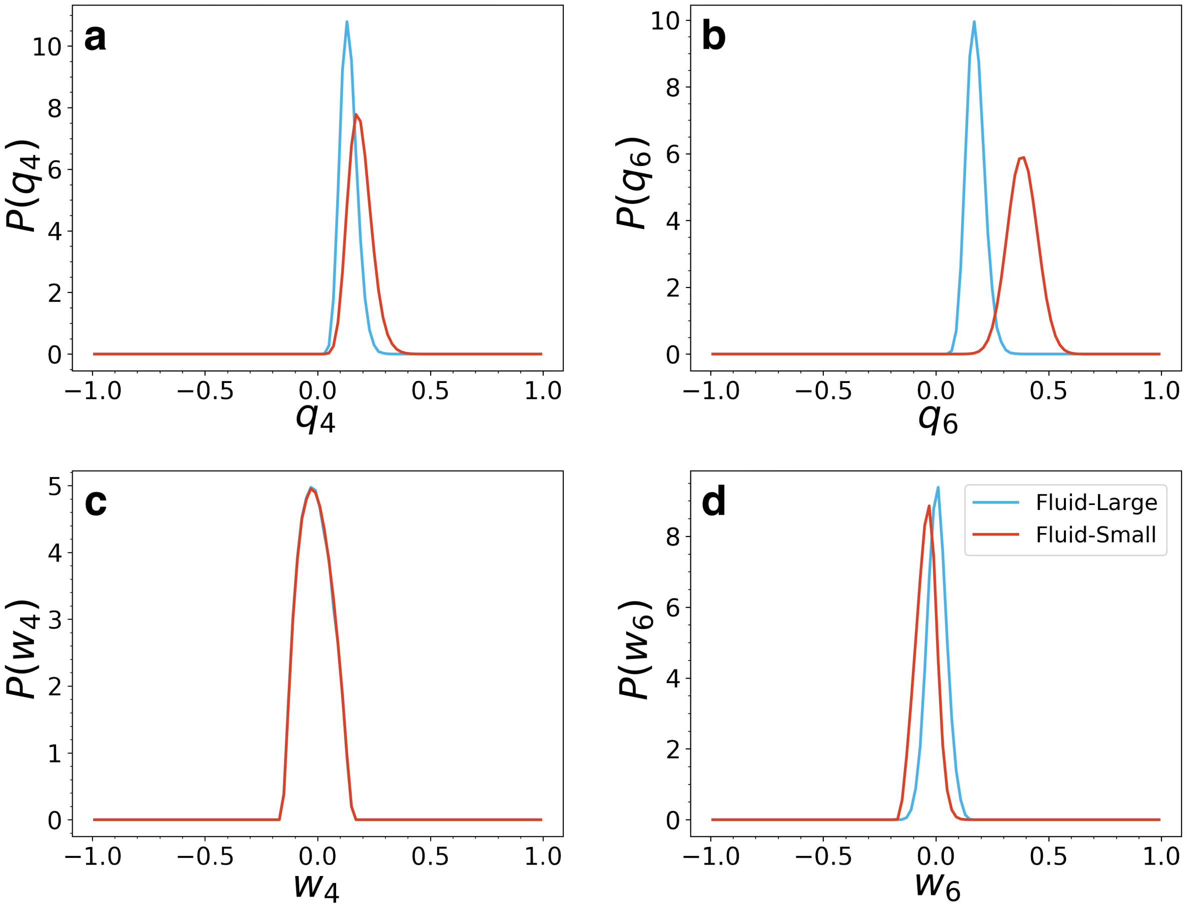


Figure 3.11: Probability distribution functions of 4 representative BOPs for large species of the fluid phase (light blue curves) and small species of the fluid phase (red curves). In particular, we plot a) $P(q_4)$, b) $P(q_6)$, c) $P(w_4)$, and d) $P(w_6)$. The majority of the BOPs show overlapping distributions (in the case of w_4 they are hardly distinguishable), but the fact that there is at least one parameter (q_6) for which the distributions are well separated guarantees that the distinction of small-species and large-species fluid particles is possible.

crystal remains unchanged. The similarity of the local environments of the small species of AB_{13} crystal and of the small species of the fluid phase is still signaled by the relatively low score of the ANN, and hence a correct classification of the small species remains a challenge.

In conclusion, we have shown that an ANN with five target classes involving a further distinction of the fluid particles is equally accurate as the ANN with four classes as employed in the present study, and represents a suitable alternative choice.

Information provided by the network

As shown by our previous analysis, we clearly observe that our trained ANN constitutes a useful tool to distinguish with high accuracy the AB_{13}, binary fluid, and fcc phase on a single-particle level by using input vectors composed of 36 (non-averaged) BOPs. Hence, the ANN can straightforwardly be used as an order parameter in nucleation studies. However, such a black box implementation of the ANN can be unsatisfactory in terms of understanding how it

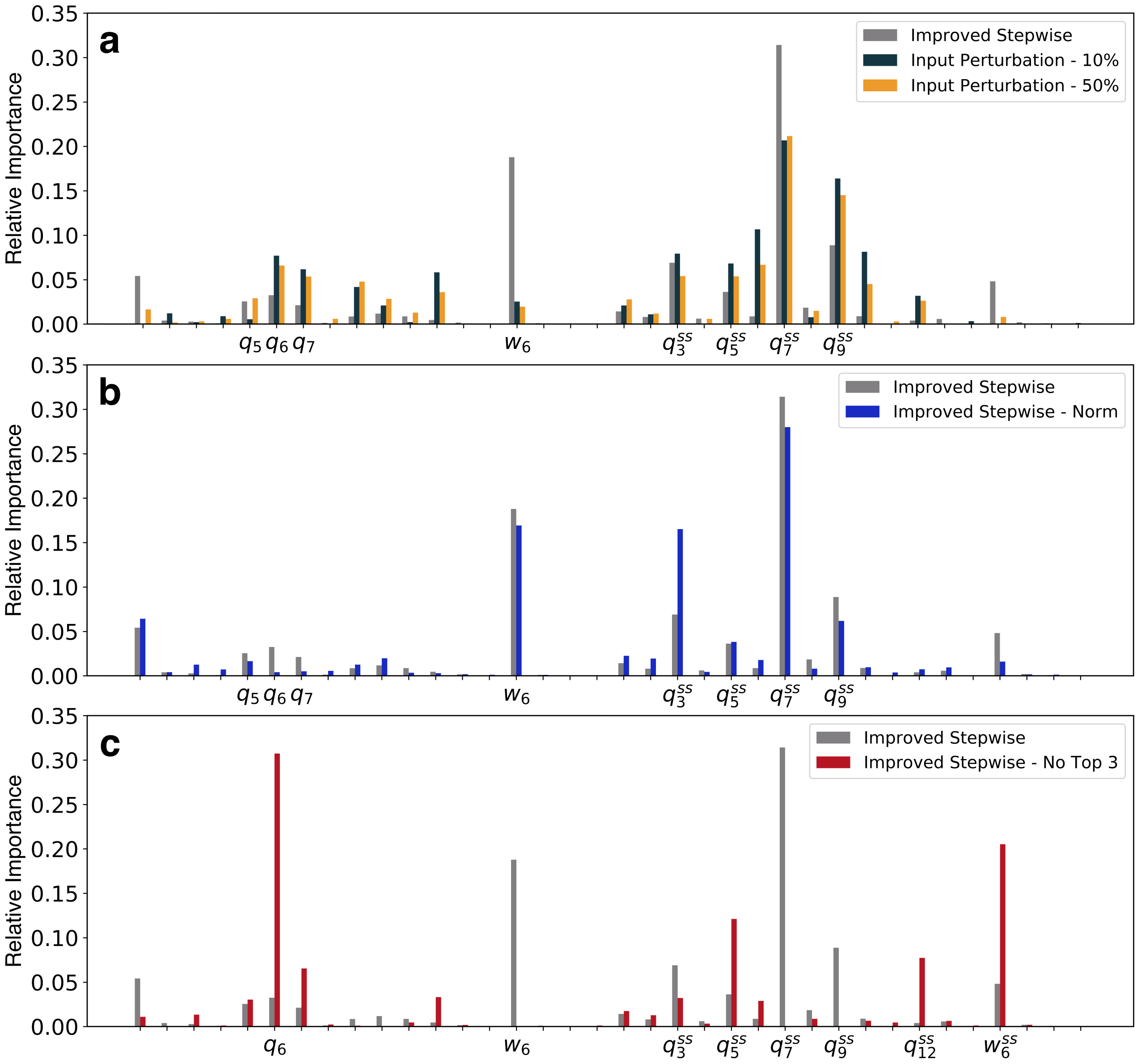


Figure 3.12: a) The relative importance RI(k) of feature k of our input vector (Eq. 5.5 in the main text) describing the local particle environment as determined by the improved stepwise technique and the input perturbation method using Gaussian white noise with a 10% and 50% variance. b) Application of the improved stepwise method on the original data set (grey) and a normalised data set (blue) such that each feature has a distribution with zero mean and unit variance. c) Application of the improved stepwise method of the original ANN (grey) and a second one trained on all the features with the exception of the top 3 most important ones (red).

makes decisions. We therefore investigate in more detail which features the ANN prioritises to distinguish the different classes. To this end, we apply the improved stepwise [172, 173] and the input perturbation [174, 175] techniques in order to determine the relative importance (RI) of all features.

In the improved stepwise technique, we determine the relative importance (RI) of a feature by replacing each value in the data set with its arithmetic mean, while keeping the remaining 35 features unchanged. Subsequently, we calculate the accuracy of the ANN on this data set.

If the change in the accuracy is large, the relative importance of this feature is high in making a prediction.

The input perturbation method works in a similar manner. Instead of replacing each value of a particular feature with its mean, we add a fixed amount of white noise, and then again assess the change in the total accuracy. The noise is sampled from a uniform distribution, centred in zero and whose interval is equal to twice a fixed percentage of the input value. We test this method twice, selecting the above percentage to be equal first to 10% and then to 50% of the input value.

In both methods, we repeat this process for all 36 input features, and calculate the relative importance of the k^{th} feature RI(k) using

$$\mathrm{RI}(k) = \frac{\Delta \mathrm{Acc}(k)}{\sum\limits_{k=1}^{36} \Delta \mathrm{Acc}(k)}, \tag{3.17}$$

where ΔAcc(k) is the difference in accuracy after applying one of the two methods to the k^{th} feature.

In Fig. 3.12a, we show our results as obtained by applying the two methods on the entire data set. It is interesting to note that the ANN does not solely prioritise even symmetries, which often play a decisive role in crystal structure recognition. In fact, odd symmetries - fivefold, sevenfold and ninefold - heavily determine the learning process of the neural network, as shown by the high RI values of the corresponding order parameters. We attribute this behaviour to the unique crystal structure of the AB$_{13}$ phase consisting of icosahedral small-sphere clusters.

However, due to the different nature of the two algorithms, the RI(k)'s vary - sometimes considerably - for the two methods. This analysis is therefore useful solely to understand which features the ANN relies the most on. As further proof, we train another ANN with the same architecture but such that the probability distribution of each feature has zero mean and unit standard deviation. Usually, this pre-processing is applied in cases where the input vector is composed of features that vary by orders of magnitude, facilitating in this way the training of the model [171]. In this work, this pre-processing has not been applied because all the BOPs vary approximately within the same range. We again apply the improved stepwise method and compare the results with the improved stepwise method applied on the original features in Fig. 3.12b. We clearly observe that the values of the relative importance vary due to differences in the numerical values of the normalised and unnormalised features, but the features that ANN recognises as the most important for the classification remain the same.

Finally, one may ask the question whether a smaller set of BOPs can also provide a satisfactory classification. We therefore trained two additional ANNs, one for which we use only the top three features picked up by the improved stepwise method as the most relevant ones, namely w_6, q_7^{ss}, and q_9^{ss}, as input vector, and one ANN for which we employ all the other features apart from these top three features as input.

In the first case using only the top 3 features we get a total accuracy of about 93%. This result demonstrates the importance of these 3 BOPs in order to obtain a satisfactory classification, even though it cannot replace the ANN trained on all 36 features.

In the second case using the remaining 33 features, the trained ANN achieves an accuracy almost identical to that of the original ANN (98.7% instead of 98.8%). This result shows that there are several ways to obtain a satisfactory classification of the different phases. By removing the top three relevant features, the ANN will give priority to other features, equally valid, as shown in Fig. 3.12c. This different set of relevant features can sometimes be highly correlated

Class	RF	ANN
AB_{13} - Large	100.0%	100.0%
AB_{13} - Small	96.6%	98.1%
Fluid	97.4%	97.8%
fcc	98.1%	99.4%

Table 3.6: Accuracies of the RF and the ANN for all four classes.

with the removed ones, as in the case of w_6^{ss}, which has a Pearson correlation value with w_6 equal to 0.87.

3.7.3 Comparison with Random Forest

In the previous sections we have demonstrated that an ANN successfully classifies particles according to the different thermodynamic phases in the system. We have also compared the performance of the ANN with a technique outside the realm of machine learning, namely the SM technique. The performance of the ANN is superior to that of the SM technique. However, the variety of machine learning techniques is vast, and the ANN is only one of them. We therefore also compare the performance of the ANN with another machine learning method.

In this section we study a second state-of-the-art machine learning algorithm, namely Random Forest (RF), and evaluate its accuracy on the same data set used for the ANN. This allows us to determine which technique works best for the purpose of this work.

RF is a machine learning algorithm with a philosophy that is very different from that of an ANN. RF, in its classification setting, consists of an ensemble of smaller classifiers called Decision Trees (DTs) [171]. Each of these DTs takes a given input, and applies a series of *if-else* conditions based on the value of the input vector features. These conditions allow the tree to branch out and finally return a result, the predicted class of the sample. In an RF, multiple DTs are trained using different parts of the training set and of the different features. For each sample to be classified, each DT votes for a certain class based on its prediction. The RF returns the class that receives the most votes. RF often succeeds not only in providing satisfactory accuracy results, but also on shedding light on the predictive ability of individual features.

To offer a comparison with the ANN used in this study, we train a RF model using 100 independent DTs. An important parameter that must be tuned is the depth of each tree, *i.e.* the number of consecutive *if-else* statements prior to a classification outcome. In fact, if this parameter is too high, RF is likely to overfit the data. We find that the maximum depth of the model for which we do not overfit the training set is equal to 9. Below we show the accuracy for each class obtained using RF, together with the results of the ANN used in this work.

Table 3.6 shows that the RF produces excellent results, although slightly lower than the ANN, particularly in the classification of fluid particles and small species of the AB_{13} crystal.

Depending on the specific application, RF may be preferred when accuracy is not the only parameter to take into account, *e.g.* with RF it is possible to obtain a very fast training of the model and a better interpretability of the classification. In this work, we require a method that is as accurate as possible as the number of solid particles in the crystal nucleus as predicted by the machine learning method determines the thermodynamics of the nucleation process. We therefore decided to employ the ANN in our nucleation study.

4

Hidden in the fluid: fivefold symmetry and polymorph selection in hard spheres

In this Chapter, we study the crystal nucleation of hard spheres for which, despite numerous experimental and theoretical studies, many aspects are far from being understood. We show that the excess of face-centred-cubic (fcc)-like particles with respect to hexagonal-close-packed (hcp)-like particles in a crystal nucleus of hard spheres as observed in simulations and experiments can be explained by the higher order structure in the fluid phase. We show that, in the metastable fluid phase, fivefold symmetry clusters – pentagonal bipyramids (PBs) – known to be inhibitors of crystal nucleation, transform into a different cluster – Siamese dodecahedra (SDs). These clusters, because of their geometric arrangement, form a bridge between the fivefold symmetric fluid and the fcc crystal, thus lowering its interfacial free energy with respect to the hcp crystal, and shedding light on the polymorph selection mechanism.

Based on: G. M. Coli, R. van Damme, C. P. Royall, and M. Dijkstra, *Hidden in the fluid: fivefold symmetry and polymorph selection in hard spheres*, Manuscript submitted (2021)

4.1 Introduction

Understanding nucleation is important in many research fields such as determining the molecular structure of proteins through crystallisation, drug design in the pharmaceutical industry, ice crystal formation in clouds — the largest unknown in the earth's radiative balance and thus crucial in the context of climate change and weather forecasts — and crystallisation of colloidal and nanoparticle suspensions with application perspectives in catalysis, opto-electronics, and plasmonics [176, 177].

However, nucleation is extremely challenging to study in molecular systems as it is a stochastic and rare process, the sizes of the crystal nuclei are often rather small, and the nuclei grow out extremely fast when they exceed their critical size. An additional obstacle is that, for most substances, different crystal polymorphs may compete during nucleation. This phenomenon is of key importance as the crystallisation of the "undesired" polymorph may for instance lead to neurodegenerative disorders such as Alzheimer's disease or eye cataract, or to reduced solubility/efficacy and even toxicity of certain drug compounds [178, 179].

Recently, impressive strides have been made in the experimental observation of early-stage crystal nucleation by using atomic-resolution in situ electron microscopy, showing the observation of different nucleation pathways to different crystal polymorphs of proteins [180], prenucleation clusters in metal organic frameworks [181], early-stage nucleation pathways of FePt nanocrystals that go beyond classical nucleation theory and current non-classical scenarios [182], amorphous precursors in protein crystallisation [183], featureless and semi-ordered clusters of NaCl nanocrystals [184], and reversible disorder-order transitions of gold crystals [185]. These recent observations differ from the current nucleation models and call for a better theoretical insight in the crystallisation pathways at the earliest stages of nucleation, when particles start to order from the metastable fluid phase and select the emerging crystal polymorphs.

Colloidal suspensions are suitable experimental systems to probe locally heterogeneous phenomena such as early-stage nucleation: the larger sizes and slower time scales of colloidal particles enable direct observation of the nucleation mechanisms [24,186]. However, even for colloidal hard spheres (HSs), undoubtedly one of the simplest colloidal model systems, the polymorph selection mechanism is yet to be revealed. In a HS system, the hcp crystal is metastable with respect to the fcc, but the free-energy difference between the two structures is tiny ($\simeq 10^{-3}k_BT$ per particle) [187, 188]. One therefore might expect to find an approximately 50% occurrence of fcc- and hcp-like particles in the crystal nucleus of hard spheres. This prediction is, however, not confirmed either in experiments [24, 186, 189, 190] nor in simulations [117, 191–194], which both show a hitherto unexplained predominance of fcc particles in the final crystal phase.

In this Chapter, we investigate, using Molecular Dynamics (MD) simulations and particle-resolved studies of colloids, the early stages of nucleation of hard spheres in order to shed light on the selection mechanism of the crystal polymorph. We study the structural transformations in the supersaturated fluid phase that finally lead to crystal nucleation. We find that the crystal embryo shows a preference towards fcc-like stacking, because of the striking similarity with local clusters present in the fluid phase. We also demonstrate that this purely geometric argument for a higher propensity to nucleate fcc is incorporated in thermodynamics by a lower interfacial free energy of fcc with respect to hcp crystals.

4.2 Simulation details

In order to generate trajectories in which we observe spontaneous nucleation, we conduct MD simulations in the isothermal-isobaric (NPT) ensemble with a constant number $N = 13500$ of nearly-hard spheres. The particles interact *via* a Weeks-Chandler-Andersen (WCA) pair potential, which can straightforwardly be employed in Molecular Dynamics (MD) simulations and which reduces to the hard-sphere potential in the limit that the temperature $T \to 0$. The WCA potential $u(r_{ij})$ reads [75]

$$u\left(r_{ij}\right) = \begin{cases} 4\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6} + \frac{1}{4}\right] & r_{ij} < 2^{\frac{1}{6}}\sigma \\ 0 & r_{ij} \geq 2^{\frac{1}{6}}\sigma, \end{cases} \tag{4.1}$$

with $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ the center-of-mass distance between particle i and j, $\mathbf{r}_i$ the position of particle i, ϵ the interaction strength, and σ the diameter of each sphere. The steepness of the repulsion between the particles can be tuned by the temperature k_BT/ϵ. We set $k_BT/\epsilon = 0.025$, which has been used extensively in previous simulation studies to mimic hard spheres [77–80].

The temperature T and pressure P are kept constant *via* the Martyna-Tobias-Klein (MTK) integrator [92], with the thermostat and barostat coupling constants $\tau_T = 1.0\ \tau_{MD}$ and $\tau_P = 1.0\ \tau_{MD}$, respectively, and $\tau_{MD} = \sigma_L\sqrt{m/\epsilon}$ is the MD time unit. The time step is set to $\Delta t = 0.004\tau_{MD}$, which is small enough to ensure stability of the simulations. We ran the simulations for $10^9\tau_{MD}$ time steps, unless specified otherwise. The simulation box is cubic and periodic boundary conditions are applied in all directions.

We select the pressure values in a region of metastability that allow us to observe nucleation phenomena on reasonable time scales. Specifically, the reduced pressure varies in the range $\beta P\sigma^3 \in [13.40, 16.00]$, which results in numerous spontaneous crystallisation events. All MD simulations are performed using the HOOMD-blue (Highly Optimised Object-oriented Many-particle Dynamics) software [195].

4.3 Classification scheme: A synergetic combination

In this Chapter, it is necessary not only to distinguish fluid-like particles from crystal-like one, but it is also crucial to monitor the behaviour of local clusters and their relationship with the crystal nucleus. In order to do so, we make use of two separate classification algorithms, which we describe in this Section.

4.3.1 Bond Order Parameters

To describe the local environment of a particle, we employ the standard bond-orientational order parameters (BOPs) introduced by Steinhardt *et al.* [86]. We first define the complex vector $q_{lm}(i)$ for each particle i

$$q_{lm}(i) = \frac{1}{N_b(i)} \sum_{j=1}^{N_b(i)} Y_{lm}(\theta(\mathbf{r}_{ij}), \phi(\mathbf{r}_{ij})), \tag{4.2}$$

where $N_b(i)$ is the number of neighbours of particle i, $Y_{lm}(\theta(\mathbf{r}_{ij}), \phi(\mathbf{r}_{ij}))$ denotes the spherical harmonics, $m \in [-l, l]$, $\theta(\mathbf{r}_{ij})$ and $\phi(\mathbf{r}_{ij})$ are the polar and azimuthal angles of the distance vector $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$, and $\mathbf{r}_i$ denotes the position of particle i.

The averaged $\bar{q}_{lm}(i)$ as introduced by Lechner and Dellago is defined as

$$\bar{q}_{lm}(i) = \frac{1}{\tilde{N}_b(i)} \sum_{j=1}^{\tilde{N}_b(i)} q_{lm}(j), \tag{4.3}$$

where $\tilde{N}_b(i)$ is the number of neighbours including particle i itself [87]. The rotationally invariant quadratic and cubic averaged bond order parameters are defined as

$$\bar{q}_l(i) \;=\; \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |\bar{q}_{lm}(i)|^2}, \tag{4.4}$$

and

$$\bar{w}_l(i) = \frac{\sum\limits_{m_1+m_2+m_3} \begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} \bar{q}_{lm_1}(i)\bar{q}_{lm_2}(i)\bar{q}_{lm_3}(i)}{\left(\sum\limits_{m=-l}^{l} |\bar{q}_{lm}(i)|\right)^{3/2}}. \tag{4.5}$$

To identify the neighbours of particle i we employ the parameter-free solid-angle-based nearest-neighbour (SANN) algorithm of Van Meel [152]. This algorithm assigns a solid angle to every potential neighbour j of i, and defines the neighbourhood of particle i to consist of the $N_b(i)$ particles nearest to i for which the sum of solid angles equals 4π.

4.3.2 Topological Cluster Classification

Differently from BOPs, in order to perform an analysis which is not solely based on local symmetries, we require an algorithm that is capable of successfully finding different topological clusters in a metastable fluid. To this end, we employ the Topological Cluster Classification (TCC) algorithm [31]. The bonds between particles are detected using a modified Voronoi construction method. The free parameter f_c, controlling the amount of asymmetry that a four-membered ring can show before being identified as two three-membered rings, is set to 0.82.

4.4 Results

4.4.1 Siamese Dodecahedra and Pentagonal Bipyramids

We perform MD simulations to study crystal nucleation in a supersaturated fluid of hard spheres. We generate many nucleation events and analyse the trajectories using two different methodologies. To follow the nucleation process, we first identify the *solid-like* particles, *i.e.* particles with a local solid-like (ordered) environment, by calculating the averaged bond order parameters [87] that are based on spherical harmonics Y_{lm}, measuring the arrangement of the neighbours around a particle, as described in Section 4.3.1. In particular, we identify particle i as solid-like if the sixfold rotational invariant $\bar{q}_6(i) \geq 0.31$, and we colour them blue in Fig. 4.1c-f. We note that this classification scheme acts on a single-particle level.

To investigate the structure of the fluid, we determine the topologies of various particle clusters present in the system using the Topological Cluster Classification (TCC) algorithm [31],

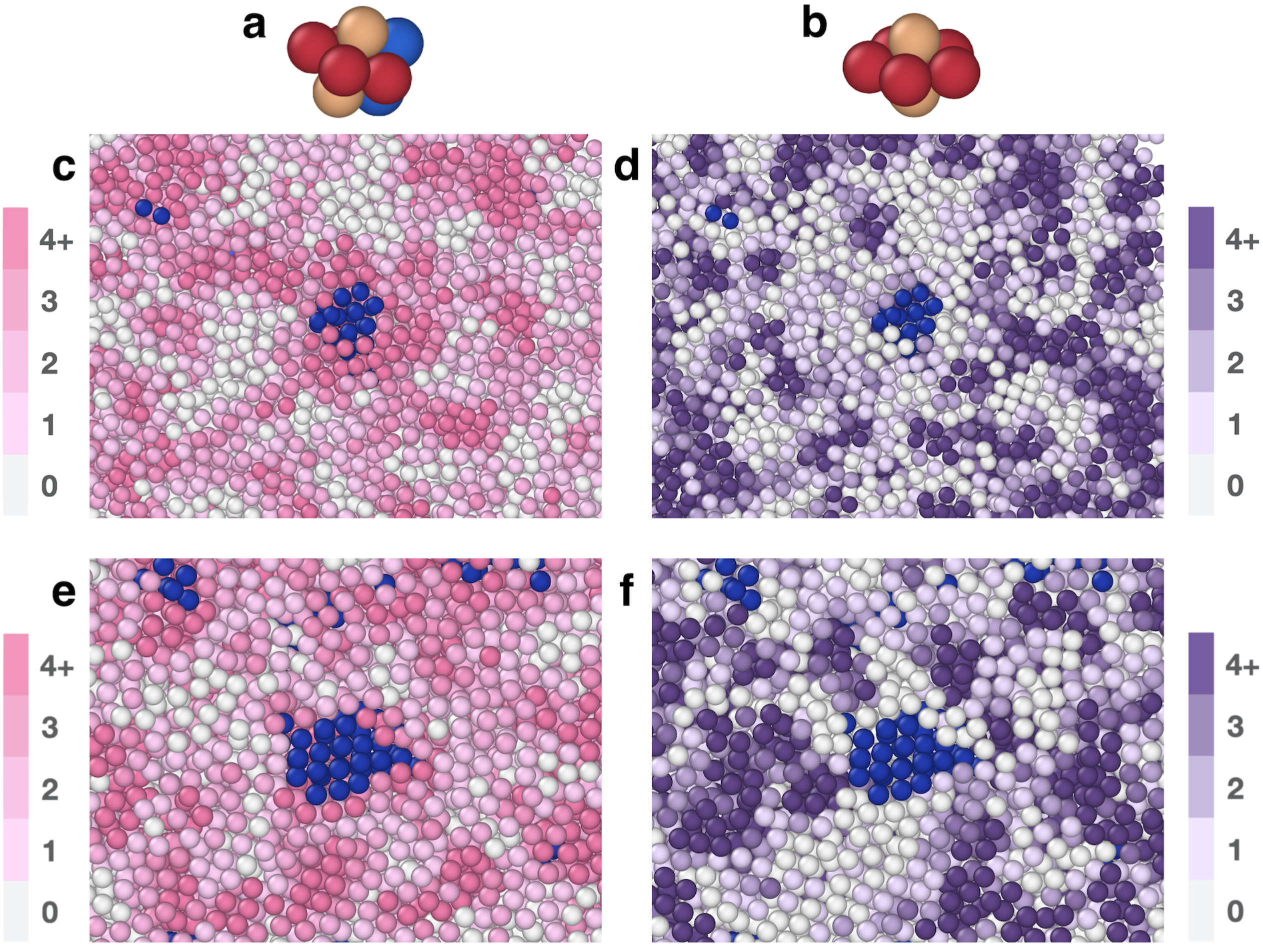


Figure 4.1: Typical configuration of a crystal nucleation event of hard spheres. (a-b) Topological arrangement of particles in (a) a Siamese Dodecahedron (SD) and (b) Pentagonal Bipyramid (PB) cluster. The colour coding is explained in the text. (c-d) Cut-through image of an early-stage nucleation event generated by MD simulations. Crystal-like particles are coloured blue, while fluid-like particles are coloured following the scale bar on the left (c) or right (d) depending on the number of SD (c) or PB (d) clusters each particle belongs to. The same colour coding of (c) and (d) is used in (e) and (f), respectively, where we analyse an experimental configuration of PMMA spheres at packing fraction $\phi = \frac{\pi}{6}\rho\sigma^3 = 0.51$, where ρ is the number density of the system, while σ is the measured diameter of the particles.

as described in Section 4.3.2. We identify local clusters of 3 up to 13 particles consisting of not only rings of three, four, and five particles with and without additional neighbouring particles, but also compounds of these basic clusters. In total, we distinguish 41 topological clusters.

We focus our attention on a specific cluster type, the Siamese Dodecahedron (SD) due to its unique behaviour in the early stages of nucleation. In addition, we also focus our attention on the Pentagonal Bipyrimid (PB) because of its abundance in the fluid phase and its geometric similarity with the SD cluster. The SD cluster consists of particles that occupy four out of the five vertices of a pentagonal planar ring which we refer to as *ring* particles (as denoted by the red particles in Fig. 4.1a). The missing particle of the pentagonal ring is replaced by two particles (denoted by the blue particles in Fig. 4.1a), which are shifted up and down with respect to the pentagonal planar ring. We refer to these particles as *shifted* particles. Finally, two *spindle* particles (gold particles in Fig. 4.1a) are placed on top and below the pentagonal ring. The PB cluster is composed of *ring* particles (red particles in Fig. 4.1b), which form a pentagonal ring with two *spindle* particles (gold particles in Fig. 4.1b) similar to the Siamese dodecahedron.

For each particle in the system, we calculate the number of SD (PB) clusters that a particle belongs to. In Fig. 4.1c and 4.1d, we colour the fluid-like particles with different shades of pink (purple), depending on the number of SD (PB) clusters they are part of, according to the scale bar on the left (right). Even though the density of SD (PB) clusters is high throughout the fluid, Fig. 4.1c and 4.1d show that the density of SD (PB) clusters is spatially heterogeneous. Specifically, we observe that the crystal nucleus is surrounded by a high density of SD clusters, whereas the opposite trend is found for the PB clusters as the PB clusters are depleted near the surface of the crystal nucleus. More remarkably, the density of PB clusters seems to be anti-correlated with the density of SD clusters.

In Fig. 4.1e and 4.1f we perform the same analysis on a experimental sample, showing a heterogeneous structure consisting of high- and low-density regions of SD and PB clusters in the fluid phase, and a crystal nucleus that is surrounded by a high density of SD clusters and a low density of PB clusters, in excellent agreement with our simulations.

The incompatibility of the fivefold clusters with crystalline order rationalises the depletion of PB clusters near the surface of the crystal nucleus. It is tempting to speculate that the SD clusters surrounding the crystal nucleus play a transient role in the formation of the crystal phase, which will be investigated in more detail below.

To better understand the role of the PB and SD clusters in the crystallisation mechanism of hard spheres, we plot in Fig. 4.2a the fraction of particles belonging to SD (pink line) and PB (purple line) clusters as a function of time during an exemplary spontaneous nucleation event along with the fraction of crystalline particles (blue line) for comparison. Fig. 4.2a shows that the fraction of crystalline particles is approximately zero in the metastable fluid phase at the beginning of this trajectory until it starts to rise when crystallisation sets in. We also observe that the populations of particles in both the SD and PB clusters are already high before crystallisation sets in, showing that the metastable fluid exhibits strong spatial correlations due to packing constraints. More surprisingly, we find an increase in the number of SD clusters during the early stages of crystallisation, which decreases to a lower value at the end of the crystallisation process since SD clusters are not present in the fcc structure. In addition, the fraction of PB clusters decreases at the onset of crystallisation.

To investigate the anti-correlation between SD and PB clusters, we also measure the combined fraction of particles belonging to either SD or PB clusters as a function of time (red line in Fig. 4.2a). The combined fraction is not only constant in the metastable fluid phase,

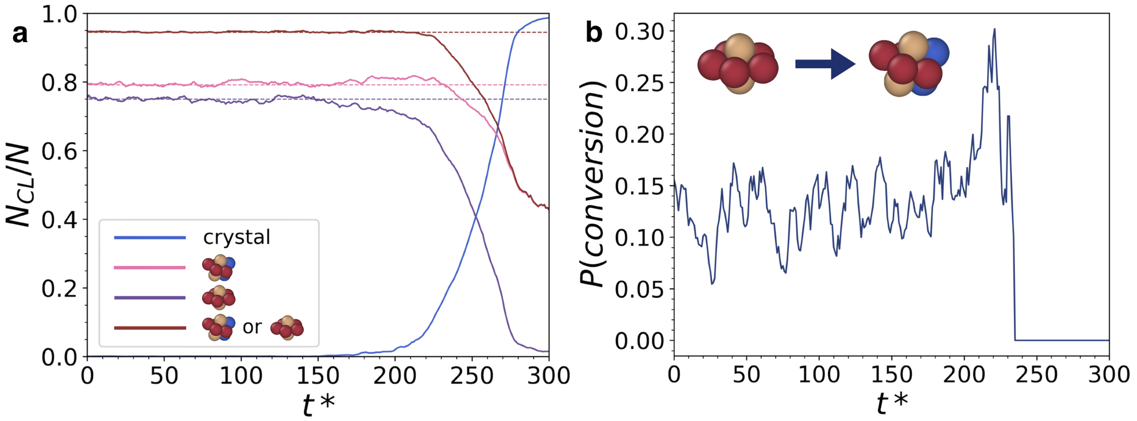


Figure 4.2: Behaviour of Siamese Dodecahedron (SD) and Pentagonal Bipyramid (PB) clusters during nucleation. (a) Fraction of particles belonging to SD (pink), PB (purple), and combined SD or PB (red) clusters along with the fraction of solid-like particles (blue) as a function of time during an exemplary spontaneous nucleation event. Note that a particle can be part of an SD cluster and a PB cluster at the same time, and therefore the corresponding fractions add up to a value which is higher than one. Also, a particle can be classified as crystal-like independently from whether it is also part of an SD cluster or not. The average values in the metastable fluid are shown by dashed lines. (b) Probability that a given PB cluster transforms into a SD cluster within a time interval of $\Delta t^* = 10$ during this nucleation event, calculated in a subcell of the system, which is centred around the center-of-mass of the biggest crystalline cluster. We set this probability to zero when the denominator, *i.e.* the number of PB clusters in the considered subcell of the system, is lower than 10 units, for poor statistics. A sketch of this conversion is shown as an inset in (b).

but also shows lesser fluctuations than the individual fractions of SD and PB clusters. More surprisingly, we observe that the combined fraction remains constant during the early stages of crystallisation, thereby demonstrating that the increase in SD clusters is a consequence of a decrease of PB clusters. The constant combined fraction of SD and PB clusters and the much smaller fluctuations suggest that there is a reversible conversion between PB and SD clusters. To this end, we calculate the probability that a PB cluster transforms into an SD cluster within a time interval $\Delta t^* = 10$ by only taking into account the subcell of the system where the first nucleus appeared. In Fig. 4.2b, we plot the conversion rate as a function of time. We find that the rate of PB into SD clusters is constant in the metastable fluid, and increases when crystallisation sets in.

We thus observe that the supersaturated fluid exhibits a heterogeneous structure of high- and low-density regions of PB and SD clusters with a continuous conversion between the two clusters. In addition, we find that the early stages of crystallisation is signaled by a higher conversion rate of PB into SD clusters, resulting in an increased fraction of SDs as shown in Fig. 4.2a. Subsequently, the number of SDs decreases when the crystal nucleus grows further, thereby demonstrating that the SD clusters represent an intermediate stage in the attachment of fluid-like particles to the crystal nucleus.

4.4.2 The nucleation mechanism

To understand the role of SD clusters in the fluid-solid transformation, we note that the four particles of the pentagonal ring of an SD cluster form a trapezoidal arrangement with two

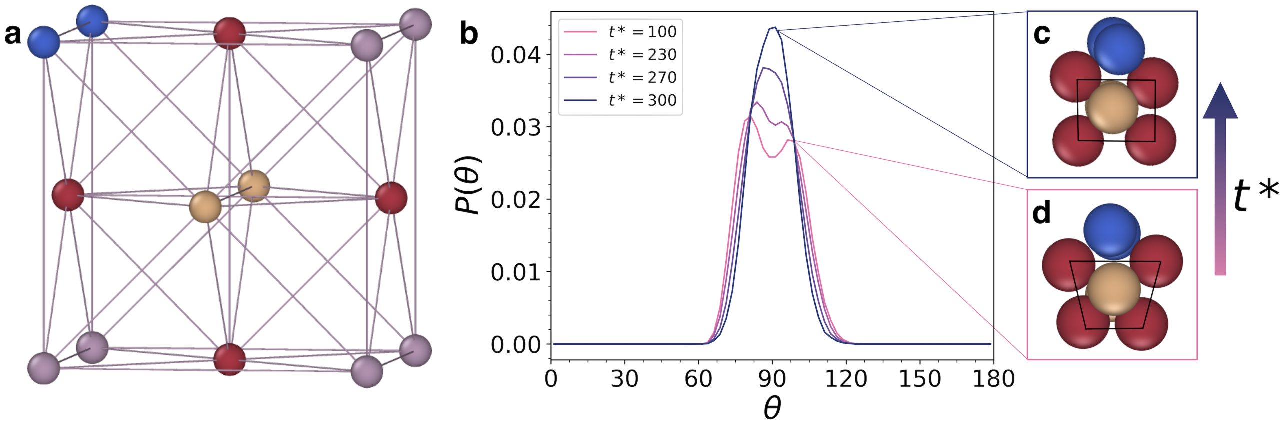


Figure 4.3: Transition from a Siamese Dodecahedron (SD) cluster to an fcc subunit. (a) Arrangement of the particles in an fcc unit cell. Red, blue and golden particles correspond to an SD cluster, while the remaining particles are coloured in lilac. (b) Probability distribution of the four angles θ of the trapezoidal arrangement of the 4 particles (red) in the pentagonal ring of all SD clusters in the system as computed at four different times during the crystallisation process. The typical arrangements of particles in SD clusters after and before nucleation are shown in (c) and (d), respectively, where the black lines connecting the centres of the *ring* particles help to better understand the transition.

acute and two obtuse angles, see Fig. 4.3d. Interestingly, in the case that these particles form a square arrangement (Fig. 4.3c), the SD cluster can be identified as a subunit of an fcc crystal as illustrated in Fig. 4.3a where the particles are denoted with the same colours to facilitate the comparison. Given this topological similarity, we speculate that the attachment of fluid-like particles to the solid nucleus proceeds *via* SD clusters where the four particles in the pentagonal ring transform from a trapezoidal to a square arrangement such that it becomes part of the fcc cluster.

To investigate this conjecture, we measure the distribution of the four angles of the trapezoidal arrangement of the 4 particles in the pentagonal ring of the SD clusters, at four different times during the crystallisation process. Fig. 4.3b shows that, in the fluid phase ($t^* = 100$), the distribution is bimodal with a peak at an angle smaller and larger than 90°, representing the trapezoidal arrangement. As crystallisation progresses, the distribution becomes unimodal with a single peak around 90°, indicating a square pattern.

Our results provide strong support that the trapezoidal arrangement of the four particles in the pentagonal ring of the SD cluster transforms into a square arrangement corresponding to a subunit of the fcc crystal. This transition is also illustrated in Fig. 4.3c and 4.3d, showing two representative SD clusters after and before the transformation, respectively. The key finding of our study is that the fivefold PB clusters – known to be inhibitors of crystal nucleation and abundant in the fluid phase – transform into SD clusters, and that the SD cluster-mediated attachment of particles to the growing nucleus proceeds *via* a simple rearrangement of particles into fcc subunits. Differently, the rearrangement of SD clusters into hcp is less straightforward and involves an additional displacement by one of the *shifted* particles (see Section 4.6.4). Hence, the propensity to grow fcc is higher than hcp, revealing that the polymorph selection mechanism in hard spheres is already hidden in the higher order structure of the fluid phase.

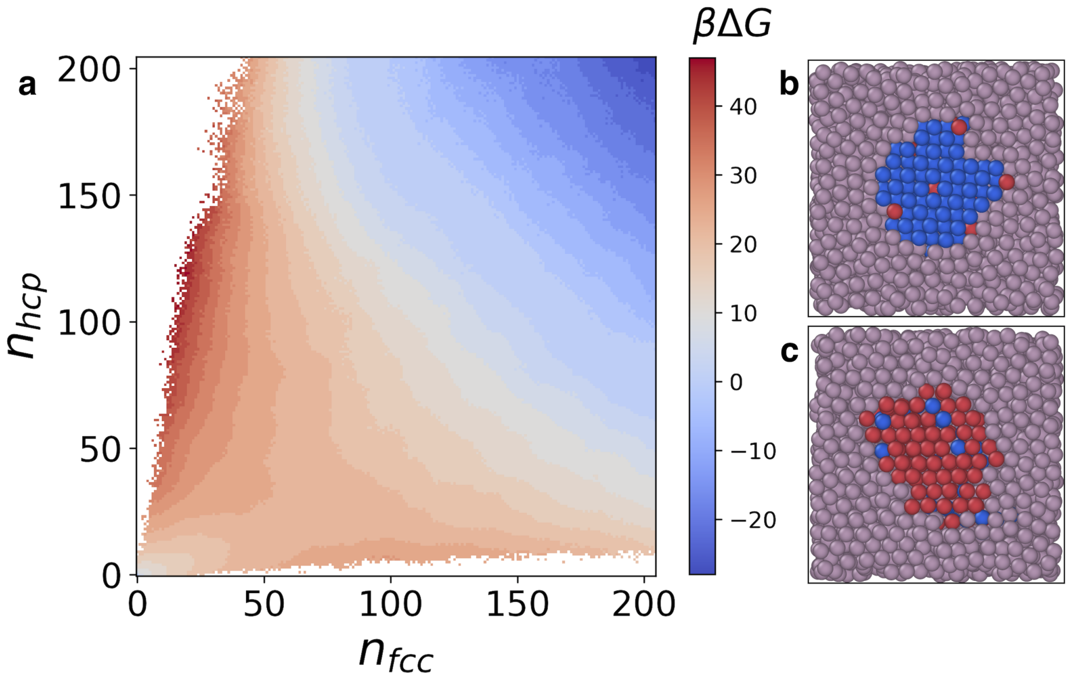


Figure 4.4: Thermodynamic propensity towards fcc-like particles in the early stages of crystal nucleation of hard spheres. (a) Gibbs free-energy barrier as a function of the number of fcc and hcp particles in the crystal nucleus. (b) A typical configuration of a nearly pure fcc crystal nucleus and of (c) a nearly pure hcp crystal nucleus as obtained from US simulations, where fcc-like particles are coloured blue, hcp-particles are red, and fluid-like particles are lilac.

4.4.3 The minimum free-energy pathway for nucleation

This finding begs the crucial question whether the polymorph selection mechanism as identified here has a kinetic or thermodynamic origin. In other words, does an fcc crystal have a lower interfacial free energy – and hence a lower Gibbs free-energy barrier – than an hcp crystal in a metastable fluid phase? To answer this question, we calculate the Gibbs free energy $\beta\Delta G(n_{fcc}, n_{hcp})$ for the formation of a crystal cluster consisting of n_{fcc} fcc-like particles and n_{hcp} hcp-like particles using the Umbrella Sampling (US) technique, see the Methods section for the technical details.

