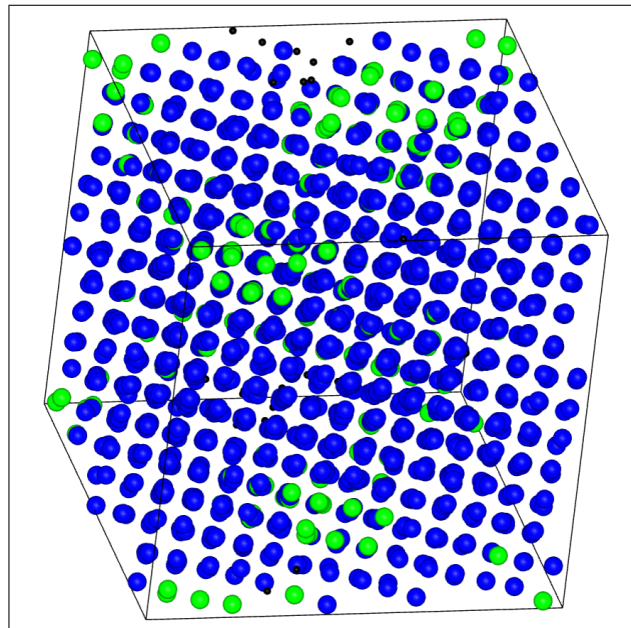**Universiteit Utrecht**

Opleiding Natuur- en Sterrenkunde

Using Machine learning to study nucleation in a model system of soft polymer colloids

BACHELOR THESIS

Steven Bos

Supervisor:

Prof. Dr. Ir. Marjolein Dijkstra
Debye Institute for Nanomaterials Science

January 22, 2021

Abstract

Despite the amount of work devoted on nucleation, the mechanism of nucleation is still not well understood. Many scenarios have been proposed such as a classical one-step nucleation mechanism or a non-classical two-step crystallization process, but both scenarios are still heavily debated. In this thesis, we investigate the crystal nucleation mechanism of Gaussian core particles using computer simulations, and quantify the results using machine learning. Using a Principal Component Analysis we will shed light on the nucleation mechanism of Gaussian core particles in the presence of different competing crystal structures.

Contents

1 Introduction

Colloidal suspensions like blood, milk and paint, are familiar to everyone. These liquids with colloidal particles suspended in it, can form various structures and exhibit corresponding transitions between these structures. A famous example is the clotting of blood, where through a complex process, the liquid blood changes to a gel, forming blood clot to prevent blood loss and heal wounds ultimately. The process of blood clotting requires endothelium lining the blood vessel after an injury to the blood vessel, making it a process which, luckily, will not happen spontaneous. There exists processes in which colloids change structure spontaneously, a process called self-organization. For example due to a temperature and or pressure change, a colloidal suspension can form a highly organized structure known as a colloidal crystal.

The formation of crystalline structures in colloidal suspensions is a heavily studied subject, but not yet well understood. What makes these types of systems self-organize is due to the forces between the particles in the system. The repulsive part of the inter-molecular forces, which becomes significant when particles are at close separation, is crucial for fluid-solid phase transitions[1, 12, 6, 4]. The particles, and their respective repulsive forces, take up volume in the system, which influences neighbouring particles, and in turn drives the system to self-organise, ultimately resulting into crystallization.

This process, known as crystal nucleation, does not always occur immediately at freezing conditions. For instance, the freezing of water into ice occurs if water is below 0° Celsius, but water cooled only slightly below this transition temperature often stay free of ice for long periods of time, sometimes not nucleating at all. Classical Nucleation Theory states that small crystalline clusters can be formed spontaneously due to thermal fluctuations in a supercooled liquid. These clusters exhibit an energetically unfavorable interface, creating a barrier between the liquid and crystal particle, stagnating further crystal growth when the crystal nucleus size is smaller than its critical size. For sufficiently large clusters, the barrier can be overcome, allowing the nucleus to further grow.

Nucleation can also occur in a non-classical way, e.g. Ostwald's step rule states that crystallization from a solution occurs in steps, in such a way that often thermodynamically unstable phases occur first, followed by the transformation to the thermodynamically stable phase[10]. This possibility was later theoretically investigated by Alexander and McTague in 1978. On the basis of a Landau theory, they argued that an expansion of the Landau free energy favors nucleation of a body-centered cubic (bcc) phase in the early stage of a weak first-order phase transition of a simple liquid.[2] Despite the focus of research on the existence of an intermediate crystal step, Kawasaki and Tanaka demonstrated that not just intermediate crystal states should be considered, but that it is equally important to consider hidden ordering in the super-cooled liquid which may be regarded as the intermediate state from the liquid phase.[5]

Simulations of particles with simple repulsive forces, like charged colloidal particles which stabilize multiple crystalline structures, can be insightful on solidification transitions in particle systems. However, simulations have so far not shown to give conclusive evidence of what type of crystals are formed during crystal nucleation.

Russo et al.[13] showed that in simulations on a system of Gaussian core particles, the nucleation pathway does not follow Ostwald's step rule of crystallization. They found the bcc phase to be favoured during nucleation despite the underlying phase diagram, this opposes Ostwald's step rule which predicts for instance at a pressure of $P\sigma^3/\epsilon = 0.05$ the nucleation to happen through the face-centered-cubic (fcc) phase before forming the final bcc phase. For a lennard-Jones fluid, Ten Wolde et al. showed that during nucleation the nucleus is primarily structured as a fcc in the core, and that the surrounding particles forming the interface between solid and liquid, is primarily structured as bcc.[16] The process of nucleation is still up for debate, therefore the nucleation process will be re-investigated in this thesis, trying to shed light on this intricate process.
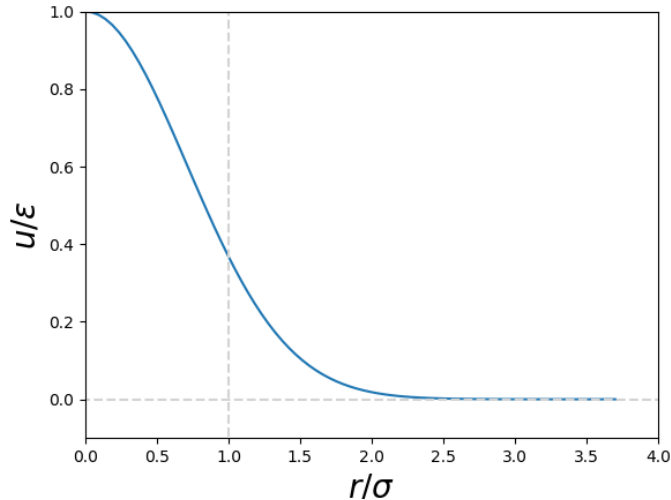
Figure 1: Gaussian core model (GCM) pair potential $u(r)$ as a function of distance r in reduced units of diameter σ and energy ϵ.

The crystal nucleation will be studied with computer simulations consisting of particles with a repulsive Gaussian core potential, from now on referred to as the Gaussian core model (GCM), originally proposed by Stillinger[15]. The GCM is less extensively studied as the popular hard-sphere model and the Lennard-Jones particle model. Despite being less studied, the Gaussian core model deserves attention for several reasons. The similarity between the stable crystal phases of the Gaussian core model and of charged colloidal particle crystals is evident when considering the cause of stability of both, being the repulsive forces in the respective models. The relative softness of the interaction compared to that of, for instance the hard-sphere model, suggests the behaviour is similar to that of polymeric blobs. Polymeric particles usually involve globular poly-atomic molecules such as camphene and cyclohexane which, when pressed together, can mutually deform and rotate to fit "bumps" into "holes".

