LU TP 20-02
January 2020

Liquid-Liquid Phase Separation of HP Lattice Proteins: A Finite-Size Scaling Analysis

Markus Ernstsson

Department of Astronomy and Theoretical Physics, Lund University

Master thesis supervised by Anders Irbäck

Abstract

Biomolecular condensates are dense droplets of proteins and nucleic acids inside living cells that form through a liquid-liquid phase separation (LLPS) process, in which intrinsically disordered proteins (IDPs) often play an important role. Furthermore, it has been shown that several of these IDPs can form similar droplets on their own. Understanding the forces driving the LLPS of IDPs, and how the process depends on the amino acid sequence, is a challenging task. One difficulty is that the systems amenable to computational modelling are limited in size. It is therefore important to analyse and understand how simulated properties depend on the system size. A finite-size scaling theory for droplet formation through phase separation exists. This thesis explores the usefulness of this theory in the study of biomolecular LLPS, using a minimal lattice-based hydrophobic-polar (HP) protein model. By Monte Carlo methods, computer simulations of two HP sequences of length 10 are conducted for a range of system sizes, with up to 640 chains. A finite-size scaling analysis of the simulation results reveals that only one of the two sequences undergoes LLPS. Furthermore, it is found that the temperature at which droplet formation sets in converges slowly to its value for infinite system size. Hence, finite-size scaling analysis is useful both in deciding the phase behaviour of the sequences and in determing the underlying phase diagram.

Populärvetenskaplig sammanfattning

I våra celler finns en rad olika organeller som samlar upp molekyler för att sedan utföra olika uppgifter — likt mikroskopiska varianter av kroppens organ. Klassiska organeller har ett membran som separerar deras inre från cellens miljö. Det finns också organeller som saknar ett omgivande membran. Dessa droppliknande anhopningar av proteiner och nukleinsyror kallas biomolekylära kondensat. De kan bildas och lösas upp spontant, till följd av exempelvis en ändring i temperatur. Vid bildandet separeras en vätskelik droppe med hög täthet, det biomolekylära kondensatet, från en omgivning som också är vätskelik men har lägre täthet — tänk en oljedroppe i vatten.

Proteiner är långa kedjor uppbyggda av olika aminosyror, och bestäms av sekvensen av aminosyror längs kedjan. Många proteiner har en kompakt och väldefinierad tredimensionell struktur, som är viktig för funktionen. Det finns dock också proteiner som i sin fungerande form är strukturellt oordnade. Vid bildandet av biomolekylära kondensat har det visat sig att strukturellt oordnade proteiner ofta spelar en central roll. Dessutom har man funnit att vissa av dessa oordnade proteiner har förmågan att bilda likartade droppar på egen hand. Denna förmåga beror på aminosyrasekvensen — om aminosyrorna kommer i annan ordning längs kedjan kan förmågan gå förlorad. Hur det kommer sig att vissa proteiner har, medan andra inte har, förmågan att bilda denna typ av droppar är viktigt att klarlägga för att förstå mekanismerna bakom bildandet av biomolekylära kondensat inuti celler.

Att modellera bildandet av proteindroppar är en utmaning p.g.a. systemens storlek och komplexitet. För att undersöka de grundläggande drivkrafterna används därför ofta förenklade modeller, där proteiner beskrivs som länkade kedjor av kulor och varje kula representerar en aminosyra, snarare än en enskild atom. I dessa modeller används också en förenklad beskrivning av hur aminosyror attraherar eller repellerar varandra. Med en modell på denna detaljnivå kan datorsimuleringar av relativt stora system utföras. Dock når man inte systemstorlekar motsvarande verkliga proteindroppar. För att kunna dra slutsatser från simuleringarna om droppbildning krävs därför att man analyserar och förstår hur resultaten beror på systemstorleken. För detta ändamål finns lyckligtvis ett teoretiskt ramverk som tidigare har testats med framgång på enkla modeller för kondensation av vanliga vätskedroppar.