To investigate the thermodynamic propensity towards fcc-like or hcp-like ordering during the nucleation process, we use Umbrella Sampling [196] to calculate the the nucleation barrier of a system of hard spheres at a pressure of $\beta P\sigma^3 = 17.0$. Similar to previous literature [193, 197] we identify the nucleus by using the dot product

$$d_l(i,j) = \frac{\sum_{m=-l}^{l} q_{lm}(i)q_{lm}^*(j)}{\left(\sum_{m=-l}^{l} |q_{lm}(i)|^2\right)^{1/2} \left(\sum_{m=-l}^{l} |q_{lm}(j)|^2\right)^{1/2}}, \tag{4.6}$$

with $l = 6$ to define solid-like bonds as those bonds between particle pairs (i, j) for which $d_6(i, j) > 0.7$, and define solid-like particles as those that have at least 7 of such solid-like bonds. Particle neighbours are defined using a distance cutoff of $r_c = 1.4\sigma$. The nucleus is then the largest set of solid-like particles that are connected by solid-like bonds. To disentangle fcc-like and hcp-like order, we subsequently classify solid-like particles as fcc-like and hcp-like based on their value of the Steinhardt bond order parameter w_4: particles with $w_4 < 0$ are fcc-like and those with $w_4 \geq 0$ are hcp-like [87]. The number of such particles are n_{fcc} and n_{hcp}, respectively, and we use these to define the US biasing potential:

$$U_b = \frac{1}{2}\lambda_{fcc}\left(n_{fcc} - n_0^{fcc}\right)^2 + \frac{1}{2}\lambda_{hcp}\left(n_{hcp} - n_0^{hcp}\right)^2, \tag{4.7}$$

where both coupling constants λ_{fcc} and λ_{hcp} are set to an equal value of $\beta\lambda_{fcc} = \beta\lambda_{hcp} = 0.05$. This allows us to sample the two-dimensional Gibbs free-energy difference $\beta\Delta G(n_{fcc}, n_{hcp})$ that is the nucleation barrier as a function of the number of fcc-like and hcp-like ordered particles. We initialise each US window (n_0^{fcc}, n_0^{hcp}) from a configuration with a nucleus with approximately $n_{fcc} \approx n_0^{fcc}$ and $n_{hcp} \approx n_0^{hcp}$. For very small nuclei up to $n = n_{fcc} + n_{hcp} \sim 20$ we measure the full cluster size distribution instead of only the size of the largest cluster, as the probability of multiple small nuclei appearing simultaneously can be significant. We implement the US scheme by adding additional Monte Carlo bias moves that accept or reject trajectories based on the bias potential of Eq. 4.7 on top of a hard-particle Monte Carlo (HPMC) simulation implemented using HOOMD-blue's HPMC module [195, 198]. Bias moves are performed every MC cycle in order to also sample regions of the free-energy landscape where the gradient is large. Finally, we reconstruct the nucleation barrier by using the Weighted Histogram Analysis Method (WHAM) [199], specifically by using the algorithm provided by Ref. 200.

In Fig. 4.4a, we plot $\beta\Delta G$ between a metastable fluid and a system with a crystalline core composed of n_{fcc} fcc-like particles and n_{hcp} hcp-like particles. The lowest free-energy path on this surface shows that the crystal nucleus has an excess of fcc-like particles in the early stages of nucleation and that the critical nucleus consists of about 70% fcc-like particles. We show two exemplary configurations of a nearly fcc-like and hcp-like cluster in Fig. 4.4b and 4.4c, respectively, demonstrating the effectiveness of our umbrella sampling method to bias towards nuclei with a certain composition.

4.5 Conclusions

In conclusion, we unveal the crystallisation and polymorph selection mechanism in a fluid of hard spheres by analysing the early stages of nucleation in MD simulations. We show that the supersaturated fluid is highly dynamic as there is a reversible conversion between fivefold Pentagonal Bipyramid and Siamese Dodecahedron clusters. The Siamese Dodecahedra have a stunning similarity with an fcc subunit, thereby explaining the as-of-yet unexplained higher propensity of fcc compared to hcp in hard spheres. Finally, we show that the polymorph selection mechanism has not only a geometric origin which is hidden in the higher-order correlations of the fluid phase, but also a thermodynamic one as the lowest free-energy path proceeds *via* a higher number of fcc-like particles with respect to hcp-like particles in the early stages of nucleation. This insight suggest ways to control the nucleation pathways and the crystal polymorphs.

Acknowledgements

I would like to thank Robin van Damme, who performed the Umbrella Sampling simulations and for a critical reading of the manuscript. Also, I would like to extend my gratitude to Prof. C. P. Royall for providing the experimental samples, and for the numerous suggestions and inputs.

4.6 Appendix

4.6.1 Crystal growth

During the early stages of nucleation of hard spheres, we observe that both in simulations and experiments the crystal nucleus is surrounded by a high density of Siamese Dodecahedron (SD) clusters and a low density of Pentagonal Bipyramid (PB) clusters.

In this Section, we show that this scenario persists even during crystal growth when the size of the nucleus is post-critical. In Fig. 4.5a and 4.5b, we show a post-critical nucleus as obtained from MD simulations. The solid-like particles are coloured blue. The fluid-like particles are coloured with different shades of pink (purple), depending on the number of SD (PB) clusters they are part of, according to the scale bar on the left (right).

The fact that the number of SD clusters around the crystal nucleus is particularly high throughout the crystal growth process is further evidence that these clusters play a transient role in the transformation of the disordered fluid to the ordered crystal phase.

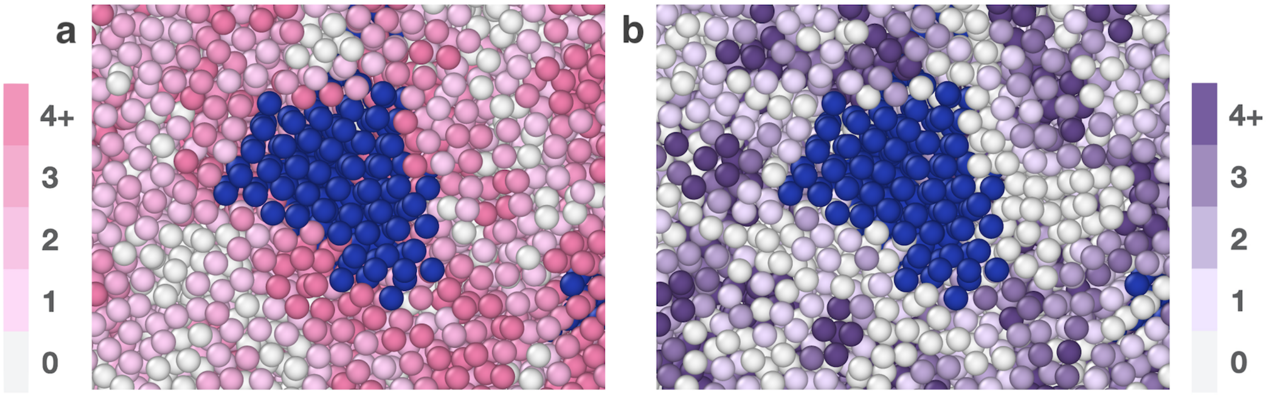


Figure 4.5: Typical configuration in the crystal growth regime of hard spheres. (a-b) Cut-through image during crystal growth as obtained from MD simulations. Crystal-like particles are coloured blue, while fluid-like particles are coloured according to the scale bar on the left (a) or right (b) depending on the number of SD (a) or PB (b) clusters each particle belongs to.

4.6.2 Subcell analysis

To obtain a better resolution on the behaviour of SD clusters during nucleation, we divide the simulation box into 64 cubic subcells of equal volume. For each subcell we compute the fraction of solid-like particles n_x, as well as the fraction of particles belonging to SD clusters n_{SD}, and to PB clusters n_{PB}. Subsequently, we measure the correlation between these number fractions by computing the Pearson correlation coefficient during the entire crystallisation trajectory. The Pearson correlation coefficient $r(x, y)$ between variables x and y reads

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{4.8}$$

where $\bar{x}$ and $\bar{y}$ represent the average values of variable x and y, respectively.

In Fig. 4.6a we plot $r(n_{SD}, n_x)$ as a function of time. We observe two distinct and noteworthy features. First, at the start of the nucleation process, n_x and n_{SD} reach a positive maximum

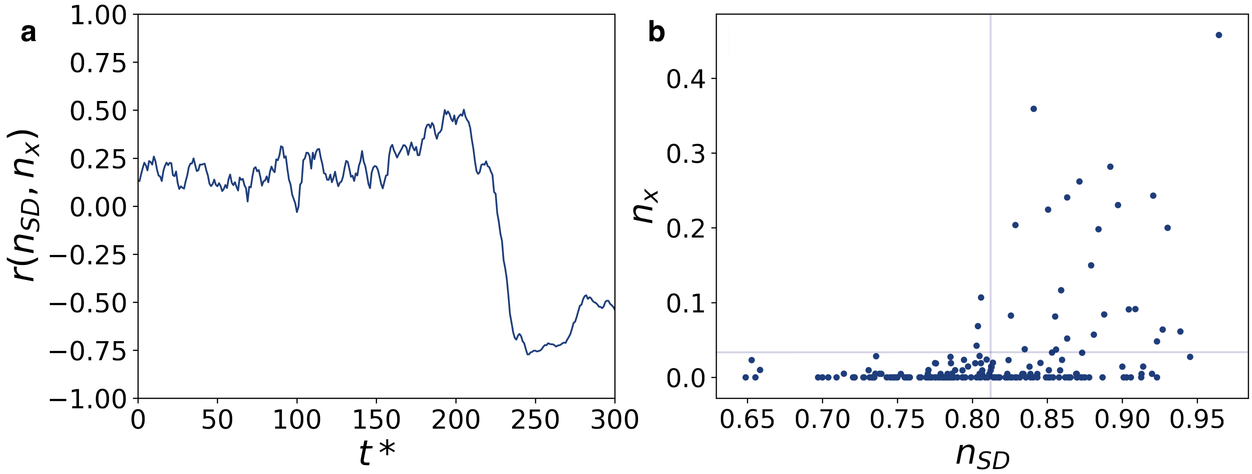


Figure 4.6: (a) Pearson correlation $r(n_{SD}, n_x)$ as a function of time t^*. (b) Fraction of crystal-like particles n_x as a function of the fraction of particles belonging to at least one SD cluster n_{SD}, calculated for each subcell in the system. The data points are from three snapshots obtained from independent MD simulations, which correspond to the maximum of the Pearson correlation function $r(n_{SD}, n_x)$. The vertical (horizontal) line indicates the average value of n_{SD} (n_x).

correlation value, demonstrating that a subcell with a high fraction of crystal-like particles also shows a high fraction of SD clusters. This feature is made explicit in Fig. 4.6b, where each point correspond to a specific subcell of the system at the peak of the Pearson correlation between n_{SD} and n_x. The second observation we make, is that during the crystal growth stage, the correlation suddenly drops to negative values, indicating that the crystal structure formed in the nucleus is incompatible with the presence of SD clusters.

4.6.3 More on Bond Order Parameters

Distinction between ring, shifted, and spindle particles

The analysis conducted in this Chapter demonstrates that SD clusters play a transient role in the attachment of fluid particles to a crystal nucleus. To investigate this transformation further, we calculate the bond order parameters (BOPs) for the particles belonging to SD clusters. In Fig. 4.7a we show the BOP values for the particles composing the SD clusters in the $\bar{q}_4 - \bar{q}_6$ plane for a typical configuration in the early stage of nucleation along with the BOPs of a typical fluid and crystal configuration for comparison. Even though a large fraction of the SD particles (bright pink points) shows a higher than average degree of fourfold and sixfold symmetry, the large spread in $\bar{q}_4$ and $\bar{q}_6$ values show that there is no clear correlation between the SD clusters and their BOP values in the fluid phase.

In order to clarify the transition process from a PB cluster to an SD cluster, in Section 4.4.1 we divided the particles composing an SD cluster into several categories – *ring*, *shifted*, and *spindle* particles. This classification is not only useful in describing the topology of the SD clusters, but also reveals additional information regarding the nucleation mechanism. By computing the probability distribution of $\bar{q}_6$ for the three different particle types of the SD clusters at the onset of crystallisation, we find that the *shifted* particles show slightly higher $\bar{q}_6$ values (blue curve in Fig. 4.7b), suggesting that the crystallisation process is largely initiated by the *shifted* particles.

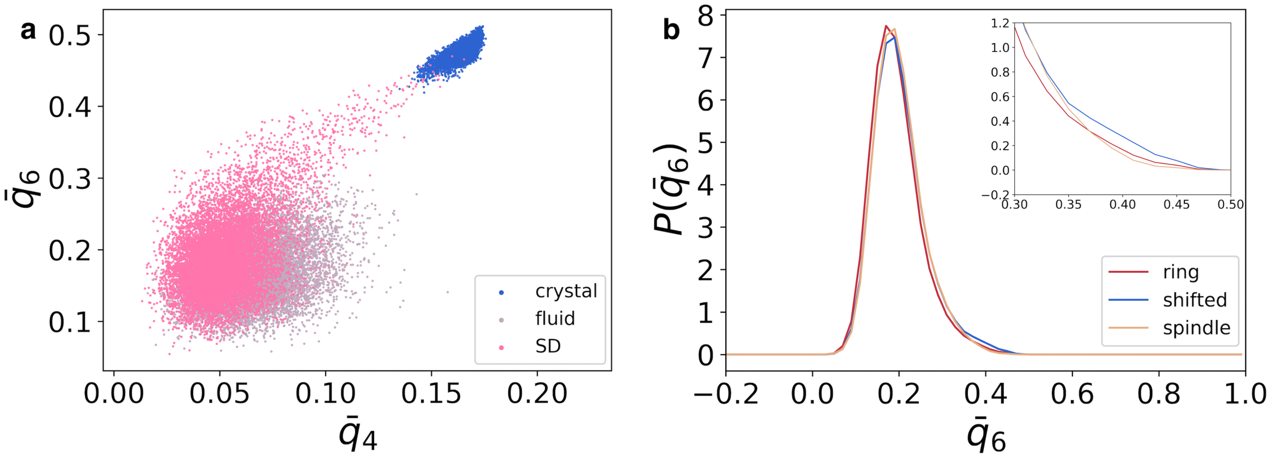


Figure 4.7: Correlation between bond order parameters (BOPs) and Siamese Dodecahedron (SD) clusters. (a) Projection of the BOP values of particles belonging to SD clusters in the early stages of nucleation on the $\bar{q}_4 - \bar{q}_6$ plane. Particles belonging to SD clusters show BOP values that are very similar to the ones of the fluid phase, which are therefore not useful in describing the behaviour of the SD particles. (b) Probability distribution $P(\bar{q}_6)$ of the *ring*, *shifted*, and *spindle* particles using the same configuration as in (a). The inset shows that the *shifted* particles show slightly higher $\bar{q}_6$ values, revealing that the nucleation process is largely initiated by the *shifted* particles.

In Section 4.4.1 we described how the particles re-arrange during the transformation of PB clusters into SD clusters. To be more specific, we showed that the transformation is initiated by the appearance of two *shifted* particles. Recalling that the disappearance of PB clusters and the subsequent excess of SD clusters enables the start of the crystallisation phenomenon, it is to be expected that the *shifted* particles of the SD cluster are indeed the first to crystallise.

Polymorph detection with different classification schemes

The results presented in this Chapter are all based on a combined use of bond order parameters (BOPs) and a Topological Cluster Classification (TCC) analysis. In particular, when using BOPs there are several choices to make, which affect the classification of the particles. These choices are related to the identification of the local neighbourhood of a particle, to the distinction of fluid-like and solid-like particles, and to the further distinction of the different crystal polymorphs. In this section, we select different criteria for all these subtasks, and check the robustness of the classification outcome. Specifically, we check that all methods record a predominance of fcc-like with respect to hcp-like ordering in the growing nucleus.

To this end, we start by using a total of three different techniques to find the local neighbourhood of a particle. The first two are based on a simple cutoff radius equal to 1.4σ and 1.5σ with σ the diameter of the particles, while the third is based on the solid-angle nearest-neighbour (SANN) algorithm.

Furthermore, in order to classify particles as solid-like or fluid-like, we use two different criteria. One is implemented *via* the criterion $\bar{q}_6 > 0.31$, as in Section 4.4.1, where we used a loose criterion, capable of detecting particles that are slightly more ordered than in the fluid phase. The second method we use here is based on dot-products of the q_{6m}, as implemented in the Umbrella Sampling calculations (see Section 4.4.3).

Finally, in order to distinguish the different crystal polymorphs, we again use different

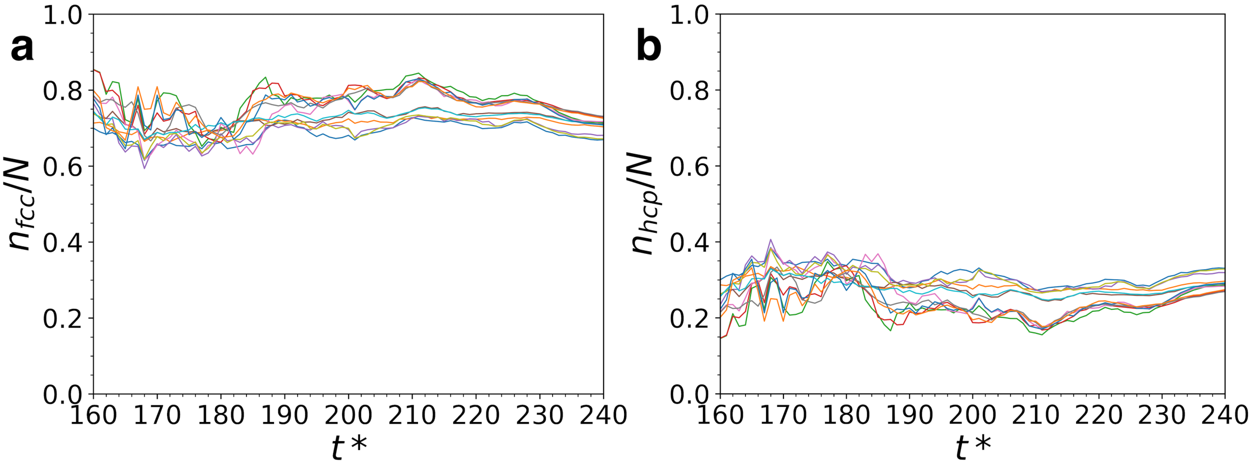


Figure 4.8: Crystal polymorphs in the growing crystalline cluster. Fraction of (a) fcc-like n_{fcc}/N and (b) hcp-like n_{hcp}/N particles in the growing nucleus consisting of N particles as identified by 12 different classification schemes described in the text.

strategies. In the first method, we employ the w_4 value, following the scheme proposed in our Umbrella Sampling calculations (see Section 4.4.3). Alternatively, we can also use the averaged bond order parameter $\bar{w}_4$.

Using all possible combinations for detecting the local environment, distinguishing between solid-like and fluid-like particles, and classifying different polymorphs, we obtain a total of 12 classification schemes. We use all of these to analyse a spontaneous nucleation trajectory and compute the fraction of fcc-like and hcp-like particles in the largest crystal nucleus during a nucleation trajectory. In Fig. 4.8a (Fig. 4.8b), we show the ratio of fcc-like (hcp-like) particles computed *via* each classification scheme. We note that the first part of the nucleation event is noisy as the denominator, *i.e.* the number of particles belonging to the main cluster, is very small.

We clearly observe that our results are robust with respect to the choice of classification scheme, thereby providing confidence that crystal nuclei are predominately composed of (60%-80%) fcc-like particles.

4.6.4 From Siamese Dodecahedra to hcp

In Fig. 4.3 we show that particles arranged in an SD cluster have a high propensity to transform to fcc due to the topological similarity between the SD cluster and a subunit of fcc. In particular, the transition between the SD cluster into an fcc subunit proceeds by changing the trapezoidal arrangement of the four *ring* particles into a square arrangement. Here, we show that such a simple transformation does not hold for the transition from an SD cluster to hcp, which involves an additional displacement by one of the *shifted* particles.

This transformation is sketched in Fig. 4.9. In Fig. 4.9a, we show the unit cell of an hcp crystal, with an additional particle on the right belonging to an imaginary adjacent unit cell. By including the latter particle, it is possible to find a pattern that resembles the SD cluster described in the text. We therefore colour the particles with the usual colour coding, thus *ring* particles are coloured red, *spindle* particles are coloured gold, and *shifted* particles are coloured blue. As one of the shifted particles of this cluster is displaced with respect to an actual SD cluster, we denote this cluster as a defective SD. Note that particles have been reduced in size

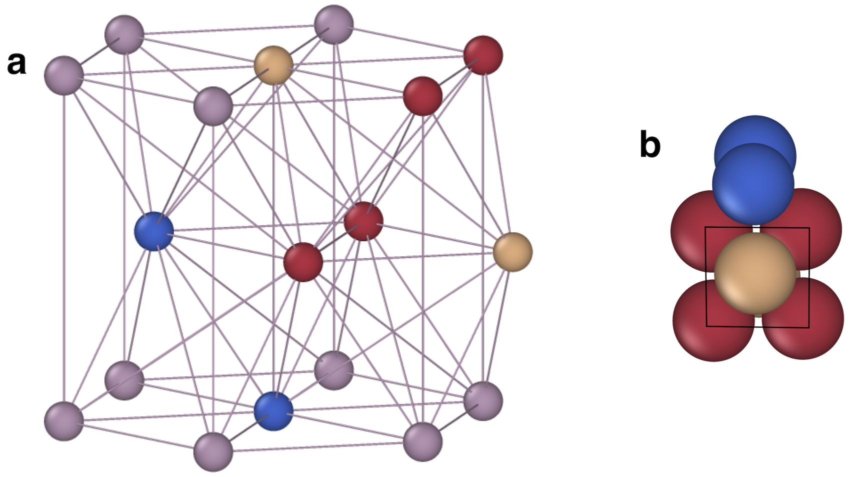


Figure 4.9: Geometrical relationship between an SD cluster and the hcp unit cell. (a) Unit cell of the hcp crystal with an additional particle belonging to the adjacent unit cell. An defective SD cluster is identified in the unit cell of an hcp phase and the particles belonging to this defective SD cluster are coloured following the colour coding explained in the text. This SD cluster is termed defective as one of the shifted particles (in blue) is displaced with respect to an actual SD cluster. (b) The defective SD cluster resulting from the pattern in (a). From this viewpoint, it can be seen that one of the *shifted* (blue) particles is displaced, as a result of the “a-b-a-b" stacking of hcp, differently from what is observed in the fcc case, where the stacking of the hexagonal planes follows the “a-b-c-a-b-c" pattern.

so that the whole unit cell is visible.

In Fig. 4.9b, we isolate only the particles belonging to this defective SD cluster and picture them with the actual colloid size. From the view point shown in this figure – and comparing it with Fig. 4.3c and 4.3d – it is evident that one of the *shifted* (blue) particles is displaced with respect to its position in the SD cluster. Hence, the transition from an SD cluster to a defective SD cluster, which resembles a subunit of hcp, involves an additional displacement of one of the *shifted* particles. We therefore conclude that this additional displacement makes the propensity of fluid-like particles to crystallise into hcp lower than that of fcc.

5

Through the order parameter filter: polymorph detection with an unsupervised learning procedure

In this Chapter we analyse the problem of finding an order parameter for classifying the polymorphs appearing during a crystal nucleation event of hard spheres. We discuss the limitations and shortcomings of many approaches in the literature, and propose a new unsupervised learning-based algorithm that combines speed of execution and interpretability, and contains a large amount of information. We finally demonstrate the effectiveness of this new order parameter by analysing hard-sphere nucleation trajectories and recognising the different polymorphs, as well as particles at the interface between the supersaturated fluid phase and the crystalline embryo.

Based on: G. M. Coli, B. Verhoef, and M. Dijkstra, *Through the order parameter filter: polymorph detection with an unsupervised learning technique*, Manuscript in preparation

5.1 Introduction

One of the problems in soft matter studies is the ability to recognise spatial patterns within a particle system. The search for an order parameter is fundamental not only to recognise, for example, a nucleation event when it occurs, but above all to be able to describe it quantitatively [24, 201, 202]. In fact, the main quantities involved in the description of a nucleation event strongly depend on the size of the nucleus, its shape, and the behaviour of the particles around it, as described in Chapters 1, 2, and 3.

In recent decades this task has been tackled through a multitude of different avenues. Order parameters based on the Common Neighbour Analysis (CNA) algorithm, which characterises the local environment of a particle through the graph constructed from its neighbours, have been proposed [30, 203]. Many other works have used Bond Order Parameters (BOPs) or their averaged counterparts, which instead expand the local environment of a particle in terms of spherical harmonics, to construct parameters that are sensitive to different spatial symmetries [86, 87, 150, 151].

In particular, in the context of BOPs, hexagonal symmetry has often been used in order to construct order parameters that distinguish a fluid and thus disordered phase from a crystalline one. However, partial information such as that obtained by looking only at sixfold symmetry can result in misclassifications, which are artefacts of the technique used.

An example is the so-called *hcp-coating* during hard-sphere nucleation, which consists of a dominant presence of hcp-like particles at the interface of a crystalline hard-sphere nucleus. It has been demonstrated that this purely depends on the fact that the typical values of the $\bar{q}_4$ and $\bar{q}_6$ order parameters for the hcp crystal are in between those of the fluid and the fcc crystal [204–206], as we show in Fig. 5.1.

Similarly, a classification which is based solely on the sixfold and the fourfold symmetries cannot satisfactorily classify the various single-component crystals such as fcc, hcp and bcc. One evidence is that the presence of the bcc crystal at the early stages of nucleation, as predicted by the theory of Alexander and McTague, is still heavily debated [207]. Depending on how one uses the same information, the bcc crystal can be detected or not detected during the classification procedure [24, 77, 121, 150, 167, 208, 209].

Recently – as well as in Chapter 3 of this thesis – it has been shown that *unconventional* symmetries possess a degree of information that can be useful, if not crucial, when combined with the sixfold and fourfold symmetries already employed, in order to distinguish between several competing phases [153, 210, 211]. For instance, in Chapter 3 we showed that an artificial neural network gives large weights to odd-symmetry BOPs like q_7 and q_9 in order to correctly distinguish the small spheres of an icosahedral AB_{13} crystal from the binary fluid phase. Alternatively, in Ref. 211 it has been shown that other BOPs like $\bar{q}_8$ and $\bar{q}_{10}$ are useful to correctly label the rich variety of phases formed by a system of hard spherotetrahedra.

In general, it would be useful to have an algorithm capable of processing autonomously a large amount of information, in order to find order parameters capable of distinguishing the phases under analysis. Recently, unsupervised learning algorithms of different nature have been proposed which aim to find order parameters for a wide range of phenomena, including self-assembly events and dynamic propensity [212–214].

In this Chapter, we use Principal Component Analysis (PCA) [215, 216] – an unsupervised learning dimensionality reduction technique – to find order parameters that can correctly classify the phases of interest during a hard-sphere nucleation event, on a single particle level. Thus, our focus is on the fcc, hcp, and bcc crystals, as well as the single-component fluid phase.

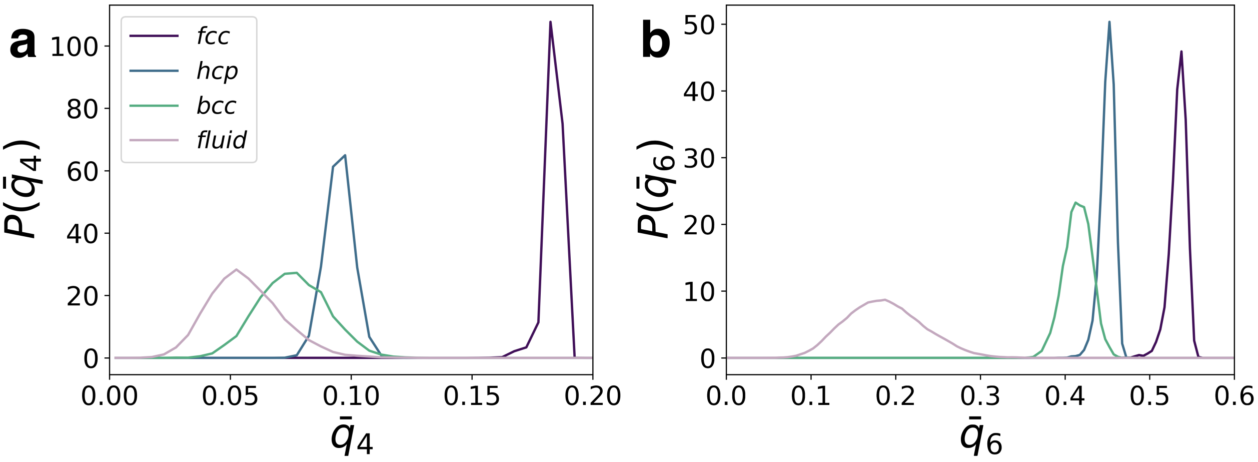


Figure 5.1: Distributions of (a) $\bar{q}_4$ and (b) $\bar{q}_6$ for the four bulk phases considered – fcc, hcp, bcc, and fluid. For both bond order parameters, the distributions of the bcc and hcp crystals lie between those of the fluid and the fcc.

This technique has the advantage of finding order parameters that contain a very high degree of information, being a linear combination of several features. Furthermore, due to the intrinsic linearity of the technique, the order parameters obtained are highly interpretable. We demonstrate the effectiveness and robustness of the algorithm by analysing nucleation trajectories obtained from numerical simulations.

Finally, we employ the Gaussian Mixture Model (GMM) to perform a clustering of the data in the new parameter space found by PCA [217–219]. In particular, depending on how the algorithm is used, this process can allow the identification of particles with local environments corresponding not only to the bulk, but also to the interface of the above phases.

5.2 Model and simulation methods

5.2.1 Simulation details

In order to build a data set where all the relevant phases are present, we simulate a single-component system where the individual particles interact *via* a Weeks-Chandler-Andersen (WCA) pair potential [75]. The interaction $u(r_{ij})$ between particles i and j at distance r_{ij} is defined as

$$u\left(r_{ij}\right) = \begin{cases} 4\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6} + \frac{1}{4}\right] & r_{ij} < 2^{\frac{1}{6}}\sigma \\ 0 & r_{ij} \geq 2^{\frac{1}{6}}\sigma, \end{cases} \tag{5.1}$$

where σ is the particle diameter, and ϵ the interaction strength. This pair potential can be straightforwardly employed in Molecular Dynamics (MD) simulations, and the softness of the repulsion can be tuned *via* the reduced temperature k_BT/ϵ. In particular, for $T \rightarrow 0$ the WCA potential reduces to the HS interaction. In this work, like in Chapters 3 and 4, we set $k_BT/\epsilon = 0.025$, which has been used extensively in previous simulation studies to mimic hard spheres [77–80].

The temperature T and pressure P are kept constant *via* the Martyna-Tobias-Klein (MTK) integrator [92], with the thermostat and barostat coupling constants $\tau_T = 1.0\ \tau_{MD}$ and $\tau_P = 1.0\ \tau_{MD}$, respectively, and $\tau_{MD} = \sigma_L\sqrt{m/\epsilon}$ is the MD time unit. The time step is set to $\Delta t = 0.004\tau_{MD}$, which is small enough to ensure stability of the simulations. We ran the simulations for $10^9\tau_{MD}$ time steps, unless specified otherwise. The simulation box is cubic and periodic boundary conditions are applied in all directions.

As in Chapter 4, we select the pressure values in a region of metastability that allow us to observe crystal nucleation on a reasonable time scale. Specifically, the reduced pressure varies in the range $\beta P\sigma^3 \in [13.40, 16.00]$, which results in numerous spontaneous crystallisation events. All MD simulations are performed using the HOOMD-blue (Highly Optimised Object-oriented Many-particle Dynamics) software [195].

In addition to MD simulations, and in order to get data of the equilibrium phases studied in this work (fcc, hcp, bcc, and single-component fluid), we carry out Monte Carlo simulations in the canonical (NVT) ensemble, with periodic boundary conditions applied in all three spatial directions. The three crystals have a number density $\rho\sigma^3 = N\sigma^3/V \simeq 0.86$ (where N is the number of particles and V the volume of the simulation box), while for the fluid phase we choose $\rho\sigma^3 \simeq 0.78$.

5.2.2 Bond Order Parameters

We characterise the local environment of each particle in the data set we built using BOPs [86, 87]. We first define the complex vector $q_{lm}(i)$ for each particle i

$$q_{lm}(i) = \frac{1}{N_b(i)} \sum_{j=1}^{N_b(i)} Y_{lm}(\theta(\mathbf{r}_{ij}), \phi(\mathbf{r}_{ij})), \tag{5.2}$$

where $N_b(i)$ is the number of neighbours of particle i, $Y_{lm}(\theta(\mathbf{r}_{ij}), \phi(\mathbf{r}_{ij}))$ denotes the spherical harmonics, $m \in [-l, l]$, $\theta(\mathbf{r}_{ij})$ and $\phi(\mathbf{r}_{ij})$ are the polar and azimuthal angles of the distance vector $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$, and $\mathbf{r}_i$ denotes the position of particle i.

The averaged $\bar{q}_{lm}(i)$ is defined as

$$\bar{q}_{lm}(i) = \frac{1}{\tilde{N}_b(i)} \sum_{j=1}^{\tilde{N}_b(i)} q_{lm}(j), \tag{5.3}$$

where $\tilde{N}_b(i)$ is the number of neighbours including particle i itself. The rotationally invariant quadratic averaged bond order parameter is defined as

$$\bar{q}_l(i) \ = \ \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |\bar{q}_{lm}(i)|^2}. \tag{5.4}$$

As in Chapter 3, to identify the neighbours of particle i we employ the parameter-free solid-angle-based nearest-neighbour (SANN) algorithm of Van Meel and co-workers [152]. This algorithm assigns a solid angle to every potential neighbour j of i, and defines the neighbourhood of particle i to consist of the $N_b(i)$ particles nearest to i for which the sum of solid angles equals 4π.

In principle, the more BOPs we use to characterise a particle, the greater the information about its local environment is. Therefore, we describe the local environment of each particle i

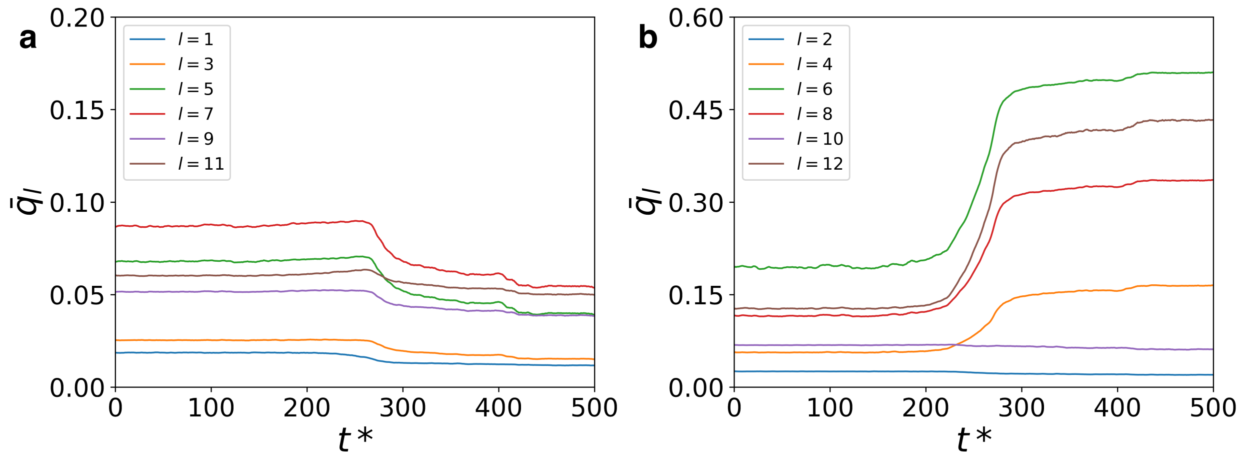


Figure 5.2: Mean value of the BOPs $\bar{q}_l$ with $l \in [1, 12]$ as a function of time $t^* = t/\tau_{MD} \cdot 10^6$, during a spontaneous nucleation trajectory performed by a system of nearly-hard spheres. In particular, in (a) we show only the BOPs with odd symmetries, while in (b) we repeat the same calculation with even symmetries BOPs.

by a 12-dimensional input vector $\mathbf{q}$ of bond order parameters defined as

$$\mathbf{q}(i) = \{\bar{q}_l(i)\}, \tag{5.5}$$

where $l \in [1, 12]$. All of these 12 BOPs contain information which can be used to distinguish between different polymorphs. As a proof of that, we calculate, as a function of time, the mean value of each BOP in a WCA system during a spontaneous nucleation trajectory. The results are reported in Fig. 5.2. Usually, only $\bar{q}_4$ and $\bar{q}_6$ are used, but also the other parameters show a change in the mean value at the onset of crystallisation, and thus demonstrate their sensitivity to the phase transition.

5.3 Feature extraction

Principal Component Analysis (PCA) belongs to the family of dimensionality reduction techniques. These techniques aim to find a space of smaller dimensionality than that in which the original data live, but at the same time they aim at preserving most of the information contained in the data [215, 216].

Generally, there are two ways in which these techniques succeed in reducing dimensionality while preserving much of the information. The first is to eliminate redundancy in the input features. In fact, some features may be highly correlated with each other, and therefore may not add extra information. The second way is to remove features that do not contain relevant information about the phenomenon one wishes to characterise, or have a negligible amount of it.

PCA is the simplest dimensionality reduction technique, as the new output features which define the new basis are a linear combination of the input features. This aspect penalises PCA in cases where inherently non-linear transformations are required to reduce the dimensionality, but at the same time gives PCA great speed of execution, and a significant interpretability of the algorithm.

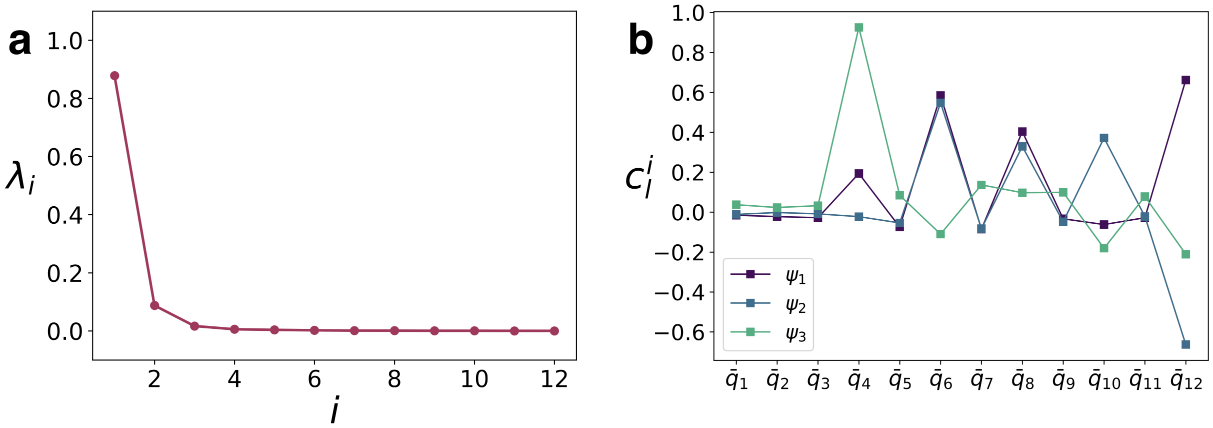


Figure 5.3: (a) The 12 eigenvalues λ_i and (b) the first three eigenvectors obtained from the application of PCA on the data set, composed by the vector $\mathbf{q}(i)$ of the four bulk phases – fcc, hcp, bcc, and fluid.

But how does PCA actually work? Suppose we have a data set with dimensionality N. PCA finds a new space with the same dimension N, defined by a new set of mutually orthogonal unit vectors $\psi_1, \psi_2, ..., \psi_N$ – eigenvectors of the transformation performed by PCA – which are defined as those on which the projection of the data is maximised. The eigenvectors ψ_i, also called *principal components*, are ranked in descending order of expressed variance, such that ψ_1 accounts for the largest variance of the data, and so on. The associated expressed fractional variance of each eigenvector is equal to the corresponding eigenvalue λ_i. By choosing a subspace of lower dimensionality $M < N$ spanned by the first M eigenvectors, we can retain most of the information of the original data while working in a lower dimensionality.

The data set on which we apply PCA is made of a total of $4 \cdot 10^3$ samples, where each sample corresponds to the vector $\mathbf{q}(i)$ defined in Eq. 5.5, computed for particle i in a specific phase. In particular, we collect 10^3 samples for each of the phases we want to classify, which are the fcc, hcp, and bcc crystals, with the addition of the fluid phase, as described in Section 5.2.1. The eigenvalues and eigenvectors resulting from the application of PCA on the data set are shown in Fig. 5.3.

In Fig. 5.3a, we show the 12 eigenvalues λ_i – which correspond to the fractional variance expressed by the corresponding eigenvector, in descending order. Interestingly, with only one eigenvector – or feature – we can already express a large fraction of the total variance of the data (more than 87%). In this work, we set the effective dimensionality of our data set to 3. In other words, we take into account only the first 3 eigenvectors, which together express more than 98% of the total variance present in the data.

Each of the 12 eigenvectors ψ_i is a linear combination of the averaged BOP $\bar{q}_l$ with $l \in [1, 12]$, and thus

$$\psi_i = \sum_{l=1}^{N} c_l^i \bar{q}_l \tag{5.6}$$

The coefficients c_l^i for the first three eigenvectors ψ_1, ψ_2, and ψ_3 are shown in Fig. 5.3b and reported in Table 5.1. If a eigenvector has a high coefficient for a specific $\bar{q}_l$, then the corresponding feature is prioritised by PCA and it contributes to a high variance of the data. Conversely, if the same coefficient is close to zero, then the corresponding $\bar{q}_l$ is essentially ignored by PCA. From Fig. 5.3b and Table 5.1 we see that the three-dimensional subspace found by

ψ_i	c_1^i	c_2^i	c_3^i	c_4^i	c_5^i	c_6^i	c_7^i	c_8^i	c_9^i	c_{10}^i	c_{11}^i	c_{12}^i
ψ_1	-0.02	-0.02	-0.03	0.19	-0.07	0.58	-0.09	0.40	-0.03	-0.06	-0.03	0.66
ψ_2	-0.01	0.00	-0.01	-0.02	-0.05	0.55	-0.08	0.33	-0.05	0.37	-0.02	-0.66
ψ_3	0.04	0.02	0.03	0.93	-0.08	-0.11	0.14	0.10	0.10	-0.18	0.08	-0.21

Table 5.1: Coefficients c_l^i (defined in Eq. 5.6) of the first three eigenvectors – ψ_1, ψ_2, and ψ_3.