Using computer simulations of Gaussian core particles spontaneous crystallization can be simulated. To study the crystal structures formed during nucleation the Steinhardt-Nelson bond orientational order parameters[14] are calculated, which is a common method for identifying and classifying the present crystal configuration. These bond order parameters describe the symmetry of the local neighbourhood of a particle in terms of spherical harmonics. Certain bond order parameters resemble the structure of a crystalline phase, therefore it is possible to identify crystalline structures using the spherical harmonics of a particle. There are an infinite amount of spherical harmonics and thus bond order parameters, however the first 12 bond order parameters are conclusive enough to describe the symmetry of the local neighborhood of a particle. To classify the crystalline structure that is present there is a set of 4 bond order parameters which are of interest, $l = 4$ distinguishes the bcc from the other configurations, $l = 6$ has minor overlap of the fluid phase with respect to crystalline ordering, in general a fluid phase has lower bond orientational order and thus lower values for the bond parameters.

Instead of looking at individual bond-orientational order parameters to do the analysis, we use a Principal Component Analysis to construct a lower dimensional representation of a much larger set of bond-orientational order parameters. By doing so we expect to accurately detect and distinguish different crystalline structures. The remainder of this thesis is structured as follows: in section 2 the simulation techniques used to simulate nucleation events of a GCM fluid are described. Afterwards, the methods used to determine the crystal structures formed during nucleation. In section 3 the results are presented and compared to earlier results of GCM investigations. Finally in section 4 we conclude and give an outlook.

4

2 Theory and method

2.1 Monte Carlo Simulations

We consider a system consisting of N particles that is described by the Hamiltonian:

$$H(\boldsymbol{r}^N, \boldsymbol{p}^N) = \sum_{i=0}^{N} (\frac{\boldsymbol{p}_i^2}{2m} + U(\boldsymbol{r}^N)),\tag{1}$$

which corresponds to the total energy of the system, being the sum of both the kinetic and potential energy. The Hamiltonian defines the behaviour of the system, describing the motion and potential energy of the particles. Macroscopic properties of the material are due to the microscopic interactions between particles. In statistical physics, canonical ensemble averages of momentum-independent observables A of a system of N particles in a volume V and temperature T can be calculated in the following manner:

$$\langle A \rangle = \frac{\int d\boldsymbol{r}^N A(\boldsymbol{r}^N) exp[-\beta U(\boldsymbol{r}^N)]}{\int d\boldsymbol{r}^N exp[-\beta U(\boldsymbol{r}^N)]},\tag{2}$$

Here $\beta = 1/k_B T$ is the inverse temperature and $U(\boldsymbol{r}^N)$ is the total potential energy of the system for N particles. Observable A, a quantity only dependent on the inter-particle distances $\boldsymbol{r}^N$ of all particles, since the inter-particle forces are velocity independent, the integration over the momenta can be ignored. Calculating these ensemble averages of the system cannot be done by numerical integration techniques, since for $N = 2000$ particles, and 10 grid points in each direction, there are $m^{DN} = 10^{6000}$ points at which the integrand has to be evaluated which is not possible within a lifetime by current day computational power. To circumvent this problem the Monte Carlo method can be employed. Using the Monte Carlo method the ensemble average is calculated by integrating over a random sampling of points instead of a regular array of points. When an infinite amount of sampling points is considered for integration, we would end up at the continuous case. Intuitively a lot of configurations will not add up to the ensemble average, due to the high probability to find a low Boltzmann weight for a configuration, especially for close-packed or overlapping configurations. Metropolis et al.[9] proposed an algorithm to simulate configurations according to their Boltzmann weight. First a random configuration of N particles is made without consideration of the Boltzmann weight, after which each particle is moved according to the following acceptance rule[3]:

$$acc(0 \rightarrow n) = min(1, \exp[-\beta(U(n) - U(0))]),\tag{3}$$

where $U(n)$ is the potential energy of configuration n. New configurations with a higher energy are to be accepted with a probability that depends on the Boltzmann weight of the new and old configuration. For a proposed configuration move with a lower total energy, the move is always accepted. Generating configurations of particles $\boldsymbol{r}_i^N$ with a probability proportional to its Boltzmann weight $exp[-\beta U(\boldsymbol{r}_i^N)]$ the ensemble average(2) reduces to:

$$\langle A \rangle = \lim_{n \rightarrow \infty} \frac{\sum_{i=0}^{N} A(\boldsymbol{r}_i^N)}{n},\tag{4}$$

where n is the number of trial moves. After each trial move in a Monte Carlo Simulation the ensemble average is calculated considering it is now a new configuration, even if the trial move is not accepted. The same acceptance rule(3) can be applied to a volume change of the simulation box after making N Monte Carlo steps, where on average every particle is moved once. By moving the particles, changing the simulation box volume and keeping the amount of particles and pressure fixed, an isobaric-isothermal NPT ensemble simulation is performed.

A simulation of a finite volume would suffer dramatically from surface effects. The fraction of particles residing at the surface of a cubic box is given by $6N^{(2/3)}$ making the fraction of surface particles significant: 60% for $N = 1000$ particles. Using periodic boundaries for the simulation box, the effect of surfaces can be avoided. Periodic boundary conditions are implemented by adding duplicates of the simulation box to all sides of the simulation box. When a particle leaves the central box, it will reappear on the opposite site in the central box, fixing the number of particles in the simulation and making the simulation volume infinitely

large. The particles are free to move over the boundaries of the box avoiding any surface effects. The distance between particle i and j is calculated with consideration of the nearest image convention stating a particle i can only interact with the nearest image of particle j, either in one of the neighbouring boxes or in the central box, therefore minimizing the distance between particles.

We consider the Gaussian core particle model, in which the particles interact with the pair potential:

$$u(r) = \epsilon \exp\left(-\frac{r^2}{\sigma^2}\right), \tag{5}$$

with σ the diameter of a particle and ϵ the energy. The potential is truncated at $R_{cut} = 3.7\sigma$. By employing reduced units the variables are dimensionless, and will never take very small or large numerical values, minimizing errors caused by rounding of the initial values of a variable, which extrapolates over several integration steps.

2.2 Simulation details

To study the different stages of nucleation, various simulations of $N = 2000$ particles in the NPT-ensemble will we conducted, using the state points shown in the phase diagram by Russo et al.[13] in figure 2.
The phase diagram is presented in the reduced temperature $T^* = \frac{k_\beta T}{\epsilon}$ and pressure $P^* = \frac{P\sigma^3}{\epsilon}$ plane. The pressure is kept at a constant value of $P^* = 0.002$ or $P^* = 0.005$ for all the simulations, and we study nucleation at varying temperatures.
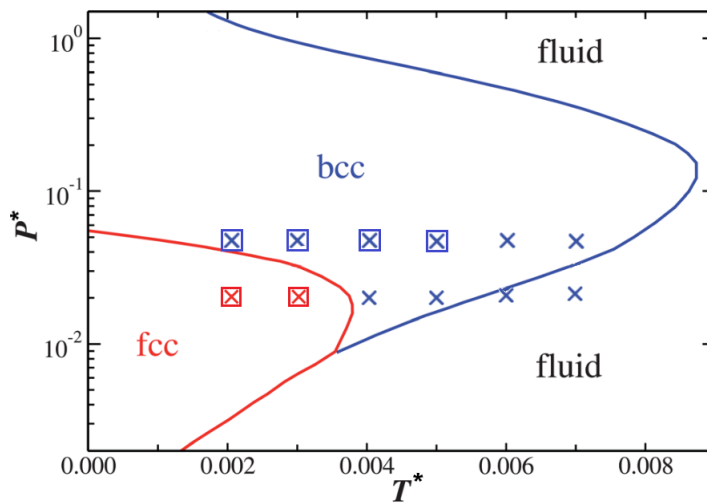

Figure 2: Phase diagram of the GCM in the pressure P^*-temperature T^* plane adapted from Russo et al.[13] reproduced from data in [11]. The state points used in the Monte Carlo Simulations are marked by the crosses, for pressures of $P^* = 0.02$ and $P^* = 0.05$, and temperatures of $T^* = 0.002$, 0.003, 0.004, 0.005, 0.006 and $T^* = 0.007$. The squares denote the state points which nucleated.