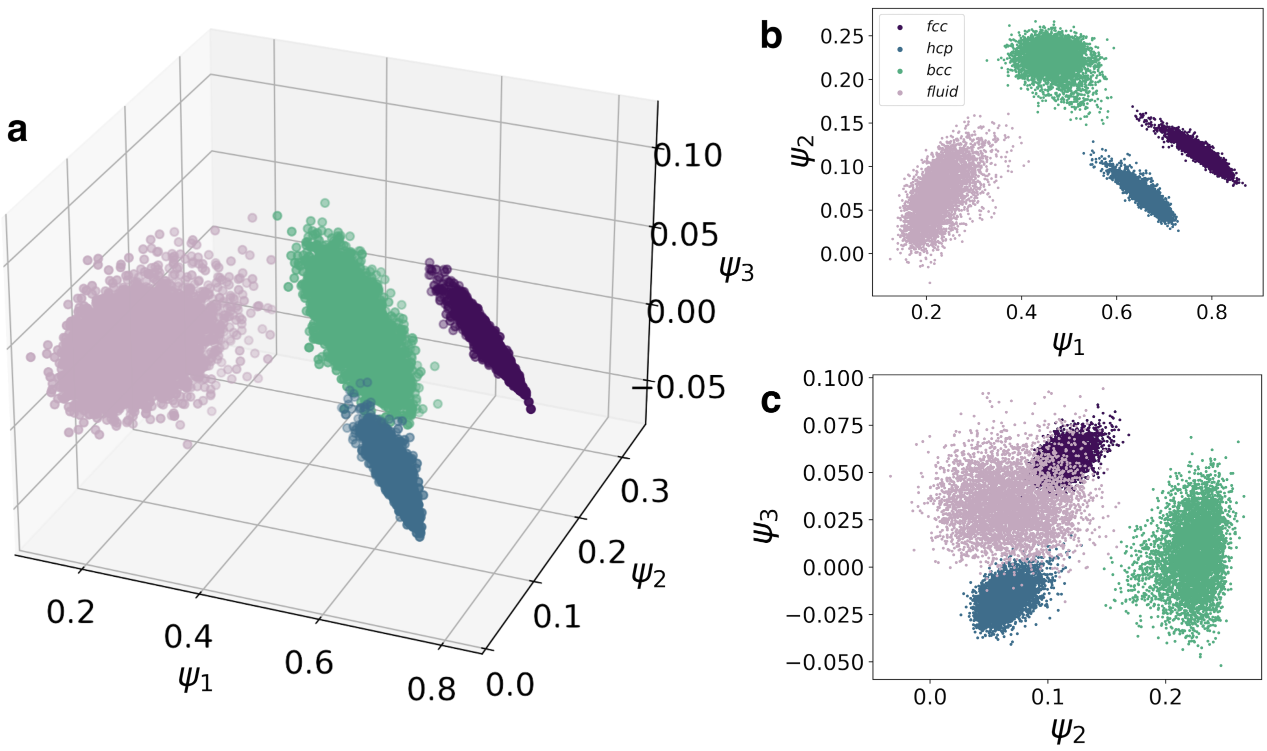


Figure 5.4: Projection of $\mathbf{q}(i)$ of the four bulk phases – fcc, hcp, bcc, and fluid – (a) onto the three-dimensional space spanned by ψ_1, ψ_2, and ψ_3, (b) onto the two-dimensional space spanned by ψ_1 and ψ_2, and (c) onto the two-dimensional space spanned by ψ_2 and ψ_3.

PCA almost completely ignores the odd symmetries. Another surprising result is that $\bar{q}_4$ is only relevant in the third eigenvector, while $\bar{q}_8$, $\bar{q}_{10}$, and $\bar{q}_{12}$, which are barely mentioned in the literature, are found to be highly prioritised for expressing the variance in the data set.

Given the new space spanned by the three selected eigenvectors, it is possible to project the data for the four bulk phases onto the new space. The results of the projection are shown in Fig 5.4. Specifically, in Fig. 5.4a we show the projection of the entire data set in the full three-dimensional space described by ψ_1, ψ_2, and ψ_3. Due to the convenience of working with 2D plots, we also show two alternative projections. The first, in Fig. 5.4b shows the data in the $\psi_1 - \psi_2$ space, while the second, in Fig. 5.4c, shows the same data in the $\psi_2 - \psi_3$ space. In the rest of the work we will analyse a spontaneous nucleation trajectory in these subspaces in order to extract information about each particles's local environment.

5.4 Results

5.4.1 Simple thresholds-based classification

Now that we have set up the new space found by PCA which enables us to distinguish the phases under investigation, we start analysing the nucleation phenomenon in this new space. We therefore consider a spontaneous nucleation trajectory at a pressure equal to $\beta P \sigma^3 = 13.80$ (the same as that analysed in Chapter 4). We then select three different configurations, the first at the very early stages of crystallisation, the second during the crystal growth phase, and the third when the crystal has almost completely spanned the whole simulation box. We calculate the vector $\mathbf{q}(i)$ for each particle i and project the results onto the PCA space. This analysis is displayed in Fig. 5.5. In particular, the time instants we take into account correspond to $t^* = 200$, $t^* = 250$, and $t^* = 300$.

In Fig. 5.5a the projection of the trajectory for these three time steps are shown on the space spanned by ψ_1 and ψ_2. The data corresponding to the particles start from the fluid phase, at the beginning of the crystallisation, and make a path in the PCA space towards the zone of the space where the clouds corresponding to the fcc and hcp crystals are present. In contrast, only a few particles intercept the bcc region, which does not seem to be involved in the crystallisation process. Finally, especially in Fig. 5.5c it is evident that the final location of many points is in a *third cloud*, in between the ones corresponding to the fcc and the hcp crystals. This is due to the stacking faults between these two structures which are present in the grown crystal, as we shall see in a successive analysis. In Fig. 5.5b we show the same analysis, but the data is projected onto the space defined by ψ_2 and ψ_3.

Given the projection of the data, it becomes crucial to establish a classification scheme to obtain quantitative results about the number of particles in the crystal nucleus, and the composition of the nucleus. To outline this algorithm, we first construct the distributions of the projection values on each of the three eigenvectors from each individual phase. The distributions of ψ_1, ψ_2, and ψ_3 of the four phases are shown in Fig. 5.6.

The outcome of these projections gives us a straightforward option in order to construct a classification algorithm, as each eigenvector alone is able to distinguish one phase from the others. In particular, ψ_1 distinguishes the fluid phase from all the three crystal phases (see Fig. 5.6a), ψ_2 separates well the bcc crystal of the rest of the phases (see Fig. 5.6b), while the final distinction between fcc and hcp is carried out through ψ_3 (see Fig. 5.6c). The vertical dashed black lines in the plots indicate the thresholds we selected in order to classify each particle.

Using the scheme described above, we analyse the same trajectory whose projections are shown in Fig. 5.5. At each snapshot, we construct the vector $\mathbf{q}$ for each particle. We then project the same vector onto ψ_1, ψ_2, and ψ_3, and based on the values of the three projection we assign a class to the particle. We repeat the process for the whole trajectory and we show the results in Fig. 5.7.

The total number of particles belonging to the main nucleus – together with the total number of fcc-like, hcp-like, and bcc-like particles – is shown in Fig. 5.7a. From the plot, it can already be seen that fcc-like particles are predominant according to this classification scheme, as also found in Chapter 4. This is confirmed in Fig. 5.7b, where we plot the composition of the nucleus as a function of time, during the crystal growth phase. We find that the fraction of fcc-like, hcp-like, and bcc-like particles in the nucleus are respectively close to 65%, 25%, and 10%.

Finally, in Fig. 5.7c, Fig. 5.7d, and Fig. 5.7e, we show three consecutive snapshots of

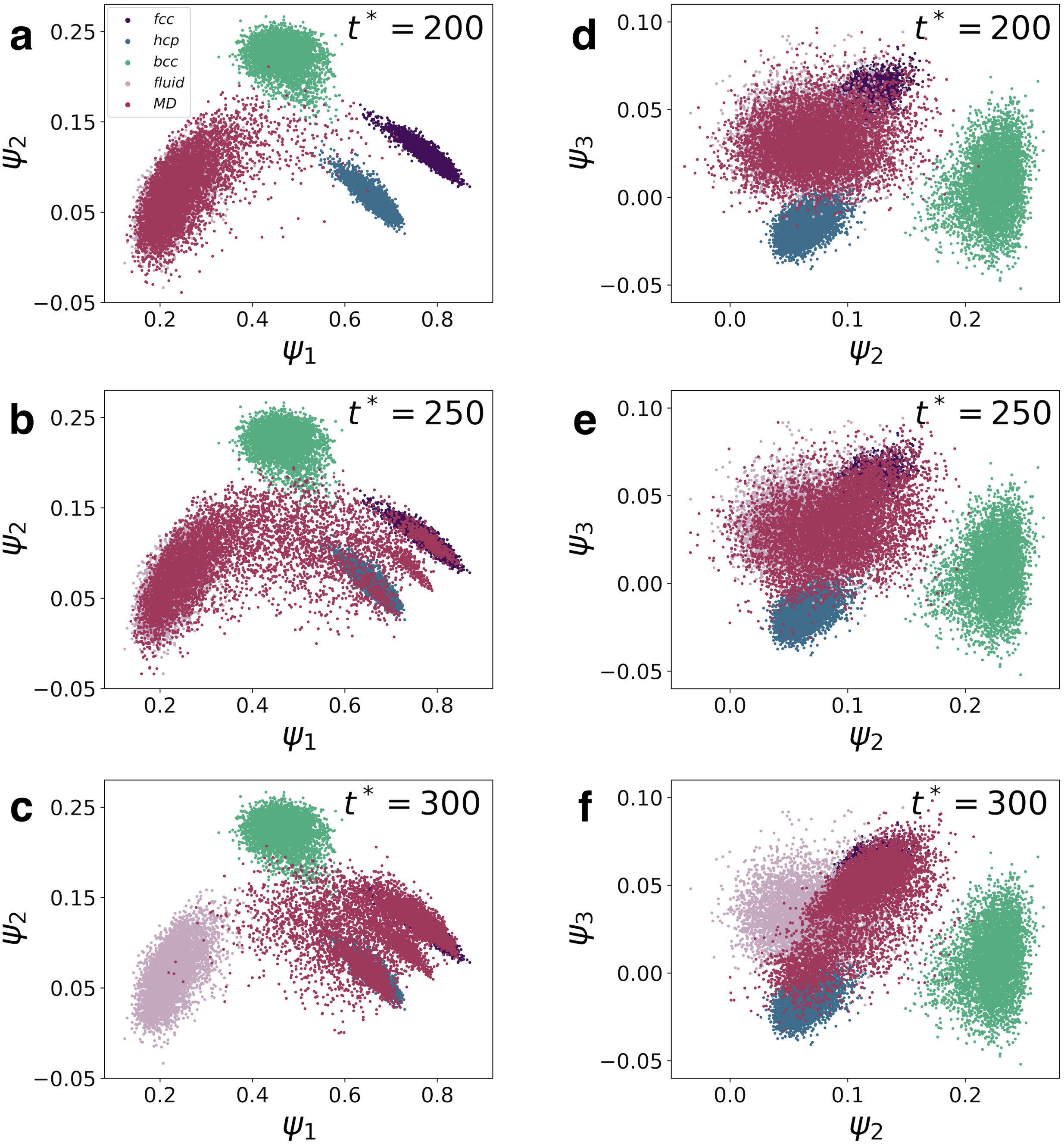


Figure 5.5: Projection of $\mathbf{q}(i)$ of the four bulk phases – fcc, hcp, bcc, and fluid – and the spontaneous nucleation trajectory of nearly hard spheres at $t^* = 200$, $t^* = 250$, and $t^* = 300$ (a,b,c) onto the two-dimensional space spanned by ψ_1, ψ_2, and (d,e,f) onto the two-dimensional space spanned by ψ_2 and ψ_3.

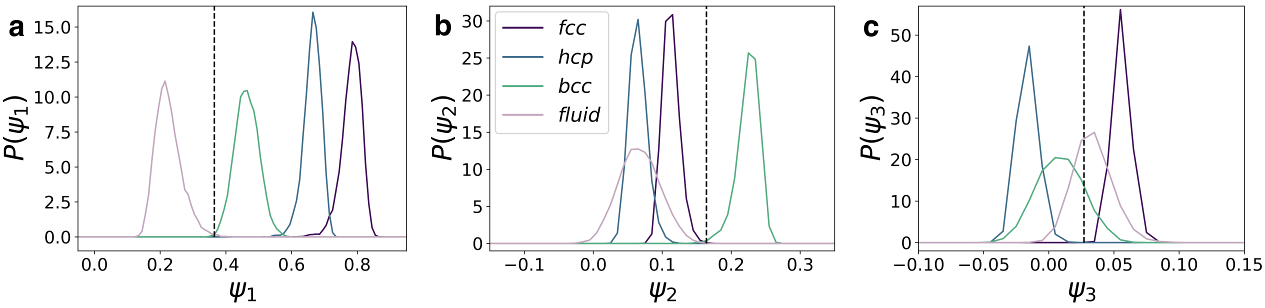


Figure 5.6: Distributions of the projection of the data of the bulk phases on top of the first three eigenvectors. The distribution of ψ_1 for each phase is shown in (a), and shows that this feature separates well the fluid phase (lilac curve) from the three crystal phases. Conversely, in (b) we show the distribution of ψ_2 for each phase, which allows us to distinguish the bcc crystal (green curve) from the other phases. Finally, the distribution of ψ_3, which helps separating the fcc-like (dark purple curve) from the hcp-like (blue curve) particles, is shown in (c).

the growing nucleus, where the colouring of the particles follow the PCA-based classification algorithm. These snapshots show that the algorithm is capable of detecting the embryo even in its very early stages, it then follows the crystallisation process during the crystal growth phase, and detects stacking faults between the fcc-like and the hcp-like stackings in the crystal.

5.4.2 Gaussian Mixture Model

In the previous Section we showed how the BOP distributions can be employed, once projected onto the first three eigenvectors, to establish a classification scheme for the phases under investigation. Obviously, this is not the only choice, and many alternatives are available.

One choice is to resort again to unsupervised learning techniques and to build a model capable of clustering the points automatically. In this Section, we use Gaussian Mixture Model (GMM) to achieve this goal. GMM is a technique that assumes that the data set is distributed according to a finite number of multivariate Gaussian distributions [217–219]. Once the model is trained on the data set and the optimal parameters of the Gaussian distributions are found, posterior probabilities can be assigned to each new data point, describing the probability of that data point being generated from each of the Gaussians in the model. Once these probabilities are computed, the simplest way to assign a class based on the output of the GMM is to choose the class which corresponds to the multivariate Gaussian distribution for which the posterior probability is the highest.

This procedure works well especially in cases where the data are actually distributed according to a discrete number of multivariate Gaussians. During a spontaneous nucleation event, however, there are time instants in which the particles do not have environments that can be easily attributed to the fluid phase, nor to a typical crystalline phase. To detect these particles – which especially in the initial instants of nucleation correspond to most of the particles composing the nucleus – we adopt a different strategy. For each particle, we first check the highest posterior probability. However, the corresponding class is actually assigned only if the data point is not further away than five times the covariance matrices from the centre of the Gaussian itself. If this criterion is not fulfilled, the particles are assigned a new label, which

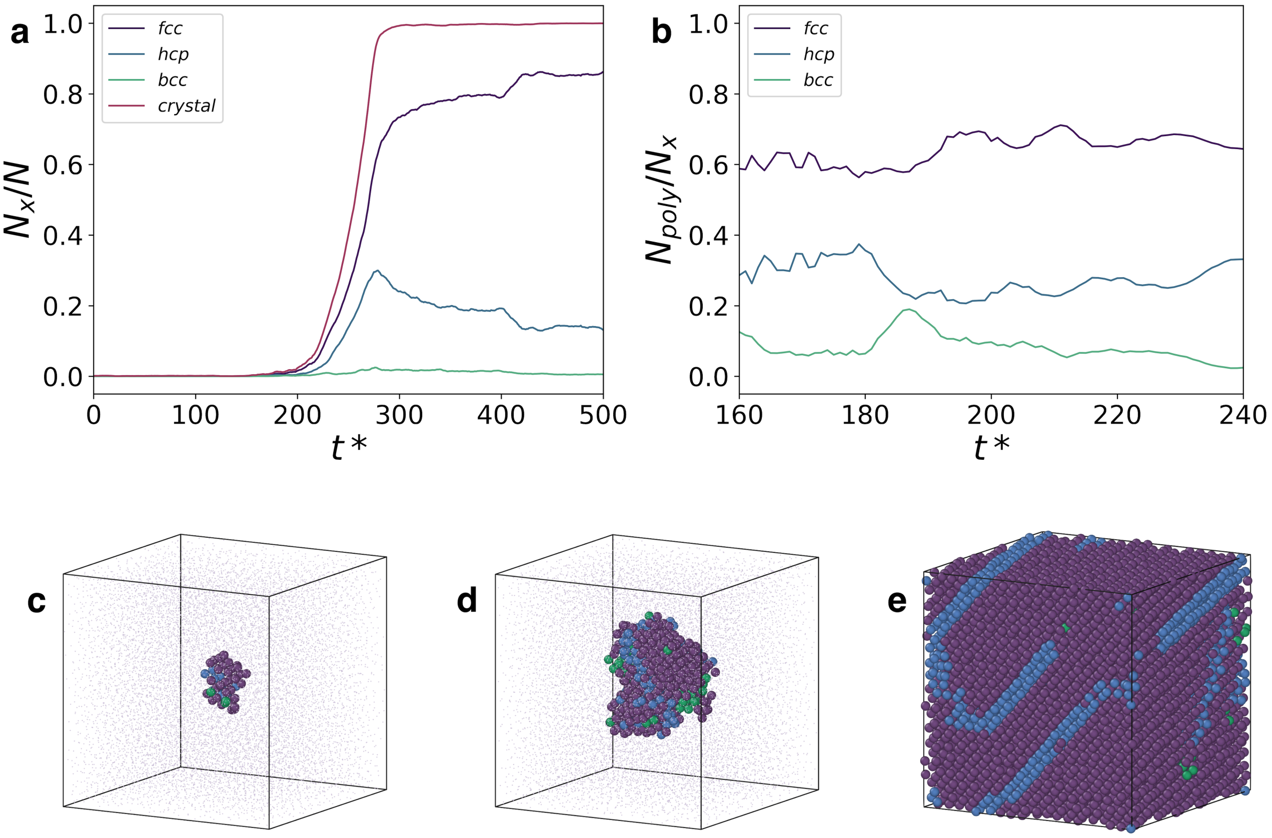


Figure 5.7: Results from the classification scheme using the first three eigenvectors given by PCA. (a) The fraction of crystalline particles, and of each crystal polymorph as a function of time, during a spontaneous nucleation trajectory. The prevalent polymorph as detected by this classification scheme is the fcc crystal. (b) The composition of the growing nucleus during the growth phase as a function of time. The fraction of fcc-like, hcp-like, and bcc-like particles in the nucleus is about 65%, 25%, and 10%, respectively. (c,d,e) Three snapshots of the system where the particles are coloured according to the colour coding in (a) and (b).

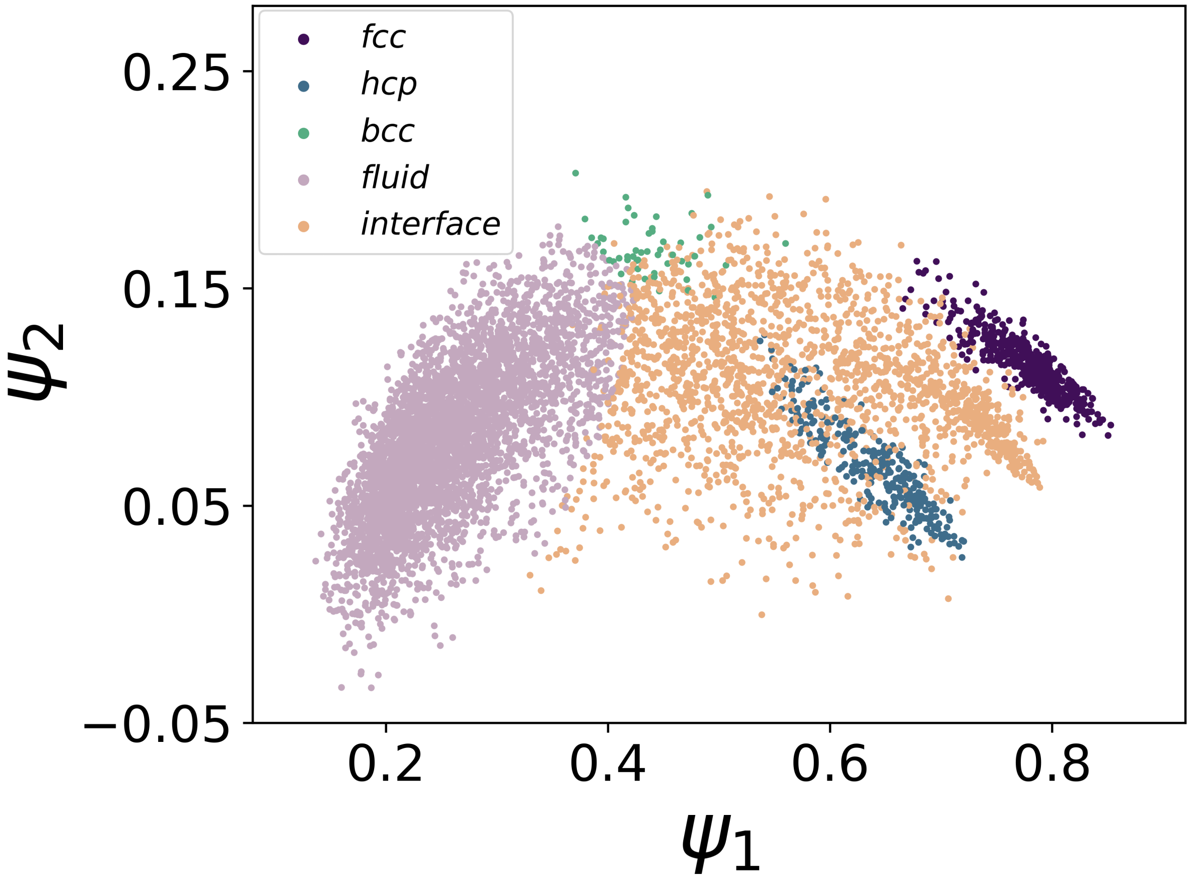


Figure 5.8: Data of a snapshot during a spontaneous nucleation event, with each particle being classified using GMM and coloured accordingly. The colour coding for the four different classes is the same as in the previous figures, while the interface particles – particles whose local environment do not correspond to one of the known bulk phases – are coloured orange. Differently from the previous figures, where also the bulk reference phases were displayed, only the data points coming from MD simulations are shown.

indicates that they are interface particles. The results of this procedure on a snapshot during a spontaneous nucleation path is shown in Fig. 5.8.

Classification strategies which do not include interface particles are often forced to assign a class even in cases where particles do not show the same local environment of the bulk phase it is assigned to. However, using the current strategy, this does not happen. Furthermore, the study of the interface of a growing nucleus is filled with open questions, and we believe that a correct identification of such particles is the first step towards a better comprehension of the crystallisation mechanism. As a proof of the ability of this technique to locate interface particles, we show in Fig. 5.9 three consecutive snapshots of the growing nucleus, where the colouring of the particles follow the GMM-based classification algorithm, as in Fig. 5.8.

In Fig. 5.8a the embryo in its very early stages is detected, and is composed of only interface particles. When the nucleus grows, we see that the orange-coloured interface particles form a coating around the bulk phase. Finally, when the crystal has formed and spanned the simulation box, most of the particles are either fcc-like or hcp-like, but interestingly the stacking faults

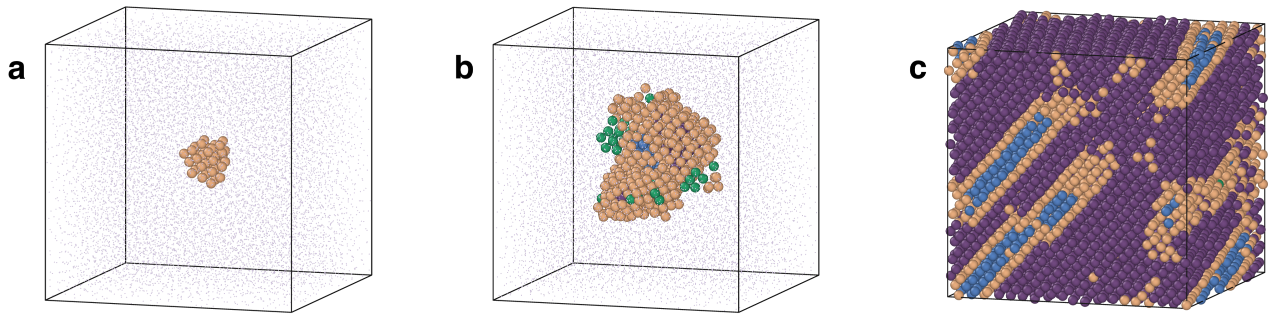


Figure 5.9: Results of the GMM-based classification scheme. In (a), (b), and (c) we plot three snapshots of the system as a result of the classification described. The colouring of the particles follow the colour coding in Fig. 5.8.

between fcc and hcp are also recognised as interface. This is a proof that this strategy not only detects fluid-crystal interfaces, but also interfaces between two crystals.

In this section we have shown that, once the dimensionality has been reduced *via* PCA, a clustering algorithm like GMM can autonomously label the particles based on their local environments, in an interpretable way. Interestingly, this technique can be implemented with the possibility of detecting the interface particles, in addition to the four bulk phases. Despite being a useful feature for a classification algorithm, in the context of the task shown in this Chapter, we note that the GMM-based classification has a disadvantage with respect to the threshold-based classification described in Section 5.4.1. In fact, using GMM, all the eigenvectors influence the classification for each particle, as the clustering takes place in three dimensions. This is not the ideal scenario, as some eigenvectors may be not particularly good to perform specific distinctions, and might therefore lead to wrong classifications. Conversely, in the scheme described in Section 5.4.1, every phase is recognised using an eigenvector which is designed for that specific task. For instance, ψ_1 works well for distinguishing fluid-like from solid-like particles, and therefore the classification of fluid particles is solely based on ψ_1. The same happens with ψ_2, which is particularly good for distinguishing bcc-like particles from fcc-like or hcp-like ones, and with ψ_3, which separates nicely fcc and hcp.

5.4.3 Incomplete data set

In this Chapter, we have seen how PCA finds a lower-dimensional linearly transformed subspace of the higher-dimensional space of the original data set, in which it is possible to distinguish well all the phases that are given to PCA as input. It is legitimate to wonder what kind of projection is obtained, if one or more phases are omitted from the initial data set.

We therefore construct an alternative data set, including the three crystal structures analysed above – fcc, hcp and bcc – but excluding the fluid phase. In this way PCA will find a new subspace that does not have to deal with finding eigenvectors that maximise the variance when considering also the BOPs of the fluid phase.

We repeat the procedure described in Section 5.3, and as a result PCA finds a new space, identified by new eigenvectors. Also here, in order to analyse a spontaneous nucleation trajectory in the new space found by PCA, we calculate the vector $\mathbf{q}$ for each particle at each time step and project the results on ψ_1 and ψ_2. The results obtained in three consecutive time steps

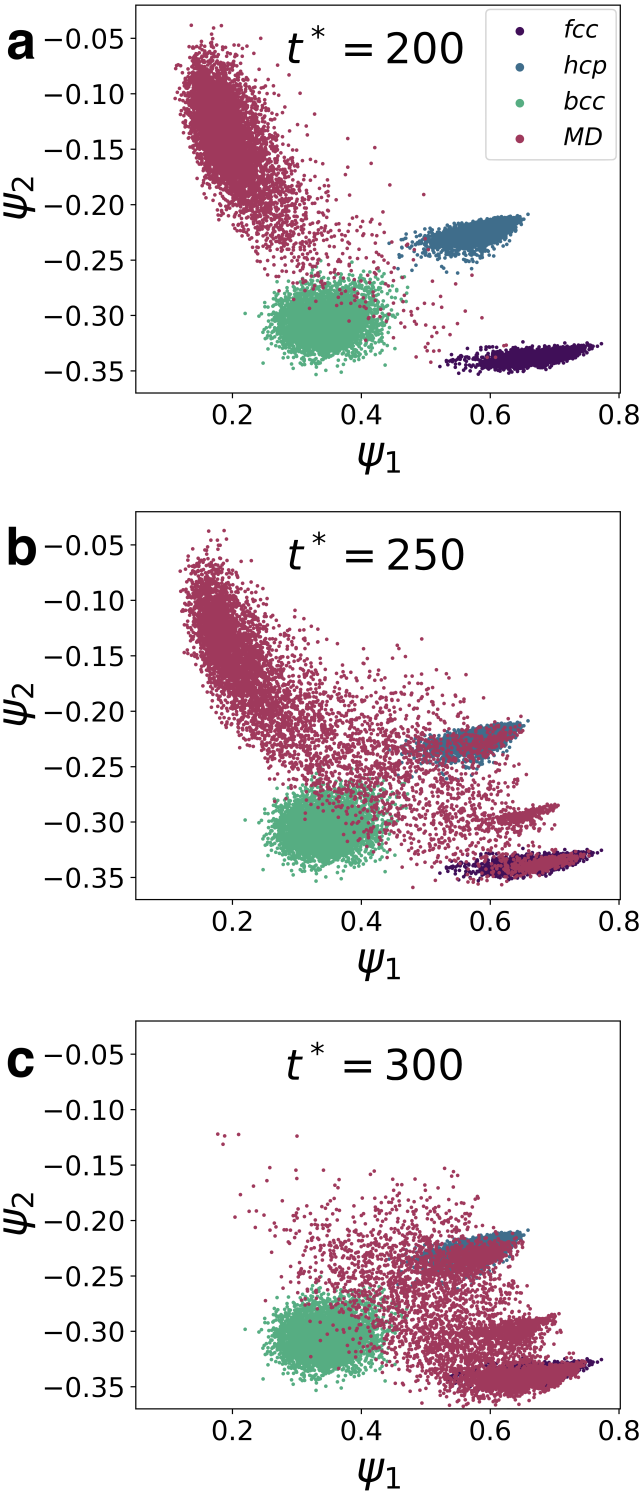


Figure 5.10: Projection of $\mathbf{q}(i)$ of the three bulk phases – fcc, hcp, and bcc – and the spontaneous nucleation trajectory of nearly hard spheres at $t^* = 200$, $t^* = 250$, and $t^* = 300$ (a,b,c) onto the two-dimensional space spanned by ψ_1 and ψ_2. Note that the subspace spanned by ψ_1 and ψ_2 is different from the one in Fig. 5.5, as the present subspace was found using PCA on a data set in which the fluid data was removed.

– $t^* = 200$, $t^* = 250$, and $t^* = 300$ – are shown in Fig. 5.10. Despite having projected the same trajectory at the same time steps, the projection in Fig. 5.10 exhibits a peculiar and different behaviour with respect to Fig. 5.5. In this space, which is not optimised for the distinction of the crystals from the fluid phase, the particles of the fluid phase seem to strongly intercept the area corresponding to the bcc crystal, before making a further transition to the hcp or the fcc crystals. This phenomenon did not happen by looking at Fig. 5.5, where the cloud corresponding to the bcc crystal was mainly avoided by the path made by the particles in the PCA subspace.

This is further evidence of the ease with which erroneous conclusions can be drawn by looking at nucleation events through order parameter filters that are either lacking information, or are not optimised to describe the phenomenon under investigation. Also, when applying PCA, it is crucial to build a data set where all the phases we wish to distinguish are present.

5.5 Conclusions

To conclude, in this Chapter we have studied the problem of a single particle-based classification of a system of nearly hard spheres during a nucleation event. We then devised a novel PCA-based classification scheme which shows a number of advantages over standard techniques which involve the use of a small number of features. First of all, PCA is able to find an optimal linear combination of several order parameters. This results in more information about the local environment of each particle being considered in the classification process. Secondly, the method is optimised for the system in question, since the subspace found by PCA is the result of applying the technique to a data set where only the phases we want to monitor are present. Finally, besides being fast and robust, the method allows us to perform a real feature extraction operation. By looking at the components of the eigenvectors, it is in fact possible to gain new insights into which features are useful to describe the nucleation phenomenon. We can, in other words, learn something ourselves from using this technique.

In general, we expect that unsupervised learning techniques can play an increasingly leading role in the identification of key features of physical phenomena, and not only of nucleation. On the other hand, in the field of soft matter, we expect that other phenomena of self-assembly or even of the glass transition can be better understood and described, starting from these kind of approaches, for both systems coming from simulations and experiments.

Acknowledgements

I would like to thanks Bryan Verhoef for useful discussions about this project.

6

Inverse design of charged colloidal particle interactions for self assembly into specified crystal structures

In this Chapter we study the inverse problem of tuning interaction parameters between charged colloidal particles interacting with a hard-core repulsive Yukawa potential, so that they assemble into specified crystal structures. Here, we target the body-centred-cubic (bcc) structure which is only stable in a small region in the phase diagram of charged colloids and is, therefore, challenging to find. In order to achieve this goal, we use the statistical fluctuations in the bond orientational order parameters to tune the interaction parameters for the bcc structure, while initialising the system in the fluid phase, using the Statistical Physics-inspired Inverse Design (SP-ID) algorithm. We also find that this optimisation algorithm correctly senses the fluid-solid phase boundaries for charged colloids. Finally, we repeat the procedure employing the Covariance Matrix Adaptation - Evolution Strategy (CMA-ES), a cutting edge optimisation technique, and compare the relative efficacy of the two methods.

Based on: R. Kumar, G. M. Coli, M. Dijkstra, and S. Sastry, *Inverse design of charged colloidal particle interactions for self assembly into specified crystal structures*, The Journal of Chemical Physics, **151 (8)**, 084109 (2019)

6.1 Introduction

In the past few years, several inverse methods have emerged that design optimal interactions in such a way that the system spontaneously assembles into a targeted structure [220–222]. These methods have received considerable attention in materials science [223–225], and have been successively used to find crystal structures for photonic band-gap applications [43], to predict crystals [44] and protein structures [45, 46], materials with optimal mechanical and transport properties [47], and for optimising the interactions for self assembly [226–230]. Though many of the methods developed are either based on black-box techniques, in which the algorithm tunes the interaction parameters without taking the statistical nature of the system into account, or are designed *ad hoc* for a particular class of systems, systematic approaches based on a statistical mechanical formulation, which are general, and allow for application tailored to specific systems of interest, have also been investigated. Recently, several methods have been proposed which take into account the statistical nature of the system in updating the inter-particle interaction parameters [231–233]. Lindquist *et al.* [231] proposed a relative entropy based inverse design approach which tunes the inter-particle interaction parameters for the self assembly of crystal structures by exploiting the statistical nature of the system. Later, Adorf *et al.* [233] modified this method by carrying out relative entropy minimisation directly in Fourier space to target the complex crystal structure. These methods have been successfully employed to design isotropic potentials to assemble a wide range of crystal structures.

In the present Chapter, we investigate the efficacy of one such method, the Statistical Physics-inspired Inverse Design (SP-ID) method developed by Miskin *et al* [234]. This method considers statistical fluctuations present in the microscopic configurations of the system for tuning the interactions between the particles. Apart from tuning the inter-particle interaction parameters, system parameters like temperature and pressure can also be tuned using this method. In order to design these interactions, we have used a quality function based on bond order parameters [235, 236] to rank the generated configurations. The same task can be faced employing numerous optimisation techniques like relative entropy based inverse design [231–233], (Adaptive) Simulated Annealing [237–239], Particle Swarm Optimisation [240], and several genetic algorithms [241]. To evaluate the effectiveness of the SP-ID algorithm, we compare it with the Covariance Matrix Adaption - Evolutionary Strategy (CMA-ES) [242–244], which we regard as a state-of-the-art optimisation technique for evolutionary computation.

We have chosen a system of colloidal particles as the model for which we wish to design the interactions in order to target a specific crystal structure. The interparticle potential of colloids offers a wide variety of functional forms. It can contain a hard-core term, a dipole-dipole term, a charge dispersion term, a screened-Coulomb (Yukawa) term and a short-ranged attractive depletion term. More specific designed colloidal particles such as patchy colloids and DNA-functionalised colloids offer even greater diversity of interactions. For all of these, the interaction parameters can be tuned. In the case of Yukawa interactions, the Debye screening length can be adjusted by changing the salt concentration, and the contact value can be tuned by altering the surface charge of the particles.

Previous studies have employed what may be termed a *forward method* in which, starting from the microscopic interaction parameters and specified thermodynamic parameters such as temperature and density, one computes the equilibrium properties and finds the stable phase of the system [245–247]. In the present work, our purpose is to employ an *inverse method*, where the target structure and equilibrium properties are the input from which we wish to design the interparticle interactions for which the particles spontaneously self assemble into the target structure.

In this study, we considered a charged colloidal system in which particles interact with a hard-core repulsive Yukawa potential. The complete phase diagram of hard-core Yukawa particles is known from earlier studies [245–247]. Because of the purely repulsive nature of the potential, this system displays only a fluid phase which can freeze into a face-centred-cubic (fcc) or a body-centred-cubic (bcc) crystal phase. The phase diagram of hard-core Yukawa particles shows one or two triple points where fcc, bcc, and fluid phases coexist. The bcc phase is only stable in a very small region in the phase diagram and, for this reason, constitutes a good test case for the reverse-engineering process. In this work, we optimise the thermodynamic conditions and the interparticle interactions which favour the crystallisation of the bcc structure. Here we show three different cases, in which we respectively tune one, two, or three parameters. In all cases, we show that both SP-ID and CMA-ES find the thermodynamic conditions and the interparticle interaction parameters that lead to the targeted bcc structure formation.

The Chapter is organised as follows: In Sec. 6.2, we define the model system studied in this work and the corresponding phase diagram and bond order parameters to identify the different phases. In Sec. 6.3, we describe the inverse design methods and the form of the quality function used in this study. Results for tuning the interactions to target the bcc structure for all three cases are discussed in Sec. 6.4. Finally, we summarise our results in Sec. 6.5.

6.2 Model and simulation methods

6.2.1 Interaction potential

We consider a hard-core repulsive Yukawa system which represents a standard model for charged colloids. The form of the potential is given by

$$\beta U(r) = \begin{cases} \beta\epsilon \frac{\exp[-\kappa\sigma(r/\sigma-1)]}{r/\sigma} & \text{for } r > \sigma \\ \infty & \text{for } r \leq \sigma, \end{cases} \tag{6.1}$$

where $\beta\epsilon$ is the contact value of the pair potential expressed in units of $k_BT = 1/\beta$, k_B is the Boltzmann constant, κ is the inverse of the Debye screening length, and σ is the hard-core diameter. In Fig. 6.1, we show the phase diagram for such a system with $\beta\epsilon = 8$ in the $(1/\kappa\sigma, \eta)$ representation and the $(1/\kappa\sigma, \beta P\sigma^3)$ plane [247]. The phase diagram exhibits a stable fluid, bcc, and fcc region, as well as two triple points at which the three phases coexist. Note that we always use scaled variables, a reduced temperature k_BT/ϵ, reduced pressure $\beta P\sigma^3$ and inverse screening length $1/\kappa\sigma$, each of which can be treated as tuning parameter for obtaining the desired behaviour.

6.2.2 Bond Order Parameters

Every optimisation algorithm, including SP-ID and CMA-ES, works on the basis of minimising a user-defined fitness function. Here, we have used the averaged bond order parameters (BOP) $\bar{q}_l$ and $\bar{w}_l$ $(l = 6)$. The first is computed as follows: [235, 236],

$$\bar{q}_l^{(i)} = \left(\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |\bar{q}_{lm}^{(i)}|^2 \right)^{1/2}, \tag{6.2}$$

where

$$\bar{q}_{lm}^{(i)} = \frac{1}{\tilde{N}_b(i)} \sum_{j=0}^{\tilde{N}_b(i)} q_{lm}^{(j)}$$

$$q_{lm}^{(i)} = \frac{1}{N_b(i)} \sum_{j=1}^{N_b(i)} Y_{lm}(\theta(\mathbf{r}_{ij}), \phi(\mathbf{r}_{ij})).$$

Here, $N_b(i)$ is the number of neighbours of particle i, $\tilde{N}_b(i)$ is the number of neighbours including particle i itself, $Y_{lm}(\theta(\mathbf{r}_{ij}), \phi(\mathbf{r}_{ij}))$ denotes the spherical harmonics with $\mathbf{r}_{ij}$ the distance vector from particle i to particle j. The second bond order parameter we used is defined as

$$\bar{w}_l^{(i)} = \frac{\sum\limits_{m_1+m_2+m_3} \begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} \bar{q}_{lm_1}^{(i)} \bar{q}_{lm_2}^{(i)} \bar{q}_{lm_3}^{(i)}}{(\sum\limits_{m=-l}^{l} |\bar{q}_{lm}^{(i)}|)^{3/2}}$$

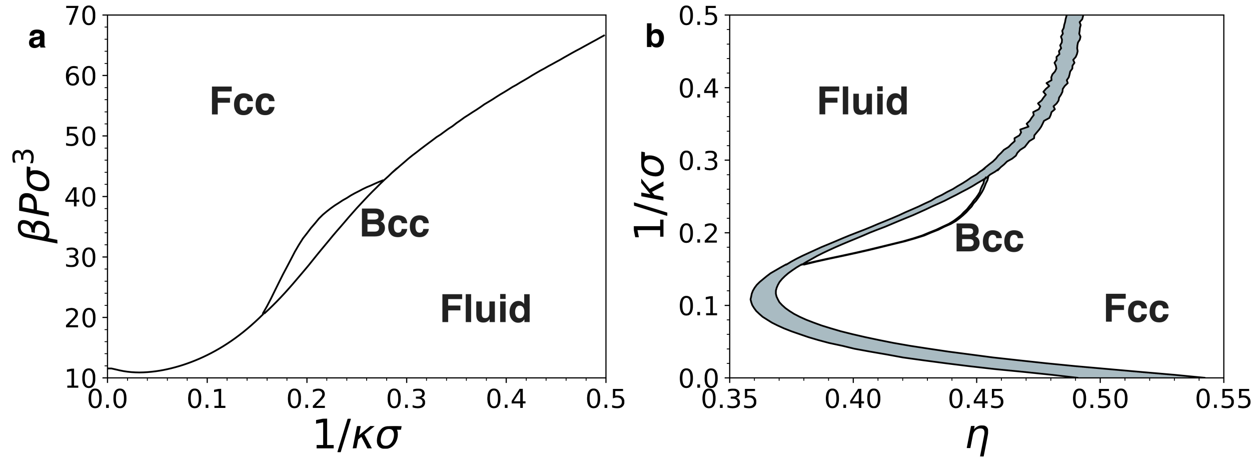


Figure 6.1: Phase diagram of a system in which the particles interact *via* a hard-core repulsive Yukawa pair potential with $\beta\epsilon = 8$ in a) the $(1/\kappa\sigma, \beta P\sigma^3)$ plane and b) the $(1/\kappa\sigma, \eta)$ representation. The phase diagram displays a stable fluid, bcc, and fcc phase [246, 247].

In order to calculate the radius of the first coordination shell for each particle, we employ the solid angle based nearest-neighbour (SANN) algorithm [248], where a nearest neighbour of a particle is identified by attributing a solid angle to each possible neighbour such that the sum of solid angles equals at least 4π.

In Fig. 6.2, we show the scatter plots of $\bar{q}_6$ *versus* $\bar{w}_6$ for the fluid, bcc, and fcc phases of a system of Yukawa particles with $\beta\epsilon = 8$ at the high-density triple point conditions. We find distinct clouds of points corresponding to the fluid, bcc, and fcc phases, and hence, $\bar{q}_6$ and $\bar{w}_6$ can be used to distinguish the three phases.

6.3 Inverse design methods

6.3.1 Statistical Physics-inspired Inverse Design method

In the SP-ID method, microscopic parameters such as the interparticle pair potential parameters are tuned by exploiting statistical fluctuations in such a way that the system will evolve to those states which correspond to the targeted macroscopic response of that system [234]. In this method, the time evolution of the probability distribution of finding the system in configuration x is written as

$$\dot{\rho}(x|\lambda_i) = \rho(x|\lambda_i)\left[f(x) - \langle f(x)\rangle\right], \tag{6.3}$$

where ρ denotes the probability of finding a system in some configuration, $\lambda_i' s$ are the adjustable parameters and $f(x)$ is a quality function which gives a weight/fitness value to each configuration based on a targeted macroscopic property. Time t is an artificial time that indexes the optimisation steps and $\langle f(x)\rangle$ is the ensemble average over all the configurations. With straightforward manipulation, the above equation can be recast as equations of motion for λ_i [234]:

$$\begin{aligned} \dot{\lambda}_i(t) &= \langle \partial_{\lambda_i} log\,(\rho)\, \partial_{\lambda_j} log\,(\rho)\rangle^{-1} \\ &\quad \times\langle[f(x) - \langle f(x)\rangle]\, \partial_{\lambda_j} log\,(\rho)\rangle, \end{aligned} \tag{6.4}$$

where $< .. >$ denotes an ensemble average at a given set of values of λ_i. To integrate Eq. 6.4, we have used a modified Euler method with a fixed time step of 4.0.