According to Classical Nucleation theory, nucleation occurs as a spontaneous process starting from a meta-stable phase. The nucleation starts with one cluster forming and growing large enough to not immediately melt back into the liquid phase due to thermal fluctuations. The nucleus has to grow large enough to overcome the nucleation barrier which is associated with a critical nucleus size. The Gibbs free energy at the top of the nucleation barrier is given by:

$$\Delta G^* = \frac{16\pi\sigma^3}{3|\Delta\mu_v|^2}, \tag{6}$$

6

with $\Delta\mu$ the chemical potential difference between the thermodynamically stable phase and the supercooled metastable phase, and σ the surface tension of the interface between the nucleus and its surroundings.

For a lower nucleation barrier, the nucleation process will obviously occur more likely opposed to a large nucleation barrier. For the brute fore simulation we require a reasonably fast nucleation rate to make sure nucleation will take place within the span of our simulation time. Therefore not all state points will be useful for this study.

2.3 Measuring crystal structure

To study the different crystals formed during nucleation, the configurations are analysed by the bond-orientational order parameters proposed by Steinhardt et al. in 1983[14]. Firstly the nearest neighbours of each particle are determined, describing the surroundings of the particles. The nearest neighbours of a particle can be defined in multiple ways, to be able to have a parameter-free calculation which can be performed on-the-run, the solid angle nearest neighbour (SANN)[8] algorithm is used. The SANN attributes to each possible neighbour of a particle, a solid angle for which the cutoff radius is calculated restricting the sum of the solid angles $\theta_{i,j}$ to be 4π. First, the particles $\{j\}$ surrounding i are to be ordered such that $r_{i,j} \leq r_{i,j+1}$ relating the number of neighbours m and the shell radius $R_i^{(m)}$ by the relation:

$$r_{i,m} \leq R_i^{(m)} < r_{i,m+1}. \tag{7}$$

Starting with the particle closest to i we calculate for each potential neighbour $\{j\}$ an angle $\theta_{i,j}$ based on the distance between the particles $r_{i,j} = |\vec{r}_j - \vec{r}_i|$ and the shell radius $R_i^{(m)}$ which is yet to be determined. The m nearest neighbours of particle i such that their solid angles associated with $\theta_{i,j}$ equals 4π:

$$4\pi = \sum_{j=1}^{m} 2\pi[1 - \cos(\theta_{i,j})] = \sum_{j=1}^{m} 2\pi(1 - r_{i,j}/R_i^{(m)}). \tag{8}$$

The number of nearest neighbours m and the shell radius $R_i^{(m)}$ are not known yet, but since they are not independent of one another, can be determined when the sum of the solid angles is calculated. The method is parameter free and computer cost efficient, making it ideal for on-the run particle simulations. A minimum amount of nearest neighbours m for a particle should be at least 3, since for a single neighbour the solid angle contribution is always less than 2π. Eqs. 7 and 8 can be combined to determine the neighbour shell radius,

$$R_i^{(m)} = \frac{1}{m-2} \sum_{j=1}^{m} r_{i,j} < r_{i,m+1} \tag{9}$$

To solve the inequality (9) and therefore finding the set of nearest neighbours $N_b(i)$, m is increased iteratively until Eq. 9 is satisfied. Nearest neighbours are not always vice versa, however this issue is to be resolved by removing particles which are not reversed nearest neighbours from the total set of nearest neighbours.

2.4 Bond-orientational order parameters

The Steinhardt bond-orientational order parameters can be calculated, given by:

$$q_{lm}(i) = \frac{1}{N_b(i)} \sum_{j=1}^{N_b(i)} Y_{lm}(\boldsymbol{r}_{ij}), \tag{10}$$

where $Y_{lm}(\boldsymbol{r}_{ij})$ denotes the spherical harmonics for l and $m \in [-l, l]$, both integers, $\boldsymbol{r}_{ij}$ the distance vector from particle i to particle j and $N_b(i)$ the set of nearest neighbours of particle i. The orientation of

the coordinate system matters for these bond-orientational order parameters $q_{lm}(i)$ but by constructing a rotational invariant bond order parameters, this is avoided:

$$q_l(i) = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |q_{lm}(i)|^2}. \tag{11}$$

Based on the Steinhardt bond-orientational order parameter, Lechner and Dellago introduced an averaged bond order parameter in 2008, to increase the accuracy of the classification of crystalline structures:

$$\bar{q}_{lm}(i) = \frac{1}{N_b(i)} \sum_{j=0}^{N_b(i)} q_{lm}(j). \tag{12}$$

By averaging the bond order parameter over the neighbouring particles, the surrounding particle structures are of influence for the particle being considered, resulting in a more accurate classification of the structure at hand. The crystalline environment around particles can be analysed with averaged and non-averaged rotational invariant bond-orientational order parameters. By comparing the bond order parameters of a configuration with those of a known crystalline structure, the configuration can be classified. Generally the q_4 and q_6 values are of interest when classifying crystals. To distinguish between solid-like and fluid-like particles, Ten Wolde et al.[16] proposed in 1996 a method where the correlation between the local symmetries of two particles is measured, by defining first a normalized complex vector for each particle as:

$$d_{6m}(i) = \frac{q_{6m}(i)}{\sqrt{\sum_{m=-6}^{6} |q_{6m}|^2}}, \tag{13}$$

and then using this complex vector to define a scalar product for two neighbouring particles i and j as:

$$S_{ij} = \sum_{m=-6}^{6} d_{6m}(i) \cdot d_{6m}^*(j). \tag{14}$$

When this scalar product S_{ij} for particle i and j exceeds a threshold of 0.7, the two particles are considered to be connected. However this still does not distinguish solid from liquid phase particles, since fluid particles also sporadically exceed this threshold value. Therefore a minimum number of connected neighbouring particles is introduced in order to classify the particle as solid-like. A reasonable minimum number of connected particles for a solid-like behaviour of a particle is set to 7, which identifies more than 99%[16] particles in an fcc structure as being solid-like. By setting this minimum, we can group solid-like and fluid-like particles and create clusters of connected particles, where solid-like particles that are neighbours belong to the same cluster. These clusters determine the phase of regions of the simulation box.

2.5   Principal component Analysis

To make sense of the bond-orientational order parameters of the simulation, Principal Component Analysis (PCA) is used to find linear combinations of parameters best describing the essence of the dataset. The axes along which the dataset has the most variation is considered to be the first principle component of the dataset, thus describing the data best. All principal components should be orthogonal. To apply PCA we need to solve an eigendecomposition problem for N particles with j parameters, forming a $N \times j$ matrix X. To center the dataset we subtract the mean value of every parameter from the dataset. The principal components of X are given by the eigenvectors of XX^T. The eigenvector corresponding to the largest eigenvalue is the first principal component. In order to classify a configuration among different crystalline structures, a PCA is performed on the main crystalline structures, bcc, fcc and hcp such that a linear space of the principal components is spanned on which the dataset is projected. The crystalline configurations used to span this linear space are first equilibrated at a temperature of $T^* = 0.002$ and pressure of $P^* = 0.05$, the fluid configuration used for the PCA is equilibrated at temperature of $T^* = 0.006$ and $P^* = 0.01$.

To make sense of the projection in the PCA space, a measure is required to what order the configuration agrees with the crystal phases. The measure used is the Mahalanobis distance first introduced by Mahalanobis in 1936[7] which is a generalization of distance, how many standard deviations away a datapoint is from the mean of the dataset. To put the distance in perspective for our dataset the Mahalanobis distance is calculated from the center of a phase cloud in the PCA space to the center of the configuration phase cloud in the same PCA space. The center of a phase cloud in the PCA space is calculated by taking the mean of all the BOP values of all the particles in the configuration, projected on the principal component in both axes. Another method that could be used is taking the circumference of all the datapoints, and to take the center of said circumference, however this method would be very sensitive to outliers in the dataset, which are common for the PCA phase cloud. The Mahalanobis distance for a datapoint X is calculated by:

$$D_M(x) = \sqrt{(x - \mu)S^{-1}(x - \mu)}, \tag{15}$$

S^{-1} is the inverse of the co-variance matrix, x the middle of the phase cloud in the principal component space which we want to compare with μ the middle of the phase cloud.