We have built our quality function $f(x)$ in the following way:

$$f(x) = \int dx' \Theta(g(x') \geq g(x))\rho(x'|\lambda) \tag{6.5}$$

where $\Theta(a \geq b)$ is equal to 1 whenever the inequality in the argument is fulfilled and zero otherwise, and $g(x)$ denotes the fitness function,

$$g(x) = (\bar{q}_6(x) - \bar{q}_6^{target})^2 + (\bar{w}_6(x) - \bar{w}_6^{target})^2 \tag{6.6}$$

with $\bar{q}_6(x) = \sum_i^N \bar{q}_6^{(i)}/N$ and $\bar{w}_6(x) = \sum_i^N \bar{w}_6^{(i)}/N$ the averages of the bond order parameters over all the particles in the system, and $\bar{q}_6^{target}$ and $\bar{w}_6^{target}$ are the corresponding quantities in the target structure (bcc). Here, the integral over x' represents a sum over a series of n different configurations as obtained from a simulation for a fixed set of parameters $(\kappa\sigma, \beta P\sigma^3, \beta\epsilon)$. The quality function $f(x)$ will have higher values for those configurations whose $\bar{q}_6$ and $\bar{w}_6$ values are closer to the target values. More precisely, $f(x)$ equals the probability of having a lower value of $g(x)$ than any other configuration drawn randomly from the equilibrium distribution. To target the bcc structure, we have chosen $\bar{q}_6^{target} = 0.395$ and $\bar{w}_6^{target} = 0.011830$ (average values of $\bar{q}_6$, $\bar{w}_6$ for the bcc structure obtained from the scatter plot shown in Fig. 6.2). For a perfect bcc structure, these values are, $\bar{q}_6 = 0.5107$ and $\bar{w}_6 = 0.013161$, but here we have targeted the bond order parameter values for a finite-temperature bcc structure. This choice of the fitness function denotes a focus in the crystal symmetry only. Generally, it strongly depends on what the user is interested in, and might include other quantities, like the lattice spacing.

6.3.2 Covariance Matrix Adaptation-Evolutionary Strategy

In order to implement CMA-ES, we draw n samples from a multivariate Gaussian distribution for each generation whose dimension D corresponds to the number of parameters we wish to tune. Subsequently, we evaluate the fitness function $g(x)$ on the generated samples, and we pick the best k samples. Using the following equations, we estimate the multivariate Gaussian distribution with mean $\vec{\mu}$ (a D-dimensional vector) and $\mathbf{\Sigma} = \sigma^2\mathbf{C}$ the covariance matrix of the Gaussian distribution for the next generation using:

$$\begin{aligned}
{\mu_i}' &= \mu_i + \sum_x w(x)(\lambda_i(x) - \mu_i) \\
{q_i}' &= (1 - c_1)q_i + c_2 \left(\sqrt{\Sigma^{-1}}\right)_{ij} ({\mu_j}' - \mu_j) \\
{p_i}' &= (1 - c_3)p_i + c_4({\mu_i}' - \mu_i) \\
{C_{ij}}' &= (1 - c_5 - c_6)C_{ij} + \\
&+ c_5 \sum_x w(x)(\frac{\lambda_i(x) - \mu_i}{\sigma}\frac{\lambda_j(x) - \mu_j}{\sigma} - C_{ij}) + c_6{p_i}'{p_j}' \\
\sigma' &= \sigma \exp\left[c_7(\frac{\parallel \vec{q'} \parallel}{\langle\parallel N(0, I) \parallel\rangle} - 1)\right]
\end{aligned} \tag{6.7}$$

where $\{x\}$ denotes the n samples consisting of multiple configurations calculated for n different parameter sets $(\kappa\sigma, \beta P\sigma^3, \beta\epsilon)$ (denoted by $\lambda_i(x)$ above) in CMA-ES, $w(x)$ is the normalised distribution of weights based on the fitness of the samples. We choose $w(x) \propto \log(k+1) - \log(i)$ where i is the rank index of sample x ($i = 1$ for the configuration with the smallest $g(x)$ value) for the best k samples, and set $w(x) = 0$ for the rest. $\vec{q}$ and $\vec{p}$ are additional D-dimensional vectors which determine, respectively, the changes in amplitude and directionality of the covariance matrix, and finally $\langle\parallel N(0, I) \parallel\rangle$ is the average length of a vector drawn from a multivariate Gaussian distribution centred in the origin and where the covariance matrix is the identity matrix. In the present work we use $n = 20$ and $k = 5$. For the first generation we initialise $\vec{q}$ and $\vec{p}$ as null vectors. Moreover, since we do not assume any *a priori* correlation between the different tuning parameters, the initial form of the covariance matrix $\mathbf{\Sigma}$ is diagonal. Finally, all the free parameters c_i of the CMA-ES are selected following the recipe in Ref. 244. It is important to note that in CMA-ES, since the samples are randomly extracted at each generation, there are often situations in which different crystal structures are simulated during the same generation. Therefore we have to make sure that the bcc crystal is always the one with the lowest fitness, and that no BOP fluctuations can reverse this order. For this reason, we found that a slightly lower value $\bar{q}_6^{target} = 0.377$ makes CMA-ES more effective in reaching the goal of targeting the bcc structure. This is due to the typical low values assumed by $\bar{w}_6$ in the phases considered here, which weigh less in the fitness function.

Finally, we would like to stress that there is a substantial difference in the use of the fitness function between the two methods. In SP-ID we rank different configurations, so $\bar{q}_6(x)$ and $\bar{w}_6(x)$ are bond order parameters computed in a single configuration. The SP-ID method is thus based on the statistical fluctuations in the bond order parameters (Fig. 6.2) in a simulation at a single set of interaction parameters in order to optimise these values for the desired bcc structure. When using the CMA-ES, we rank samples, so $\bar{q}_6(x)$ and $\bar{w}_6(x)$ are computed as ensemble averages of these bond order parameters over multiple configurations for distinct sets of parameters. The CMA-ES method is thus based on a ranking of the different samples as obtained for different interaction parameter sets in order to optimise the parameter values.

6.3.3 Simulation details

In order to evaluate the ensemble averages and to generate distinct configurations for the SP-ID method, we perform constant pressure and constant temperature (NPT) Monte Carlo

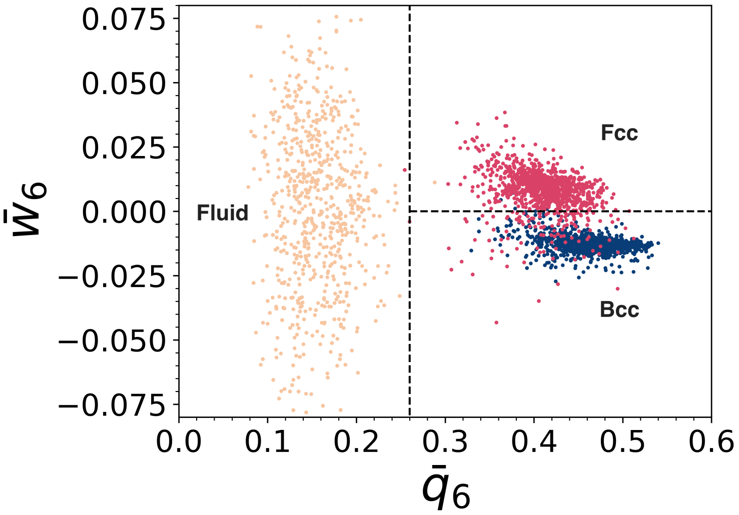


Figure 6.2: Scatter plot of the averaged bond order parameters $\bar{q}_6$ *versus* $\bar{w}_6$ for the fluid, bcc, and fcc phase of a system of Yukawa particles with contact value $\beta\epsilon = 8$ at the high-density triple point. Each point corresponds to a single particle. In total 2000 points were chosen randomly from each structure.

(MC) simulations on systems consisting of $N = 250$ hard-core repulsive Yukawa particles. We initialise the simulations by placing the particles randomly in a cubic simulation box. We equilibrate the system up to 10^5 MC cycles. One Monte Carlo cycle corresponds to N particle moves and one volume move. The particle and volume moves are adjusted in such a way that 45% of the particle moves and 20% of the volume moves are accepted. We save 10^3 uncorrelated samples and evaluate the ensemble averages in Eq. 6.4. These uncorrelated samples are used to calculate the new parameters. Once the parameters have been changed, we repeat the whole procedure starting the simulation from the last configuration generated with the previous parameters. We repeat this process until the target structure is reached.

In the CMA-ES algorithm, we evaluate the ensemble averages by performing simulations at different parameter sets at each generation. At every next generation, we take the last configuration of the fittest sample as the starting point for all the new samples.

6.4 Results

6.4.1 Tuning parameters with the SP-ID method

We start from the case in which we tune only one inter-particle interaction parameter, the inverse Debye screening length $1/\kappa\sigma$ of the interaction potential (Eq. 6.1) to target the bcc structure. At this stage, reduced pressure $\beta P\sigma^3$ and the contact value $\beta\epsilon$ are kept constant at $\beta P\sigma^3 = 33$ and $\beta\epsilon = 8$. The initial value of $1/\kappa\sigma$ is 0.4, to make sure the system starts from a fluid configuration. Given the form discussed in Sec. 6.3, the quality function gives higher weights to those configurations whose $\bar{q}_6$ and $\bar{w}_6$ values are closer to the target values. The probability of finding the system in some configuration x, in the NPT ensemble is given by, $\rho(x|\lambda_i) \propto \exp(-\beta H(x))$. When only one parameter is tuned ($1D$ case), the equation of motion (Eq. 6.4) becomes,

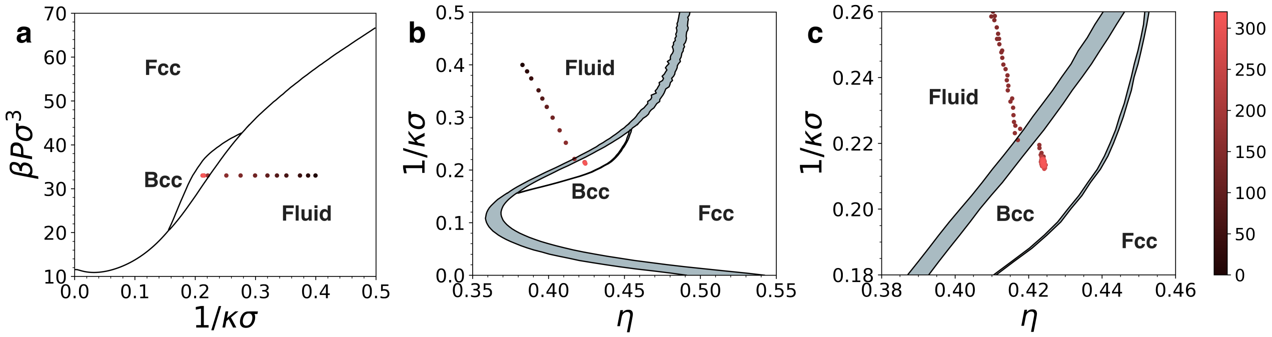


Figure 6.3: Evolution of the parameters in a) the $(1/\kappa\sigma, \beta P\sigma^3)$ plane and b) the $(1/\kappa\sigma, \eta)$ plane when the system is initialised in the fluid phase at reduced pressure $\beta P\sigma^3 = 33$, inverse Debye screening length $1/\kappa\sigma = 0.4$ and contact value $\beta\epsilon = 8$. $\beta P\sigma^3$ and $\beta\epsilon$ are kept fixed, and only $1/\kappa\sigma$ is tuned. In c) we show an enlarged view of the data near the fluid-bcc coexistence region. In a) and b), for visual clarity, we draw only 1 every 20 points, while in c) all points are displayed.

$$\frac{d}{dt}(\kappa\sigma) = -Cov[\frac{\partial(\beta H)}{\partial(\kappa\sigma)}, \frac{\partial(\beta H)}{\partial(\kappa\sigma)}]^{-1}Cov[\frac{\partial(\beta H)}{\partial(\kappa\sigma)}, f], \tag{6.8}$$

where $H = U + PV$ is the Hamiltonian of the system.

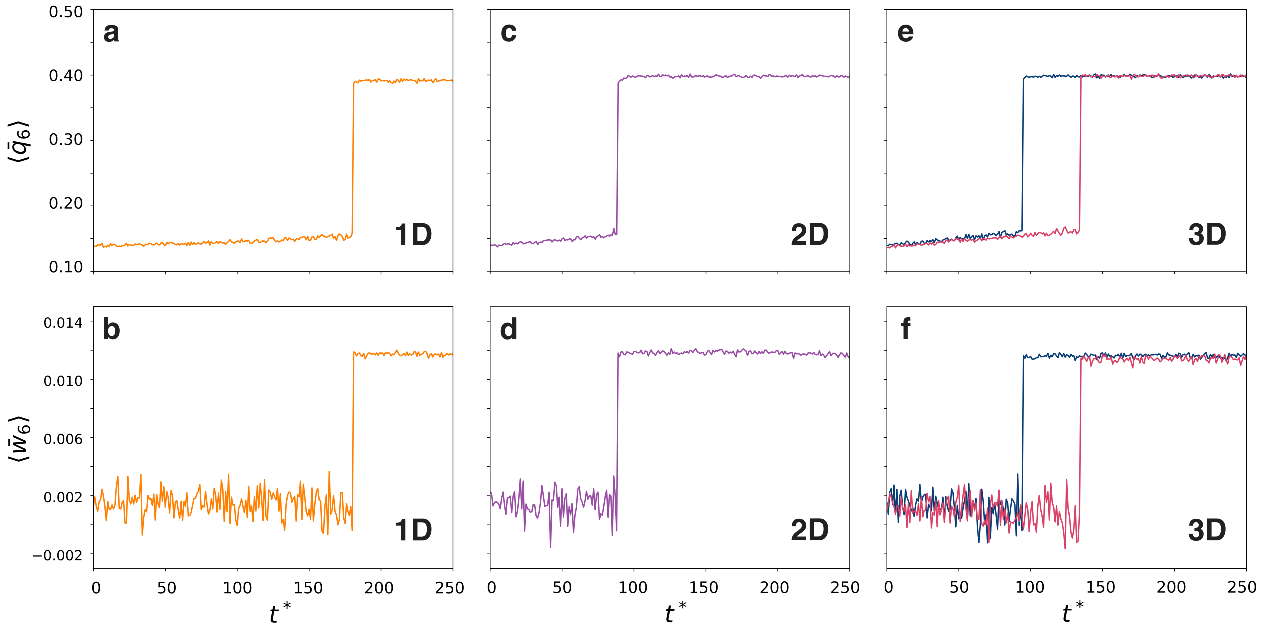


Figure 6.4: Evolution of $\langle \bar{q}_6 \rangle$ and $\langle \bar{w}_6 \rangle$ as a function of simulation time during the optimisation using the SP-ID algorithm for the $1D$ case in a) and b), the $2D$ case in c) and d), and for the $3D$ case in e) and f), respectively. The system is initialised in the fluid phase at reduced pressure $\beta P\sigma^3 = 33$, inverse screening length $1/\kappa\sigma = 0.4$ and contact value $\beta\epsilon = 8$. For the $3D$ case, blue coloured symbols represent the $\langle \bar{q}_6 \rangle$ and $\langle \bar{w}_6 \rangle$ values when the system is initialised at (i) $1/\kappa\sigma = 0.4$, $\beta P\sigma^3 = 33$, $\beta\epsilon = 8$, while red coloured symbols represent bond order parameter values for the case (ii) $1/\kappa\sigma = 0.4$, $\beta P\sigma^3 = 25$, $\beta\epsilon = 6$. The simulation time indicates the number of simulations performed at distinct sets of interaction parameters.

By solving Eq. 6.8, a new value of $1/\kappa\sigma$ is obtained and the algorithm keeps on optimising this interaction parameter until the goal is reached, *i.e.*, $\bar{q}_6 = \bar{q}_6^{target}$ and $\bar{w}_6 = \bar{w}_6^{target}$. The path of the parameters is shown in Fig. 6.3a and Fig. 6.3b in the $(1/\kappa\sigma, \beta P\sigma^3)$ and $(1/\kappa\sigma, \eta)$ planes. In both, the optimiser correctly tunes $1/\kappa\sigma$ to reach the bcc structure which can also be verified by examining the evolution of the average $\bar{q}_6$ and $\bar{w}_6$ values as the simulation proceeds. In Fig. 6.4a and Fig. 6.4b, we plot the average $\bar{q}_6$ and $\bar{w}_6$ values as a function of the simulation time. At the very beginning of the simulation, $\bar{q}_6$ and $\bar{w}_6$ values show that the system is in the fluid phase and as the algorithm optimises the interactions, there is a sharp transition in both of these values which exactly happens at the fluid-bcc phase boundary. Once the system reaches the bcc phase, it remains in the bcc phase. We also find that as the system reaches the phase boundaries, there is a sudden change in the slope of the parameter's trajectory as shown in Fig. 6.3c. In other words, the optimiser (Eq. 6.4) correctly recognises the phase boundaries present in the phase diagram.

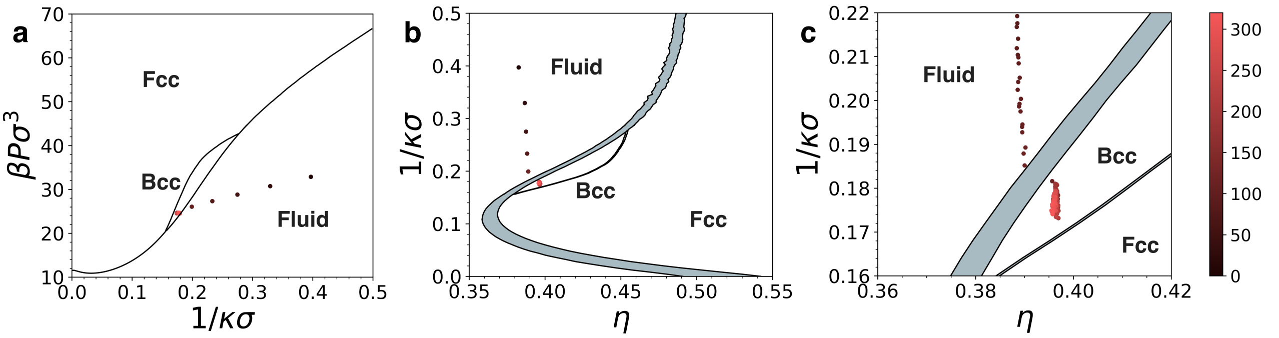


Figure 6.5: Evolution of the parameters in a) the $(1/\kappa\sigma, \beta P\sigma^3)$ plane and b) the $(1/\kappa\sigma, \eta)$ plane when the system is initialised in the fluid phase at reduced pressure $\beta P\sigma^3 = 33$, inverse Debye screening length $1/\kappa\sigma = 0.4$ and contact value $\beta\epsilon = 8$. $\beta\epsilon$ is kept fixed, and two parameters $1/\kappa\sigma$ and $\beta P\sigma^3$ are tuned. In c) we show an enlarged view of the data near the fluid-bcc coexistence region. In a) and b), for visual clarity, we draw only 1 every 20 points, while in c) all points are displayed.

We now analyse the case in which we tune two parameters simultaneously ($2D$ case), one inter-particle interaction parameter, the inverse Debye screening length $1/\kappa\sigma$ and one system parameter, reduced pressure $\beta P\sigma^3$ while $\beta\epsilon = 8$ is kept constant. Note that we use σ and k_BT as our unit of length and energy, respectively, which we keep fixed and hence $\beta\epsilon$ and $\beta P\sigma^3$ can be varied independently from each other. Here, we initialise the system again in the fluid phase at $1/\kappa\sigma = 0.4$ and $\beta P\sigma^3 = 33$. The equations of motion (Eq. 6.4) for the two parameters become

$$\frac{d}{dt}\begin{bmatrix} \beta P\sigma^3 \\ \kappa\sigma \end{bmatrix} = -\begin{bmatrix} Cov[\frac{\partial(\beta H)}{\partial(\beta P\sigma^3)}, \frac{\partial(\beta H)}{\partial(\beta P\sigma^3)}] & Cov[\frac{\partial(\beta H)}{\partial(\beta P\sigma^3)}, \frac{\partial(\beta H)}{\partial(\kappa\sigma)}] \\ Cov[\frac{\partial(\beta H)}{\partial(\beta P\sigma^3)}, \frac{\partial(\beta H)}{\partial(\kappa\sigma)}] & Cov[\frac{\partial(\beta H)}{\partial(\kappa\sigma)}, \frac{\partial(\beta H)}{\partial(\kappa\sigma)}] \end{bmatrix}^{-1} \begin{bmatrix} Cov[\frac{\partial(\beta H)}{\partial(\beta P\sigma^3)}, f] \\ Cov[\frac{\partial(\beta H)}{\partial(\kappa\sigma)}, f] \end{bmatrix}. \quad (6.9)$$

In Fig. 6.5a and Fig. 6.5b, we show the path of the parameters in the $(1/\kappa\sigma, \beta P\sigma^3)$ and $(1/\kappa\sigma, \eta)$ planes. As the simulation time proceeds, SP-ID successfully optimises both the interaction parameters in such a way that the final structure formed is the bcc crystal. The form of the quality function is the same as we have used for the one parameter case. Variations of $\bar{q}_6$ and $\bar{w}_6$ values also verify the formation of the bcc structure from the fluid phase (Fig. 6.4c

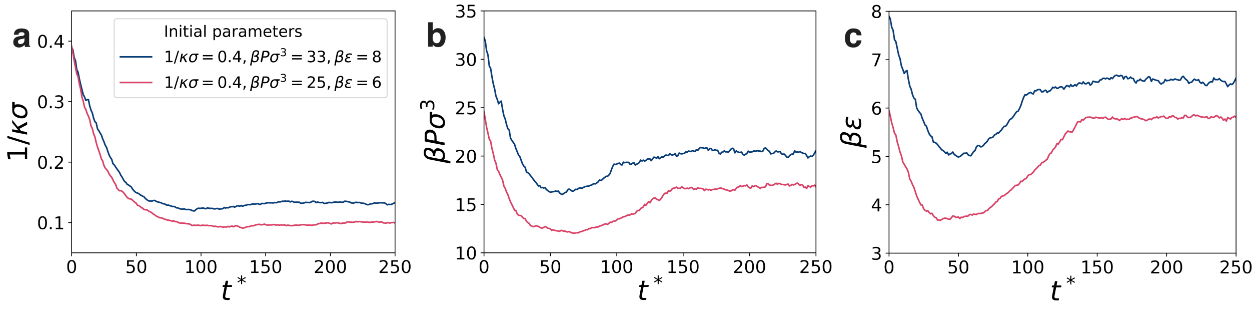


Figure 6.6: Evolution of a) $1/\kappa\sigma$, b) $\beta P\sigma^3$, and c) $\beta\epsilon$ as a function of simulation time when the system is initialised in the fluid phase at (i) $\beta P\sigma^3 = 33$, $1/\kappa\sigma = 0.4$ $\beta\epsilon = 8$ (blu line), and (ii) $\beta P\sigma^3 = 25$, $1/\kappa\sigma = 0.4$ $\beta\epsilon = 6$ (red line). Three parameters are tuned to target the bcc structure, namely $\beta P\sigma^3$, $1/\kappa\sigma$ and $\beta\epsilon$.

and Fig. 6.4d). We find from Fig. 6.5c that the optimiser recognises the phase boundaries very well as also found for the one parameter case.

Finally, we investigate the case in which we tune three parameters simultaneously ($3D$ case), two inter-particle interaction parameters, the inverse Debye screening length $1/\kappa\sigma$ and the contact value $\beta\epsilon$, and one system parameter, reduced pressure $\beta P\sigma^3$. We perform two independent simulations by initialising the system at two different state points in the fluid phase at (i) $1/\kappa\sigma = 0.4$, $\beta P\sigma^3 = 33$ and $\beta\epsilon = 8$, (ii) $1/\kappa\sigma = 0.4$, $\beta P\sigma^3 = 25$ and $\beta\epsilon = 6$ and optimise all three parameters, $\kappa\sigma$, $\beta P\sigma^3$ and $\beta\epsilon$ for the bcc phase. The equations of motion (Eq. 6.4) when three parameters are tuned become

$$\frac{d}{dt}\begin{bmatrix} \beta P\sigma^3 \\ \kappa\sigma \\ \beta\epsilon \end{bmatrix} = -\begin{bmatrix} Cov[\frac{\partial(\beta H)}{\partial(\beta P\sigma^3)}, \frac{\partial(\beta H)}{\partial(\beta P\sigma^3)}] & Cov[\frac{\partial(\beta H)}{\partial(\beta P\sigma^3)}, \frac{\partial(\beta H)}{\partial(\kappa\sigma)}] & Cov[\frac{\partial(\beta H)}{\partial(\beta P\sigma^3)}, \frac{\partial(\beta H)}{\partial(\beta\epsilon)}] \\ Cov[\frac{\partial(\beta H)}{\partial(\kappa\sigma)}, \frac{\partial(\beta H)}{\partial(\beta P\sigma^3)}] & Cov[\frac{\partial(\beta H)}{\partial(\kappa\sigma)}, \frac{\partial(\beta H)}{\partial(\kappa\sigma)}] & Cov[\frac{\partial(\beta H)}{\partial(\kappa\sigma)}, \frac{\partial(\beta H)}{\partial(\beta\epsilon)}] \\ Cov[\frac{\partial(\beta H)}{\partial(\beta\epsilon)}, \frac{\partial(\beta H)}{\partial(\beta P\sigma^3)}] & Cov[\frac{\partial(\beta H)}{\partial(\beta\epsilon)}, \frac{\partial(\beta H)}{\partial(\kappa\sigma)}] & Cov[\frac{\partial(\beta H)}{\partial(\beta\epsilon)}, \frac{\partial(\beta H)}{\partial(\beta\epsilon)}] \end{bmatrix}^{-1} \times$$
$$\times \begin{bmatrix} Cov[\frac{\partial(\beta H)}{\partial(\beta P\sigma^3)}, f] \\ Cov[\frac{\partial(\beta H)}{\partial(\kappa\sigma)}, f] \\ Cov[\frac{\partial(\beta H)}{\partial(\beta\epsilon)}, f] \end{bmatrix}. \tag{6.10}$$

In Fig. 6.6, we plot the path of the tuned parameter trajectories as a function of simulation time when the system is initialised in the fluid phase and the desired goal is to reach the targeted bcc structure. Initially, all three parameter values decrease while the system is in the fluid phase and once it crosses the phase boundary between the fluid and bcc phase, they start to saturate. As the simulation time proceeds, the optimiser successfully optimises the parameters in such a way that the final structure becomes the bcc phase. Here we also use the same form of the quality function as we have used earlier for the one and two parameter cases.

To confirm that the final structure is a bcc phase, we also plot the evolution of $\langle\bar{q}_6\rangle$ and $\langle\bar{w}_6\rangle$ as a function of the simulation time in Fig. 6.4e and Fig. 6.4f, which indeed confirms that the simulation reaches the optimal bond order parameter values for the bcc phase.

6.4.2 Tuning parameters with the CMA-ES method

We now employ the CMA-ES algorithm to analyse all the cases already studied using the SP-ID method, *i.e.* the tuning of one, two and three parameters, highlighting the differences between the two inverse design optimisers. We use the parameters corresponding to the initial state point employed in the SP-ID algorithm as the initial mean vectors of the multivariate Gaussian distribution in the CMA-ES algorithm. In addition, we start the CMA-ES algorithm with a diagonal covariance matrix with a standard deviation of 10% around its mean value for each parameter. Finally, the initial values of each component of the vectors $\vec{q}$ and $\vec{p}$ are set to zero.

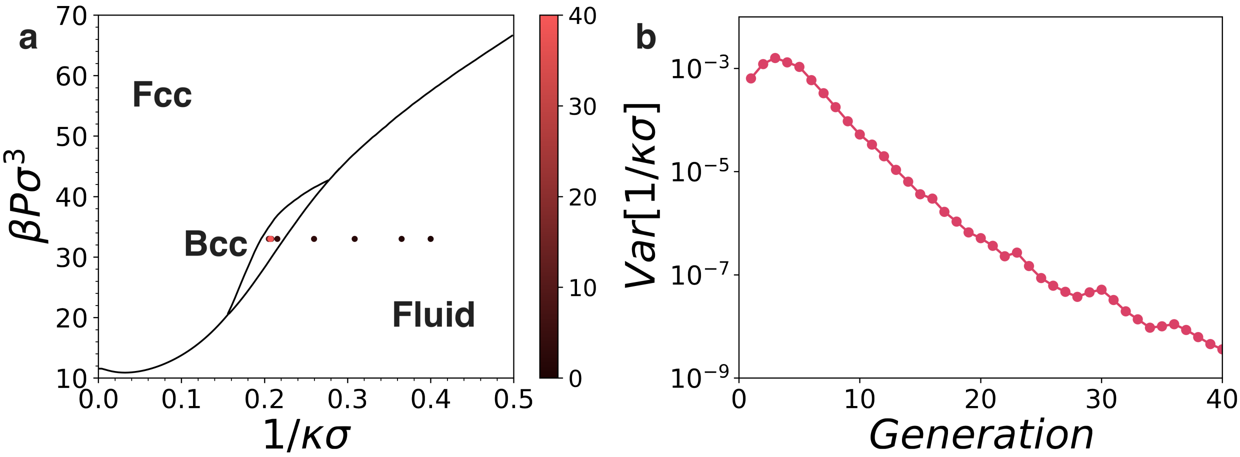


Figure 6.7: a) Evolution of the mean value of the Gaussian distribution for $1/\kappa\sigma$ in the $(1/\kappa\sigma, \beta P\sigma^3)$ plane when the system is initialised in the fluid phase at reduced pressure $\beta P\sigma^3 = 33$, inverse Debye screening length $1/\kappa\sigma = 0.4$ and contact value $\beta\epsilon = 8$. The parameters $\beta\epsilon$ and $\beta P\sigma^3$ are kept fixed, and $1/\kappa\sigma$ is tuned using the CMA-ES method. b) The variance of the Gaussian distribution for $1/\kappa\sigma$ at each generation.

In Fig. 6.7 we show the results for the one parameter case. The points displayed in Fig 6.7a represent the mean value of the Gaussian distribution in each generation. At the beginning, the algorithm tries to decrease the fitness function value by decreasing the inverse Debye screening length $1/\kappa\sigma$. Since all the steps point in the same direction, the variance of the Gaussian distribution increases (Fig 6.7b) and the mean value is found inside the bcc region for the first time at the 5^{th} generation. After this, the algorithm learns to perform much smaller steps in order to explore the vicinity of the phase diagram and, at the same time, the covariance starts to shrink. From this generation onwards the updates are in random directions inside the bcc region, leading to a further exponential decrease of the covariance of the Gaussian distribution. At the 12^{th} generation, all the 20 simulations have a $k\sigma$ value for which the structure corresponds to the stable bcc phase.

The cases in which we tune two or three parameters do not present significantly different behaviour of the CMA-ES algorithm with respect to the one parameter case. In the two parameter case, $\vec{\mu}$ enters the bcc region for the first time at the 6^{th} generation, while at the 20^{th} generation all the samples have entered the bcc region. The results for the $2D$ case are shown in Fig 6.8. We note that the probability of performing big jumps in the parameters space is lower when more of these parameters are varied at the same time, since they all contribute to the increase or decrease of the quality function, and the updates of the mean values of

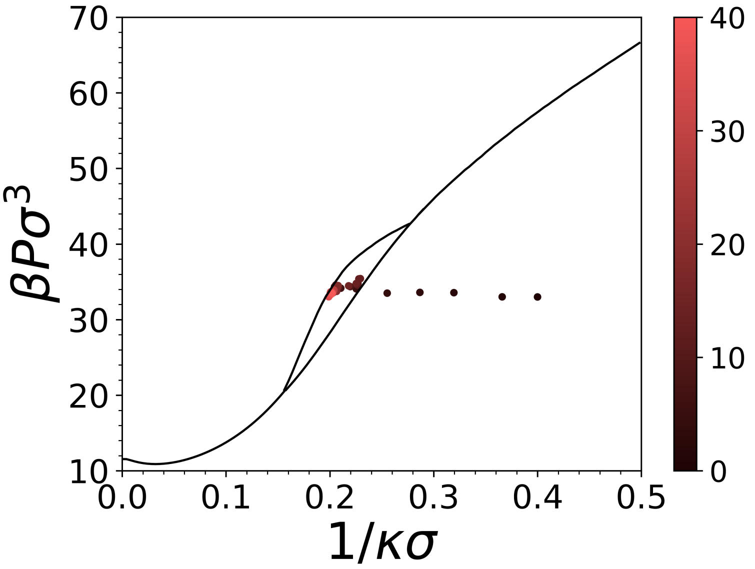


Figure 6.8: Evolution of the the mean value of the multivariate Gaussian distribution for $1/\kappa\sigma$ and $\beta P\sigma^3$ in the $(1/\kappa\sigma, \beta P\sigma^3)$ plane when the system is initialised in the fluid phase at reduced pressure $\beta P\sigma^3 = 33$, inverse Debye screening length $1/\kappa\sigma = 0.4$ and contact value $\beta\epsilon = 8$. The parameter $\beta\epsilon$ is kept fixed, and $1/\kappa\sigma$ and $\beta P\sigma^3$ are tuned using the CMA-ES method.

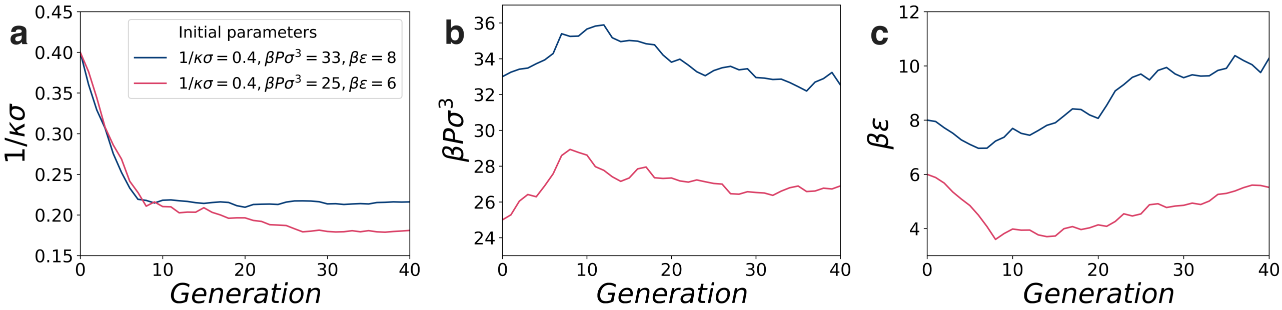


Figure 6.9: Evolution of the the mean value of the multivariate Gaussian distribution for inverse Debye screening length $1/\kappa\sigma$, reduced pressure $\beta P\sigma^3$ and contact value $\beta\epsilon$ when the system is initialised in the fluid phase at (i) $\beta P\sigma^3 = 33$, $1/\kappa\sigma = 0.4$ $\beta\epsilon = 8$ (blue line), and (ii) $\beta P\sigma^3 = 25$, $1/\kappa\sigma = 0.4$ $\beta\epsilon = 6$ (red line). Three parameters are tuned to target the bcc structure, namely $\beta P\sigma^3$, $1/\kappa\sigma$ and $\beta\epsilon$ using the CMA-ES method.

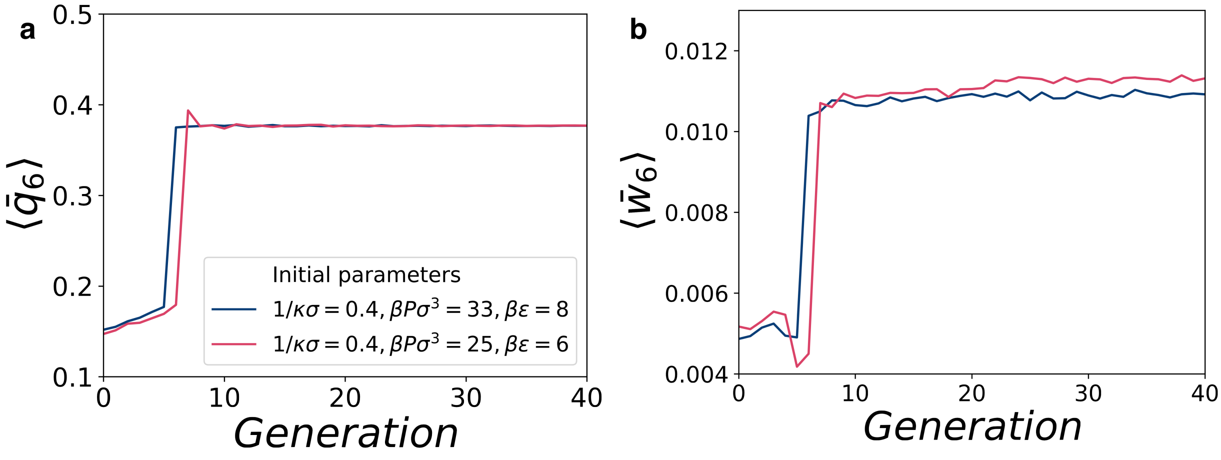


Figure 6.10: Evolution of a) $\langle \bar{q}_6 \rangle$ and b) $\langle \bar{w}_6 \rangle$ of the sample with the highest value of the fitness at each generation during the $3D$ optimisation with the CMA-ES method for both the investigated cases. We observe a jump of $\bar{q}_6$ and $\bar{w}_6$ when, for the first time, at least one of the generated samples is in the bcc region, even if the mean value of the multivariate Gaussian distribution at the same generation does not lie in the same region.

the multivariate Gaussian distribution may therefore not always be in the same direction. This prevents the covariance matrix from growing too fast. Finally, the results for both the investigated $3D$ cases are shown in Fig 6.9. To confirm that the final structure is a bcc phase, we plot the evolution of $\langle \bar{q}_6 \rangle$ and $\langle \bar{w}_6 \rangle$ as a function of the simulation time in Fig. 6.10.

These analyses provide benchmarks for how fast the CMS-ES algorithm converges in finding the target structure, as compared to the SP-ID algorithm. Since each generation of the CMA-ES requires n times (20 in this work) the computational effort of SP-ID, a comparison of run lengths reveals that SP-ID performs better than CMA-ES. Although their efficiencies are still comparable, the difference becomes more marked when the dimensionality of the problem increases. Moreover, in SP-ID, one has information on when the phase boundary is crossed. In CMA-ES the size of the steps from one generation to the other varies, depending on the current form of the covariance matrix. This makes the algorithm explore the landscape in an optimal way, sacrificing information about the phase boundaries. Tuning the free parameters of CMA-ES (c_i) one can force the algorithm to perform small steps and gain the same information on the phase boundaries, but this would result in a huge rise of the computational effort required, which would make CMA-ES an algorithm of poor efficiency and use. Finally, it should be noted that in CMA-ES all the simulations belonging to the same generation can be run simultaneously, which can be a great advantage, depending on the computational resources of the user.

6.5 Conclusions

We studied the inverse problem of tuning interaction parameters between charged colloids interacting *via* a hard-core repulsive Yukawa potential, so that they self-assemble into a targeted crystal structure. We targeted the bcc structure which occupies a narrow region in the phase diagram of the above system and is therefore challenging to find. We showed a comparison between two different optimisation algorithms in order to achieve our goal: Statistical Physics-inspired Inverse Design (SP-ID) and Covariance Matrix Adaption - Evolutionary Strategy (CMA-ES). The first makes use of the statistical fluctuations in the bond order parameters to iteratively change the interaction parameters of the system. In addition to effectively tuning the interaction parameters for obtaining the target structure, the SP-ID method correctly identifies the fluid-solid phase boundaries present in the phase diagram. The CMA-ES algorithm generates samples from a multivariate Gaussian distribution at each generation and evaluates the fitness of these samples in order to evolve the interaction parameters of the distribution. The number of generations needed to reach the goal is on average lower in the case of the CMA-ES, and the steps in parameter space are usually larger. This advantage is offset by the need to simulate multiple samples, and we find that the computational effort required in the two methods is comparable. On the other hand, because of the larger step sizes of the parameters, probing phase equilibrium with CMA-ES can be less straightforward than with SP-ID. Most importantly, we showed that both of these inverse design methods lead to the targeted bcc structure by tuning the interactions between the particles. Thus, our results demonstrate both methods to be effective search algorithms that may be employed in other design tasks. Although the quality function used here is strictly structural, one may in principle also include quantifiers of dynamics, which may be useful in optimising the kinetic self-assembly pathways, for example, to account for and exclude glassy dynamics, as well as other kinetic factors of self assembly. Finally, we stress how exploiting thermal fluctuations of a statistical system results in a more effective search in the fitness function space. In the future, it would be useful and fascinating to compare SP-ID to other related inverse design methods [231–233].

Acknowledgements

I would like to thank Dr. Rajneesh Kumar, who implemented the SP-ID method, and Prof. Srikanth Sastry for the uncountable inputs and ideas provided during the project.

7

Inverse design of soft materials via a deep-learning-based evolutionary strategy

In this Chapter, we introduce a generic inverse design method to efficiently reverse-engineer crystals, quasicrystals, and liquid crystals by targeting their diffraction patterns. Our algorithm relies on the synergetic use of an evolutionary strategy for parameter optimisation, and a convolutional neural network as an order parameter, and provides a new way forward for the inverse design of experimentally feasible colloidal interactions, specifically optimised to stabilise the desired structure.

Based on: G. M. Coli / E. Boattini, L. Filion, and M. Dijkstra, *Inverse design of soft materials via a deep-learning-based evolutionary strategy*, Manuscript submitted (2021)

7.1 Introduction

Self-assembly of colloidal particles is ubiquitous in nature and is considered to be of paramount importance for the design of novel functional materials. For example, viruses, lipid bilayers, tissues, atomic and molecular crystals, liquid crystals, and nanoparticle superlattices are all self-assembled from smaller components in a highly intricate way. The structure of such an assembly is determined by the interactions of the building blocks and by the thermodynamic conditions, e.g. pressure, temperature, or composition. Understanding the relation between building blocks and self-assembled arrangements is essential for materials design as the physical properties of materials are intimately related to the structure.

On the other hand, huge progress has been made over the past decades in the synthesis and fabrication of colloidal particles, resulting in a spectacular variety of novel colloidal building blocks to the point where particles with any shape and interaction potential can be made on demand [34–38]. Traditionally, tremendous efforts have been devoted to the "forward design" problem: Which structures with what properties are formed for a given colloidal building block under what circumstances? A major drawback of this approach is that the number of possible building blocks and thermodynamic conditions is limitless, making a systematic exploration of these design spaces intractable.

The true challenge in materials science is to develop a robust, versatile algorithm for solving the "inverse design" problem and to design building blocks that self-assemble into a target structure. The lack of such an inverse design method (IDM) forms a significant obstacle for the full exploitation of colloidal self-assembly in the development of tomorrow's materials [39–42].

In this work, we present a general inverse design method based on deep-learning techniques to reverse engineer a multitude of thermodynamic phases, ranging from crystals, to liquid crystals and even quasicrystals. A novel machine-learning-based order parameter is combined with an evolutionary strategy which searches the multi-dimensional parameter space to optimise the colloidal interactions and thermodynamic conditions (density, temperature, etc.) for the self-assembly of a target phase.

Designing an IDM to reverse engineer phases, from crystals, to liquid crystals, and quasicrystals, generally requires two ingredients. First, one should define an order parameter that is sensitive to the global structure of a multitude of phases and can be exploited as a fitness function indicating how "close" one is to the desired outcome. Secondly, one has to devise a mathematical scheme to update the design parameters based on the chosen fitness function.