The results of the projections and calculations will be further discussed in the next section where all the results and findings of the simulations are laid out.

3   Results

During the simulation we kept track of the particle configurations every $10k$ Monte Carlo steps, this way we were able to calculate the number of solid-like particles over time. A visualization of the fraction of solid-like particles in the system is shown in figure 3, from which it is clear which state points nucleated or stayed fluid like.
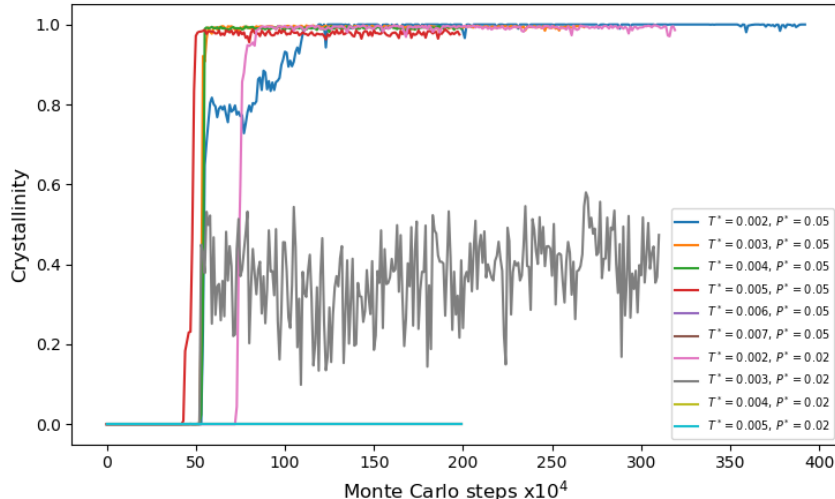

Figure 3: Fraction of solid-like particles as a function of Monte Carlo steps.

Spontaneous crystallization did not occur for all the state points at which the Monte Carlo simulation was running, when looking at the crystallinity of the system at the end of the simulation time. For pressure $P^* = 0.02$, only for temperatures $T^* = 0.002$ and $T^* = 0.003$ crystallization occurred, for the other state points at this pressure the amount of solid-like particles remained negligible and thus these state points will not be considered for the analysis of the crystal nucleation. For a reduced pressure $P^* = 0.05$ and temperatures of $T^* = 0.002$, 0.003, 0.004 and $T^* = 0.005$ the number of solid-like particles increased during the simulation, and thus some sort of nucleation occurred which we will further investigate. At the remainder of the state points no nucleation occurred and thus these state points are not taken into consideration for the analysis of crystal nucleation.

The state points at which no nucleation occurred within the time-frame of the simulation, the phase diagram by Russo[13] did predict a stable crystal phase, except for state points ($T^* = 0.006$ and $0.007, P^* = 0.02$). However there is a slight difference for a phase diagram to predict a stable phase and for the system to transform to this state, in the case of crystal nucleation the aforementioned nucleation barrier 6 may prevent spontaneous nucleation. When after roughly 3 million Monte Carlo steps, we concluded that the nucleation rate at these state points is very low. The nucleation barrier is too large to observe nucleation in a brute force Monte Carlo simulation. In future studies we could make use of Umbrella sampling, where the system is biased towards a nucleation event, enabling us to analyse nucleation at state points with a relatively high nucleation barrier.
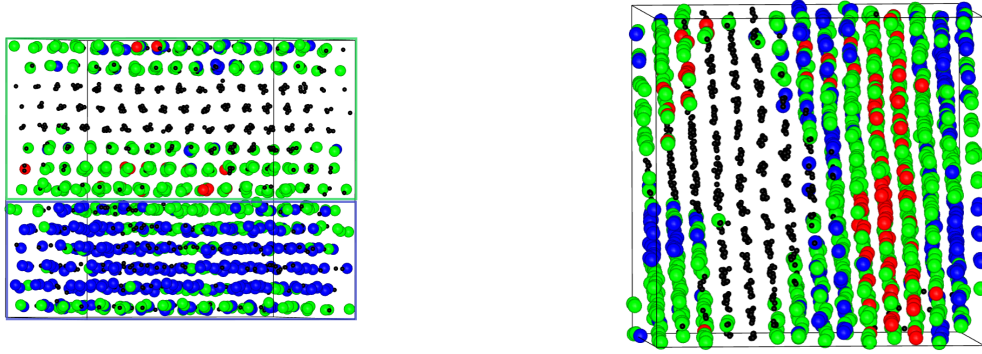
Figure 4: Snapshots of the Gaussian core model particle configurations during the Monte Carlo simulation. The figure on the left is a configuration of the state point ($T^* = 0.003, P^* = 0.02$) after 3 million Monte Carlo steps. The figure on the right is a configuration of the state point ($T^* = 0.002, P^* = 0.05$) after 1 million Monte Carlo steps. The fluid like particles are displayed as small black spheres, together with the fcc-like (red) particles, bcc-like particles (blue) and the hcp-like (green) particles. The green region is dominated by hcp-like particles and the blue region by bcc-like particles. Due to the different orientation of the fluid-like particle clusters compared to the rest of the simulation volume they are not considered to be solid-like due to equation 14.

Looking at the state points where solid-like particles did form, the graph for state point ($T^* = 0.003, P^* = 0.02$) immediately stands out from the rest. For the state point crystallization did not happen at once, the configuration does not go near complete crystallinity, and the amount of solid-like particles seems to fluctuate between roughly 400 and 1200. An equilibrium is reached at which the system is not completely turned into solid-like particles. To get a better understanding of what happens during the simulation, a snapshot of the configuration is visualized in figure 4 on the left. Notably a large cluster of fluid-like particles is formed which is not considered solid-like, due to the averaging over the q_6 value of the neighbouring particles with a different orientation by equation 14. Two regions are formed with one being predominantly bcc-like and one being hcp-like suggesting two separate clusters of solid-like particles started forming during the simulation, instead of one cluster forming and growing as would be expected in Classical Nucleation theory. Therefore this state point might not be appropriate to study crystal nucleation.

The simulation at state point $T^* = 0.002, P^* = 0.05$ seemed to nucleate at once but abruptly stopped nucleating after reaching a crystallinity of about 80%, after which it gradually but slowly continued forming solid-like particles towards complete nucleation. To gain a better understanding of what made the system stop nucleating, the particle configuration is plotted in figure 4 on the right, displaying the simulation after 1 million Monte Carlo Steps which is during the period of stuttering nucleus growth. The fluid-like particles show a second nucleus started growing with a slightly different orientation from the main nucleus which is a predominantly a mixture of bcc-like and hcp-like particles, slowing down the initial nucleus growth. This again is not in line with what is expected from nucleation growth, as described in Classical Nucleation theory, rendering this state point not appropriate for further crystal nucleation studies.

For the remaining state points spontaneous complete nucleation did occur using the brute force Monte Carlo simulation, all in about the same time-frame after $400k$ Monte Carlo steps. These state points will be the focus of the remainder of this work. To better describe the nucleation process the averaged bond order parameters of the configurations are calculated for which Principal component Analysis is done.