The latter requirement can be easily satisfied by choosing among several techniques, either borrowed from classical optimisation algorithms [48, 49, 249] or inspired by statistical physics [50, 51, 249]. Our inverse design method (IDM) uses the Covariance Matrix Adaptation (CMA) evolutionary strategy for parameter optimisation [249, 250]. Conversely, the choice of an effective fitness function represents the real bottleneck for any IDM to succeed. In the last decade, a plethora of order parameters has been used to define fitness functions for all kinds of phases. For instance, free-energy or chemical-potential differences with respect to the competing structures have been employed to reverse engineer 3D crystal lattices starting from (non)spherical colloids [52, 53]. Often, full knowledge of the target crystal has been translated into a fitness function by computing the mean square displacements of the particles with respect to their target lattice points [39], or through the radial distribution function [54–56]. The sometimes unrealistic resulting potentials have been explicitly filtered by Adorf *et al.* in order to obtain smooth and short-range interactions [57].

Although all these fitness function definitions brilliantly achieve their goals, they often lack generality, and most importantly, they are not able to simultaneously and equally penalise competing phases. In other words, they do not have the ability to create an approximately flat fitness landscape, where the design engine can move smoothly, with only one preferred region corresponding to the target phase. Moreover, in the case of quasicrystals, in spite of the certified need of two inherent length scales in the system [251, 252], the actual positions of the constituent particles remain unknown, therefore representing a significant challenge to the above strategies.

Inspired by the highly successful history of identifying phases by their scattering patterns in combination with advances in machine learning (ML), we attack the problem from a new avenue and directly use an encoding of the structure factor as the order parameter. To this end, we train a convolutional neural network (CNN) to classify different phases from their diffraction pattern, and use the result to construct a fitness function, such that configurations with a higher likelihood of being classified as the target phase, will be scored with a higher fitness. A sketch of the final algorithm is shown in Fig. 7.1.

This algorithm turns out to be extremely robust and versatile, facilitating the inverse design of, not only crystal and liquid crystalline phases, but also quasicrystals – which due to their non-periodicity are notoriously difficult to inverse design.

7.2 Inverse Design Method

7.2.1 General framework

Our IDM combines the CMA evolutionary strategy for parameters optimisation, and a CNN for the fitness evaluation, which are both described in detail in the following subsections. The goal is to optimise the free parameters of a given model in order to favour the formation of a target phase.

The method proceeds in generations, or iterations, consisting of essentially three steps: a) sampling, b) fitness evaluation, and c) update. In the following, we give a general overview of these three steps, which are sketched in Fig. 7.1.

In the first step (Fig. 7.1a), we draw a fixed number of candidate sets of parameters from a multivariate Gaussian distribution. The dimension of this multivariate Gaussian distribution is determined by the number of design parameters that we wish to tune. For each candidate set of parameters, we then perform a simulation of the system and save a number of representative configurations. In the second step (Fig. 7.1b), we score and rank the samples based on their fitness f. In general, the fitness is a measure of similarity between a sample and a specific target, and it is maximised when the target is reached. Here, we introduce a new fitness function based on CNNs that are trained to classify different phases based on their diffraction patterns. We use this CNN to process the configurations saved during each simulation, and assign a larger fitness to samples with a higher probability of being classified as the target phase. Finally, based on this score, the mean and the covariance matrix of the multivariate Gaussian distribution are updated using the CMA equations, which are designed to facilitate an efficient exploration of parameter space. As sketched in Fig. 7.1c, the update not only allows the mean of the distribution to move towards regions with a higher fitness, but it also speeds up sampling by stretching the distribution when several updates are in the same direction, and then shrinking it once the fitness is maximised. This whole procedure is repeated multiple times until the fitness is maximised and/or a predetermined convergence criterion is met.

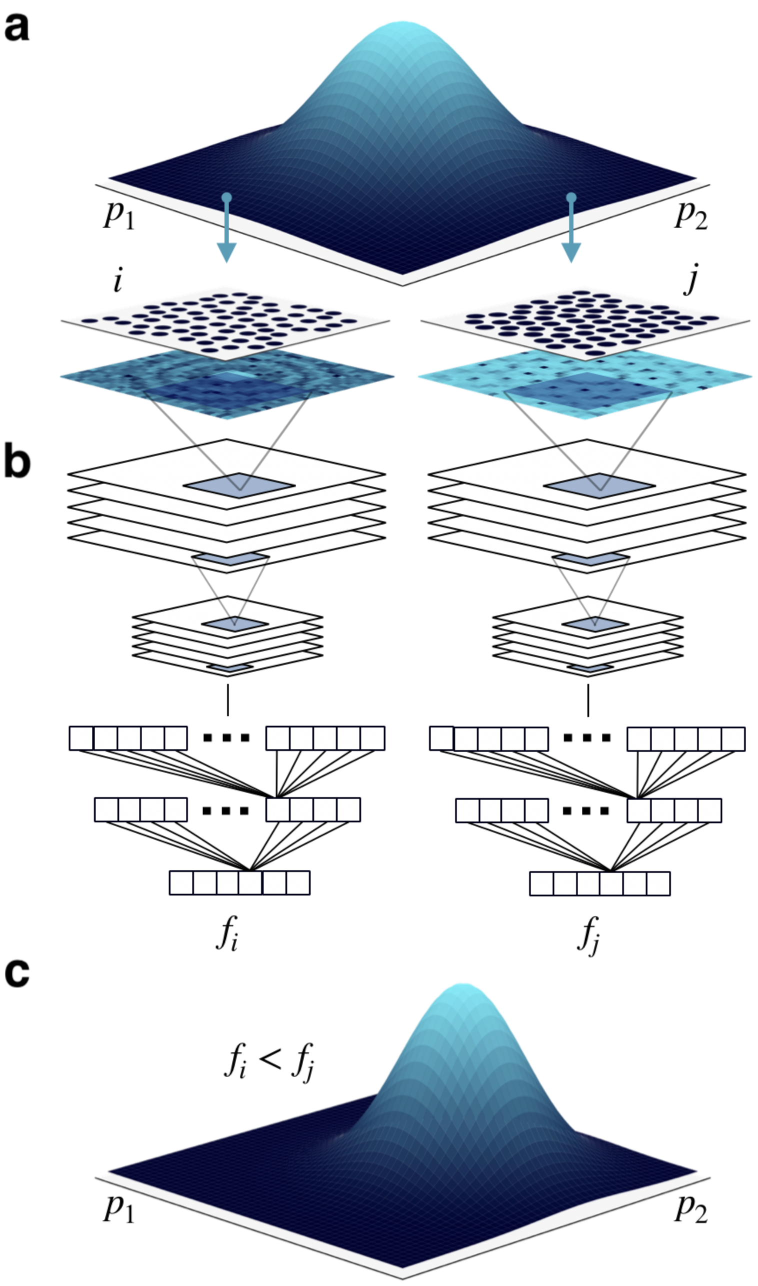


Figure 7.1: Schematic representation of the three steps performed at each generation. (a) In the first step, we draw candidate sets of parameters (p_1 and p_2 in the figure) from a multivariate Gaussian distribution. For each set, or sample, we then perform a simulation. (b) In the second step, samples are ranked and scored based on their fitness f, which is evaluated using a convolutional neural network trained to classify phases according to their diffraction patterns. Samples with a higher likelihood of being classified as the target phase will be scored with a higher fitness. (c) In the third and final step, the Gaussian distribution is updated in order to move towards regions of the parameter space where the fittest samples have been encountered.

7.2.2 Convolutional neural networks as a fitness function

CNNs are a particular type of deep neural networks specifically designed to handle tensorial inputs, such as images. For a detailed description of CNNs see e.g. Ref. ?. In this work, we train a CNN to classify different phases from their diffraction patterns, which are either 2D or 3D images. The output of the CNN is then used to define a fitness function f for the evolutionary strategy.

More specifically, the CNN takes as input the diffraction pattern of a given configuration, and outputs a vector of real numbers with as many components as the number of phases to distinguish. Each number in the output is indicative of the probability that the given input corresponds to one of the phases. We use this CNN to process the configurations saved during each simulation, and define the fitness of a given sample as

$$f = \bar{P}_{\text{target}} \tag{7.1}$$

where P_{target} is the probability that the diffraction pattern of a given configuration is classified as the target phase by the CNN, and the bar indicates an average taken over representative configurations saved during the simulation of that sample.

7.2.3 Training the convolutional neural networks

To train the convolutional neural networks to recognise different phases, we need to perform a number of different steps. Specifically, we first generate a number of real space equilibrium configurations for each phase, and then generate the associated diffraction patterns. In order to reduce computational time and memory usage, these diffraction patterns are preprocessed before being used to train the convolutional neural networks. Each of these steps is described in detail in the remainder of this section.

Generating the training configurations

The configurations for training the CNNs are generated by performing MC simulations of the 2D HCSS model [252–255] and the softened-core-shoulder model of spherocylinders in three dimensions (3D) [256].

In 2D, simulations are performed in the isobaric-isothermal ensemble (NPT) of a system of $N = 256$ particles in a square box of side-length L with periodic boundary conditions. For each of the six phases considered (fluid, hexagonal (HEX) and square (SQ) crystals, dodecagonal (QC12), decagonal (QC10), and octadecagonal (QC18) quasicrystals), we run simulations at different state points and collect 10^4 independent configurations.

In 3D, simulations are performed in the canonical ensemble (NVT) of a system of $N = 432$ particles in a rectangular box elongated in the z direction (*i.e.*, $L_x = L_y = L$ and $L_z > L$), and with periodic boundary conditions. For each of the five phases considered (isotropic (I), smectic (SM), 3D hexagonal (3DHEX) and 3D square (3DSQ) crystals, and 3D twelvefold (3DQC12) quasicrystal), we run simulations at different state points and collect $5 \cdot 10^3$ independent configurations.

Generating the diffraction patterns

Diffraction patterns for each configuration are evaluated using

$$S(\mathbf{k}) = \frac{1}{N}\rho(\mathbf{k})\rho(-\mathbf{k}), \tag{7.2}$$

where $\rho(\mathbf{k}) = \sum_{j=1}^{N} e^{-i\mathbf{k}\cdot\mathbf{r}_j}$ is the Fourier transform of the density, $\mathbf{r}_j$ is the position of particle j, and $\mathbf{k}$ is a wave vector. In 2D, the $\mathbf{k}$ vectors are chosen by $\mathbf{k} = \frac{2\pi}{L}(n_x, n_y)$, where n_x and n_y are two integers in the interval $[-64, 64]$. As a result, the 2D diffraction patterns considered in this work are built on a 129 × 129 grid. In 3D, the $\mathbf{k}$ vectors are chosen by $\mathbf{k} = 2\pi(\frac{n_x}{L}, \frac{n_y}{L}, \frac{n_z}{L_z})$, with $n_x, n_y, n_x \in [-32, 32]$, resulting in a $65 \times 65 \times 65$ grid.

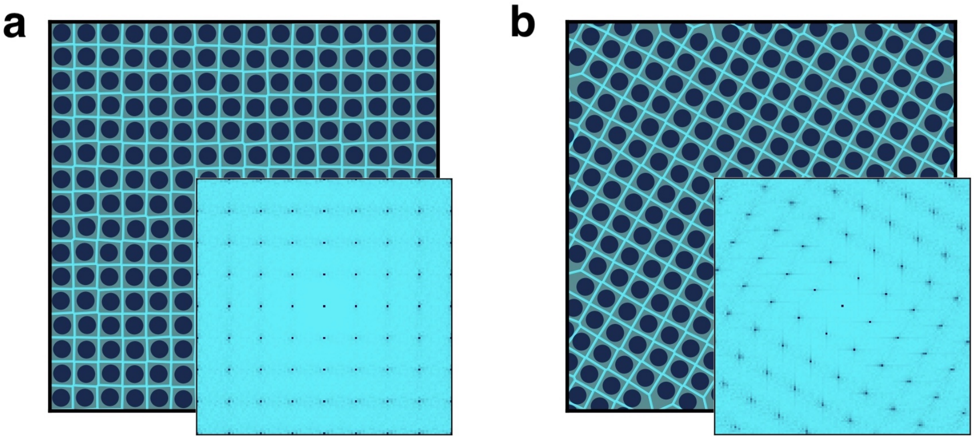


Figure 7.2: Data transformation. (a) Snapshot along with its Voronoi tessellation and diffraction pattern of a square crystal in its original orientation. (b) Same snapshot along with its Voronoi tessellation and diffraction pattern as in (a) after a rotation by a $\pi/6$ angle. Note that the rotation is performed in real space.

While diffraction patterns are by definition translationally invariant, they are not invariant to rotations. However, we must ensure that the CNNs are able to classify the desired phases regardless of their orientation. To this end, each training configuration is rotated by a random angle before evaluating its diffraction pattern. A representation of this transformation in the 2D case is shown in Fig. 7.2. In the 3D case, given the inherent symmetry of the model of spherocylinders considered, we randomly rotate each configuration around the z axis (which always corresponds to the elongated axis of the box). In a more general case, one could perform random rotations around a randomly selected axis. Note that, to rotate a configuration, we first create a larger copy of the system by copying the original simulation box in all directions. We then rotate this larger copy of the system, and finally take a portion of it of the same size as the original simulation box.

The sets of diffraction patterns obtained after having rotated each configuration are finally used to build the data sets for training the CNNs.

Preprocessing

In order to increase the overall efficiency, the diffraction patterns undergo a final preprocessing step before being used as the input of the CNNs. In particular, each diffraction pattern passes through a MaxPooling filter, that effectively reduces the input size by a factor 4 in each dimension. The effect of this transformation is shown in Fig. 7.3 for both the (a) 2D, and (b) 3D cases. Note that this is not a necessary step of the algorithm and its only purpose is to increase the efficiency of the method in terms of computational time and memory usage. With such a

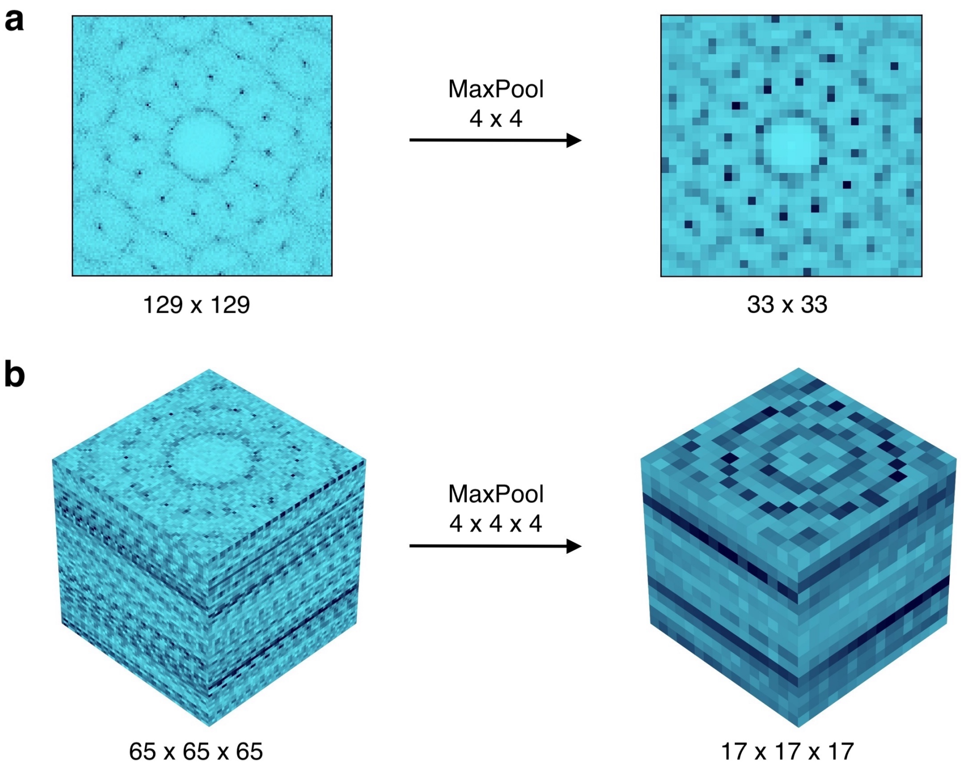


Figure 7.3: Preprocessing. The size of the diffraction pattern of a QC12 in (a) 2D and (b) 3D is reduced through a MaxPooling filter.

preprocessing, the CNNs used here can be trained within one hour on the CPU of a modern laptop.

Neural network architecture

The CNNs used in this work are composed of two convolutional layers for feature extraction, and a fully-connected part with one hidden layer for the final classification. The architecture of the 2D CNN is shown in Fig. 7.4. As shown in the figure, each convolutional layer performs three operations on the input: a convolution, a non-linear transformation through a ReLU activation function, and a downsampling operation through a 2×2 MaxPooling layer. In the following, we give all the details about the network parameters.

The first convolutional layer has one input channel (*i.e.*, the diffraction pattern to process) and nine output channels (*i.e.*, the extracted features). As indicated in Fig. 7.4, the kernels used in this layer have a size $s = 4 \times 4$, padding $p = 1$, and stride $s = 1$. The second convolutional layer has nine input channels and four output channels, and the kernels of this layer have a size $s = 3 \times 3$, padding $p = 1$, and stride $s = 1$. The output of the second convolutional layer is stacked and flattened, in order to be used as the input of the fully-connected part of the network. The latter consists of a hidden layer of dimension 20 with a ReLU activation function, and an output layer with a SoftMax activation function. The size of the output layer is equal to the number of phases we wish to distinguish, which is 6 in the 2D case.

The 3D CNN has almost the same structure as the 2D one, with the only exception being that the convolutional kernels are extended to three dimensions (e.g. a 3×3 kernel in 2D

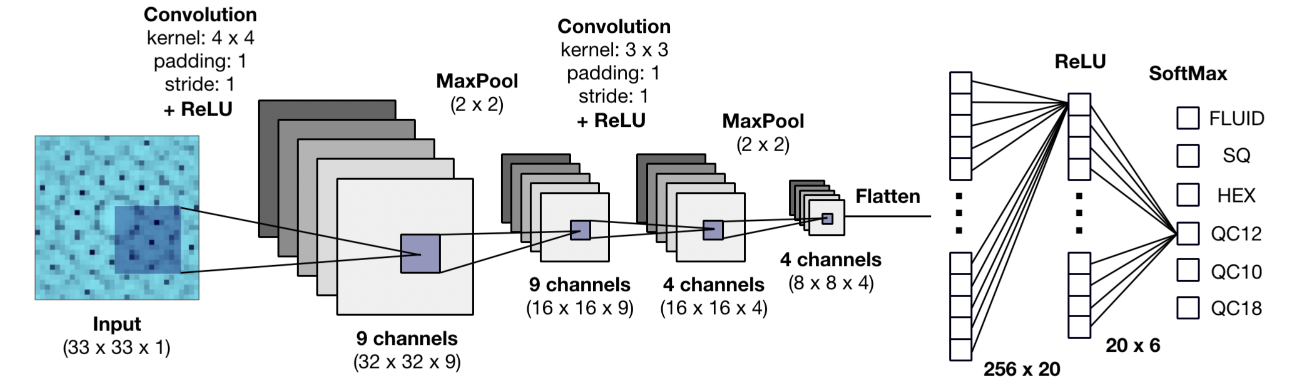


Figure 7.4: Representation of the 2D convolutional neural network. The network is composed of two convolutional layers for feature extraction, and a fully-connected part with one hidden layer for the final classification. All details about kernels, layer size, and activation functions are also shown.

becomes a $3 \times 3 \times 3$ kernel in 3D), and the output layer has a dimension of 5 (we consider 5 phases in the 3D system).

Training

The parameters of the CNNs are optimised by minimising the cross-entropy loss with the addition of a weight decay regularisation term [257, 258]. Specifically, the loss is minimised with the Adam optimiser [259], a learning rate of 10^{-4}, and a PyTorch implementation [260]. Early stopping is also applied in order to prevent overfitting.

7.2.4 Workflow of the CMA evolutionary strategy

The CMA evolutionary strategy optimises iteratively the design parameters across successive generations. At each generation, we draw n samples from a multivariate Gaussian distribution, whose dimension D corresponds to the number of parameters we wish to optimise. Subsequently, we evaluate the fitness function f of the generated samples, we order the samples in ascending order based on their fitness, and we pick the k best samples. Finally, the mean $\vec{\mu}$ (a D-dimensional vector) and the covariance matrix $\mathbf{\Sigma} = \sigma^2\mathbf{C}$ of the Gaussian distribution are updated using the following equations:

$$\begin{aligned}
{\mu_i}' &= \mu_i + \sum_{x\in\mathbf{X}} w(x)(\lambda_i(x) - \mu_i) \\
{q_i}' &= (1-c_1)q_i + c_2\left(\sqrt{\Sigma^{-1}}\right)_{ij}({\mu_j}' - \mu_j) \\
{p_i}' &= (1-c_3)p_i + c_4({\mu_i}' - \mu_i) \\
{C_{ij}}' &= (1-c_5-c_6)C_{ij} + c_6{p_i}'{p_j}' \\
&\quad + c_5\sum_{x\in\mathbf{X}} w(x)\left(\frac{\lambda_i(x)-\mu_i}{\sigma}\frac{\lambda_j(x)-\mu_j}{\sigma} - C_{ij}\right) \\
\sigma' &= \sigma\exp\left[c_7\left(\frac{\parallel \vec{q'} \parallel}{\langle\parallel N(0,I)\parallel\rangle} - 1\right)\right]
\end{aligned} \tag{7.3}$$

where $\mathbf{X}$ denotes the set of the k best samples consisting of multiple configurations obtained for k different parameter sets (denoted by $\lambda_i(x)$), $w(x)$ is the normalised distribution of weights based on the fitness of the samples, and c_i's are free parameters. We choose $w(x) \propto \log(k+1) - \log(m)$, where m is the rank index of sample x ($m = 1$ for the configuration with the largest f value). $\vec{q}$ and $\vec{p}$ are additional D-dimensional vectors that control, respectively, the changes in amplitude and directionality of the covariance matrix. Additionally, $\langle \| N(0, I) \| \rangle$ is the average length of a vector drawn from a multivariate Gaussian distribution centred in the origin and where the covariance matrix is the identity matrix. In the present work, we use $n = 10$ and $k = 5$ for all cases where we optimise two parameters, *i.e.* $D = 2$. When optimising three parameters ($D = 3$), we use instead $n = 20$ and $k = 8$ in order to guarantee a faster exploration of the phase space. For the first generation, we initialise $\vec{q}$ and $\vec{p}$ as null vectors. Moreover, since we do not assume any *a priori* correlation between the different tuning parameters, the initial form of the covariance matrix $\mathbf{\Sigma}$ is diagonal. Finally, all the free parameters c_i of the CMA-ES are set equal to 0.2, as proposed in Ref. 250.

7.3 Results

7.3.1 Setting up the IDM in two dimensions

The first model we consider is a two-dimensional system in which the particles interact with a hard-core square-shoulder (HCSS) potential:

$$\beta u(r) = \begin{cases} \infty, & r < \sigma \\ \beta\epsilon, & \sigma \leq r \leq \delta \\ 0, & r > \delta, \end{cases} \tag{7.4}$$

with r the centre-of-mass distance between two particles, ϵ the interaction strength, σ the core diameter, δ the interaction range, and $\beta = 1/k_B T$ with k_B Boltzmann's constant and T the temperature. This model has been shown to self-assemble into a variety of phases [252–255], including several crystal structures and various quasicrystals (QCs), which makes it an ideal playground for setting up and testing our IDM. The three QCs we consider here, which are the dodecagonal (QC12), the decagonal (QC10), and the octadecagonal (QC18) quasicrystals, are found to be stable for different values of the interaction range δ, and only in a tiny range of densities ρ and temperatures T. In all cases we explore the competing stable phases, which include the fluid, the hexagonal (HEX) crystal, and the square (SQ) crystal phase.

To set up our IDM we trained a CNN to classify the aforementioned phases based on their two-dimensional diffraction patterns, as described above. Specifically, the CNN takes as input the diffraction pattern of a given configuration, and outputs a vector of real numbers with as many components as the number of phases to distinguish. Each number in the output is indicative of the likelihood that the given input corresponds to one of the phases. This output is then used to define the fitness function to target a specific phase.

The data set for training the CNN is built by performing MC simulations of the HCSS model in the NPT ensemble. For each phase, we perform simulations at different state points, and collect a large number of independent configurations. The set of diffraction patterns generated from these configurations constitutes the data set on which the CNN is trained and validated. Overall, we find the CNN to be highly effective and able to classify all phases with 100% accuracy.

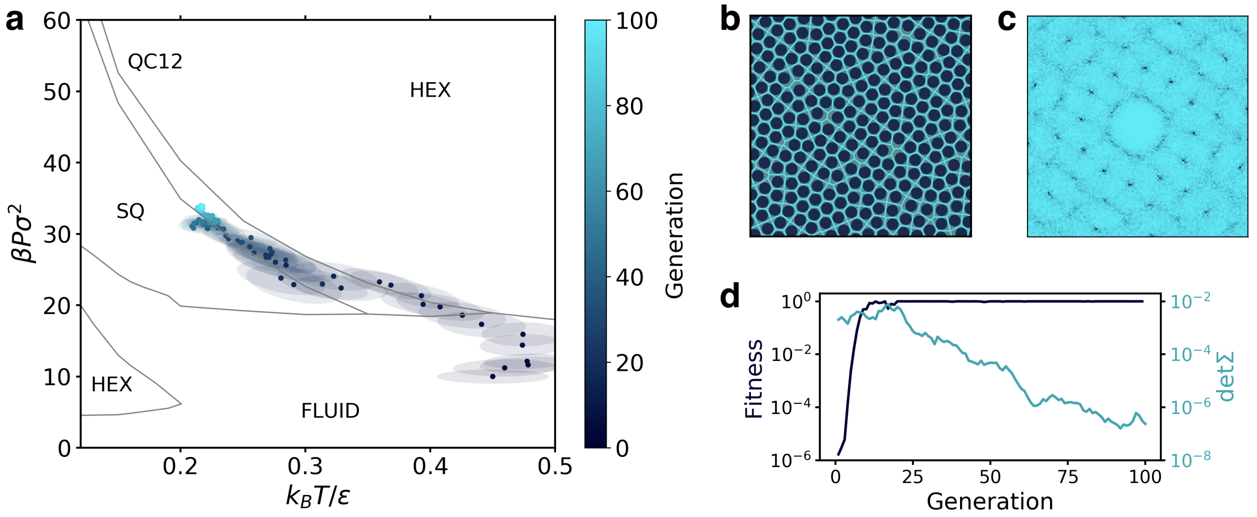


Figure 7.5: Reverse engineering of the QC12 in the HCSS model. (a) Evolution of the Gaussian distribution in the $k_BT/\epsilon - \beta P\sigma^2$ plane. Points and ellipses represent the mean and the covariance matrix (within one standard deviation) of the distribution. The phase diagram in the background is adapted from Ref. 253. (b) Representative snapshot of the QC12 obtained during the last generation. The hard cores are shown in a dark colour, while lines show their Voronoi tessellation. (c) Diffraction pattern of the snapshot in (b). (d) Evolution of the mean fitness and the determinant of the covariance matrix.

7.3.2 Reverse engineering of the QC12 in the HCSS model

We start our investigation by considering the HCSS model with a fixed value of the shoulder width $\delta = 1.4\sigma$, at which the QC12 phase has been shown to be stable [253, 254]. The phase diagram as a function of temperature and pressure (adapted from Ref. 253) is reported in Fig. 7.5a.

The goal here is to reverse engineer the QC12 phase by letting the evolutionary strategy find the narrow region in the phase diagram where the QC12 phase is stable by tuning the system parameters pressure P and temperature T. In other words, we keep the interaction parameters fixed, while trying to optimise the thermodynamics variables to favour the formation of the QC12. Our knowledge of the phase diagram allows us to easily asses and monitor the performance of the reverse engineering process.

To explicitly target the QC12, we use the output of the trained CNN to define the fitness function f for the evolutionary strategy. In particular, for any sample, *i.e.* for any simulation, we define the fitness as $f = \bar{P}_{\mathrm{QC12}}$, where P_{QC12} is the probability that the diffraction pattern of a given configuration is classified as a QC12 by the CNN, and the bar indicates an average taken over representative configurations visited during the simulation.

The results of the reverse engineering process are summarised in Fig. 7.5. Starting the reverse engineering process with a Gaussian centred in the region of stability of the fluid phase, the algorithm reaches the region where the target QC12 is stable in approximately 25 generations. Fig. 7.5a shows the evolution of the multivariate Gaussian distribution in the temperature k_BT/ϵ-pressure $\beta P\sigma^2$ plane across successive generations. A representative snapshot obtained in the last (100th) generation is shown in Fig. 7.5b, while the corresponding diffraction pattern, characterised by twelvefold rotational symmetry, is shown in Fig.7.5c.

The success of the algorithm heavily relies on the ability of the CNN to spot even small structural variations in the system. At the early stages of the reverse engineering process, when the system is in the fluid phase, the algorithm already finds it convenient to increase the pressure, and hence the density, in order to increase the overall structural order. This can clearly be seen in Fig. 7.5d, where we plot the evolution of the mean fitness averaged over all samples. Although the variations of the fitness in the early generations are very tiny, they are sufficient to guide the evolutionary strategy in the right direction.

An efficient exploration of phase space is then made possible by the CMA equations, which evolve the Gaussian distribution at each generation. This not only allows the mean of the distribution to move towards regions with a higher fitness, but it also allows the covariance to stretch when several updates are in the same direction, and then shrink once the fitness is maximised. This is shown in Fig. 7.5d, where we plot the evolution of both the mean fitness and the determinant of the covariance matrix. The determinant becomes larger when the fitness improves, and it decays exponentially once the fitness is maximised.

Note that here we initialised the mean of the Gaussian distribution at a specific state point within the region of stability of the fluid phase, but we find the algorithm to be largely robust to changes in the initial conditions. In the Appendix, we show additional trajectories of the reverse engineering of the QC12 obtained by starting with a Gaussian distribution centred at different state points, *i.e.* in the fluid phase, the SQ phase, the HEX phase at relatively high temperature and low pressure, and the HEX phase at relatively low temperature and high pressure. In all cases, the mean of the parameters distribution converges to the region of stability of the target QC12, clearly showing that the performance is not affected by the particular choice made for the initial conditions.

Furthermore, we would like to stress a crucial aspect that demonstrates the versatility of the algorithm. In fact, the same method, and the exact same CNN, can be used to target any phase that was included in the training data set, simply by changing the definition of the fitness. For instance, to reverse engineer the hexagonal crystal phase, it is sufficient to impose $f = \bar{P}_{\mathrm{HEX}}$. A trajectory of the reverse engineering of the hexagonal crystal is shown in the Appendix.

7.3.3 Reverse engineering of QC12, QC10, and QC18 in the HCSS model

As already discussed, in addition to the QC12, the HCSS model exhibits two other quasicrystalline structures, which are stabilised for different values of the shoulder width δ. As a natural next test, we now explore whether we can reverse engineer all the three stable quasicrystals (QC12, QC10, and QC12) considered in this work. To this end, we fix the temperature to $k_BT/\epsilon = 0.17$, a temperature for which all three QCs are stable, and let the evolutionary strategy optimise the shoulder width δ and the pressure P for each specific QC. In all three cases, we start the reverse engineering process from the same state point in the fluid phase ($\delta = 1.5\sigma$ and $\beta P\sigma^2 = 30$), and choose the fitness function appropriate for the target phase. The results of the reverse engineering process are summarised in Fig. 7.6. In particular, Figs. 7.6a-c show the evolution of the multivariate Gaussian distribution when targeting (a) the QC12, (b) the QC10, and (c) the QC18. Depending on the QC to be found, the distribution evolves in different directions, and eventually converges to different state points. In all cases, the final values of pressure and shoulder width obtained are in excellent agreement with those at which the three

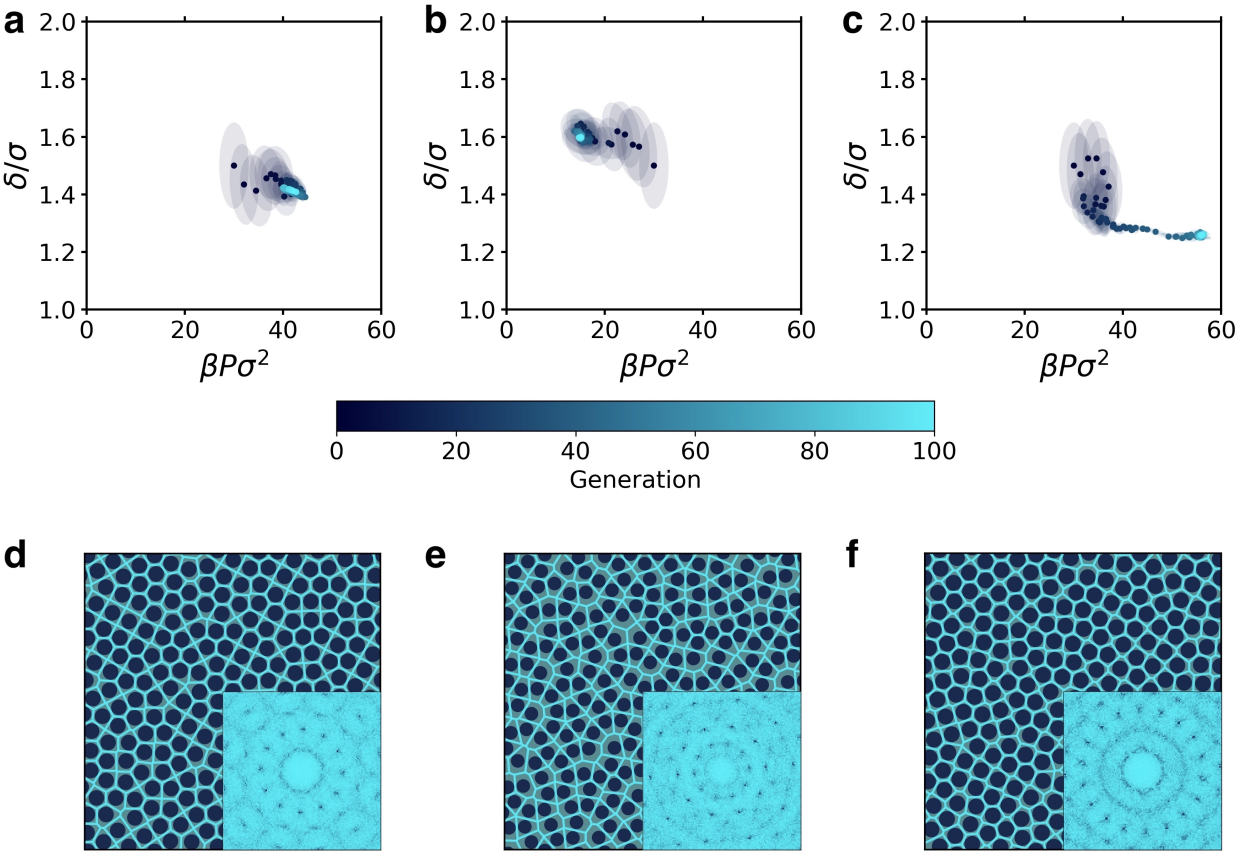


Figure 7.6: Reverse engineering of QC12, QC10, and QC18 in the HCSS model. (a-c) Evolution of the Gaussian distribution in the $\beta P\sigma^2 - \delta/\sigma$ plane during the reverse engineering of (a) the QC12, (b) the QC10, and (c) the QC18 phases. Points and ellipses represent the mean and the covariance matrix (within one standard deviation) of the distribution. (d-f) Representative snapshots of the (d) QC12, (e) QC10, and (f) QC18 obtained in the last generation, along with their diffraction patterns and Voronoi tessellations.

QCs have been shown to be stable [252–255]. Representative snapshots of the QCs obtained and their diffraction patterns are shown in Fig. 7.6d-f. Each diffraction pattern immediately confirms the presence of the correct quasicrystalline structure.

7.3.4 Application to a new model interaction

Thus far we have only addressed the model that was used for training the CNN. A natural next question is whether the method is general enough to work on other model systems, *without* having to retrain the CNN for the specific model under consideration. To answer this question, we now consider a two-dimensional softened-core-shoulder (SCS) model with an interaction potential given by:

$$u(r)/\epsilon = \left(\frac{\sigma}{r}\right)^{14} + \frac{1 - \tanh[k(r-\delta)]}{2}, \tag{7.5}$$

where ϵ is the energy scale, σ represents the typical core diameter, and k and δ are two parameters that, respectively, control the steepness and the characteristic interaction range. Similar to the HCSS, the QC12 has been shown to be stable in a limited range of densities and temperatures with a shoulder width of $\delta = 1.35\sigma$ and $k\sigma = 10$ [261, 262].

To test the ability of our method to be effective on new types of interactions, we use the same CNN that was trained on the HCSS model in order to reverse engineer the QC12 in the SCS model. Similar to the HCSS case, we keep the interaction parameters fixed, *i.e.* $\delta = 1.35\sigma$ and $k\sigma = 10$, and let the evolutionary strategy find the region of densities and temperatures in which the QC12 is stable. The phase diagram in Fig. 7.7a is used as a reference to asses and monitor the performance of the method. Note that, since this phase diagram is in terms of density and temperature, simulations are now performed in the canonical ensemble. Moreover, in contrast to the HCSS case, there are now stable coexistence regions between multiple phases (indicated with a grey background in Fig. 7.7a). As the CNN was not trained on configurations with a phase coexistence, this represents a further robustness test for our method.

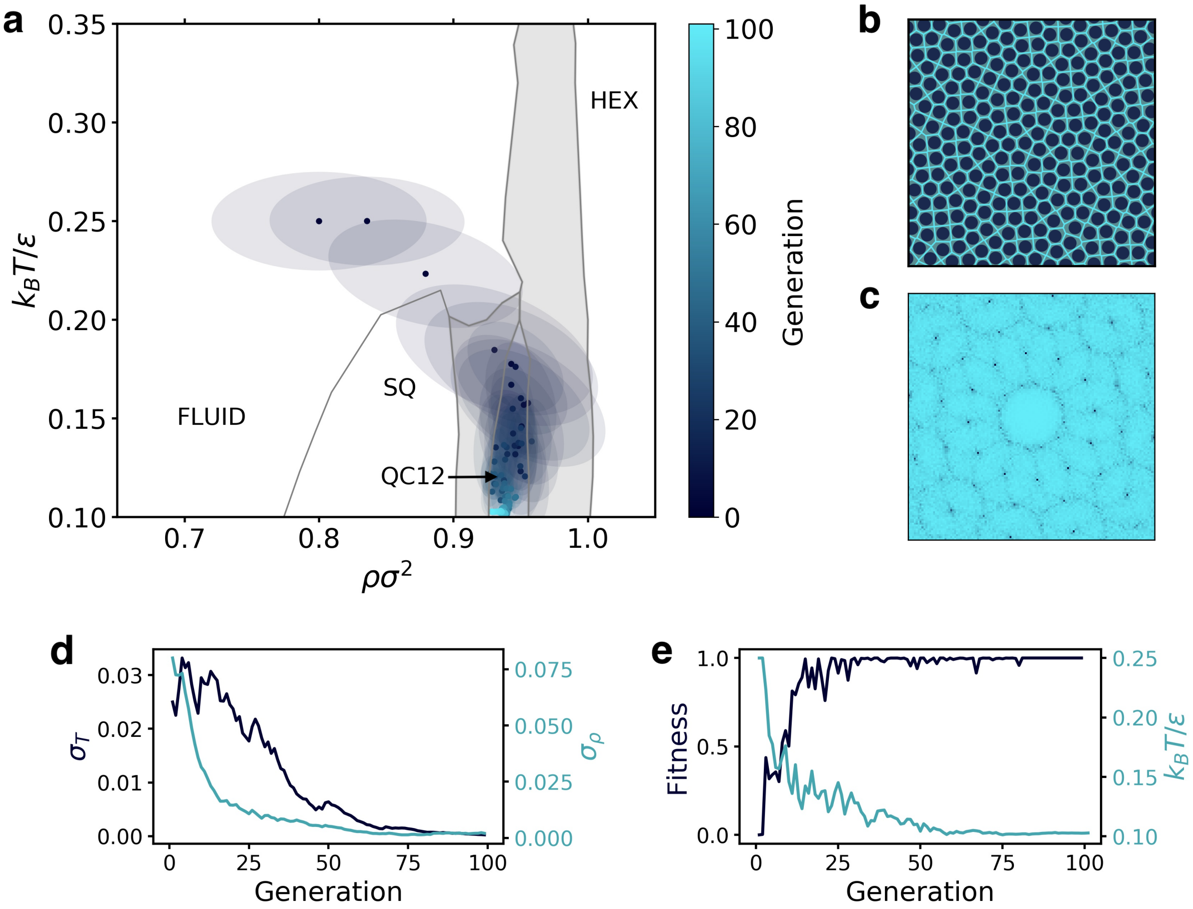


Figure 7.7: Reverse engineering of the QC12 in the SCS model. (a) Evolution of the Gaussian distribution in the $\rho\sigma^2 - k_BT/\epsilon$ plane. Points and ellipses represent the mean and the covariance matrix (within one standard deviation) of the distribution. The phase diagram in the background is adapted from Ref. 262. Coexistence regions are indicated in light grey. (b) Representative snapshot of the QC12 obtained during the last generation and its Voronoi tessellation. (c) Diffraction pattern of the snapshot in (b). (d) Evolution of the square root of the covariance matrix's diagonal elements, which correspond to the standard deviations along the temperature (σ_T) and density (σ_ρ) directions. (e) Evolution of the mean fitness and the mean temperature in (a).

The results of the reverse engineering process are summarised in Fig. 7.7. Specifically, Fig 7.7a shows the evolution of the multivariate Gaussian distribution in the temperature-density plane. Starting with a distribution centred in the fluid region, the algorithm immediately starts to increase the density and lower the temperature in order to increase the overall order. Impressively, after only 5 generations, the mean of the distribution is already inside the region of stability of the QC12, demonstrating the robustness of the CNN to changes in the interaction potential. In the remaining generations, the covariance of the distribution shrinks, and the mean moves towards lower temperatures in the phase diagram. A representative snapshot of the QC12 obtained during the last generation and its diffraction pattern are shown in Figs. 7.7b and 7.7c, respectively.

Looking more closely at the evolution of the model parameters, it is interesting to observe the different behaviour of the temperature and density components. After the first 5 iterations, the density simply oscillates in the tiny range of stability of the QC12, while a large exploration keeps happening in temperature. This can be seen also by looking at the evolution of the standard deviations of temperature (σ_T) and density (σ_ρ) in Fig. 7.7d. While σ_ρ decays almost monotonously from the very beginning, σ_T oscillates for about 20 generations before starting its decay.

We would also like to stress that the reason why the algorithm seems to prefer lower temperatures, despite being already in the stability region of the target phase, is mainly related to a decrease in the thermal fluctuations. With a lower amount of thermal noise, the CNN is presented with cleaner configurations, which, on average, are scored with a higher fitness. This can be seen in Fig. 7.7e, where we plot the evolution of both the mean fitness and the mean temperature. As the temperature goes down, the fluctuations of the fitness become smaller.

7.3.5 Phase discovery

The fundamental ability of the algorithm to generalise to different interaction potentials opens up the possibility of discovering quasicrystals in new model systems. For instance, given the similarities between the SCS and the HCSS models, we might ask whether also the SCS model stabilises different quasicrystals for different values of the shoulder width δ. We note that, compared to the HCSS model, much less is known about the phase behaviour of the two-dimensional SCS system.

Here, we explore the possibility of the SCS model to form a QC10. To this end, we fix $k\sigma = 10$ as in the previous case, and let the evolutionary strategy optimise three parameters: shoulder width δ, temperature T, and pressure P. Note that, by varying these three parameters simultaneously, the algorithm might encounter phases that were not included in the data set for training the CNN. We do not expect this to be a problem, as long as no phase is misclassified as the target phase. This could possibly cause the algorithm to get stuck and eventually converge to the wrong phase. While this problem did not occur in our test, a simple solution would be to include the newly found phase in the training data set, and retrain the CNN.

The results of the reverse engineering process are summarised in Fig. 7.8. Starting from a fluid phase, the evolutionary strategy decreases the temperature, and increases both the pressure and shoulder width in order to maximise the fitness (see Fig. 7.8c-f), discovering the not-yet-predicted QC10 phase for this system. As a further confirmation that the algorithm has indeed found a QC10, Fig. 7.8b shows a representative snapshot obtained during the last generation, along with the corresponding diffraction pattern. Hence, our algorithm has suc-

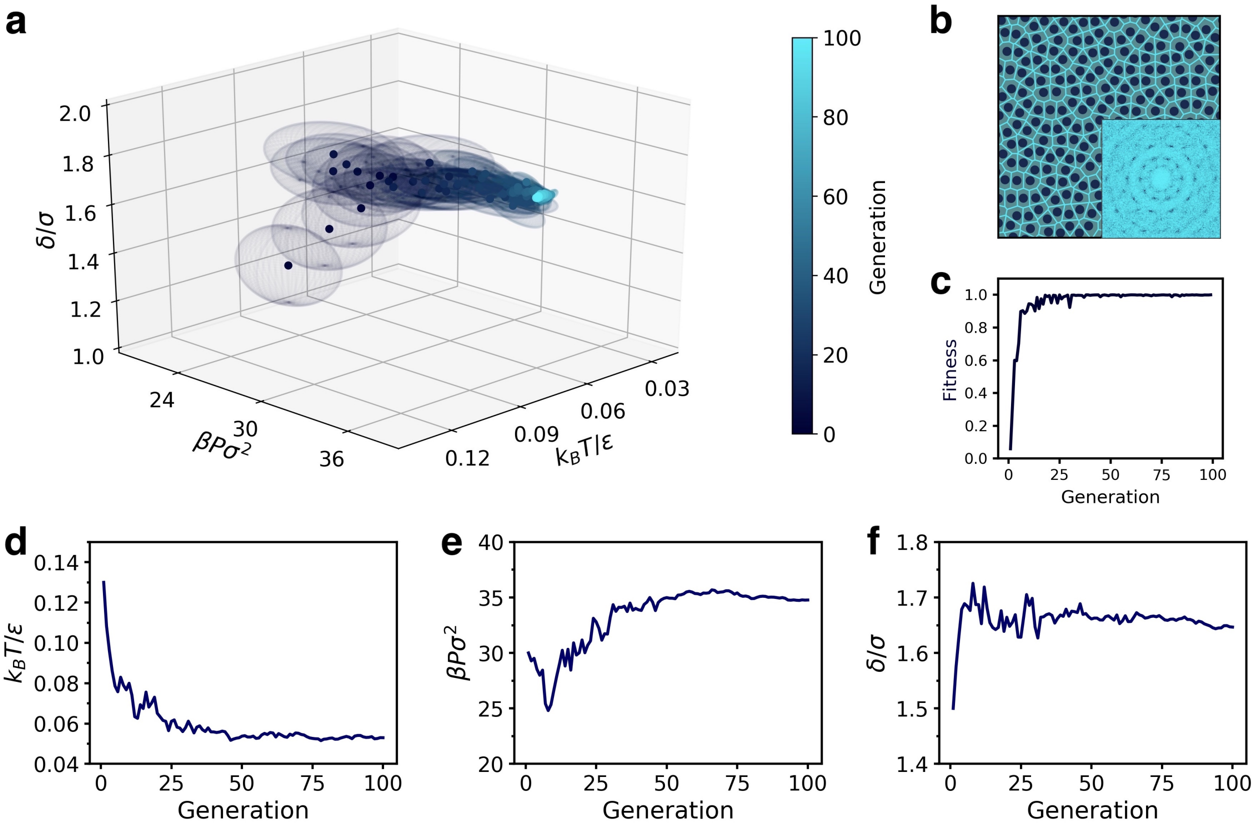


Figure 7.8: Discovery of the QC10 in the SCS model. (a) Evolution of the Gaussian distribution in $k_BT/\epsilon - \beta P\sigma^2 - \delta/\sigma$ space. Points and ellipsoids represent the mean and the covariance matrix (within one standard deviation) of the distribution. (b) Representative snapshot of the QC10 obtained during the last generation, along with its diffraction pattern and Voronoi tessellation. (c) Evolution of the mean fitness. (d-f) Evolution of the three parameters in (a) optimised in the reverse engineering process: (d) temperature k_BT/ϵ, (e) pressure $\beta P\sigma^2$, and (f) shoulder width δ/σ.

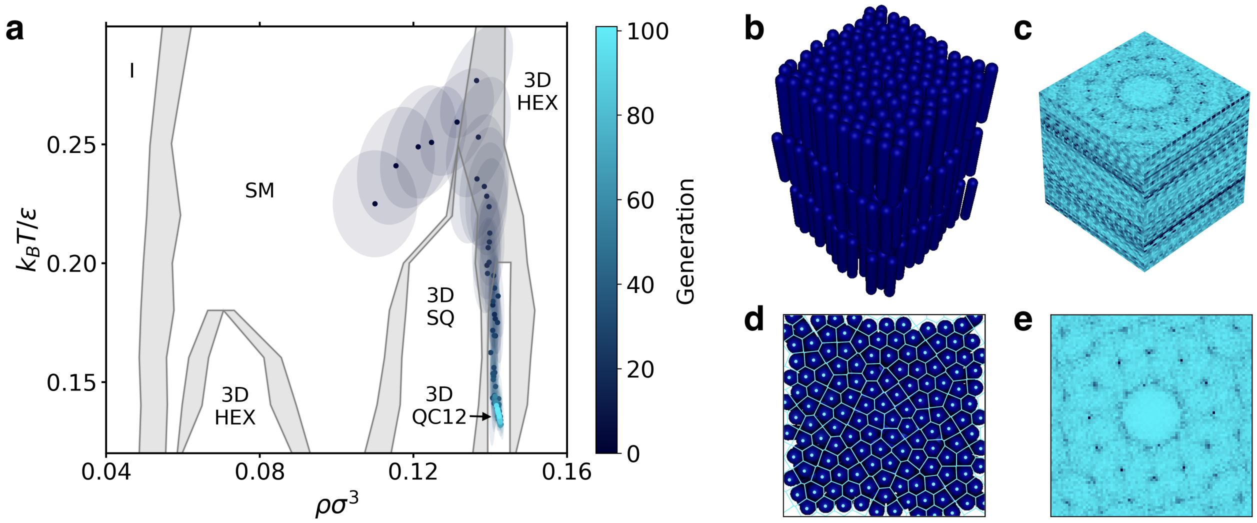


Figure 7.9: Reverse engineering of the QC12 in a three-dimensional model of soft spherocylinders. (a) Evolution of the Gaussian distribution in the $\rho\sigma^3 - k_BT/\epsilon$ plane. Points and ellipses represent the mean and the covariance matrix (within one standard deviation) of the distribution. The phase diagram in the background is adapted from Ref. 256. Coexistence regions are indicated in light grey. (b) Representative snapshot of the 3DQC12 obtained during the last generation, and (c) its three-dimensional diffraction pattern. (d) Top view of the snapshot in (b). The centres of mass and the corresponding Voronoi tessellation are highlighted in a light colour. (e) In-layer diffraction pattern of the top view in (d).

cessfully located a new phase in the SCS model.

7.3.6 Extension to three-dimensional systems

Up to this point, we have shown the efficacy of our method for two-dimensional systems where the scattering pattern is simply a 2D image. Finally, we extend and test our approach on 3D systems. To do so, we consider a 3D system of rod-like particles, modelled as hard-core spherocylinders with a soft deformable corona. We consider spherocylinders with a length-to-diameter ratio $L/\sigma = 5$, interacting *via* the pair potential in Eq. 7.5, where the centre-of-mass distance r is replaced by the minimum distance between two rods d_m, which depends on both the centre-of-mass distance and the relative orientation of the two rods. Note that this model underestimates the repulsive force in the case of parallel rods.

The phase behaviour of this system with $k\sigma = 10$ and $\delta = 1.35\sigma$ has been recently studied in Ref. 256. In addition to the standard isotropic (I) and smectic (SM) phases, this model has been shown to stabilise phases consisting of quasi-two-dimensional layers with unconventional symmetries, including square (3DSQ) and hexagonal (3DHEX) crystals, and a three-dimensional twelvefold quasicrystal (3DQC12). The phase diagram in terms of density and temperature is reported in Fig. 7.9a.

As done in the 2D case, in order to set up our IDM, we train a CNN to classify all the stable phases of this system. Note, however, that the inputs of the CNN are now three-dimensional diffraction patterns. Again, we find the CNN to be highly effective and able to classify all phases with 100% accuracy. The output of the trained CNN is then used to define the fitness for the evolutionary strategy where we target the 3DQC12 phase.

The results of the reverse engineering process are summarised in Fig. 7.9. In particular, Fig. 7.9a shows the evolution of the multivariate Gaussian distribution in the density-temperature plane. Starting with a distribution centred in the SM phase, the mean of the distribution evolves *via* the coexistence region of the SM and 3DHEX phase, to the 3DSQ-3DQC12 phase coexistence region, until it converges in the stability region of the 3DQC12 phase. We note that, although the shortest path in parameter space requires the distribution to cross the 3DSQ region, the algorithm actually avoids it, preferring to enter the coexistence region at high temperature and then move downwards in temperature, where samples with higher fitness are encountered. Surprisingly, this pathway for the formation of QC12 phases was also identified in Ref. 255.

A representative snapshot of the 3DQC12 obtained during the last generation along with its 3D diffraction pattern is shown in Figs. 7.9a and 7.9b, respectively. As a further confirmation of the in-layer QC12 arrangement, Figs. 7.9c and 7.9d show a top view of the same snapshot and the corresponding in-layer 2D diffraction pattern.

The extension of our method to the 3D case is of particular interest from a practical point of view. While a 2D diffraction pattern immediately provides structural information that is easy to read even by eye, the 3D counterpart is much harder to interpret. For this reason, in order to deal with 3D systems, it is often necessary to project the particles coordinates onto the planes with the relevant symmetries. This aspect becomes irrelevant when using a CNN that naturally processes the full 3D information thanks to its inherent architecture.

7.4 Conclusions

Diffraction patterns are used across a multitude of areas in materials science, to understand what structure one is dealing with. In general, this information constitutes a unique signature of each structure, whether it is a crystal, a fluid, a liquid crystal, or a quasicrystal, and shows significant robustness to changes in density and interaction potentials. This can efficiently incorporate all the relevant information of a target phase, and therefore provides a natural order parameter for IDMs.

With the present work we have shown how the use of CNNs as diffraction patterns classifiers, can provide a useful order parameter for the reverse engineering of a multitude of phases. For the above reason, an IDM built on such an order parameter, is not restricted to a specific class of materials, but is instead naturally tailored to reverse engineer multiple colloidal phases, ranging from crystals and quasicrystals, to liquid crystals.

Our results pave the way to structure optimisation and discovery, especially with binary and ternary systems, where the design space becomes even larger due to new system parameters such as size ratio and composition. In these cases, where the present knowledge of phase diagrams and emerging phases is limited, IDMs can prove extremely precious and efficient.

Acknowledgements

I would like to thank Emanuele Boattini, who equally contributed to this work. Also, I would like to extend my gratitude to Dr. Laura Filion for her precious contributions to this work.

7.5 Appendix

7.5.1 Reverse engineering of the QC12 starting from different initial conditions

In the main text, we have shown a trajectory of the reverse engineering of the QC12 in the HCSS model, starting from a specific state point in the region of stability of the fluid phase. Here, we explore whether the performance of the method is affected by a different choice of the initial conditions. To this end, we perform additional trajectories of the reverse engineering of the QC12, starting with a Gaussian distribution centred at different state points, *i.e.* in the fluid phase, the SQ phase, the HEX phase at relatively high temperature and low pressure, and the HEX phase at relatively low temperature and high pressure.

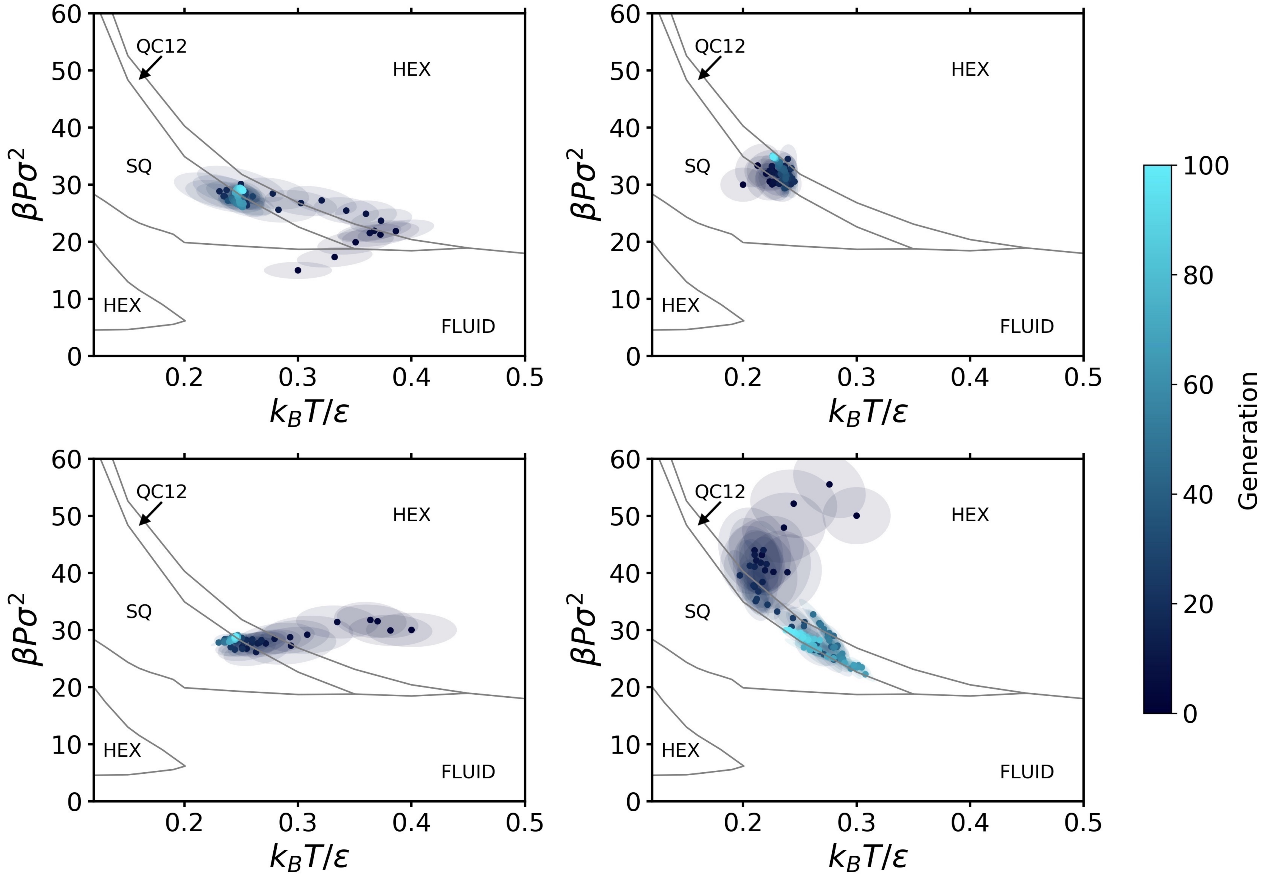


Figure 7.10: Reverse engineering of the QC12 starting from different initial conditions. Four different trajectories showing the evolution of the Gaussian distribution in the $k_BT/\epsilon - \beta P\sigma^2$ plane. Each trajectory, is initialised with the Gaussian centred at different state points, *i.e.* in the fluid phase (left top), the SQ phase (right top), the HEX phase at relatively high temperature and low pressure (left bottom), and the HEX phase at relatively low temperature and high pressure (right bottom). Points and ellipses represent the mean and the covariance matrix (within one standard deviation) of the distribution. The phase diagram in the background is adapted from Ref. 253.

Figure 7.10 shows the results of the reverse engineering process obtained from four, distinct, initial state points. In all cases, the mean of the parameters distribution converges within the region of stability of the target QC12, clearly showing that the performance is not affected by the particular choice made for the initial conditions.

7.5.2 Reverse engineering of the hexagonal crystal in the HCSS model

In the main text, we focused on reverse engineering QCs. However, in principle, the exact same method can be used to reverse engineer any phase that was included in the data set for training the CNN, simply by changing the definition of the fitness. As an example, in Figure 7.11 we report the results of the reverse engineering of the HEX phase in the HCSS model.

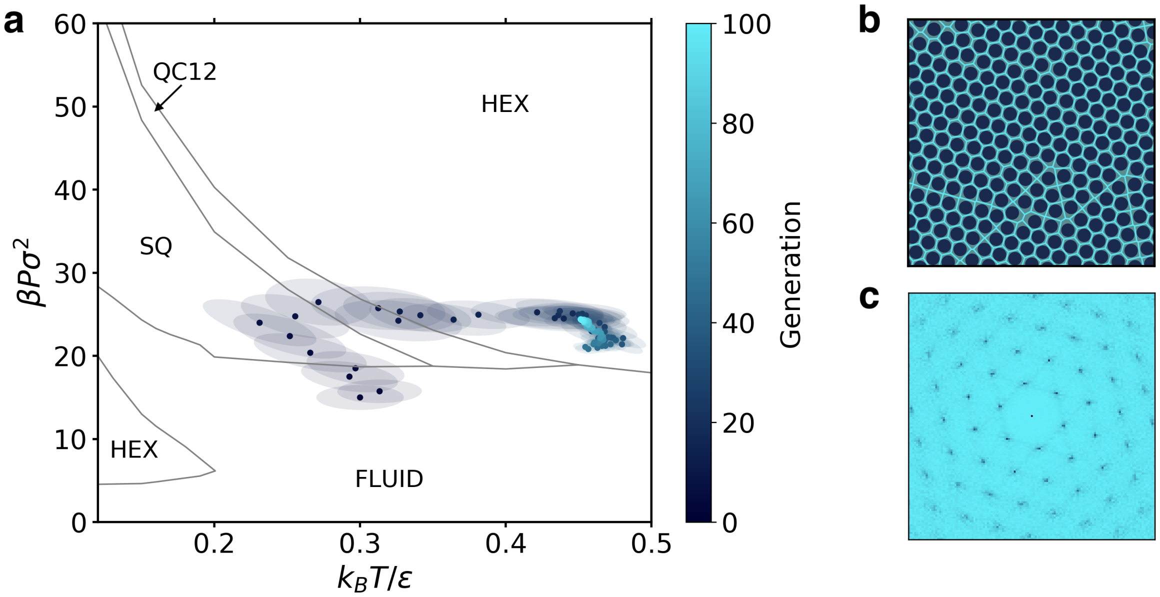


Figure 7.11: Reverse engineering of the HEX phase in the HCSS model. (a) Evolution of the Gaussian distribution in the $k_BT/\epsilon - \beta P\sigma^2$ plane. Points and ellipses represent the mean and the covariance matrix (within one standard deviation) of the distribution. The phase diagram in the background is adapted from Ref. 253. (b) Representative snapshot of the HEX crystal obtained during the last generation, and (c) its diffraction pattern.

8

Towards a full photonic bandgap colloidal crystal: inverse design techniques leading the way

In this concluding Chapter we build on what has been described in this thesis, and show how machine learning-based inverse design methods (IDMs) can be used to solve problems related to the realisation of a colloidal crystal with a bandgap in the visible region.

8.1 Overview

In this thesis, we investigated crystal nucleation for various colloidal systems. This is an extremely large field of research with many open questions. Attempts to answer some of these problems, which have been presented in the current work, may sometimes seem unrelated. It is the aim of this Chapter to try to link these findings together, and highlight some potential outlets for the work presented in this thesis.

In Chapters 2 and 3 we investigated the nucleation process for several colloidal binary crystals – the three types of Laves Phases (LPs) and the icosahedral AB_{13} phase – with the goal of finding for the first time spontaneous nucleation events for such systems using brute force numerical simulations. In both cases the seeding approach was used (see Sections 2.4 and 3.4). The results of this technique are important for two distinct reasons. The first is that it provides valuable information about the thermodynamic quantities involved, such as the Gibbs free-energy barrier, the interfacial free energy and the nucleation rate, useful also for potential experimental investigations. Secondly, the seeding technique tells us in what range of supersaturation – and therefore pressure or density – the free-energy barrier to overcome is low enough to observe spontaneous nucleation. Despite a similar motivation behind the study, Chapters 2 and 3 then evolved independently and differently. In Chapter 2 we studied how, through particle softness, it is possible to influence the concentration of fivefold symmetry clusters, which in turn dictate the onset of kinetic arrest of the system. On the other hand, in Chapter 3 we showed how the use of machine learning techniques helped considerably the classification of the particles in the system, and allowed us to use more local descriptors, even if intrinsically less accurate, thanks to the large model capacity of neural networks.

In Chapters 4 and 5, we turned to single-component systems, for which much more is known about their nucleation process, but for which crucial questions are still open. In particular, in Chapter 4 we dived into the hard-sphere nucleation mechanism, with the goal of explaining the polymorph selection that takes place when witnessing a nucleation event. We found, thanks to a combined use of Bond Order Parameters (BOP) and Topological Cluster Classification (TCC), that in the highly heterogeneous and structured fluid phase, local clusters whose concentration is very high determine a preference towards fcc crystal nucleation for purely geometric reasons. All this leads to a lower interfacial free energy of fcc compared to hcp and therefore to the above-mentioned phenomenon of polymorph selection. Conversely, in Chapter 5, we address the problem of a proper classification of multiple single-component crystals (fcc, hcp, and bcc). We analyse how some classification schemes which have already been used in the past can lead to misclassification, and we propose a new way to perform the same classification based on unsupervised learning. In fact, we use Principal Component Analysis (PCA) on a large data set composed of multiple BOPs – and therefore containing a lot of information about the reference structures – finding a new space where the different polymorphs are easily identifiable, using a few eigenvectors, which in turn are linear combinations of the BOPs used. This keeps the classification easily interpretable.

In Chapters 6 and 7 we focused our attention on inverse design methods (IDMs) as a way to find the optimal conditions at which the desired colloidal crystal, quasicrystal, or liquid crystal, can form. In Chapter 6, we reverse engineered a bcc crystal that, in a repulsive Yukawa particle system, is only stable in a narrow region of the phase diagram. This task allowed us to test two algorithms of different nature – the first based on the statistical fluctuations of the system under investigation and the second derived from classical optimisation techniques. Building upon the knowledge acquired in Chapter 6, in Chapter 7 we develop a new IDM, based on the

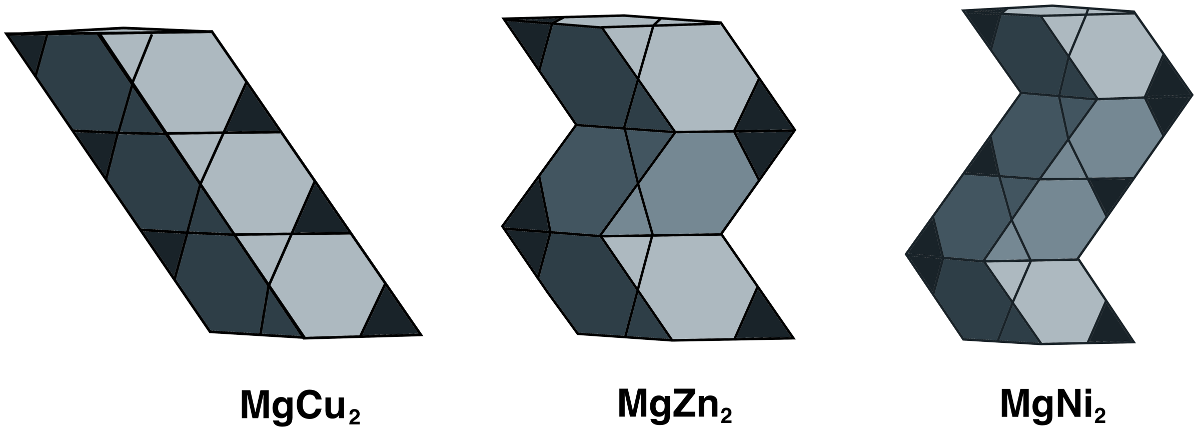


Figure 8.1: The three typical stacking patterns of the Friauf polyhedra, or truncated tetrahedra, which determine the resulting LP. Each Friauf polyhedron describes the environment of a large sphere in a LP, which is placed at the centre of the polyhedron. The 4 large neighbours sit at the centre of the four facets, in a tetrahedral pattern, while the 12 small neighbours sit at the vertices of the polyhedron. Note that the $MgNi_2$ stacking is a combination of the other two.

use of a convolutional neural network (CNN) as an order parameter and showing extremely interesting and promising results for future work. In fact, this IDM not only manages to reverse engineer a large number of phases, including phases notoriously difficult to classify, as for instance quasicrystals, but is equally applicable to two-dimensional and three-dimensional cases, and enabled us to find a quasicrystal not yet reported in the literature for a given model.

Finally, in this Chapter, we build on what we learned about Laves Phase nucleation in Chapter 2, and investigate the possibility of finding a self-assembly route to obtain a crystal with a photonic bandgap, through the IDM described in Chapter 7.

8.2 Photonic crystals and Laves Phases

At the beginning of this thesis, in Chapter 1, we discussed various reasons to study colloidal systems. Specifically, we mentioned that these systems are particularly suitable for the fabrication of photonic crystals, *i.e.*, crystals with a bandgap in the visible region. These materials can control the propagation of visible light due to their periodic modulations of the dielectric constant which define the photonic crystal.

So why are we interested in the Laves Phase? The main reason behind this interest lies in the extraordinary optical properties of the sublattices of these crystal structures. The lattice of all LPs is composed of large and small spheres, but in the case of $MgCu_2$ the respective sublattices correspond to the diamond and the pyrochlore structures, which, if realised with colloidal particles, would both enable a bandgap in the visible region. In other words, the colloidal $MgCu_2$ Laves Phase is an ideal precursor for a photonic crystal [9].

LPs have been shown to be thermodynamically stable for a binary mixture of hard spheres over a specific range of density, stoichiometry and size ratio. Unfortunately though, as shown in Chapter 2, using a short range, isotropic and purely repulsive interaction, the LP with the lowest bulk free energy and therefore the one that is thermodynamically stable with respect to the others is $MgZn_2$, and not $MgCu_2$. Any brute force simulation or experiment performed in

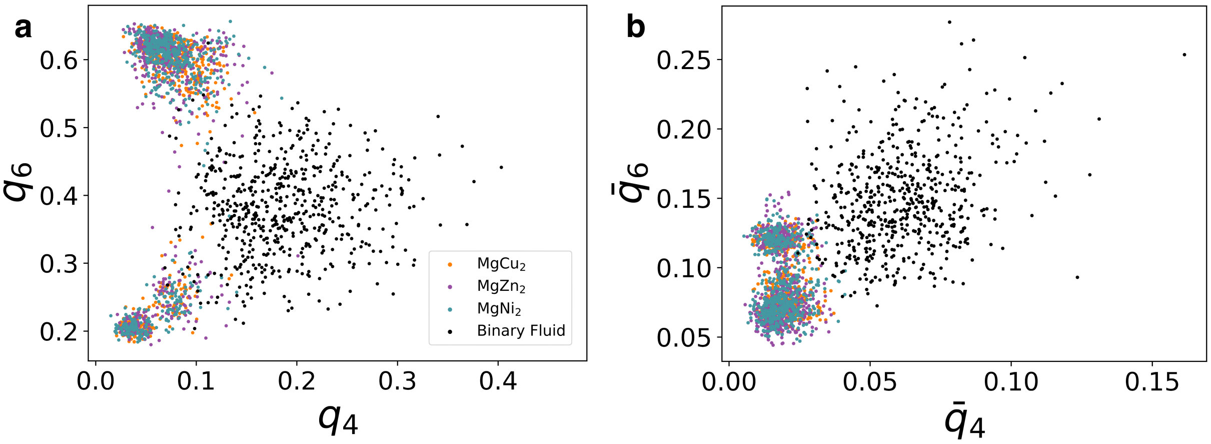


Figure 8.2: (a) Projection on the q_4-q_6 plane of the large and small particles which compose the four phases we wish to distinguish, *i.e.* $MgCu_2$, $MgZn_2$, $MgNi_2$, and the binary fluid [86]. In (b) we show the same analysis, using the averaged $\bar{q}_4$ and $\bar{q}_6$ instead [87]. For each particle, the neighbours have been identified using the SANN algorithm and are considered irrespective of their species [152]. In both cases, LPs can be distinguished from the binary fluid, but not among themselves.

bulk, and using particles interacting as described above, will form an LP that does not have the desired optical properties.

One strategy to overcome this problem is to use a system which is different from a binary hard-sphere mixture. In particular, one may include anisotropy in the interaction potential, so that some predefined stacking of the Friauf polyhedra – which geometrically describe the local environment of a large particle in the three LPs, and whose stacking determines the resulting LP, see Fig. 8.1 – can be preferred over the others, resulting in the $MgCu_2$ structure.

At this point two main problems arise. The first is that, using a model for which we do not have phase diagrams, it is extremely difficult to guess the optimal interaction parameters, as well as the optimal thermodynamic conditions to give rise to a self-assembly process that leads to the formation of the $MgCu_2$ crystal. It is precisely to solve these problems that IDMs were born. However, in the context of an IDM, we saw in Chapters 6 and 7 that an essential requirement for the successful implementation of such methods is a suitable order parameter. In other words, it is necessary to be able to distinguish not only the LPs from the (binary) fluid phase, but it is also crucial to be able to distinguish the different LPs from each other.

Most of the LP classification algorithms developed so far are indeed able to distinguish the crystalline phase from the fluid phase, but fail in their intent to recognise individual LPs. An example was given in Chapter 2, where LPs were classified *via* an algorithm based on dot-products d_{ij} between q_{6m} of the different particles in the system, similar to what has already been done for single component crystals [160,197]. In particular, the distributions of d_{ij} between two large spheres, two small spheres, and between a large and a small sphere, are different in the case of LPs than in the distributions computed for a binary fluid. This analysis is shown in Fig. 2.7. A similar result was obtained by Bommineni and co-workers, who classified particles as belonging to LPs based on the values of the q_4 and q_6 BOPs [86,149]. With these techniques, it is relatively straightforward to distinguish between crystal-like particles belonging to LPs, and fluid-like particles, as shown in Fig. 8.2.

The attempt which is certainly closer to the need of classifying individual LPs independently, was made by Boattini and co-workers [153]. Here, the authors, using a neural network without hidden layers and therefore without any source of non-linearity, distinguished a total of 7 local environments, of which 6 belong to the different particles in the LPs, plus one local environment corresponding to the fluid-like particles. In particular, in the $MgCu_2$ crystal, the authors distinguished 2 local environments, in the $MgZn_2$ crystal 3 local environments, and in the $MgNi_2$ crystal even 5. Although this classification represents an obvious step forward with respect to the algorithms described previously, the fact that some local environments are shared by more than one LP, makes it more complex to target a single crystal structure – the $MgCu_2$ in our case.

In the following Section, we show a deep learning-based classification scheme which correctly classifies the three LPs with great accuracy. This, on its turn, allows a true inverse design strategy to be implemented and thus opens the door to a wide variety of self-assembly opportunities.

8.3 Convolutional neural network-based order parameter

In order to correctly recognise the $MgCu_2$, $MgZn_2$, $MgNi_2$, and the binary fluid phase, we face the problem from a different perspective with respect to the methods mentioned above. In fact, instead of making a prediction for each particle in the system, we classify the system as a whole. This choice has a clear advantage: such a classification scheme is not deceived by the fact that the same local environment is present in multiple LPs.

As in Chapter 7, we encode the structure factor of these phases in a 3-dimensional diffraction pattern and use a convolutional neural network (CNN) to correctly classify the corresponding systems. More specifically, the CNN takes a 3-dimensional diffraction pattern as input, and gives as output an array of real numbers, corresponding to the probability that the input diffraction pattern corresponds to each of the phases to be distinguished.

The reason behind the choice of such an architecture is to be found in the great capacity of a CNN to handle tensorial inputs, like images [263].

The configurations for training the CNN are generated by performing MC simulations of a binary hard-sphere mixture in the canonical ensemble (NVT). The stoichiometry has been selected to be equal to the one of the Laves Phase ($x_L = N_L/(N_L + N_S) = 1/3$), while the size ratio σ_S/σ_L has been set to 0.78. The number of particles was set to $N = 576$ for all phases, and periodic boundary conditions were applied in all three dimensions. We then run simulations and collect 10^3 independent configurations for each phase.

The architecture we selected for the CNN, apart from the dimension of the output layer, which corresponds to the number of phases to classify, is exactly the same as in Chapter 7. Also, the generation of the diffraction patterns, the preprocessing applied to the input data, as well as all the training parameters, can all be found in Section 7.2.3. Examples of typical three-dimensional diffraction patterns are given in Fig. 8.3.

Training the CNN for a total of 200 epochs, we find that the loss function stabilises, showing a plateau, and assumes approximately the same values for both the training and the test sets, indicating that the model did not overfit the training data (see Fig. 8.4). More importantly, the accuracy of the model on the 4 classes – calculated on the test set – is equal to 100% for each of them (Table 8.1). This fundamental result shows how, thanks to a technique belonging to the class of deep learning algorithms, it is possible to distinguish with extreme precision even *similar* phases, such as the three LP crystals, in addition to the binary fluid.

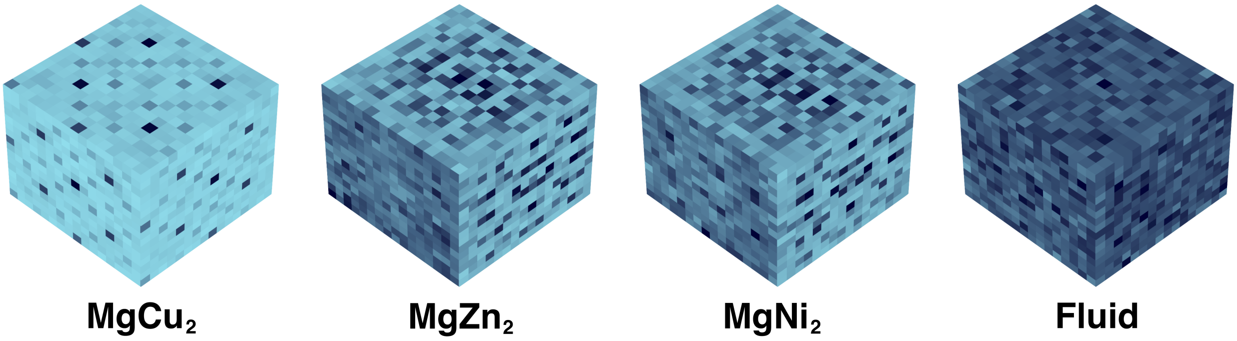


Figure 8.3: Examples of the preprocessed three-dimensional diffraction patterns generated by the four phases we wish to classify. Note that the symmetry planes of the $MgZn_2$ and of the $MgNi_2$ do not coincide with the x-, y- and z-axis. For this reason, while the $MgCu_2$ crystal displays a diffraction pattern with clear peaks, as expected for a crystal structure, the other two LPs would need a precise rotation in order to be interpretable as a crystal structure by eye. However, here lies the great capabilities of a CNN, which is able to correctly classify the three LPs regardless their orientation, even when it becomes extremely challenging to do that through human-engineered features.

Class	Accuracy
$MgCu_2$	100%
$MgZn_2$	100%
$MgNi_2$	100%
Fluid	100%

Table 8.1: Accuracies of the CNN on the test set calculated for all four classes.

However, the importance of this result lies mainly in the possibilities it opens up. The existence of an order parameter that can distinguish LPs from each other is in fact a fundamental ingredient for the set up of an inverse design strategy, able to efficiently explore the parameter space with the aim to find the optimal region where the $MgCu_2$ is the stable phase.

In the next section, we list some potential models to achieve this goal.

8.4 Candidate models

We have seen that a binary mixture of isotropic spheres tends to stabilise the $MgZn_2$ structure, and that including a degree of directionality in the interaction potential is likely to be a crucial ingredient in order to make the $MgCu_2$ crystal stable.

In order to understand what might be a system that has the capability to stabilise a $MgCu_2$ phase, we take inspiration from a work which has been recently published by He and co-workers, in which they show that a colloidal diamond structure may be obtained by using nanoparticles with attractive tetrahedral patches [38]. Since the large spheres of the $MgCu_2$ crystal are arranged in such a pattern, one possibility is to use a patchy model for the large spheres, like the Kern-Frenkel model [264].

For the small spheres there are two possibilities. It is obviously possible to think of a patchy particle model for the small spheres as well, with a non-tetrahedral arrangement of patches, in order to stabilise the pyrochlore structure, which corresponds to the pattern generated by

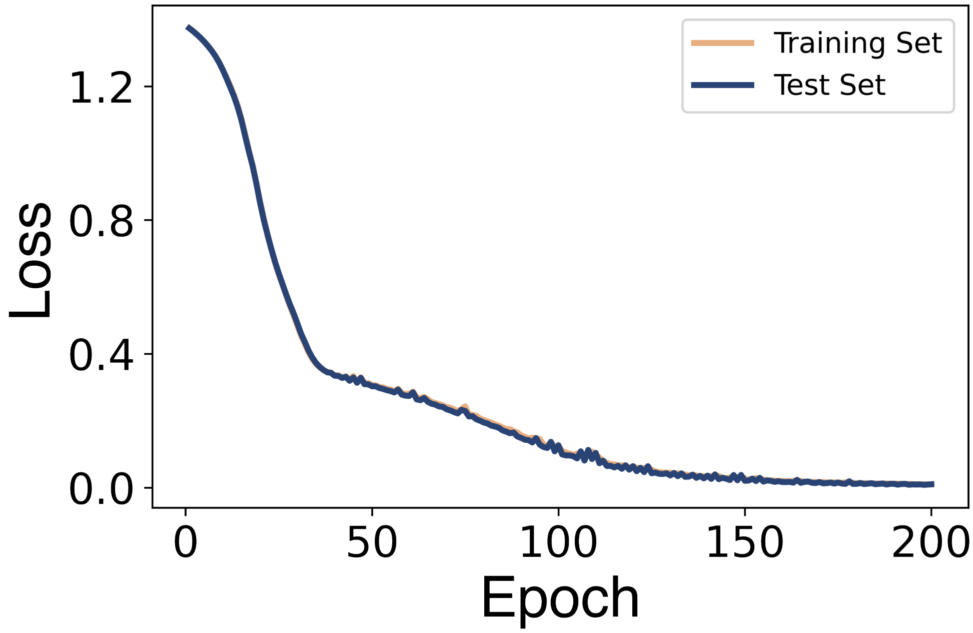


Figure 8.4: Loss function as a function of the epochs for both the training set (gold curve) and the test set (blue curve). The two curves are nearly identical, which shows that the CNN is not overfitting the training data, and is therefore able to predict the four classes reliably on new unseen samples.

small spheres in the $MgCu_2$ crystal. However, a simpler choice would be to use an isotropic and repulsive interaction potential, such as a hard-sphere potential, for the small spheres. It may be possibile that, if the large spheres self-assemble, due to the attractive patches, into a diamond structure, the small spheres could position themselves according to the lattice sites of a pyrochlore crystal simply by maximising the packing of the system. Obviously, in such a model, many parameters are at stake, and this is why a CNN-based IDM would be fundamental. Parameters such as the opening angle of the patches, as well as the ideal stoichiometry and size ratio, and of course temperature and pressure (or density), are all fundamental parameters capable of altering the stability relationships between all the phases involved.

We close this section by adding that, of course, there are other particle models that seem promising for self-assembling into such crystals, such as a system of dumbbells [265], or diblock polymer-homopolymer blends [266]. Again, the actual ability of these models to make the $MgCu_2$ stable could be tested through the CNN-based IDM we have shown, however an accurate description of these models goes beyond the scope of this thesis.

8.5 Conclusions

All the chapters covered in this thesis revolve around the concept of nucleation in soft matter systems. As already explained in Chapter 1, the importance of the study of nucleation in soft matter systems is twofold.

First, it is crucial to understand the mechanisms of nucleation in such systems, because of the strong analogy that colloids share with atomic and molecular systems.

Chapters 2, 3, 4, and 5 go into this direction. In fact, in binary systems, the relationship between crystallisation and glass transition has been thoroughly studied, highlighting the role of fivefold symmetry clusters and of softness degree in the particles interaction. Still in binary systems the relationship between particles of very different sizes during nucleation has been analysed. For single component systems, the mechanism of nucleation for hard spheres has been studied in the light of the behaviour of several smaller topological clusters but whose concentration in the fluid is very high. The classification algorithm proposed in Chapter 5 inherently contains a high degree of information about the system, and could therefore inspire future work.

Secondly, the study of nucleation in soft matter systems is also important because it is the formation mechanism of many structures with advanced functional properties, which may form the basis of new technologies.

In this case, Chapters 2, 3, 6, and 7 may be of interest to the reader. Here we have studied the physical conditions necessary for the formation of several binary crystals, including the $MgCu_2$ Laves Phase, known for its extraordinary optical properties. Finally, we have addressed inverse design methods, which constitute a family of remarkably promising techniques in order to optimise the self-assembly conditions for a given system, or even in order to discover new phases not yet reported in the literature. An example of the power and versatility of such systems is given in this very last Chapter.

Being an extremely broad topic, obviously many questions regarding nucleation remain open. As a matter of fact, the biggest problem in soft matter is related to nucleation, which is the gigantic discrepancy between the nucleation rates of hard spheres as measured by experiments with respect to simulations [78,160,197,267,268].

More specifically, some questions arise directly from the research work presented in this thesis. For example, the nucleation mechanism presented for hard spheres is based on the presence of specific particle clusters that exist as the system is very dense. How does the mechanism change when studying the nucleation of Yukawa particles, which are often employed as a model for real-world colloids, and which show crystallisation phenomena also at lower densities? Are there other clusters which act as transient clusters during nucleation? More in general, is it possible to combine a large set of topological descriptors, and use a machine learning-based – and therefore automated – technique to find these clusters for several different systems? Differently, taking inspiration from the last chapters, is our IDM effective on lower density structure, such as open crystals? And what can we achieve for 3D quasicrystals which are not layered, as the one presented in Chapter 7? Can we use IDMs to also deduce something about quasicrystal formation?

These are obviously challenging questions, but also fascinating ones. We hope that some of the conclusions drawn by the work presented in this thesis, as well as some of the analysis and methods we developed, can help future researchers to answer some of these questions, in order to broaden the knowledge we already have about this exciting field of research.

Acknowledgements

I would like to thank Emanuele Boattini for providing useful suggestions about training the CNN.

References

[1] A. D. McNaught et al., *Compendium of chemical terminology*, volume 1669, Blackwell Science Oxford, 1997.

[2] T. Graham, *Liquid diffusion applied to analysis*, Philos. Trans. R. Soc. Lond. , 183 (1861).

[3] R. Brown, *A brief account of microscopical observations made in the months of june, july and august 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies*, Phil. Mag. **4**, 161 (1828).

[4] W. Sutherland, *A dynamical theory of diffusion for non-electrolytes and the molecular mass of albumin*, Lond. Edinb. Dubl. Phil. Mag. **9**, 781 (1905).

[5] A. Einstein, *Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen*, Annalen der physik **4** (1905).

[6] G. M. Whitesides and B. Grzybowski, *Self-assembly at all scales*, Science **295**, 2418 (2002).

[7] Y. A. Vlasov, X.-Z. Bo, J. C. Sturm, and D. J. Norris, *On-chip natural assembly of silicon photonic bandgap crystals*, Nature **414**, 289 (2001).