3.1 PCA projection

The method used for classifying the crystalline structures in the configurations during simulations is Principal Component Analysis of the Bond order parameters. First a PCA is done of the basic crystal structures, fcc, bcc and hexagonal close packed (hcp) and the fluid phase configuration to gain a better understanding of how to classify the different crystal phases formed during nucleation. The results of the PCA of the fluid, fcc, hcp and bcc configuration of the GCM particles are displayed in 5, where the absolute values of the coefficients of the first and second principal components are displayed. Clearly the 4th and 6th averaged bond order parameter are of importance in describing the crystalline structures, as expected, however the 8th and 12th averaged bond order parameter also have a high absolute value for the PCA thus these components are also of significance when describing the data set. As can be seen in the bar plot the first three averaged bond order parameters are negligible when describing the crystalline structures, as expected since the first three spherical harmonics are too simple to accurately describe the positions of point particles. Using the coefficients of the eigenvectors of the data set, or principal components, the four configurations are projected on the two-dimensional principal component space as can be seen in figure 5a.
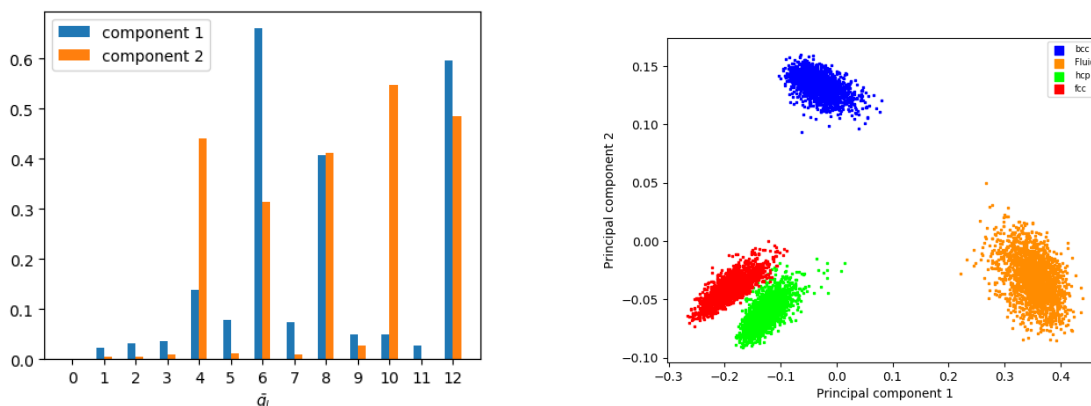
(a) Absolute values of the coefficients of the first and second principal component for the averaged bond order parameter.

(b) $\bar{q}_l$ of the fluid, hcp, fcc and bcc configurations projected on the first two principal components of the dataset.

Figure 5: Principal Components analysis of averaged bond order parameters of the fluid, fcc, bcc and hcp GCM particle configurations.

The projections of the phase clouds are well separated in the PCA space, which makes it possible to distinguish the different crystalline structures for the configurations obtained from simulations. The fluid phase is completely separated from the crystalline phases, and also the bcc phase is nicely separated from the hcp and fcc phase. Therefore the PCA plane of the first two principal components is suitable for identifying crystal structures from the fluid phase during nucleation, and also the bcc crystal structure can be identified from the other crystal phases. However to identify whether a hcp or fcc crystal structure is present the projection should be more seperated. By introducing the 3rd principal component, and therefore projecting the configurations on the second and third principal component, the fcc phase and hcp phase are seperated by the third component in figure: 6.
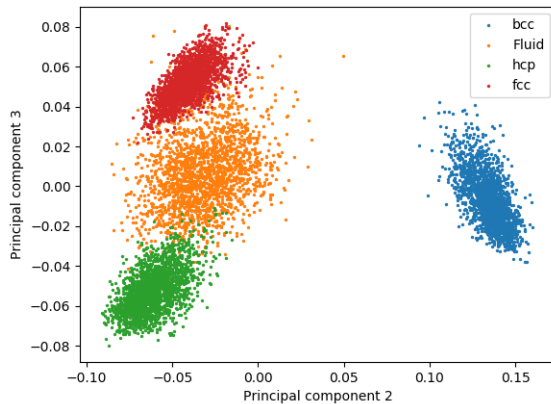
Figure 6: $\bar{q}_l$ of the fluid, hcp, fcc and bcc configurations projected on the second and third principal component of the dataset.

Using the second and third principal component the fcc phase is nicely seperated from the hcp space in the PCA space making the third component more suitable for classifying hcp structures from the fcc structures. To map the overlap between the different principal components we calculated the probability distribution of the first four principal components displaying the results in figure: 7.
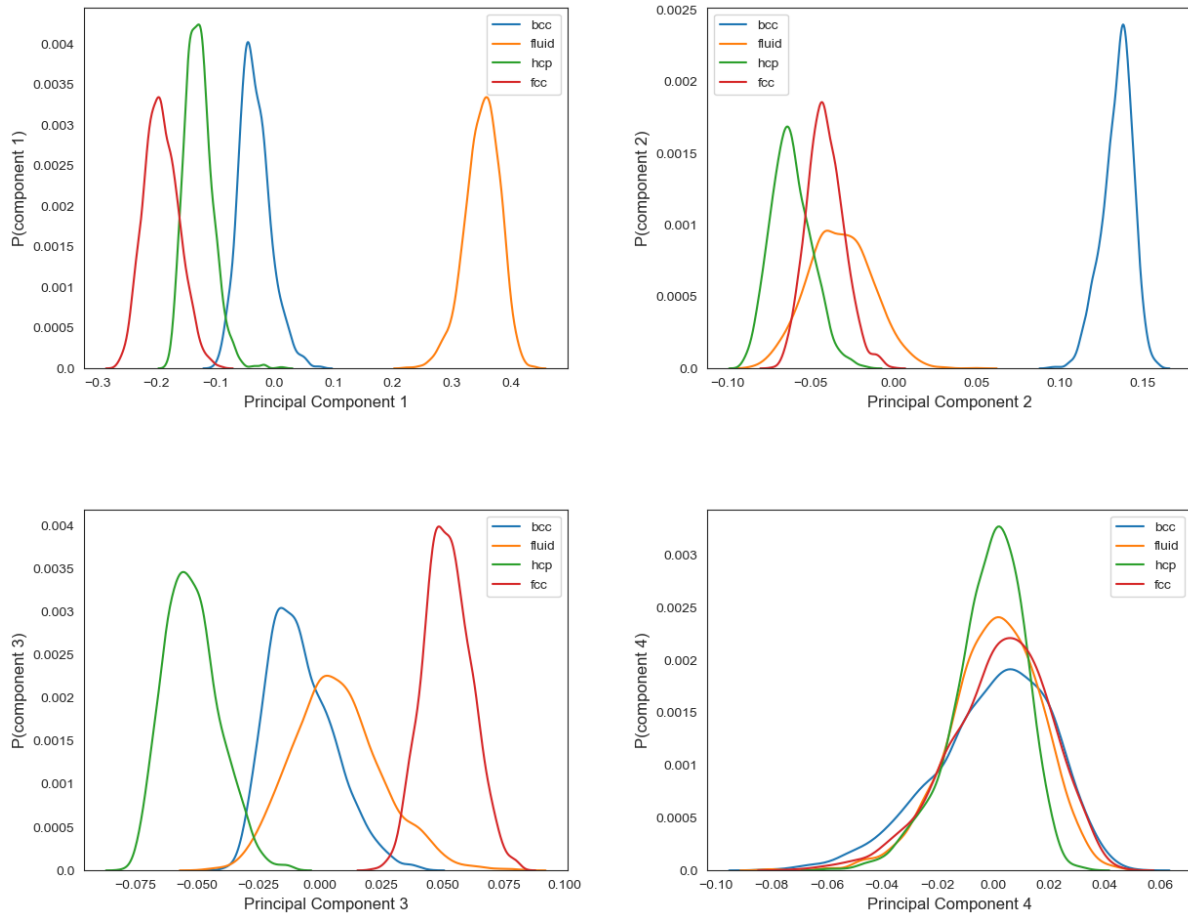
Figure 7: Probability distributions of the first four Principal components using the averaged bond order parameters of the fluid, fcc, bcc and hcp phase of Gaussian core particles. From left top to right bottom in order, the first, second, third and fourth Principal component.

The first principal component has very steep and sharp peeks in the probability distribution for all the different configurations. The fluid phase has no overlap with the crystal phases, as expected. Also the bcc and fcc phase show little overlap when using the first principal component. We can conclude that the fluid, bcc and fcc phase are different in principle. The fcc and hcp phase are very similar in structure according to the first principal component. When looking at the second principal component, the bcc phase has the most affiliation with this component, distinguishing itself from the other phases. The graph for the fluid phase shows a relatively wider peak making it more difficult to distinguish the fluid phase from the other phases just by the second principal component. The aforementioned similarity in structure between the hcp and fcc crystal structure is resolved by the third principal component where still relatively sharp peeks in the probability distribution for both structures are well seperated. To accurately distinguish the hcp structures from the fcc structures we predict the third principal component will be crucial. The bcc and fluid phase fail to show characteristics in the third principal component probability graph. Each phase is predicted to be able to accurately be distinguished using the right set of principal components, however this does not stop us from looking further into the probability distributions of the other principal components. The other principal components, for example the fourth, are all not interesting when describing the crystalline structures. The phases almost completely overlap and get wider for each iteration of higher principal component. For the remainder of this work we will focus on the first three principal components.
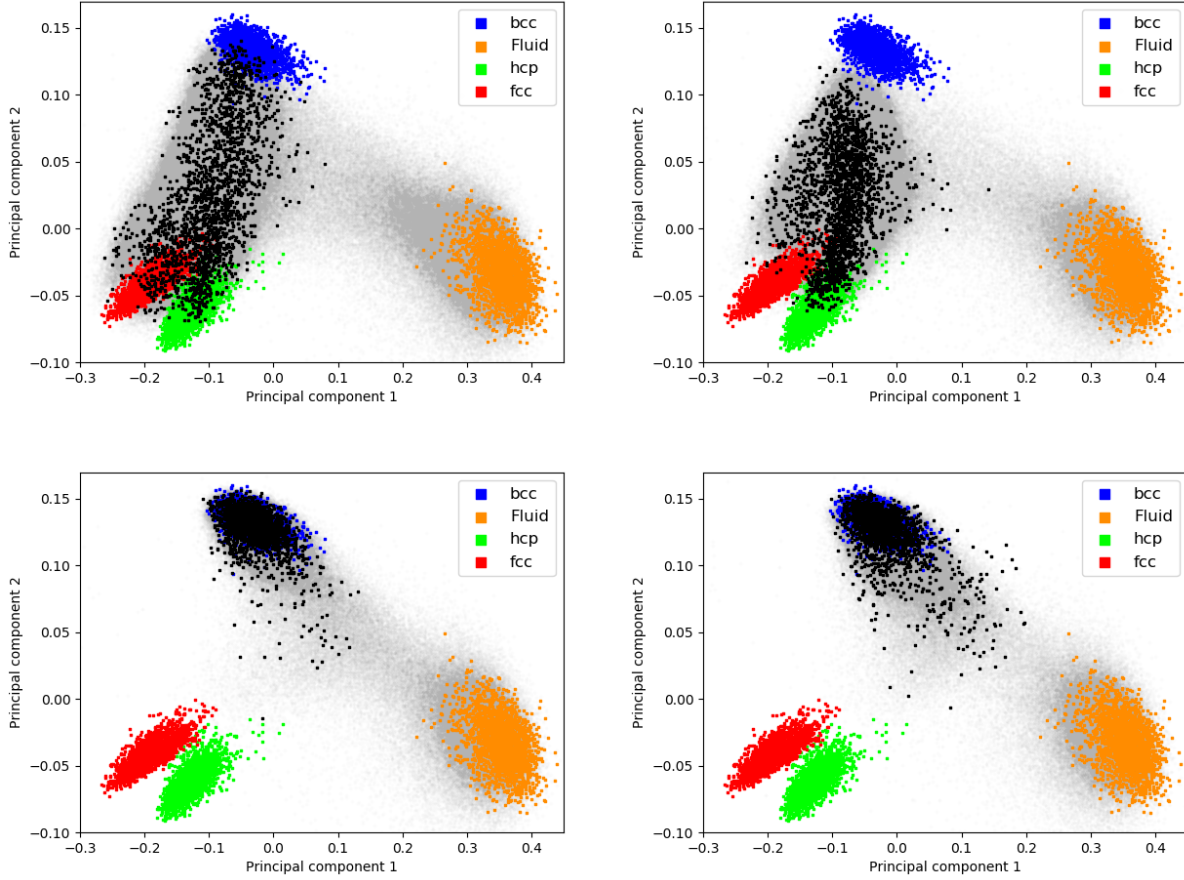
Figure 8: Projection of averaged bond order parameters of simulations onto PCA space of first two principal components. The PCA space is spanned by running Principal Component Analysis on configurations of fluid, fcc, bcc and hcp ordered particles. Left-top state point $T = 0.002, P = 0.02$ after 3640k Monte Carlo steps, right-top state point $T = 0.003, P = 0.05$ after 3290k Monte Carlo steps, Left-bottom state point $T = 0.004, P = 0.05$ after 1990k Monte Carlo steps, right-bottom state point $T = 0.005, P = 0.05$ after 1990k Monte Carlo steps. The light grey cloud denotes all the configurations during the duration of the simulation, to get an understanding of the regions of phase space the configuration passed through. The dark points are the configuration at the final snapshot of the simulation.

The principal components of the averaged bond order parameters can distinguish the different crystal phases for the configurations. To track what type of crystal structures form during the nucleation of our simulations the averaged bond order parameters of the configurations are projected onto the PCA space spanned by the first two principal components. This way it is possible to accurately distinguish solid-like from fluid-like particles, and to distinguish the bcc phase from the fcc and hcp phase. Afterwards the hcp and fcc phase can be distinguished using the third principal component.

The result of the projection of the state point $T^* = 0.002, P^* = 0.02$ is shown in figure 8 top-left where snapshots of the configuration during the nucleation are projected on the PCA space. From the snapshot it is not clear at what crystalline phase the configuration ended at, the projected PCA cloud ends up between the bcc and fcc,hcp clouds, only slightly overlapping with the crystal phases. The projected cloud of this configuration is relatively large and sparse compared to the other phase clouds. This indicates the structure for the configuration consists of a wide range of BOP values which is not as expected for a bulk crystal structure, but the configuration does consist of solid-like particles. The expected behaviour for the state point $T^* = 0.002$ and $P^* = 0.02$ is to nucleate to a bcc crystal first before nucleating as a fcc crystal like Russo et al.[**russo**] and others found.

15

The snapshot for state point $T^* = 0.003, P^* = 0.05$ for which the bcc phase is the stable phase, displayed in 8 shows a similar behaviour, the PCA cloud does not overlap well in the first and second principal component space. Running the simulation for a longer period of time does not seem to change the structure of the configuration when looking at the PCA cloud which does not seem to change anymore. To make a better distinction between the hcp and fcc like particles in the configuration, the set of averaged BOP values is projected onto the first and third principal component in figure 9:
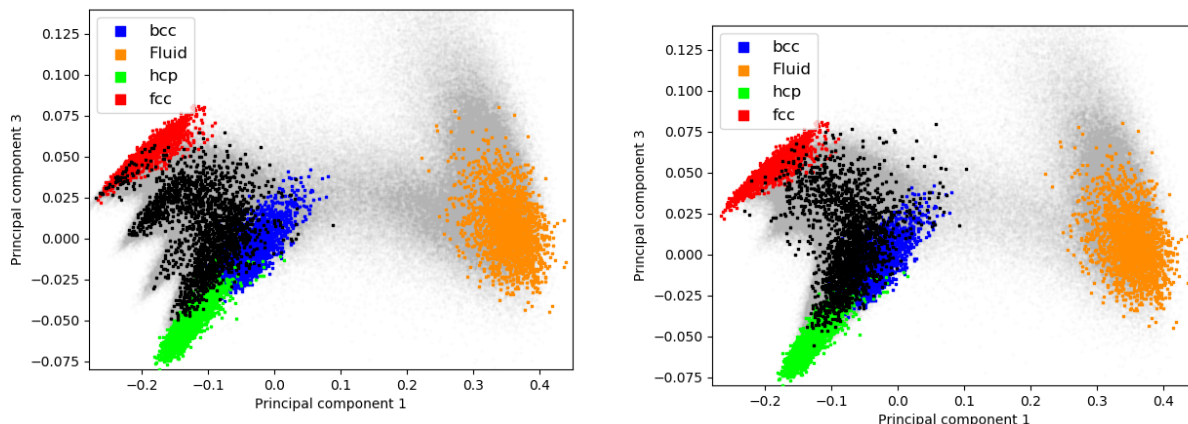

Figure 9: Projection of averaged BOP onto the first and third Principal Component. State point $T = 0.002, P = 0.02$ (left) and $T = 0.003, P = 0.05$ (right). The light grey cloud denotes all the configurations during the duration of the simulation, to get an understanding of the regions of phase space the configuration passed through. The dark points are the configuration at the final snapshot the simulation.