[8] E. A. Kamenetzky, L. G. Magliocco, and H. P. Panzer, *Structure of solidified colloidal array laser filters studied by cryogenic transmission electron microscopy*, Science **263**, 207 (1994).

[9] A.-P. Hynninen, J. H. Thijssen, E. C. Vermolen, M. Dijkstra, and A. Van Blaaderen, *Self-assembly route for photonic crystals with a bandgap in the visible region*, Nat. Mater **6**, 202 (2007).

[10] J. F. Galisteo-López, M. Ibisate, R. Sapienza, L. S. Froufe-Pérez, Á. Blanco, and C. López, *Self-assembled photonic structures*, Adv. Mater. **23**, 30 (2011).

[11] J. Sanders, *Close-packed structures of spheres of two different sizes i. observations on natural opal*, Phil. Mag. A **42**, 705 (1980).

[12] M. Murray and J. Sanders, *Close-packed structures of spheres of two different sizes ii. the packing densities of likely arrangements*, Phil. Mag. A **42**, 721 (1980).

[13] P. M. Chaikin, T. C. Lubensky, and T. A. Witten, *Principles of condensed matter physics*, volume 10, Cambridge university press Cambridge, 1995.

[14] D. G. Fahrenheit, *Barometri novi descriptio*, Philos. Trans. Royal Soc. Lond. **33**, 179 (1724).

[15] F. Szabadváry, *Joseph louis gay-lussac (1778–1850) and analytical chemistry*, Talanta **25**, 611 (1978).

[16] J. W. Gibbs, *On the equilibrium of heterogeneous substances*, Am. J. Sci. Arts **16**, 441 (1878).

[17] M. Volmer and A. Weber, *Keimbildung in übersättigten gebilden*, Zeitschrift für physikalische Chemie **119**, 277 (1926).

[18] N. G. Van Kampen, *Stochastic processes in physics and chemistry*, volume 1, Elsevier, 1992.

[19] R. Becker and W. Döring, *The kinetic treatment of nuclear formation in supersaturated vapors*, Ann. Phys **24**, 752 (1935).

[20] D. W. Oxtoby, *Homogeneous nucleation: theory and experiment*, J. Phys. Condens. Matter **4**, 7627 (1992).

[21] J. Zeldovich, *A theory of the ignition on incandescent surfaces*, J. Exp. Theor. Phys **12**, 525 (1942).

[22] F. C. Frank, *Supercooling of liquids*, Proc. R. Soc. Lond., A Math. Phys. Sci. **215**, 43 (1952).

[23] D. R. Nelson, *Defects and geometry in condensed matter physics*, Cambridge University Press, 2002.

[24] U. Gasser, E. R. Weeks, A. Schofield, P. Pusey, and D. Weitz, *Real-space imaging of nucleation and growth in colloidal crystallization*, Science **292**, 258 (2001).

[25] N. C. Karayiannis, R. Malshe, J. J. de Pablo, and M. Laso, *Fivefold symmetry as an inhibitor to hard-sphere crystallization*, Phys. Rev. E **83**, 061505 (2011).

[26] J. Taffs, S. R. Williams, H. Tanaka, and C. P. Royall, *Structure and kinetics in the freezing of nearly hard spheres*, Soft Matter **9**, 297 (2013).

[27] J. D. Bernal, *A geometrical approach to the structure of liquids*, Nature **183**, 141 (1959).

[28] J. Finney, *Random packings and the structure of simple liquids. i. the geometry of random close packing*, Proc. R. Soc. Lond., A Math. phys. sci. **319**, 479 (1970).

[29] J. Taffs and C. P. Royall, *The role of fivefold symmetry in suppressing crystallization*, Nat. Commun. **7**, 1 (2016).

[30] A. Stukowski, *Structure identification methods for atomistic simulations of crystalline materials*, Model. Simul. Mater. Sci. Eng. **20**, 045021 (2012).

[31] A. Malins, S. R. Williams, J. Eggers, and C. P. Royall, *Identification of structure in condensed matter with the topological cluster classification*, J. Chem. Phys. **139**, 234506 (2013).

[32] A.-P. Hynninen and M. Dijkstra, *Phase diagrams of hard-core repulsive yukawa particles*, Phys. Rev. E **68**, 021407 (2003).

[33] M. Chiappini, T. Drwenski, R. Van Roij, and M. Dijkstra, *Biaxial, twist-bend, and splay-bend nematic phases of banana-shaped particles revealed by lifting the "smectic blanket"*, Phys. Rev. Lett. **123**, 068001 (2019).

[34] S. C. Glotzer and M. J. Solomon, *Anisotropy of building blocks and their assembly into complex structures*, Nat. Mater **6**, 557 (2007).

[35] S. Sacanna and D. J. Pine, *Shape-anisotropic colloids: Building blocks for complex assemblies*, Curr. Opin. Colloid Interface Sci. **16**, 96 (2011).

[36] K. Miszta et al., *Hierarchical self-assembly of suspended branched colloidal nanocrystals into superlattice structures*, Nat. Mater **10**, 872 (2011).

[37] M. A. Boles, M. Engel, and D. V. Talapin, *Self-assembly of colloidal nanocrystals: From intricate structures to functional materials*, Chem. Rev. **116**, 11220 (2016).

[38] M. He, J. P. Gales, É. Ducrot, Z. Gong, G.-R. Yi, S. Sacanna, and D. J. Pine, *Colloidal diamond*, Nature **585**, 524 (2020).

[39] M. C. Rechtsman, F. H. Stillinger, and S. Torquato, *Optimized interactions for targeted self-assembly: application to a honeycomb lattice*, Phys. Rev. Lett. **95**, 228301 (2005).

[40] M. Florescu, S. Torquato, and P. J. Steinhardt, *Designer disordered materials with large, complete photonic band gaps*, Proc. Natl. Acad. Sci. U.S.A **106**, 20658 (2009).

[41] M. Z. Miskin and H. M. Jaeger, *Adapting granular materials through artificial evolution*, Nat. Mater **12**, 326 (2013).

[42] A. Jain, J. A. Bollinger, and T. M. Truskett, *Inverse methods for material design*, AIChE J. **8**, 2732 (2014).

[43] K. Ho, C. T. Chan, and C. M. Soukoulis, *Existence of a photonic gap in periodic dielectric structures*, Phys. Rev. Lett. **65**, 3152 (1990).

[44] A. R. Oganov and C. W. Glass, *Crystal structure prediction using ab initio evolutionary techniques: Principles and applications*, J. Chem. Phys. **124**, 244704 (2006).

[45] B. I. Dahiyat and S. L. Mayo, *De novo protein design: fully automated sequence selection*, Science **278**, 82 (1997).

[46] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, *Design of a novel globular protein fold with atomic-level accuracy*, science **302**, 1364 (2003).

[47] S. Hyun and S. Torquato, *Designing composite microstructures with targeted properties*, J. Mater. Res. **16**, 280 (2001).

[48] A. W. Long and A. L. Ferguson, *Rational design of patchy colloids via landscape engineering*, Mol. Syst. Des. Eng. **3**, 49 (2018).

[49] M. R. Khadilkar, S. Paradiso, K. T. Delaney, and G. H. Fredrickson, *Inverse design of bulk morphologies in multiblock polymers using particle swarm optimization*, Macromolecules **50**, 6702 (2017).

[50] M. Z. Miskin, G. Khaira, J. J. de Pablo, and H. M. Jaeger, *Turning statistical physics models into materials design engines*, Proc. Natl. Acad. Sci. U.S.A **113**, 34 (2016).

[51] Y. Geng, G. van Anders, P. M. Dodd, J. Dshemuchadse, and S. C. Glotzer, *Engineering entropy for the inverse design of colloidal crystals from hard shapes*, Sci. Adv. **5** (2019).

[52] A. Jain, J. R. Errington, and T. M. Truskett, *Inverse design of simple pairwise interactions with low-coordinated 3d lattice ground states*, Soft Matter **9**, 3866 (2013).

[53] G. van Anders, D. Klotsa, A. S. Karas, P. M. Dodd, and S. C. Glotzer, *Digital alchemy for materials design: Colloids and beyond*, ACS Nano **9**, 9542 (2015).

[54] R. B. Jadrich, J. A. Bollinger, B. A. Lindquist, and T. M. Truskett, *Equilibrium cluster fluids: Pair interactions via inverse design*, Soft Matter **11**, 9342 (2015).

[55] W. D. Piñeros, B. A. Lindquist, R. B. Jadrich, and T. M. Truskett, *Inverse design of multicomponent assemblies*, J. Chem. Phys. **148**, 104509 (2018).

[56] Z. M. Sherman, M. P. Howard, B. A. Lindquist, R. B. Jadrich, and T. M. Truskett, *Inverse methods for design of soft materials*, J. Chem. Phys. **152**, 140902 (2020).

[57] C. S. Adorf, J. Antonaglia, J. Dshemuchadse, and S. C. Glotzer, *Inverse design of simple pair potentials for the self-assembly of complex structures*, J. Chem. Phys. **149**, 204102 (2018).

[58] M. Dijkstra and E. Luijten, *From predictive modelling to machine learning and reverse engineering of colloidal self-assembly*, Nat. Mater **20**, 762 (2021).

[59] A. M. Kalsin, M. Fialkowski, M. Paszewski, S. K. Smoukov, K. J. Bishop, and B. A. Grzybowski, *Electrostatic self-assembly of binary nanoparticle crystals with a diamond-like lattice*, Science **312**, 420 (2006).

[60] K. J. Bishop, N. R. Chevalier, and B. A. Grzybowski, *When and why like-sized, oppositely charged particles assemble into diamond-like crystals*, J. Phys. Chem. Lett. **4**, 1507 (2013).

[61] E. V. Shevchenko, D. V. Talapin, N. A. Kotov, S. O'brien, and C. B. Murray, *Structural diversity in binary nanoparticle superlattices*, Nature **439**, 55 (2006).

[62] W. H. Evers, B. D. Nijs, L. Filion, S. Castillo, M. Dijkstra, and D. Vanmaekelbergh, *Entropy-driven formation of binary semiconductor-nanocrystal superlattices*, Nano Lett. **10**, 4235 (2010).

[63] S. Yoshimura and S. Hachisu, *Order formation in binary mixtures of monodisperse latices*, in *Frontiers in Colloid Science In Memoriam Professor Dr. Bun-ichi Tamamushi*, Springer: New York, 1983; pp 59-70.

[64] M. Hasaka, H. Nakashima, and K. Oki, *Structure of the laves phase observed in polystyrene latexes*, Trans. Jpn. Inst. Met. **25**, 65 (1984).

[65] G. H. Ma, T. Fukutomi, and N. Morone, *Preparation and analysis of ordered structure of binary mixtures composed of poly (4-vinylpyridine) and polystyrene microgels*, J. Colloid Interface Sci. **168**, 393 (1994).

[66] J.-P. Gauthier, E. Fritsch, B. Aguilar-Reyes, A. Barreau, and B. Lasnier, *Phase de laves dans la première opale ct bidisperse*, C. R. Geosci. **336**, 187 (2004).

[67] B. Cabane, J. Li, F. Artzner, R. Botet, C. Labbez, G. Bareigts, M. Sztucki, and L. Goehring, *Hiding in plain view: Colloidal self-assembly from polydisperse populations*, Phys. Rev. Lett. **116**, 208001 (2016).

[68] N. Schaertl, D. Botin, T. Palberg, and E. Bartsch, *Formation of Laves phases in buoyancy matched hard sphere suspensions*, Soft Matter **14**, 5130 (2018).

[69] P. K. Bommineni, M. Klement, and M. Engel, *Growing binary hard sphere crystals*, arXiv:1912.06251. arXiv.org e-Print archive. https://arxiv.org/pdf/1912.06251.pdf (accessed January 22, 2020) (2019).

[70] S. Jungblut and C. Dellago, *Crystallization of a binary Lennard-Jones mixture*, J. Chem. Phys. **134**, 104501 (2011).

[71] K. Zhao and T. G. Mason, *Shape-designed frustration by local polymorphism in a near-equilibrium colloidal glass*, Proc. Natl. Acad. Sci. U.S.A. **112**, 12063 (2015).

[72] P. Ronceray and P. Harrowell, *Suppression of crystalline fluctuations by competing structures in a supercooled liquid*, Phys. Rev. E **96**, 042602 (2017).

[73] J. Russo, F. Romano, and H. Tanaka, *Glass forming ability in systems with competing orderings*, Phys. Rev. X **8**, 021040 (2018).

[74] E. G. Teich, G. van Anders, and S. C. Glotzer, *Identity crisis in alchemical space drives the entropic colloidal glass transition*, Nat. Commun. **10**, 1 (2019).

[75] J. D. Weeks, D. Chandler, and H. C. Andersen, *Role of repulsive forces in determining the equilibrium structure of simple liquids*, J. Chem. Phys. **54**, 5237 (1971).

[76] D. Wang et al., *Binary icosahedral quasicrystals of hard spheres in spherical confinement*, arXiv:1906.10088. arXiv.org e-Print archive. https://arxiv.org/pdf/1906.10088.pdf (accessed January 24, 2020) (2019).

[77] T. Kawasaki and H. Tanaka, *Formation of a crystal nucleus from liquid*, Proc. Natl. Acad. Sci. U.S.A. **107**, 14036 (2010).

[78] L. Filion, R. Ni, D. Frenkel, and M. Dijkstra, *Simulation of nucleation in almost hard-sphere colloids: The discrepancy between experiment and simulation persists*, J. Chem. Phys. **134**, 134901 (2011).

[79] D. Richard and T. Speck, *Crystallization of hard spheres revisited. i. extracting kinetics and free energy landscape from forward flux sampling*, J. Chem. Phys. **148**, 124110 (2018).

[80] D. Richard and T. Speck, *Crystallization of hard spheres revisited. ii. thermodynamic modeling, nucleation work, and the surface of tension*, J. Chem. Phys. **148**, 224102 (2018).

[81] J. M. Polson, E. Trizac, S. Pronk, and D. Frenkel, *Finite-size corrections to the free energies of crystalline solids*, J. Chem. Phys. **112**, 5339 (2000).

[82] C. Vega, E. Sanz, J. Abascal, and E. Noya, *Determination of phase diagrams* via *computer simulation: Methodology and applications to water, electrolytes and proteins*, J. Phys.: Condens. Matter **20**, 153101 (2008).

[83] D. Frenkel and B. Smit, *Understanding Molecular Simulation: from Algorithms to Applications*, Second Edition; Computational Science Series; Elsevier: Cambridge, MA, 2002, Vol. 1, pp. 241-245.

[84] N. Wood, J. Russo, F. Turci, and C. P. Royall, *Coupling of sedimentation and liquid structure: Influence an hard sphere nucleation*, J. Chem. Phys. **149**, 204506 (2018).

[85] P. Crowther, F. Turci, and C. P. Royall, *The nature of geometric frustration in the kob-andersen mixture*, J. Chem. Phys. **143**, 044503 (2015).

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

[87] W. Lechner and C. Dellago, *Accurate determination of crystal structures based on averaged local bond order parameters*, J. Chem. Phys. **129**, 114707 (2008).

[88] J. R. Espinosa, C. Vega, C. Valeriani, and E. Sanz, *Seeding approach to crystal nucleation*, J. Chem. Phys. **144**, 034501 (2016).

[89] J. R. Espinosa, P. Sampedro, C. Valeriani, C. Vega, and E. Sanz, *Lattice mold technique for the calculation of crystal nucleation rates*, Faraday Discuss. **195**, 569 (2017).

[90] J. A. Anderson, C. D. Lorenz, and A. Travesset, *General purpose molecular dynamics simulations fully implemented on graphics processing units*, J. Comput. Phys. **227**, 5342 (2008).

[91] J. Glaser, T. D. Nguyen, J. A. Anderson, P. Lui, F. Spiga, J. A. Millan, D. C. Morse, and S. C. Glotzer, *Strong scaling of general-purpose molecular dynamics simulations on gpus*, Comput. Phys. Commun. **192**, 97 (2015).

[92] G. J. Martyna, D. J. Tobias, and M. L. Klein, *Constant pressure molecular dynamics algorithms*, J. Chem. Phys. **101**, 4177 (1994).

[93] S. Auer and D. Frenkel, *Crystallization of weakly charged colloidal spheres: A numerical study*, J. Phys.: Condens. Matter **14**, 7667 (2002).

[94] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, Second Edition; Elsevier: Cambridge, MA, 1990.

[95] Y. Rosenfeld and N. Ashcroft, *Theory of simple classical fluids: Universality in the short-range structure*, Phys. Rev. A **20**, 1208 (1979).

[96] M. Ross, *Generalized lindemann melting law*, Phys. Rev. **184**, 233 (1969).

[97] U. R. Pedersen, L. Costigliola, N. P. Bailey, T. B. Schrøder, and J. C. Dyre, *Thermodynamics of freezing and melting*, Nat. Commun. **7**, 12386 (2016).

[98] D. Richard and T. Speck, *The role of shear in crystallization kinetics: From suppression to enhancement*, Sci. Rep. **5**, 14610 (2015).

[99] J. A. Barker and D. Henderson, *Perturbation theory and equation of state for fluids. ii. a successful theory of liquids*, J. Chem. Phys. **47**, 4714 (1967).

[100] H. C. Andersen, J. D. Weeks, and D. Chandler, *Relationship between the hard-sphere fluid and fluids with realistic repulsive forces*, Phys. Rev. A **4**, 1597 (1971).

[101] W. Götze, *Recent tests of the mode-coupling theory for glassy dynamics*, J. Phys.: Condens. Matter **11**, A1 (1999).

[102] G. Brambilla, D. El Masri, M. Pierno, L. Berthier, L. Cipelletti, G. Petekidis, and A. B. Schofield, *Probing the equilibrium dynamics of colloidal hard spheres above the mode-coupling glass transition*, Phys. Rev. Lett. **102**, 085703 (2009).

[103] R. Ni, M. A. C. Stuart, and M. Dijkstra, *Pushing the glass transition towards random close packing using self-propelled hard spheres*, Nat. Commun. **4**, 2704 (2013).

[104] C. P. Royall, A. Malins, A. J. Dunleavy, and R. Pinney, *Strong geometric frustration in model glassformers*, J. Non-Cryst. Solids **407**, 34 (2015).

[105] B. A. Lindquist, R. B. Jadrich, and T. M. Truskett, *Communication: From close-packed to topologically close-packed: Formation of Laves phases in moderately polydisperse hard-sphere mixtures*, J. Chem. Phys. **148**, 191101 (2018).

[106] P. K. Bommineni, N. R. Varela-Rosales, M. Klement, and M. Engel, *Complex crystals from size-disperse spheres*, Phys. Rev. Lett. **122**, 128005 (2019).

[107] G. Avvisati, T. Dasgupta, and M. Dijkstra, *Fabrication of colloidal Laves phases* via *hard tetramers and hard spheres: Bulk phase diagram and sedimentation behavior*, ACS Nano **11**, 7702 (2017).

[108] É. Ducrot, M. He, G.-R. Yi, and D. J. Pine, *Colloidal alloys with preassembled clusters and spheres*, Nat. Mater. **16**, 652 (2017).

[109] J. Mattsson, H. M. Wyss, A. Fernandez-Nieves, K. Miyazaki, Z. Hu, D. R. Reichman, and D. A. Weitz, *Soft colloids make strong glasses*, Nature **462**, 83 (2009).

[110] S. Sengupta, F. Vasconcelos, F. Affouard, and S. Sastry, *Dependence of the fragility of a glass former on the softness of interparticle interactions*, J. Chem. Phys. **135**, 194503 (2011).

[111] J. Krausser, K. H. Samwer, and A. Zaccone, *Interatomic repulsion softness directly controls the fragility of supercooled metallic melts*, Proc. Natl. Acad. Sci. U.S.A. **112**, 13762 (2015).

[112] C. E. Pueblo, M. Sun, and K. Kelton, *Strength of the repulsive part of the interatomic potential determines fragility in metallic liquids*, Nat. Mater. **16**, 792 (2017).

[113] H. J. Schöpe, G. Bryant, and W. van Megen, *Two-step crystallization kinetics in colloidal hard-sphere systems*, Phys. Rev. Lett. **96**, 175701 (2006).

[114] D. Erdemir, A. Y. Lee, and A. S. Myerson, *Nucleation of crystals from solution: Classical and two-step models*, Acc. Chem. Res. **42**, 621 (2009).

[115] T. Schilling, H. J. Schöpe, M. Oettel, G. Opletal, and I. Snook, *Precursor-mediated crystallization process in suspensions of hard spheres*, Phys. Rev. Lett. **105**, 025701 (2010).

[116] R. P. Sear, *The non-classical nucleation of crystals: Microscopic mechanisms and applications to molecular crystals, ice and calcium carbonate*, Int. Mater. Rev. **57**, 328 (2012).

[117] J. Russo and H. Tanaka, *The microscopic pathway to crystallization in supercooled liquids*, Sci. Rep. **2**, 505 (2012).

[118] S. Karthika, T. Radhakrishnan, and P. Kalaichelvi, *A review of classical and nonclassical nucleation theories*, Cryst. Growth Des. **16**, 6663 (2016).

[119] J. Russo and H. Tanaka, *Nonclassical pathways of crystallization in colloidal systems*, MRS Bull. **41**, 369 (2016).

[120] J. T. Berryman, M. Anwar, S. Dorosz, and T. Schilling, *The early crystal nucleation process in hard spheres shows synchronised ordering and densification*, J. Chem. Phys. **145**, 211901 (2016).

[121] M. Li, Y. Chen, H. Tanaka, and P. Tan, *Revealing roles of competing local structural orderings in crystallization of polymorphic systems*, Sci. Adv. **6**, eaaw8938 (2020).

[122] M. E. Leunissen, C. G. Christova, A.-P. Hynninen, C. P. Royall, A. I. Campbell, A. Imhof, M. Dijkstra, R. Van Roij, and A. Van Blaaderen, *Ionic colloidal crystals of oppositely charged particles*, Nature **437**, 235 (2005).

[123] E. V. Shevchenko, D. V. Talapin, N. A. Kotov, S. O'Brien, and C. B. Murray, *Structural diversity in binary nanoparticle superlattices*, Nature **439**, 55 (2006).

[124] M. Dijkstra, *Entropy-driven phase transitions in colloids: From spheres to anisotropic particles*, Adv. Chem. Phys **156**, 35 (2015).

[125] E. V. Shevchenko, D. V. Talapin, S. O'Brien, and C. B. Murray, *Polymorphism in ab13 nanoparticle superlattices: An example of semiconductor- metal metamaterials*, J. Am. Chem. Soc. **127**, 8741 (2005).

[126] M. Eldridge, P. Madden, and D. Frenkel, *Entropy-driven formation of a superlattice in a hard-sphere binary mixture*, Nature **365**, 35 (1993).

[127] L. Filion and M. Dijkstra, *Prediction of binary hard-sphere crystal structures*, Phys. Rev. E **79**, 046714 (2009).

[128] J. Sanders and M. Murray, *Ordered arrangements of spheres of two different sizes in opal*, Nature **275**, 201 (1978).

[129] S. Hachisu and S. Yoshimura, *Optical demonstration of crystalline superstructures in binary mixtures of latex globules*, Nature **283**, 188 (1980).

[130] P. Bartlett, R. Ottewill, and P. Pusey, *Freezing of binary mixtures of colloidal hard spheres*, J. Chem. Phys. **93**, 1299 (1990).

[131] P. Bartlett, R. Ottewill, and P. Pusey, *Superlattice formation in binary mixtures of hard-sphere colloids*, Phys. Rev. Lett. **68**, 3801 (1992).

[132] N. Hunt, R. Jardine, and P. Bartlett, *Superlattice formation in mixtures of hard-sphere colloids*, Phys. Rev. E **62**, 900 (2000).

[133] F. X. Redl, K.-S. Cho, C. B. Murray, and S. O'Brien, *Three-dimensional binary superlattices of magnetic nanocrystals and semiconductor quantum dots*, Nature **423**, 968 (2003).

[134] K. Overgaag, W. Evers, B. de Nijs, R. Koole, J. Meeldijk, and D. Vanmaekelbergh, *Binary superlattices of pbse and cdse nanocrystals*, J. Am. Chem. Soc. **130**, 7833 (2008).

[135] Z. Chen and S. O'Brien, *Structure direction of ii- vi semiconductor quantum dot binary nanoparticle superlattices by tuning radius ratio*, ACS Nano **2**, 1219 (2008).

[136] M. I. Bodnarchuk, M. V. Kovalenko, W. Heiss, and D. V. Talapin, *Energetic and entropic contributions to self-assembly of binary nanocrystal superlattices: Temperature as the structure-directing factor*, J. Am. Chem. Soc. **132**, 11967 (2010).

[137] Y. Kang et al., *Engineering catalytic contacts and thermal stability: Gold/iron oxide binary nanocrystal superlattices for co oxidation*, J. Am. Chem. Soc. **135**, 1499 (2013).

[138] M. I. Bodnarchuk, R. Erni, F. Krumeich, and M. V. Kovalenko, *Binary superlattices from colloidal nanocrystals and giant polyoxometalate clusters*, Nano Lett. **13**, 1699 (2013).

[139] Y. Sakamoto, Y. Kuroda, S. Toko, T. Ikeda, T. Matsui, and K. Kuroda, *Electron microscopy study of binary nanocolloidal crystals with ico-ab13 structure made of monodisperse silica nanoparticles*, J. Phys. Chem. C **118**, 15004 (2014).

[140] M. A. Boles and D. V. Talapin, *Many-mody effects in nanocrystal superlattices: Departure from sphere packing explains stability of binary phases*, J. Am. Chem. Soc. **137**, 4494 (2015).

[141] X. Ye, C. Zhu, P. Ercius, S. N. Raja, B. He, M. R. Jones, M. R. Hauwiller, Y. Liu, T. Xu, and A. P. Alivisatos, *Structural diversity in binary superlattices self-assembled from polymer-grafted nanocrystals*, Nat. Commun. **6**, 1 (2015).

[142] J. Wei, N. Schaeffer, and M.-P. Pileni, *Ligand exchange governs the crystal structures in binary nanocrystal superlattices*, J. Am. Chem. Soc. **137**, 14773 (2015).

[143] S. Punnathanam and P. Monson, *Crystal nucleation in binary hard sphere mixtures: a monte carlo simulation study*, J. Chem. Phys. **125**, 024508 (2006).

[144] S. R. Ganagalla and S. N. Punnathanam, *Free energy barriers for homogeneous crystal nucleation in a eutectic system of binary hard spheres*, J. Chem. Phys. **138**, 174503 (2013).

[145] P. K. Bommineni and S. N. Punnathanam, *Molecular simulation of homogeneous crystal nucleation of ab2 solid phase from a binary hard sphere mixture*, J. Chem. Phys. **147**, 064504 (2017).

[146] E. Sanz, C. Valeriani, D. Frenkel, and M. Dijkstra, *Evidence for out-of-equilibrium crystal nucleation in suspensions of oppositely charged colloids*, Phys. Rev. Lett. **99**, 055501 (2007).

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

[148] T. Dasgupta, G. M. Coli, and M. Dijkstra, *Tuning the glass transition: Enhanced crystallization of the laves phases in nearly hard spheres*, ACS Nano **14**, 3957 (2020).

[149] P. K. Bommineni, M. Klement, and M. Engel, *Spontaneous crystallization in systems of binary hard sphere colloids*, Phys. Rev. Lett. **124**, 218003 (2020).

[150] P. R. Ten Wolde, M. J. Ruiz-Montero, and D. Frenkel, *Numerical evidence for bcc ordering at the surface of a critical fcc nucleus*, Phys. Rev. Lett. **75**, 2714 (1995).

[151] P. R. Ten Wolde, M. J. Ruiz-Montero, and D. Frenkel, *Numerical calculation of the rate of crystal nucleation in a lennard-jones system at moderate undercooling*, J. Chem. Phys. **104**, 9932 (1996).

[152] J. A. van Meel, L. Filion, C. Valeriani, and D. Frenkel, *A parameter-free, solid-angle based, nearest-neighbor algorithm*, J. Chem. Phys. **136**, 234107 (2012).

[153] E. Boattini, M. Ram, F. Smallenburg, and L. Filion, *Neural-network-based order parameters for classification of binary hard-sphere crystal structures*, Mol. Phys. **116**, 3066 (2018).

[154] J. R. Espinosa, C. Vega, C. Valeriani, and E. Sanz, *The crystal-fluid interfacial free energy and nucleation rate of nacl from different simulation methods*, J. Chem. Phys. **142**, 194709 (2015).

[155] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, New York, NY, 2006.

[156] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning representations by back-propagating errors*, Nature **323**, 533 (1986).

[157] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, *On the importance of initialization and momentum in deep learning*, in *Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013*, pages 1139–1147.

[158] G. M. Torrie and J. P. Valleau, *Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling*, J. Comput. Phys. **23**, 187 (1977).

[159] A. Warmflash, P. Bhimalapuram, and A. R. Dinner, *Umbrella sampling for nonequilibrium processes*, J. Chem. Phys. **127**, 114109 (2007).

[160] L. Filion, M. Hermes, R. Ni, and M. Dijkstra, *Crystal nucleation of hard spheres using molecular dynamics, umbrella sampling, and forward flux sampling: A comparison of simulation techniques*, J. Chem. Phys. **133**, 244115 (2010).

[161] A. Cuetos and M. Dijkstra, *Kinetic pathways for the isotropic-nematic phase transition in a system of colloidal hard rods: A simulation study*, Phys. Rev. Lett. **98**, 095701 (2007).

[162] R. J. Allen, P. B. Warren, and P. R. Ten Wolde, *Sampling rare switching events in biochemical networks*, Phys. Rev. Lett. **94**, 018104 (2005).

[163] A. Laio and M. Parrinello, *Escaping free-energy minima*, Proc. Natl. Acad. Sci. U.S.A. **99**, 12562 (2002).

[164] F. Trudu, D. Donadio, and M. Parrinello, *Freezing of a lennard-jones fluid: From nucleation to spinodal regime*, Phys. Rev. Lett. **97**, 105701 (2006).

[165] D. Moroni, P. R. Ten Wolde, and P. G. Bolhuis, *Interplay between structure and size in a critical crystal nucleus*, Phys. Rev. Lett. **94**, 235703 (2005).

[166] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, *Transition path sampling: Throwing ropes over rough mountain passes, in the dark*, Annu. Rev. Phys. Chem. **53**, 291 (2002).

[167] W. Lechner, C. Dellago, and P. G. Bolhuis, *Role of the prestructured surface cloud in crystal nucleation*, Phys. Rev. Lett. **106**, 085701 (2011).

[168] J. A. Anderson, C. D. Lorenz, and A. Travesset, *General purpose molecular dynamics simulations fully implemented on graphics processing units*, J. Comput. Phys. **227**, 5342 (2008).

[169] J. Glaser, T. Dac Nguyen, J. A. Anderson, P. Lui, F. Spiga, J. A. Millan, D. C. Morse, and S. C. Glotzer, *Strong scaling of general-purpose molecular dynamics simulations on gpus*, Comput. Phys. Commun. **192**, 97 (2015).

[170] A. S. Keys, C. R. Iacovella, and S. C. Glotzer, *Characterizing complex particle morphologies through shape matching: Descriptors, applications, and algorithms*, J. Comput. Phys. **230**, 6438 (2011).

[171] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, volume 1, Springer series in statistics New York, 2001.

[172] M. Gevrey, I. Dimopoulos, and S. Lek, *Review and comparison of methods to study the contribution of variables in artificial neural network models*, Ecol. Modell. **160**, 249 (2003).

[173] J. D. Olden, M. K. Joy, and R. G. Death, *An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data*, Ecol. Modell. **178**, 389 (2004).

[174] J. Yao, N. Teng, H.-L. Poh, and C. L. Tan, *Forecasting and analysis of marketing data using neural networks*, J. Inf. Sci. Eng. **14**, 843 (1998).

[175] M. Scardi and L. W. Harding Jr, *Developing an empirical model of phytoplankton primary production: A neural network case study*, Ecol. Modell. **120**, 213 (1999).

[176] R. P. Sear, *Nucleation: theory and applications to protein solutions and colloidal suspensions*, J. Condens. Matter Phys. **19**, 033101 (2007).

[177] T. Palberg, *Crystallization kinetics of colloidal model suspensions: recent achievements and new perspectives*, J. Condens. Matter Phys. **26**, 333101 (2014).

[178] T. Ohm, M. Kirca, J. Bohl, H. Scharnagl, W. Groβ, and W. März, *Apolipoprotein e polymorphism influences not only cerebral senile plaque load but also alzheimer-type neurofibrillary tangle formation*, Neuroscience **66**, 583 (1995).

[179] J. Bauer, S. Spanton, R. Henry, J. Quick, W. Dziki, W. Porter, and J. Morris, *Ritonavir: an extraordinary example of conformational polymorphism*, Pharm. Res. **18**, 859 (2001).

[180] A. E. Van Driessche, N. Van Gerven, P. H. Bomans, R. R. Joosten, H. Friedrich, D. Gil-Carton, N. A. Sommerdijk, and M. Sleutel, *Molecular nucleation mechanisms and control strategies for crystal polymorph selection*, Nature **556**, 89 (2018).

[181] J. Xing, L. Schweighauser, S. Okada, K. Harano, and E. Nakamura, *Atomistic structures and dynamics of prenucleation clusters in mof-2 and mof-5 syntheses*, Nat. Commun. **10**, 1 (2019).

[182] J. Zhou et al., *Observing crystal nucleation in four dimensions using atomic electron tomography*, Nature **570**, 500 (2019).

[183] L. Houben, H. Weissman, S. G. Wolf, and B. Rybtchinski, *A mechanism of ferritin crystallization revealed by cryo-stem tomography*, Nature **579**, 540 (2020).

[184] T. Nakamuro, M. Sakakibara, H. Nada, K. Harano, and E. Nakamura, *Capturing the moment of emergence of crystal nucleus from disorder*, J. Am. Chem. Soc. **143**, 1763 (2021).

[185] S. Jeon et al., *Reversible disorder-order transitions in atomic crystal nucleation*, Science **371**, 498 (2021).

[186] P. Pusey, W. Van Megen, P. Bartlett, B. Ackerson, J. Rarity, and S. Underwood, *Structure of crystals of hard colloidal spheres*, Phys. Rev. Lett. **63**, 2753 (1989).

[187] P. G. Bolhuis, D. Frenkel, S.-C. Mau, and D. A. Huse, *Entropy difference between crystal phases*, Nature **388**, 235 (1997).

[188] E. G. Noya and N. G. Almarza, *Entropy of hard spheres in the close-packing limit*, Mol. Phys. **113**, 1061 (2015).

[189] C. Dux and H. Versmold, *Light diffraction from shear ordered colloidal dispersions*, Phys. Rev. Lett. **78**, 1811 (1997).

[190] Z. Cheng, J. Zhu, W. B. Russel, W. V. Meyer, and P. M. Chaikin, *Colloidal hard-sphere crystallization kinetics in microgravity and normal gravity*, Appl. Opt. **40**, 4146 (2001).

[191] V. Luchnikov, A. Gervois, P. Richard, L. Oger, and J. Troadec, *Crystallization of dense hard sphere packings: Competition of hcp and fcc close order*, J. Mol. Liq. **96**, 185 (2002).

[192] B. O'malley and I. Snook, *Crystal nucleation in the hard sphere system*, Phys. Rev. Lett. **90**, 085702 (2003).

[193] L. Filion, M. Hermes, R. Ni, and M. Dijkstra, *Crystal nucleation of hard spheres using molecular dynamics, umbrella sampling, and forward flux sampling: A comparison of simulation techniques*, J. Chem. Phys. **133**, 244115 (2010).

[194] F. Leoni and J. Russo, *Non-classical nucleation pathways in stacking-disordered crystals*, arXiv preprint arXiv:2105.05506 (2021).

[195] J. A. Anderson, J. Glaser, and S. C. Glotzer, *HOOMD-blue: A Python package for high-performance molecular dynamics and hard particle Monte Carlo simulations*, Comput. Mater. Sci. **173**, 109363 (2020).

[196] G. Torrie and J. Valleau, *Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling*, J. Comput. Phys. **23**, 187 (1977).

[197] S. Auer and D. Frenkel, *Prediction of absolute crystal-nucleation rate in hard-sphere colloids*, Nature **409**, 1020 (2001).

[198] J. A. Anderson, M. Eric Irrgang, and S. C. Glotzer, *Scalable Metropolis Monte Carlo for simulation of hard shapes*, Comput. Phys. Commun. **204**, 21 (2016).

[199] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, *The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method*, J. Comput. Chem. **13**, 1011 (1992).

[200] A. Grossfield, *WHAM: the weighted histogram analysis method.*

[201] S. Auer and D. Frenkel, *Suppression of crystal nucleation in polydisperse colloids due to increase of the surface free energy*, Nature **413**, 711 (2001).

[202] E. Sanz, C. Valeriani, T. Vissers, A. Fortini, M. E. Leunissen, A. Van Blaaderen, D. Frenkel, and M. Dijkstra, *Out-of-equilibrium processes in suspensions of oppositely charged colloids: liquid-to-crystal nucleation and gel formation*, J. Condens. Matter Phys. **20**, 494247 (2008).

[203] J. D. Honeycutt and H. C. Andersen, *Molecular dynamics study of melting and freezing of small lennard-jones clusters*, J. Phys. Chem. **91**, 4950 (1987).

[204] C. Desgranges and J. Delhommelle, *Molecular mechanism for the cross-nucleation between polymorphs*, J. Am. Chem. Soc. **128**, 10368 (2006).

[205] C. Desgranges and J. Delhommelle, *Molecular simulation of the crystallization of aluminum from the supercooled liquid*, J. Chem. Phys. **127**, 144509 (2007).

[206] T. Kawasaki and H. Tanaka, *Structural origin of dynamic heterogeneity in three-dimensional colloidal glass formers and its link to crystal nucleation*, J. Condens. Matter Phys. **22**, 232102 (2010).

[207] S. Alexander and J. McTague, *Should all crystals be bcc? landau theory of solidification and crystal nucleation*, Phys. Rev. Lett. **41**, 702 (1978).

[208] X. Ji, Z. Sun, W. Ouyang, and S. Xu, *Crystal nucleation and metastable bcc phase in charged colloids: A molecular dynamics study*, J. Chem. Phys. **148**, 174904 (2018).

[209] W. Ouyang, B. Sun, Z. Sun, and S. Xu, *Entire crystallization process of lennard-jones liquids: A large-scale molecular dynamics study*, J. Chem. Phys. **152**, 054903 (2020).

[210] G. M. Coli and M. Dijkstra, *An artificial neural network reveals the nucleation mechanism of a binary colloidal ab13 crystal*, ACS Nano **15**, 4335 (2021).

[211] R. van Damme, G. M. Coli, R. van Roij, and M. Dijkstra, *Classifying crystals of rounded tetrahedra and determining their order parameters using dimensionality reduction*, ACS Nano **14**, 15144 (2020).

[212] C. S. Adorf, T. C. Moore, Y. J. Melle, and S. C. Glotzer, *Analysis of self-assembly pathways with unsupervised machine learning algorithms*, The J. Phys. Chem. B **124**, 69 (2019).

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

[214] E. Boattini, S. Marín-Aguilar, S. Mitra, G. Foffi, F. Smallenburg, and L. Filion, *Autonomously revealing hidden local structures in supercooled liquids*, Nature Commun. **11**, 1 (2020).

[215] H. Hotelling, *Analysis of a complex of statistical variables into principal components.*, J Educ. Psychol. **24**, 417 (1933).

[216] K. Pearson, *On lines and planes of closest fit to systems of points in space*, Lond. Edinb. Dubl. Phil. Mag. **2**, 559 (1901).

[217] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the em algorithm*, J. R. Stat. Soc. Series B Stat. Methodol. **39**, 1 (1977).

[218] G. Schwarz, *Estimating the dimension of a model*, Ann. Stat. , 461 (1978).

[219] J.-P. Baudry, A. E. Raftery, G. Celeux, K. Lo, and R. Gottardo, *Combining mixture components for clustering*, J Comput. Graph. Stat. **19**, 332 (2010).

[220] M. C. Rechtsman, F. H. Stillinger, and S. Torquato, *Optimized interactions for targeted self-assembly: application to a honeycomb lattice*, Phys. Rev. Lett. **228301**, 95 (2005).

[221] M. C. Rechtsman, H. C. Jeong, P. M. Chaikin, S. Torquato, and P. J. Steinhardt, *Optimized structures for photonic quasicrystals*, Phys. Rev. Lett. **101**, 073902 (2008).

[222] M. C. Rechtsman, F. H. Stillinger, and S. Torquato, *Negative poisson's ratio materials via isotropic interactions*, Phys. Rev. Lett. **101**, 085501 (2008).

[223] S. Torquato, *Inverse optimization techniques for targeted self-assembly*, Soft Matter **5**, 1157 (2009).

[224] A. Jain, J. A. Bollinger, and T. M. Truskett, *Inverse methods for material design*, AIChE Journal **8**, 2732 (2014).

[225] J. M. Jaeger, *Celebrating soft matter's 10th anniversary: Toward jamming by design*, Soft Matter **11**, 12 (2015).

[226] A. F. Hannon, K. W. Gotrik, C. A. Ross, and A. Alexander-Katz, *Inverse design of topographical templates for directed self-assembly of block copolymers*, ACS Macro Lett. **2**, 251 (2013).

[227] J.-B. Chang, H. K. Choi, A. F. Hannon, A. Alexander-Katz, C. A. Ross, and K. K. Berggren, *Design rules for self-assembled block copolymer patterns using tiled templates*, Nat. Commun. **5**, 1 (2014).

[228] M. C. Rechtsman, F. H. Stillinger, and S. Torquato, *Designed interaction potentials via inverse methods for self-assembly*, Phys. Rev. E **73**, 011406 (2006).

[229] A. F. Hannon, Y. Ding, W. Bai, C. A. Ross, and A. Alexander-Karz, *Optimizing topographical templates for directed self-assembly of block copolymers via inverse design simulations*, Nano Lett. **14**, 318 (2014).

[230] S. Hormoz and M. P. Brenner, *Design principles for self-assembly with short-range interactions*, Proc. Natl. Acad. Sci. U.S.A. **108**, 5193 (2011).

[231] B. A. Lindquist, R. B. Jadrich, and T. M. Truskett, *Communication: Inverse design for self-assembly via on-the-fly optimization*, J. Chem. Phys. **145**, 111101 (2016).

[232] R. B. Jadrich, B. A. Lindquist, and T. M. Truskett, *Probabilistic inverse design for self-assembling materials*, J. Chem. Phys. **146**, 184103 (2017).

[233] C. S. Adorf, J. Antonaglia, J. Dshemuchadse, and S. C. Glotzer, *Inverse design of simple pair potentials for the self-assembly of complex structures*, J. Chem. Phys. **149**, 204102 (2018).

[234] M. Z. Miskin, G. Khaira, J. J. de Pablo, and H. M. Jaeger, *Turning statistical physics models into materials design engines*, Proc. Natl. Acad. Sci. U.S.A. **113**, 34 (2016).

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

[236] W. Lechner and C. Dellago, *Accurate determination of crystal structures based on averaged local bond order parameters*, J. Chem. Phys. **129**, 114707 (2008).

[237] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Optimization by simulated annealing*, Science **220**, 671 (1983).

[238] L. Ingber, *Adaptive simulated annealing (asa)*, J. Global Optim., Caltech Alumni Association, Pasadena, CA (1993).

[239] L. Ingber, *Adaptive simulated annealing (asa): Lessons learned*, arXiv preprint cs/0001018 (2000).

[240] M. A. M. De Oca, T. Stützle, M. Birattari, and M. Dorigo, *A comparison of particle swarm optimization algorithms based on run-length distributions*, in *International Workshop on Ant Colony Optimization and Swarm Intelligence*, pages 1–12, Springer, 2006.

[241] M. Mitchell, *An introduction to genetic algorithms*, MIT press, 1998.