In projecting onto the third component, the distinction between the fcc and hcp structure became more clear. However the projections of the state points did not have a better overlap with the phase clouds. For the state point $(T^* = 0.003, P^* = 0.05)$ the configuration cloud is more compact than that of state point $(T^* = 0.002, P^* = 0.02)$, suggesting the BOP values are more representative than the more sparse phase cloud. Still the results from the projections for these state points have very few overlap with the known stable crystal structures, and most of the projected BOP values ended up in between the predicted phase clouds. We suggest that the used stable structures of the PCA space are too different in structure than the formed crystal configuration of the simulations, since the overlap of the configuration is minimal. With the present PCA projection it is not suitable to quantify the degree of bcc, hcp or fcc ordering for these two state points. Considering we had only one simulation in the stable fcc region to successfully nucleate, we suggest more simulations should be done less close to the border between the fcc and bcc stable region. For example a state point $T^* = 0.001, P^* = 0.01$ could be insightful.

The state points $T^* = 0.004, P^* = 0.05$ and $T^* = 0.005, P^* = 0.05$ display similar behaviour in the PCA space as displayed in figure 8 bottom-left and bottom-right resp. The configuration has a good overlap with the bcc phase cloud in the PCA space after nucleating from the fluid phase. After 2 million Monte Carlo steps we conclude the state points have nucleated to the meta stable bcc phase. The fluid phase quickly nucleated to the bcc phase without much of a diversion, but rather a straight line towards the bcc phase. To better understand how much the different configurations overlap with the phases, the Mahalanobis distance is calculated from the center of the configuration PCA cloud towards the center of the respective phase clouds. The distance is displayed over time in figure 10.
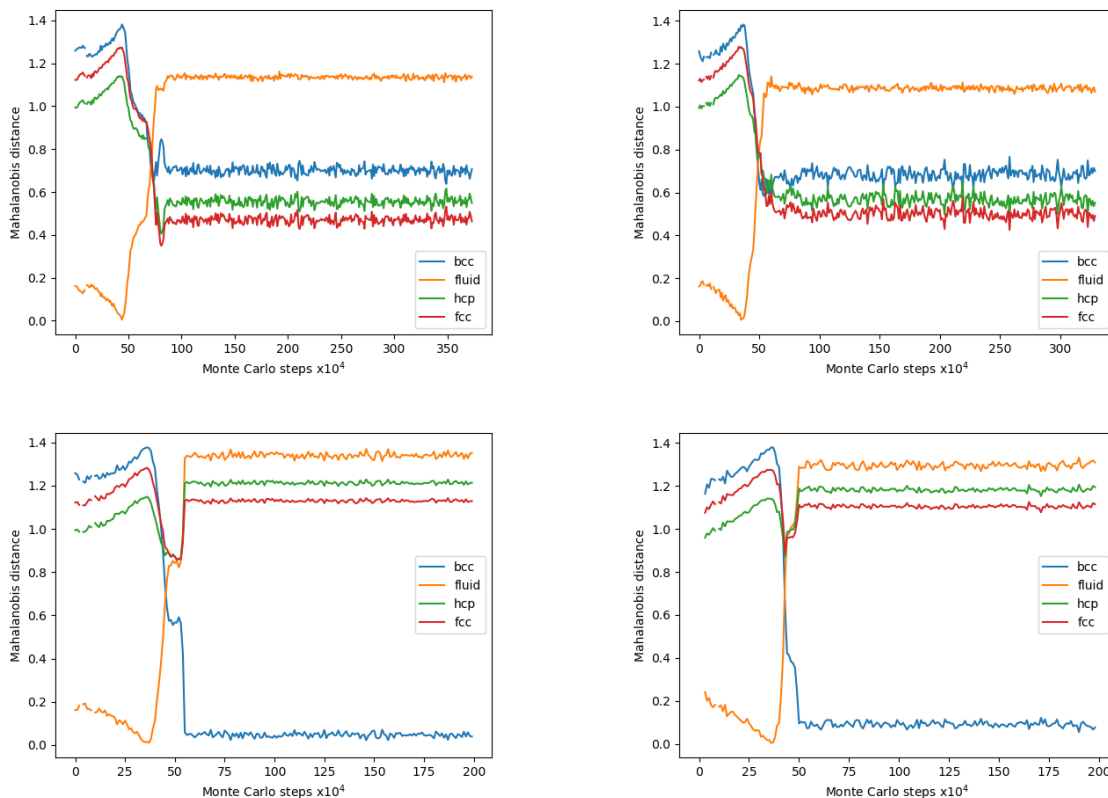
3.2 Mahalanobis distance

Figure 10: Mahalanobis distance in the first and second Principal Component space. The distance is calculated from the middle of the relatives phase clouds. State points $T = 0.002, P = 0.02$ (top-left), $T = 0.003, P = 0.05$ (top-right), $T = 0.004, P = 0.05$ (bottom-left) and $T = 0.005, P = 0.05$ (bottom-right)

For the state points $T^* = 0.002, P^* = 0.02$ and $T^* = 0.003, P^* = 0.05$ the Mahalanobis distance in figure: 10 does not clarify any further what type of crystal structure formed during the simulation. The values for the Mahalanobis distance relative to the bcc, fcc and hcp phase do not variate much although the fcc and hcp phase do seem to be closer than the bcc crystal phase. Noticeable is the Mahalanobis distance to the fluid phase decreasing over the first 500k Monte Carlo steps of the simulation, where we expect it to be zero to the fluid phase at the start, since the nucleation of the configuration starts from the fluid phase. This means that the averaged BOP values of the fluid phase and the starting configuration do differ apparently, this discrepancy is probably due to the starting configuration which has a relatively large volume compared to the fluid phase used in the PCA projection.

The fluid used for the simulations is not an equilibrated metastable fluid at the state point which the simulation is started at. By running the simulation the fluid is first equilibrated at the state point, making the the configuraiton better overlap with the fluid phase in the PCA space. While the simulation is running, the configuration is compressed, decreasing the volume of the simulation box. After the configuration is compressed enough, the particles start ordering itself in a lower free energy state at which the phase starts nucleating, and thus the ordering increasing which is reflected in the Mahalanobis distance to the fluid phase rapidly increasing.