[242] N. Hansen and A. Ostermeier, *Completely derandomized self-adaptation in evolution strategies*, Evol. Comput. **9**, 159 (2001).

[243] N. Hansen, S. D. Müller, and P. Koumoutsakos, *Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es)*, Evol. Comput. **11**, 1 (2003).

[244] N. Hansen, *The cma evolution strategy: a comparing review*, in *Towards a new evolutionary computation*, pages 75–102, Springer, 2006.

[245] E. J. Meijer and F. El Azhar, F.El Azhar, *Novel procedure to determine coexistence lines by computer simulation. application to hard-core yukawa model for charge-stabilized colloids*, J. Chem. Phys. **106**, 4678 (1997).

[246] F. El Azhar, M. Baus, J. P. Ryckaert, and E. J. Meijer, *Line of triple points for the hard-core yukawa model: A computer simulation study*, J. Chem. Phys. **112**, 5121 (2000).

[247] A. P. Hynninen and M. Dijkstra, *Phase diagrams of hard-core repulsive yukawa particles*, Phys. Rev. E **68**, 021407 (2003).

[248] J. A. van Meel, L. Filion, C. Valeriani, and D. Frenkel, *A parameter-free, solid-angle based, nearest-neighbor algorithm*, J. Chem. Phys. **136**, 234107 (2012).

[249] R. Kumar, G. M. Coli, M. Dijkstra, and S. Sastry, *Inverse design of charged colloidal particle interactions for self assembly into specified crystal structures*, J. Chem. Phys. **151**, 084109 (2019).

[250] N. Hansen, *Towards a new evolutionary computation*, Stud. Fuzziness Soft Comput. **192**, 75 (2006).

[251] K. Barkan, H. Diamant, and R. Lifshitz, *Stability of quasicrystals composed of soft isotropic particles*, Phys. Rev. B **83**, 172201 (2011).

[252] T. Dotera, T. Oshiro, and P. Ziherl, *Mosaic two-lengthscale quasicrystals*, Nature **506**, 208 (2014).

[253] H. Pattabhiraman, A. P. Gantapara, and M. Dijkstra, *On the stability of a quasicrystal and its crystalline approximant in a system of hard disks with a soft corona*, J. Chem. Phys. **143**, 164905 (2015).

[254] H. Pattabhiraman and M. Dijkstra, *Phase behaviour of quasicrystal forming systems of core-corona particles*, J. Chem. Phys. **146**, 114901 (2017).

[255] H. Pattabhiraman and M. Dijkstra, *On the formation of stripe, sigma, and honeycomb phases in a core–corona system*, Soft Matter **13**, 4418 (2017).

[256] G. Campos-Villalobos, M. Dijkstra, and A. Patti, *Nonconventional phases of colloidal nanorods with a soft corona*, Phys. Rev. Lett. **126**, 158001 (2021).

[257] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Inc., New York, NY, USA, 1995.

[258] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[259] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980 (2014).

[260] A. Paszke et al., *Pytorch: An imperative style, high-performance deep learning library*, in *Advances in neural information processing systems*, pages 8026–8037, 2019.

[261] N. P. Kryuchkov, S. O. Yurchenko, Y. D. Fomin, E. N. Tsiok, and V. N. Ryzhov, *Complex crystalline structures in a two-dimensional core-softened system*, Soft Matter **14**, 2152 (2018).

[262] L. A. Padilla and A. Ramírez-Hernández, *Phase behavior of a two-dimensional core-softened system: new physical insights*, J. Phys. Condens. Matter **32**, 275103 (2020).

[263] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT press, 2016.

[264] N. Kern and D. Frenkel, *Fluid–fluid coexistence in colloidal systems with short-ranged strongly directional attraction*, J. Chem. Phys. **118**, 9882 (2003).

[265] I. Zubieta, M. V. Del Saz, P. Llombart, C. Vega, and E. G. Noya, *Nucleation of pseudo hard-spheres and dumbbells at moderate metastability: appearance of a15 frank–kasper phase at intermediate elongations*, Phys. Chem. Chem. Phys. **21**, 1656 (2019).

[266] A. J. Mueller, A. P. Lindsay, A. Jayaraman, T. P. Lodge, M. K. Mahanthappa, and F. S. Bates, *Emergence of a c15 laves phase in diblock polymer/homopolymer blends*, ACS Macro Lett. **9**, 576 (2020).

[267] K. Schätzel and B. J. Ackerson, *Density fluctuations during crystallization of colloids*, Phys. Rev. E **48**, 3766 (1993).

[268] J. Harland and W. Van Megen, *Crystallization kinetics of suspensions of hard colloidal spheres*, Phys. Rev. E **55**, 3054 (1997).

Summary

This thesis has undisputed protagonists that are discussed from the first to the last page: colloidal systems, *i.e.* colloidal particles dispersed in a solvent. What distinguishes colloids from other particles such as atoms and molecules? Although a very common definition of colloids, and also the most intuitive one, concerns their typical size range, which goes from tens of nanometres to micrometers, the key to understand what we are dealing with lies in their dynamics. Colloids, in fact, due to the continuous interaction with small solvent particles, show a random dynamics that goes by the name of Brownian motion. The most important feature of Brownian motion is that it allows the colloidal system to efficiently explore a very large number of configurations of positions and velocities, *i.e.* the phase space. This way, colloids are able to spatially organise themselves spontaneously, through a process called self-assembly, and to form the thermodynamically stable phase.

In a very large number of cases, the process of self-assembly occurs through the phenomenon of nucleation. This process begins with the random formation of a nucleus of the so-called child phase in a small region of the system. This phenomenon has a cost in terms of free energy, due to the creation of an interface between the nucleus and the surrounding parent phase. If the nucleus of the child phase is smaller than a certain threshold, then it will have even greater difficulty in growing, and will therefore melt back. Conversely, if the nucleus is larger than a certain threshold, the system will gain a thermodynamic advantage from the indefinite growth of the new phase, which therefore becomes the macroscopic phase of the system. Nucleation in colloidal systems is the major theme whose study pervades every Chapter of this thesis. Why are we so interested in it? There are mainly two reasons. The first, more intuitive, lies in the extraordinary properties of the products of nucleation. Through this phenomenon it is possible to create new materials, such as photonic crystals, which are considered to be key in the development of new generation technologies. The second reason of interest is more subtle, and concerns the analogy that colloidal systems have with atomic and molecular systems. In fact, nucleation phenomena can also be observed in the latter systems during phase transitions. However, these occur at too small length scales and too high speeds to be studied in detail. Conversely, colloids are large enough and slow enough to be observed by conventional optical techniques such as light microscopes, or to be modelled in numerical simulations without the inclusion of quantum effects, which typically slow down simulations by order of magnitudes.

We begin our investigation by studying the nucleation of the colloidal analogue of three binary crystal structures called Laves Phases (LPs). In particular, the three crystals under investigation are the $MgCu_2$, $MgZn_2$, and $MgNi_2$ crystals. Through free-energy calculations, it was found in 2007 that LPs can be self-assembled from a binary mixture of hard spheres with a specific size ratio range. In other words, under certain thermodynamic conditions, they are the stable phase for the described particle system. Despite this, LPs have never been observed to spontaneously crystallise from such a fluid mixture in simulations nor in experiments of micron-sized colloids. To understand why, we study LP nucleation in Chapter 2, using various particle systems in which we tune the softness of the particles. We use the seeding technique to find all the essential variables to describe nucleation, and to find the range of supersaturation for observing spontaneous nucleation events. Although thermodynamics is found to be invariant with respect to the softness of the particles, these events occur only with the softest particles. Analysing both the higher order spatial correlations in the fluid, and the dynamics of such

systems, we finally find that harder systems correspond to a higher concentration of fivefold symmetry clusters, known to be inhibitors of crystal nucleation. These clusters in turn cause a critical slowdown in the dynamics, which inhibits crystal nucleation. This study not only addresses the difficulties in finding spontaneous nucleation events of LPs with systems of hard spheres, but provides detailed indications of the interaction required to form these structures in simulations and, most importantly, in experiments.

We continue our journey studying the nucleation phenomenon of another binary crystal, and undoubtedly the most curious one, which is the colloidal analogue of $NaZn_{13}$ (to which we refer as AB_{13} crystal). This crystal was experimentally observed for the first time by Sanders and co-workers in 1987 in natural gem opals. Many numerical studies followed, starting from Eldridge, Madden and Frenkel in 1993, which demonstrated the stability of the AB_{13} crystal, and provided a bulk phase diagram of this highly exotic crystal. In spite of its proven thermodynamic stability, just like the LPs, the AB_{13} has never been observed to spontaneously nucleate in numerical simulations. Thus, in Chapter 3, we study crystal nucleation of the AB_{13} phase in a binary mixture of colloidal hard spheres. To follow the nucleation process, we employ an artificial neural network to identify the AB_{13} phase from the binary fluid phase and competing crystal structures at a single-particle level with unprecedented accuracy. We show that standard techniques fail in identifying the different phases and that machine learning is indispensable in achieving this task. Moreover, investigating in more detail how the neural network makes decisions enables us to learn from the machine learning algorithm which features and symmetries are the most important ones for the classification. Employing the trained neural network as an order parameter, we investigate the nucleation process of the AB_{13} crystal, clarifying the role of local structures in the supersaturated fluid, and shedding light on the relationship between large and small spheres during the very first stages of the cryatallisation event. First, we use the seeding approach, which enables us to calculate all relevant physical observables regarding nucleation, and permits to draw a quantitative comparison with nucleation of other hard-sphere crystals, namely the face-centred-cubic structure and the colloidal LPs. Secondly, we conduct brute force Molecular Dynamics (MD) simulations, which constitute the first events of AB_{13} spontaneous nucleation reported in the literature. Finally, we use a different classification algorithm based on the recognition of local motifs, and demonstrate that binary nucleation of the AB_{13} phase proceeds *via* a co-assembly process of large spheres and icosahedral small-sphere clusters attaching to the nucleus, and is well-described by a classical nucleation process.

In Chapter 4, we turn our attention to the simplest and most studied system in soft matter: the hard-sphere system. Despite numerous experimental and theoretical studies, many aspects of the nucleation process such as the polymorph selection mechanism in the early stages of nucleation are far from being understood. A clear example is the giant discrepancy between nucleation rates obtained from simulations or measured in experiments, which differ by several orders of magnitude. In this Chapter, we focus on the excess of face-centred-cubic (fcc)-like particles with respect to hexagonal-close-packed (hcp)-like particles in a crystal nucleus of hard spheres as observed in simulations and experiments. To explain this phenomenon, we simulate a large number of spontaneous nucleation events using nearly hard spheres. In addition to following the formation and growth of a crystal nucleus, we also focus on the behaviour of specific topological clusters, which are characterised by a large concentration in the fluid phase, which is thus shown to be structurally heterogeneous. We show that, in the metastable fluid phase, fivefold symmetry clusters - pentagonal bipyramids - known to be inhibitors of crystal nucleation, transform into a different cluster - Siamese dodecahedra. These clusters, because of their geometric arrangement, form a bridge between the fivefold symmetric fluid and the fcc crystal, thus lowering its interfacial free energy with respect to the hcp crystal, and shedding light on the polymorph selection mechanism. This mechanism is analysed in simulations and experiments, which show analogous behaviour.

In Chapter 5 we keep our focus on the same system as in Chapter 4, but with the task of finding a classification algorithm for the different crystal polymorphs which intrinsically contains a large amount of information about the local environment of each particle. In order to achieve this, we resort to the use of a technique belonging to the family of unsupervised learning algorithms: Principal Component Analysis (PCA). In particular, for each of the phases of interest – fcc, hcp, and body-centred-crystal (bcc) crystals, and the monodisperse fluid phase – we simulate the equilibrium bulk structures and compute a large number of features to describe the local neighbourhood of each particle. These are chosen so that different features are sensitive to different symmetries shown by the local environment. Using PCA on the entire data set transforms the parameter space, autonomously finding new order parameters along which the projection of the original data is maximised. In this way, even using a small number of new order parameters, it is possible to maintain a very high level of information about each particle. PCA offers several advantages, all related to the fact that the new order parameters are a linear combination of the original features. In particular, PCA has a high speed of execution, and provides easily interpretable results, from which it is possible to gain new insights on which features are useful to distinguish the different polymorphs. Once PCA is used, it is possible to project data from nucleation events onto the new space and construct classifiers directly on the new space. In Chapter 5 we define two of them. The first is based on simple thresholds on the new order parameters, while the second involves the use of an additional unsupervised learning technique, Gaussian Mixture Model (GMM). In particular, by clustering the data with GMM it is possible to label particles not only as belonging to the known phases, but also as belonging to the interface between two phases. This study shows how dimensionality reduction and clustering algorithms can play a decisive role in classification tasks. In general, we expect that unsupervised learning techniques will play an increasingly leading role in the identification of key features of physical phenomena, and not only of nucleation. On the other hand, in the field of soft matter, we expect that other phenomena of self-assembly or even of the glass transition can be better understood and described, starting from these kinds of approaches, for systems coming both from simulations and experiments.

Unlike what we have just discussed, in many cases one is not interested in the nucleation mechanism itself, but only in the nucleation products. It is for this reason that, in Chapters 6, 7, and 8, we focus our attention on Inverse Design Methods (IDMs). IDMs are considered to be the holy grail of materials design *via* colloid self-assembly. The question we try to answer is: given a desired material structure (e.g. s specific crystalline phase), how can we tune the colloidal building blocks to obtain the desired self-assembly product? On the synthesis side, chemists have made incredible strides to the point where colloidal building blocks can nearly be made with interactions defined on demand. However, while state-of-the-art computational and theoretical statistical mechanics methods are extremely capable of solving the so-called “forward” design problem, where we predict the structures formed by an a priori defined set of colloidal building blocks, a robust, versatile algorithm for solving the inverse problem remains a significant challenge. The lack of such an IDM forms a significant obstacle for the full exploitation of colloidal self-assembly in the development of the next generation of materials.

As a first approach, in Chapter 6, we reverse-engineer a bcc crystal that, in a repulsive Yukawa particle system, is only stable in a narrow region of the phase diagram. This task allows us to test two algorithms of different nature – the first based on the statistical fluctuations of the system under investigation and the second derived from classical optimisation techniques – and to evaluate their similarities and differences. Building on the knowledge acquired in Chapter 6, in Chapter 7 we introduce a general IDM to efficiently reverse-engineer desired crystals, quasicrystals, and liquid crystals by targeting their diffraction patterns. Our algorithm relies on the synergetic use of an evolutionary strategy for parameter optimisation, and a convolutional neural network that classifies phases from their diffraction patterns as an order parameter. The robustness and versatility of this new order parameter allowed us to successfully reverse-engineer various crystal, quasicrystal and liquid crystal structures in two- and three-dimensional model systems, thus providing a new way forward for the inverse design of experimentally feasible colloidal interactions, specifically optimised to stabilise a desired structure. Finally, in Chapter 8, we build on what we learned about LP nucleation in Chapter 2, and investigate the possibility of finding a self-assembly route to obtain a crystal with a photonic bandgap through the IDM described in Chapter 7.

Being an extremely broad topic, obviously many questions regarding nucleation still remain open. Most of them are obviously challenging questions, but also fascinating ones. We hope that some of the conclusions drawn by the work presented in this thesis, as well as some of the analysis and methods we developed, can help future researchers to answer some of these questions, in order to broaden the knowledge we already have about this exciting field of research.

Samenvatting

Dit proefschrift heeft onbetwiste hoofdrolspelers die van de eerste tot de laatste bladzijde worden besproken: colloïdale systemen, d.w.z. colloïdale deeltjes gedispergeerd in een oplosmiddel. Wat onderscheidt colloïden van andere deeltjes zoals atomen en moleculen? Hoewel een zeer gebruikelijke definitie van colloïden, en ook de meest intuïtieve, betrekking heeft op hun typische grootte, die gaat van tientallen nanometers tot micrometers, ligt de sleutel om te begrijpen waar we mee te maken hebben in hun dynamica. Colloïden vertonen namelijk, als gevolg van de voortdurende interactie met kleine deeltjes van het oplosmiddel, schijnbaar willekeurige dynamica die Brownse beweging wordt genoemd. Het belangrijkste kenmerk van de Brownse beweging is dat het colloïdale systeem door zijn willekeurigheid op efficiënte wijze een zeer groot aantal configuraties van posities en snelheden, de faseruimte, kan verkennen. Op deze manier kunnen colloïden zich spontaan ruimtelijk organiseren, via een proces dat zelforganisatie wordt genoemd, en zo de thermodynamisch stabiele fase vormen.

In een zeer groot aantal gevallen verloopt het proces van zelforganisatie via het verschijnsel nucleatie. Dit proces begint met de willekeurige vorming van een kern van de zogenaamde kindfase in een klein gebied van het systeem. Dit verschijnsel heeft een prijs in termen van vrije energie, door het ontstaan van een grensvlak tussen de kern en de omringende moederfase. Als de kern van de kindfase kleiner is dan een bepaalde drempelwaarde, zal deze nog moeilijker kunnen groeien, en dus weer gaan smelten. Omgekeerd, als de kern groter is dan een bepaalde drempelwaarde, zal het systeem thermodynamisch voordeel behalen door onbeperkte groei van de nieuwe fase, die daardoor de macroscopische fase van het systeem wordt. Nucleatie in colloïdale systemen is het hoofdthema dat in elk hoofdstuk van dit proefschrift een rol speelt. Waarom zijn we er zo in geïnteresseerd? Er zijn hoofdzakelijk twee redenen. De eerste, meer intuïtieve, reden ligt in de buitengewone eigenschappen van de producten van nucleatie. Door dit verschijnsel is het mogelijk nieuwe materialen te creëren, zoals fotonische kristallen, die worden gezien als de sleutel tot de ontwikkeling van nieuwe technologieën. De tweede reden is subtieler en betreft de analogie die colloïdale systemen hebben met atomaire en moleculaire systemen. In feite kunnen nucleatieverschijnselen ook in de laatstgenoemde systemen worden waargenomen tijdens faseovergangen. Deze doen zich echter voor op te kleine lengteschalen en bij te hoge snelheden om in detail te kunnen worden bestudeerd. Omgekeerd zijn colloïden groot en langzaam genoeg om te worden waargenomen met conventionele optische technieken zoals lichtmicroscopen, of om te worden gemodelleerd in numerieke simulaties waarbij geen rekening wordt gehouden met kwantumeffecten, die simulaties gewoonlijk met ordes van grootte vertragen.

We beginnen ons onderzoek met de nucleatie van de colloïdale analoog van drie binaire kristalstructuren, de zogenaamde Laves-fasen (LP's). De drie onderzochte kristallen zijn in het bijzonder $MgCu_2$, $MgZn_2$ en $MgNi_2$. Door middel van vrije-energieberekeningen werd in 2007 vastgesteld dat LP's zich kunnen vormen uit een binair mengsel van harde bollen met specifieke grootteverhoudingen. Met andere woorden, onder bepaalde thermodynamische condities zijn ze thermodynamisch stabiel voor dit desbetreffende systeem. Desondanks is nog nooit waargenomen dat LPs spontaan kristalliseren uit zo'n vloeistofmengsel, noch in simulaties, noch in experimenten met harde bolletjes van micron-grootte. Om te begrijpen waarom, bestuderen we in Hoofdstuk 2 LP nucleatie in een variëteit systemen waarin we de zachtheid van de deeltjes variëren. We gebruiken de seeding techniek om alle essentiële variabelen te vinden om nucleatie te beschrijven, en om uit te vinden welk bereik van oververzadiging geschikt is om spontane nucleatie waar te nemen. Hoewel de zachtheid van de deeltjes de thermodynamica niet blijkt te veranderen, treedt nucleatie enkel op bij de zachtste deeltjes. Door analyse van zowel de hogere orde ruimtelijke correlaties in de vloeistof, als de dynamica van dergelijke systemen, vinden we uiteindelijk dat hardere systemen overeenkomen met een hogere concentratie van vijfvoudig symmetrische clusters, waarvan bekend is dat ze de kristallisatie belemmeren. Deze clusters veroorzaken op hun beurt een kritieke vertraging in de dynamica, die de nucleatie verhindert. Dit onderzoek behandelt niet alleen de moeilijkheden bij het vinden van spontane nucleatie van LPs met systemen van harde bollen, maar geeft ook gedetailleerde aanwijzingen over de interacties die nodig zijn om deze structuren te reproduceren in simulaties en, het belangrijkst, in experimenten.

We zetten onze reis voort met het bestuderen van de nucleatie van een ander, ongetwijfeld het meest merkwaardige, binair kristal, namelijk het colloïdale analoog van $\mathrm{NaZn_{13}}$ (waarnaar we verwijzen als $\mathrm{AB_{13}}$). Dit kristal werd voor het eerst experimenteel waargenomen door Sanders en zijn medewerkers in 1987 in natuurlijk opaal. Talrijke numerieke studies volgden, te beginnen met Eldridge, Madden en Frenkel in 1993, die de stabiliteit van het $\mathrm{AB_{13}}$ kristal aantoonden en bulk fasediagrammen van dit zeer exotische kristal leverden. Ondanks de bewezen thermodynamische stabiliteit van $\mathrm{AB_{13}}$ is het nog nooit waargenomen dat $\mathrm{AB_{13}}$ spontaan nucleëert in een numerieke simulatie (net zoals bij de LP's). Daarom bestuderen we in Hoofdstuk 3 de nucleatie van de $\mathrm{AB_{13}}$ fase in een binair mengsel van colloïdale harde bollen. Om het nucleatieproces te volgen, gebruiken we een neuraal netwerk om de $\mathrm{AB_{13}}$ fase te onderscheiden van de binaire vloeistoffase en concurrerende kristalstructuren, op het niveau van één deeltje en met een ongekende nauwkeurigheid. We tonen aan dat standaardtechnieken falen in het onderscheiden van de verschillende fasen en dat machine learning onontbeerlijk is om deze taak te volbrengen. Door in meer detail te onderzoeken hoe het netwerk beslissingen neemt, kunnen we bovendien van het machine-learning algoritme leren welke eigenschappen en symmetrieën het belangrijkst zijn voor de classificatie. Met het geoptimaliseerde neurale netwerk als ordeparameter onderzoeken we het nucleatieproces van $\mathrm{AB_{13}}$, verduidelijken we de rol van lokale structuren in de oververzadigde vloeistof en werpen we licht op de relatie tussen grote en kleine bollen tijdens de allereerste stadia van het nucleatie proces. Eerst gebruiken we de seeding-benadering, die ons in staat stelt om alle relevante fysische variabelen met betrekking tot de nucleatie te berekenen en een kwantitatieve vergelijking te maken met de nucleatie van andere harde-bolkristallen, namelijk de face-centred-cubic structuur en de colloïdale LP's. Ten tweede voeren we brute-force Moleculaire Dynamica (MD) simulaties uit, die de eerste spontane nucleatiekernen van $\mathrm{AB_{13}}$ vormen die in de literatuur worden vermeld. Ten slotte gebruiken we een ander classificatiealgoritme gebaseerd op de herkenning van lokale motieven en tonen we aan dat de binaire nucleatie van de $\mathrm{AB_{13}}$ fase verloopt via een co-organisatieproces van grote bollen en icosaëdrische kleine bolclusters die zich vasthechten aan de kern, en dat het goed wordt beschreven door het klassieke nucleatieproces.

In Hoofdstuk 4 richten wij ons op het eenvoudigste en meest bestudeerde systeem in de zachte materie: het system van harde bollen. Ondanks talrijke experimentele en theoretische studies zijn vele aspecten van het nucleatieproces, zoals het selectiemechanisme van polymorfen in de vroege stadia van nucleatie, nog lang niet begrepen. Een duidelijk voorbeeld is de enorme discrepantie tussen de nucleatiesnelheden berekend uit simulaties en gemeten in experimenten, die meerdere ordes van grootte van elkaar verschillen. In dit hoofdstuk richten we ons op het overschot aan face-centered-cubic (fcc)-achtige deeltjes ten opzichte van hexagonal-close-packed (hcp)-achtige deeltjes in een kristalkern van harde bollen, zoals waargenomen in simulaties en experimenten. Om dit fenomeen te verklaren, simuleren we een groot aantal spontane nucleatie gebeurtenissen van harde bollen. Naast het volgen van de vorming en groei van kristalkernen gedurende nucleatie, richten we ons ook op het gedrag van specifieke topologische clusters, welke al in grote concentratie aanwezig zijn in de vloeistof en structureel heterogeen blijkt te zijn. Wij tonen aan dat, in de metastabiele vloeistoffase, vijfvoudig symmetrische clusters - vijfhoekige bipyramiden - waarvan bekend is dat zij remmers zijn van nucleatie, transformeren in een andere cluster - Siamese dodecaëders. Deze clusters vormen door hun geometrische rangschikking een brug tussen de vijfvoudig symmetrische vloeistof en het fcc-kristal, waardoor de oppervlakte-energie verlaagd wordt ten opzichte van het hcp-kristal, en licht geworpen wordt op het selectiemechanisme van de polymorf. Dit mechanisme wordt geanalyseerd in simulaties en experimenten, die een analoog gedrag laten zien.

In Hoofdstuk 5 richten we ons op hetzelfde systeem als in Hoofdstuk 4, en zelfs op dezelfde nucleatie gebeurtenissen, maar met de taak om een classificatie algoritme te vinden voor de verschillende polymorfen dat intrinsiek een grote hoeveelheid informatie bevat over de lokale omgeving van elk deeltje. Om dit te bereiken, maken wij gebruik van een techniek die behoort tot het gebied van unsupervised learning: Principal Component Analysis (PCA). In het bijzonder simuleren wij voor elk van de interessante fasen - fcc-, hcp- en bcc-kristallen (body-centred-crystal) en de monodisperse vloeistoffase - de evenwichtsstructuren en berekenen wij een groot aantal kenmerken om de omgeving voor elk deeltje te beschrijven. Deze worden zo gekozen dat verschillende kenmerken gevoelig zijn voor verschillende symmetrieën, en daarom verhoogt het gebruik van vele kenmerken de mate van informatie over de lokale omgeving van elk deeltje dat we gebruiken als input voor PCA. Het gebruik van PCA op de volledige dataset transformeert de parameterruimte, waarbij automatisch nieuwe ordeparameters worden gevonden waarlangs de projectie van de oorspronkelijke gegevens is gemaximaliseerd. Op deze manier is het mogelijk, zelfs met gebruikmaking van een klein aantal nieuwe ordeparameters, een zeer hoog niveau van informatie over elk deeltje te behouden. PCA biedt verschillende voordelen, die alle verband houden met het feit dat de nieuwe ordeparameters een lineaire combinatie zijn van de oorspronkelijke kenmerken. PCA heeft met name een hoge uitvoeringssnelheid en levert gemakkelijk interpreteerbare resultaten op, waaruit nieuwe inzichten kunnen worden verkregen over welke kenmerken nuttig zijn om de verschillende polymorfen van elkaar te onderscheiden. Als PCA eenmaal gebruikt is, is het mogelijk om de gegevens van nucleatiekernen te projecteren op de nieuwe ruimte en direct te classificeren in de nieuwe ruimte. In Hoofdstuk 5 definiëren we er twee. De eerste is gebaseerd op eenvoudige drempelwaardes voor de nieuwe ordeparameters, terwijl de tweede een aanvullende techniek voor unsupervised learning gebruikt, het Gaussian Mixture Model (GMM). In het bijzonder is het door het clusteren van de gegevens met GMM niet alleen mogelijk om deeltjes te labelen als behorend tot de bekende fasen, maar ook als behorend tot het grensvlak tussen twee fasen. Deze studie toont aan hoe dimensiereductie en clusteringalgoritmen een beslissende rol kunnen spelen bij classificatietaken. In het algemeen verwachten we dat unsupervised learning technieken een steeds grotere rol kunnen spelen bij de identificatie van belangrijke kenmerken van fysische verschijnselen, en niet alleen van nucleatie. Anderzijds verwachten we dat op het gebied van de zachte materie andere fenomenen van zelforganisatie of zelfs de glasovergang beter begrepen en beschreven kunnen worden, uitgaande van deze benadering, voor zowel systemen afkomstig van simulaties als van experimenten.

In tegenstelling tot wat we zojuist besproken hebben, waar de nadruk volledig lag op het nucleatieproces, is nucleatie op zichzelf vaak niet interessant als mechanisme, maar is men meer geïnteresseerd in het verkrijgen van de producten ervan. Het is om deze reden dat wij in de Hoofdstukken 6, 7 en 8 onze aandacht richten op Inverse Design Methods (IDMs). IDMs worden beschouwd als de heilige graal van materiaalontwerp via colloïde zelforganisatie. De vraag die we proberen te beantwoorden is: gegeven een gewenste materiaalstructuur (bv. een specifieke kristallijne fase), hoe kunnen we de colloïdale bouwstenen aanpassen om het gewenste product te verkrijgen? Wat de synthese betreft, hebben chemici ongelooflijk veel vooruitgang geboekt tot op het punt waar colloïdale bouwstenen bijna kunnen worden gemaakt met op verzoek gedefinieerde interacties. Hoewel de modernste computationele en theoretische statistisch-mechanische methodes uitermate geschikt zijn om het zogenaamde "Forward Design" probleem op te lossen, waarbij we de structuren voorspellen die gevormd worden door een a priori gedefinieerde set van colloïdale bouwstenen, blijft een robuust, veelzijdig algoritme om het “Inverse Design” probleem op te lossen een belangrijke uitdaging. Het ontbreken van een dergelijk IDM vormt een belangrijke hindernis voor de volledige benutting van colloïdale zelforganisatie bij de ontwikkeling van de volgende generatie materialen.

Als eerste benadering, in Hoofdstuk 6, reverse-engineeren we een bcc kristal dat, in een repulsief Yukawa deeltjes systeem, alleen stabiel is in een klein gebied van het fasediagram. Deze taak stelt ons in staat twee algoritmen van verschillende aard te testen - de eerste gebaseerd op de statistische fluctuaties van het onderzochte systeem en de tweede afgeleid van klassieke optimalisatietechnieken - en hun overeenkomsten en verschillen te evalueren. Voortbouwend op de kennis die in Hoofdstuk 6 is opgedaan, introduceren we in Hoofdstuk 7 een algemeen IDM om op efficiënte wijze gewenste kristallen, quasikristallen en vloeibare kristallen te reverse-engineeren door ons te richten op hun diffractiepatronen. Ons algoritme berust op het synergetische gebruik van een evolutionaire strategie voor parameteroptimalisatie, en een convolutioneel neuraal netwerk dat fasen classificeert op basis van hun diffractiepatronen als ordeparameter. De robuustheid en veelzijdigheid van deze nieuwe ordeparameter stelde ons in staat om met succes verschillende kristal-, quasikristal- en vloeibare kristalstructuren in twee- en driedimensionale modelsystemen te reverse-engineeren, en zo een nieuwe weg in te slaan voor het inverse design van experimenteel haalbare colloïdale interacties, specifiek geoptimaliseerd om een gewenste structuur te stabiliseren.

Tenslotte, in Hoofdstuk 8, bouwen we voort op wat we geleerd hebben over LP nucleatie in Hoofdstuk 2, en onderzoeken we de mogelijkheid om een zelforganisatie route te vinden om een kristal met een fotonische band gap te verkrijgen door middel van de IDM beschreven in Hoofdstuk 7.

Aangezien dit een zeer breed onderwerp is, blijven er uiteraard nog vele vragen over nucleatie open. De meeste van deze vragen zijn natuurlijk uitdagend, maar ook fascinerend. Wij hopen en geloven dat een aantal van de conclusies die getrokken zijn uit het werk dat in dit proefschrift gepresenteerd is, en een aantal van de analyses en methoden die we ontwikkeld hebben, onderzoekers kunnen helpen om een aantal van deze vragen te beantwoorden, om zo de kennis die we al hebben over dit spannende onderzoeksgebied te verbreden.

Acknowledgements

"All happy families are alike; each unhappy family is unhappy in its own way.". So begins *Anna Karenina*, the famous novel by Lev Tolstoj, one of the greatest and most influential writers of the 19th century. In 1997, this incipit was used as the basis for the so-called "principle of Anna Karenina" by geographer, historian, ornithologist, and author Jared Diamond, in his masterpiece (and Pulitzer price winner!) *Guns, Germs and Steel*. According to this principle, a complex result can only be achieved by satisfying a defined set of requirements, and it only takes one unsatisfied constraint for the final result not to be achieved. This concept can be applied in economics, in personal relationships, in biology – the author uses it to explain why zebras are not tameable, and more in general why very few animals are – and in a person's professional life. I strongly believe that it is also applicable to the incredible journey that the PhD is, for it is a complex route that can only be walked smoothly if many conditions are met. This short section therefore is dedicated to all the people who made sure that every piece of the puzzle was in its right place along the way.

First and foremost, I want to thank my supervisor. Marjolein, I remember very well the day we met for the first time. I had asked about the research carried out at SCM, and you received me in your office to present me at least half a dozen ongoing projects. More than by the content of the projects themselves - of which I actually understood very little at that time - I was deeply impressed by your enthusiasm and the passion for research you immediately showed me. The human aspect was decisive in my choice, and it is the same aspect on which I want to focus in these few lines. It is unnecessary to underline your scientific achievements, as they are well known to everyone in our field and not only. What I would like to emphasise, however, is your ability to create a wonderful environment in which to carry on research. I know you pay a lot of attention to this aspect of your job, and I want to express all my gratitude for this. It is unfortunately not a common skill in Academia, but one of the most precious ones.

Not only my supervisor, but many other people in the SCM staff deserve to be thanked. In one way or another, during coffee breaks, in the secretariat office, or in the work discussion room, you have all contributed to making these four years easier, richer, and more stimulating. In particular, thank you Laura for being so friendly from day one, and for always being up for a chat when we ended up staying late in the office. I think it's a real blessing to have the chance to have casual scientific discussions with you for anyone tackling a PhD in nucleation. Thank you Alfons for all the times you listened to my Friday morning presentations with the critical eye that distinguishes you. Your frequent inputs have significantly improved every project I have worked on. Thank you René, your enthusiasm for research is contagious, and that for teaching, a true inspiration. It has been a pleasure to be teaching assistant for your courses.

If I started this journey in SCM, I owe a lot to my previous supervisor, Francesco Sciortino. Thank you Francesco for introducing me to the field of soft matter physics, and for having continuously and patiently guided and stimulated me during the months I had the pleasure of working with you. Today, as then, I consider it a real honour to have learned from you.

Speaking of past masters, I would like to spend a few lines to thank my high school maths and physics teacher, Gualtiero Grassucci. Thank you Gualtiero, your way of looking at mathematics and physics with the wonder of a child profoundly influenced my decision to start this journey to become a scientist. The fact that, in your classes, the number of students enrolling in mathematics and physics at university is so unusually high is no coincidence, and is something

you should be extremely proud of.

A notoriously exciting aspect of a PhD journey is the chance to work together with people from all over the world. During these four years I have had the pleasure of working with fantastic scientists from whom I have learned a lot, and thanks to whom I became a better researcher myself. Rajneesh, Giulia, Robin, René, Da, Ernest, Alfons, Emanuele, Laura, and Paddy, thank you. I would like to particularly thank Tonnishtha and Srikanth. Toni, I was lucky to carry on my first project with an experienced and skilled researcher like you, thanks for teaching me so much and for your attitude. Sri, your knowledge has no limits and your inputs in each meeting have been extraordinary. Thanks for your dedication to the projects and for being such an exceptional host in Bangalore: I will never forget the drinks with you, Mark Miller, and Daan Frenkel.

Thanks to my paranymphs, former office mates, and dear friends Massi and Emanuele. Leaving aside the irony that I enrolled in a PhD abroad in search of an international environment, and found myself in the office with two people from my own region in Italy, I can safely say that you have been the best surprise of my PhD. Thank you because your successes have pushed me to improve day after day, thank you for your sincere advice, professional or not, and above all thank you because I am sure I will never have so much fun in any other office I will happen to work, unless it's with you again. We should probably apologise to the people in the neighbouring offices, but I would still do it all over again.

Speaking about neighbouring offices, I feel the need of expressing all my gratitude to the wonderful colleagues I met at SCM through the years. In this respect too, I have been more than fortunate to have met a group of people who are so passionate about their work and from whom I have learned a lot during the weekly work discussions on Friday morning. It was a pity we didn't get to meet so often during the pandemic, but on every occasion we did, be it a borrel at the coffe-corner, at the park, or downtown, or even a group outing or a karaoke night, I just had so much fun with all of you.

One of the factors that played a major role in making me feel at home in the Netherlands is certainly all the incredible people I met on my way, and with whom I shared unforgettable moments. Phil, you have been my first friend in Utrecht, thanks for being the chillest guy I have ever met and for stimulating my French. Naveed, thanks for being a great friend and such a kind person. I loved to see how you turn into the most irritating teaser when playing foosball. Rama, thanks for inviting me to your wedding on my very second day of work, it has been the most wonderful journey of my life. Merijn and Marleen, thanks for having been such great neighbours, and for the evenings spent together, especially during the pandemic. Jacques, Rohith, Logan, I feel so grateful for having met you guys. The time with you is always filled with laughter and I am looking forward to all the future moments we will share. Finally, a gigantic thank you to the Italian crew, the best I could have ever hoped for. You have truly made me feel at home from day one, I will never forget this. Elena, Emanuele, Silvia, Massi, Laura (you are at least half Italian now), Raf, Eleonora, Vania, Nino, Irene, Gabri, Enrico, Greta, Fabrizio, Francesco A., Francesco P., Jerry (you also count very much as an Italian), Giuliana, and Giovanni, any amount of words will never be enough to thank you.

Un grazie enorme va agli amici di una vita. Non vivere più la mia quotidianità in vostra compagnia è ciò che più di tutto mi fa rimpiangere di aver lasciato casa. Ogni volta che torno a Latina, prima di rivedervi, ho sempre un po' di timore che qualcosa nel nostro rapporto possa essere cambiato a causa della distanza. Eppure, ogni volta, bastano due minuti per rendermi conto ancora una volta che nulla è cambiato e cambierà mai, che la nostra amicizia è solida come la roccia, e che ci saremo sempre l'uno per l'altro.

Grazie alla mia famiglia, a chi c'è, e a chi purtroppo non c'è più. Siete e sarete per sempre un punto di riferimento imprescindibile nella mia vita. Grazie Mamma e Papà per l'amore incondizionato e per il supporto costante. Grazie soprattutto perché sentire la vostra fiducia nei miei confronti in ogni mia scelta è il dono più bello che potessi ricevere. Grazie Pierpaolo per i consigli su praticamente qualsiasi aspetto della mia vita, e per essere contemporaneamente, da sempre, una guida preziosa e un amico sincero.

Infine, grazie Serena. Grazie per la tua straordinaria forza interiore e per la tua altrettanto straordinaria sensibilità. Grazie per essere così brillante, per tutte le risate che mi fai fare dalla mattina alla sera, per essere un vulcano di energia. Se guardo al futuro con ottimismo ed impazienza, è per tutte le avventure che vivremo insieme, mano nella mano, con l'entusiasmo di due ragazzini.

Gabriele

List of publications

This thesis is based on the following publications:

- T. Dasgupta/G. M. Coli, M. Dijkstra, *Tuning the glass transition: enhanced crystallization of Laves phases in nearly hard spheres*, ACS Nano **14 (4)**, 3957-3968 (2020). Chapter 2.

- G. M. Coli, M. Dijkstra, *An artificial neural network reveals the nucleation mechanism of a binary colloidal AB_{13} crystal*, ACS Nano **15 (3)**, 4335-4346 (2021). Chapter 3.

- G. M. Coli, R. van Damme, C. P. Royall, M. Dijkstra, *Hidden in the fluid: fivefold symmetry and polymorph selection in hard spheres*, manuscript submitted (2021). Chapter 4.

- G. M. Coli, B. Verhoef, M. Dijkstra, *Through the order parameter filter: polymorph detection with an unsupervised learning procedure*, manuscript in preparation. Chapter 5.

- R. Kumar, G. M. Coli, M. Dijkstra, S. Sastry, *Inverse design of charged colloidal particle interactions for self assembly into specified crystal structures*, The Journal of Chemical Physics **151 (8)**, 084109 (2019). Chapter 6.

- G. M. Coli/E. Boattini, L. Filion, M. Dijkstra, *Inverse design of soft materials via a deep-learning-based evolutionary strategy*, manuscript submitted (2021). Chapter 7.

Other publications by the author:

- G. Fiorucci, G. M. Coli, J. T. Padding, M. Dijkstra, *The effect of hydrodynamics on the crystal nucleation of nearly hard spheres*, The Journal of Chemical Physics **152 (6)**, 064903 (2020).

- D. Wang, T. Dasgupta, E.B. van der Wee, D. Zanaga, T. Altantzis, Y. Wu, G.M. Coli, C.B. Murray, S. Bals, M. Dijkstra and A. van Blaaderen, *Binary icosahedral clusters of hard spheres in spherical confinement*, Nature Physics **17**, 128–134 (2020).

- R. van Damme, G. M. Coli, R. van Roij, M. Dijkstra, *Classifying crystals of rounded tetrahedra and determining their order parameters using dimensionality reduction*, ACS Nano **14 (11)**, 15144-15153 (2020).

- P. Cats, S. Kuipers, S. de Wind, R. van Damme, G.M. Coli, M. Dijkstra and R. van Roij, *Machine learning free-energy functionals using density profiles from simulations*, APL Materials **9**, 031109 (2021).

Oral and poster presentations

Part of the work of this thesis was presented at:

- Physics@Veldhoven, Veldhoven, The Netherlands (2019)
  Poster: *Hard times for hard spheres: enhanced crystallisation of the Laves phase from soft colloids* - Best Poster Prize Nomination

- International Soft Condensed Matter Conference, Edinburgh, Scotland (2019)
  Talk: *Hard times for hard spheres: enhanced crystallisation of the Laves phase from soft colloids*

- Nanoseminar, Utrecht, The Netherlands (2019)
  Talk: *Hard times for hard spheres: enhanced crystallisation of the Laves phase from soft colloids*

- Physics@Veldhoven, Veldhoven, The Netherlands (2020)
  Poster: *Quasicrystals, the Rev.Eng.(e): topological reverse engineering through evolutionary computation*

- Physics@Veldhoven (virtual), Veldhoven, The Netherlands (2021)
  Talk: *Neural network reveals the nucleation mechanism of a binary colloidal* AB_{13} *crystal*

- CECAM Workshop (virtual): Local structure meets machine learning in soft matter systems, Lausanne, Switzerland (2021)
  Talk: *Inverse design of soft materials via a deep-learning-based evolutionary strategy*

- 11^{th} Liquid Matter Conference (virtual), Prague, Czech Republic (2021)
  Poster: *The role of fivefold symmetry in the polymorph selection of hard-sphere crystal nucleation*

- 11^{th} Liquid Matter Conference (virtual), Prague, Czech Republic (2021)
  Talk: *Inverse design of soft materials via a deep-learning-based evolutionary strategy*

About the author

Gabriele Maria Coli was born on February 22, 1993 in Rome, Italy. After graduating *cum laude* from high school in his home town, Latina, in 2012, he moved to Rome and studied Physics at the University of Rome “La Sapienza”. In 2015 he obtained his Bachelor’s degree *cum laude* with a thesis entitled “Stochastic optimisation with Simulated Annealing method and application to Random Hypergraph Bicolouring model.”, supervised by Prof. Federico Ricci-Tersenghi. After his Bachelor’s studies, he studied Theoretical Physics at University of Rome “La Sapienza”. For his final Master’s project, he worked on his thesis entitled “Numeric study of the dynamics of a vitrimeric DNA gel” under the supervision of Prof. Francesco Sciortino. He obtained his Masrer’s degree *cum laude* in 2017. In November 2017 he started as a PhD candidate at the Soft Condensed Matter group of Utrecht University under the supervision of Prof. Marjolein Dijkstra. The results of his research have been presented in several national and international conferences, and constitute the subject of this thesis. During his PhD, he has been a member of the DAC Committee for a total of 2 years, thus contributing to uniting the research groups at the Debye Institute of Nanomaterials Science by organising activities and events.