A snapshot of both state points is displayed in figure: 11 as the top two configurations. The hcp-like and fcc-like particles are much more present than in the bottom two configurations where the bcc-like particles are dominant. Although the simulation volume does seem to be less homogeneously nucleated, with cluster of hcp-like and fcc-like particles distributed in clusters over the simulation volume.
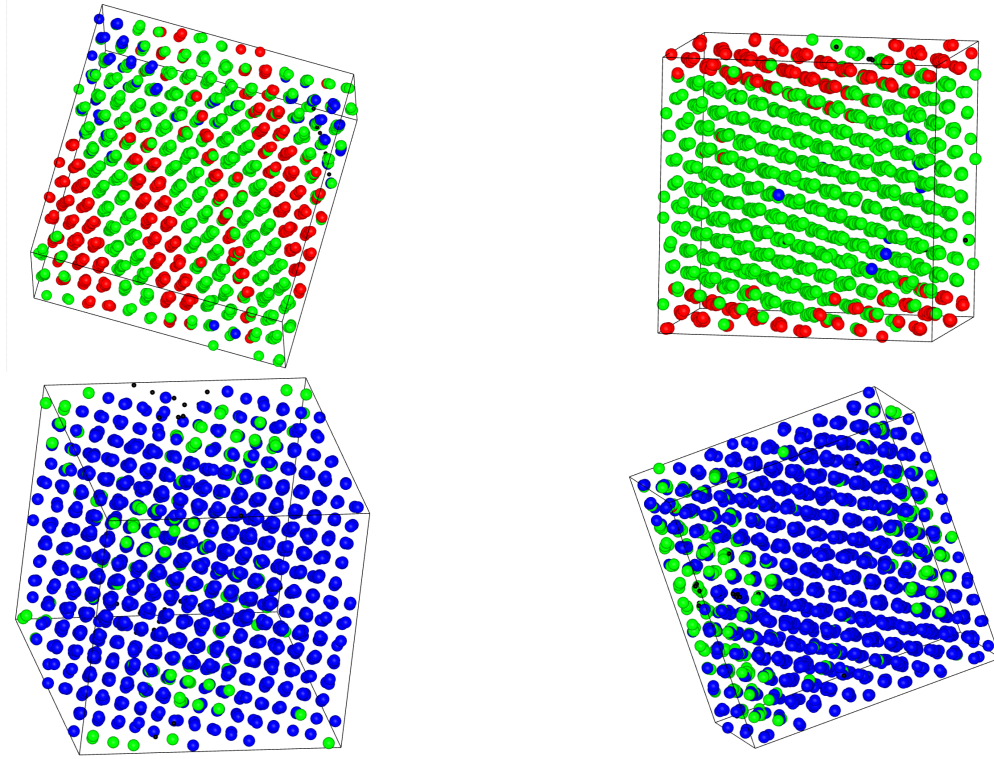
Figure 11: Snapshots of the configurations at the end of the simulation. State points $T = 0.002, P = 0.02$ (top-left), $T = 0.003, P = 0.05$ (top-right), $T = 0.004, P = 0.05$ (bottom-left) and $T = 0.005, P = 0.05$ (bottom-right). The fluid like particles are displayed as small black spheres, together with the fcc-like (red) particles, bcc-like particles (blue) and the hcp-like (green) particles. Two top two snapshots showing a mixture of mostly hcp-like and fcc-like particles, where the bottom two snapshots are dominated by bcc-like particles.

For the state points $T^* = 0.004, P^* = 0.05$ and $T^* = 0.005, P^* = 0.05$ the Mahalanobis distance in figure:10 shows a much clearer picture of the nucleating phase than just the PCA projection of the configuration. It is clearly visible, after the formation of solid-like particles, thus the rapid decrease of fluid-like particles, the bcc-like particles have formed. The Mahalanobis distance of both the configurations to the bcc phase remain stable, with little fluctuations, but not completely zero. The simulated configuration probably will never completely overlap with the bcc phase and thus the distance will remain non-zero.

The configurations are displayed in the bottom half of figure 11, with the bcc-like particles being dominantly present over the other particles. The bcc-like particles seem to homogeneously have distributed itself over the simulation volume, giving the indication that the majority of the crystal formed is bcc stacked. we conclude that the configuration for this state point nucleated to a bcc crystal in the bcc stable phase space without Ostwald step rule[10]. This conclusion is in line with the findings of Russo et al.[**russo**] who found in their experiment the particles in the stable bcc phase to nucleate to the bcc phase.

4 Conclusion and outlook

The understanding of the crystallization process of the GCM particles is important, since these type of particles are very common in nature and every day life, and should be better understood.

Based on literature we expected the GCM fluid to nucleate via a bcc order structure into a fcc crystal, and for the nucleation in the bcc phase region to happen straight into a bcc crystal without Ostwald's step rule. A principal Component Analysis was performed to distinguish fluid, bcc, fcc and hcp like particles, which deemed possible when looking at the probability distributions of the different principal components. However when projecting the simulated configurations for the different state points, only the state points nucleating in the bcc stable phase region gave convincing results, where the resulting phase cloud overlapped nicely with the bcc cloud in the PCA space. For the nucleation in the fcc stable phase region, the PCA projection showed fewer promising results, forming a projection between the known stable crystal structures in the PCA space, suggesting a mixture of the different crystal structures had formed.

To quantify the results of the PCA projection, the Mahalanobis distance in the PCA space was calculated between the middle of the configuration cloud and the respective stable phase clouds. Using the results of the Mahalanobis calculation we concluded the GCM fluid nucleated into a bcc structured crystal in the bcc stable phase region. Although the PCA projection method deemed insightful in quantifying the classification of the bcc ordered crystal, the difficulty in identifying the mixture of crystals formed, showed evidently. More investigation needs to be done to asses the problems and obstacles of crystal ordering during crystal nucleation. Also more simulations should be run at lower pressure, deeper in the stable fcc phase region to be more certain it was not a borderline case between the bcc and fcc stable phase region that was discussed. The problems we faced during the analysis of the results show that the crystallization process should be assesed with great care, taking into consideration different analysis models to accurately distinguish the different crystal orderings.

References

[1] B. J. Alder and T.E. Wainwright. "Phase Transition in Elastic Disks". In: *Phys. Rev.* 127 (1962), p. 359.

[2] S. Alexander and J. McTague. "Should all crystals be bcc? Landau Theory of Solidification and Crystal Nucleation". In: 41.10 (1978), pp. 702–705.

[3] D. Frenkel and B. Smit. "Chapter 5 - Monte Carlo Simulations in Various Ensembles". In: *Understanding Molecular Simulation (Second Edition)*. Academic Press, 2002, pp. 111–137.

[4] J.P. Hansen and D. Schiff. "Influence of interatomic repulsion on the structure of liquids at melting". In: *Mol. Phys.* 25 (1973), p. 1281.

[5] T. Kawasaki and H. Tanaka. "Formation of a crystal nucleus from liquid". In: *Proceedings of the National Academy of Sciences* 107.32 (2010), pp. 14036–14041. ISSN: 0027-8424.

[6] H. C. Longuet-Higgins and B Widom. "A rigid sphere model for the melting of argon". In: *Mol. Phys.* 8 (1964), pp. 549–556.

[7] P. C. Mahalanobis. "On the generalised distance of statistics." In: 2 (1936), pp. 49–55.

[8] J. A. van Meel et al. "A parameter-free, solid-angle based, nearest-neighbor algorithm". In: *The Journal of Chemical Physics* 136.23 (2012), p. 234107.

[9] N. Metropolis et al. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092.

[10] W. Ostwald. "Studien uber die bildung und umwandlung fester korper". In: *Z. Phys. Chem* 22 (1897), pp. 289–330.

[11] S. Prestipino, F. Saija, and P. V. Giaquinta. "Phase diagram of the Gaussian-core model". In: *Phys. Rev. E* 71 (5 May 2005), p. 050102.

[12] C.N.R. Rao and K.J. Rao. "Phase transformations in solids". In: *Progress in Solid State Chemistry* 4 (1951). ISSN: 0079-6786.

[13] J. Russo and H. Tanaka. "Selection mechanism of polymorphs in the crystal nucleation of the Gaussian core model". In: *Soft Matter* 8 (15 2012), pp. 4206–4215.

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

[15] F. H. Stillinger. "Phase transitions in the Gaussian core system". In: *The Journal of Chemical Physics* 65.10 (1976), pp. 3968–3974.

[16] P.R. ten Wolde, M. J. Ruiz-Montero, and D Frenkel. "Simulation of homogeneous crystal nucleation close to coexistence". In: *Faraday Discuss.* 104 (0 1996), pp. 93–110.