I detta examensarbete används detta teoretiska ramverk för att analysera simuleringsresultat för två korta aminosyrasekvenser i en enkel proteinmodell, erhållna för en rad olika systemstorlekar. Resultaten för små system antyder möjligheten att båda dessa sekvenser bildar droppar liknande biomolekylära kondensat. En systematisk analys av systemstorleksberoendet visar dock att så är fallet bara för den ena av de två sekvenserna. För denna sekvens äger vidare droppbildning rum bara när temperaturen understiger ett visst tröskelvärde. Denna tröskeltemperatur visar sig bero relativt starkt på systemstorleken. De teoretiska förutsägelserna gör det möjligt att från simuleringsresultaten på ett kontrollerat sätt uppskatta tröskeltemperaturens värde i stora system.

Contents

List of Figures

1 Introduction

In eukaryotic cells, there are specific types of structures known as biomolecular condensates. These structures can be described as membraneless organelles that serve the purpose of concentrating proteins and nucleic acids for a multitude of different cell processes [1, 2]. Biomolecular condensates are spherical structures which are able to coalesce and also exchange contents rapidly with their surroundings. They are liquid-like droplets of proteins and nucleic acids, and form through a liquid-liquid phase separation (LLPS) process. In many cases, it turns out that intrinsically disordered proteins (IDPs[1]) play a major role in this process. Furthermore, it has been shown that several different IDPs can go through a LLPS process on their own [3–5]. However, the forces driving the LLPS of IDPs, and the dependence of the process on the amino acid sequence, remain elusive.

Understanding the underlying reasons for the observed phase behaviour of IDPs and their role in LLPS is of great importance. There are several theoretical and computational methods for studying and modeling LLPS. A widely used analytical method is the Flory-Huggins mean field theory [6, 7]; however, it does not take into account the amino acid ordering along the chain. A different field of methods which do consider the ordering is molecular simulation with explicit chains. These computational methods remove the need of certain approximations made in analytical models while making structural properties easily available. Therefore, despite being computationally costly, an increasing number of explicit-chain simulations of biomolecular LLPS has been reported in recent years [8–14]. Another recent method involves the use of statistical field theory. By transforming the polymers into fields, the formation of biomolecular condensates can be explored by means of field theory simulations [15, 16]. This approach may alleviate the system size constraint present in explicit-chain simulations.

However, no matter how one simulates a system, it must be finitely sized. Hence, understanding how properties of the system depend on its size is of utmost importance. A quantitative approach to this problem is offered by finite-size scaling theory [17–19]. The usefulness of this approach in the study of biomolecular LLPS was recently explored by Nilsson and Irbäck, using a simple off-lattice hydrophobic-polar (HP) protein model [20].

This thesis investigates droplet formation in the minimal lattice-based HP protein model of Lau and Dill [21], using finite-size scaling theory. In this coarse-grained model, proteins are realised as chains of beads, with each bead being either H or P. Using Monte Carlo (MC) methods, the phase behaviour of two distinct sequences is examined: an alternating sequence, called Φ (HPHPHPHPHP), and a "blocky" sequence, called Θ (HHHHHPPPPP). For each sequence, several different system sizes (up to 640 chains) are simulated, keeping the monomer concentration, ρ, fixed (1%). To determine whether or not these sequences undergo LLPS, the finite-size scaling properties of the specific heat are investigated. Additionally, the finite-size scaling analysis makes it possible to investigate the shape of the underlying phase diagram in a systematic and controlled manner. Through the finite-size scaling analysis, it is found that one of the sequences, Φ, undergoes LLPS, while the other, Θ, does not. The same conclusion was reached by Nilsson and Irbäck [20], who

[1] IDPs are proteins that do not have to fold into a specific conformation to function.

studied the same two sequences using a different model.

If one considers chains of length one where the single bead is an H, then the HP protein model reduces to the so-called lattice gas (LG) [22]. The LG system is equivalent to the standard Ising model with ferromagnetic nearest-neighbour interactions, which has been extensively studied. This equivalence makes the LG system an ideal testbed for studies of droplet formation. Indeed, Zierenberg and Janke recently performed a finite-size scaling analysis of droplet formation in the LG system [23].

Therefore, to test our methods developed for HP proteins, a finite-size scaling analysis of the 3D LG model is performed and the results are compared to those of Zierenberg and Janke [23] as well as to existing data for the Ising model. Generally, good agreement is found with the results of Zierenberg and Janke, with a minor difference concerning the scaling of the maximum specific heat.

This thesis is organised as follows. Section 2 describes the biophysical model and numerical methods used, and outlines the finite-size scaling theory for droplet formation developed by Refs. [17–19]. Section 3 presents the results obtained for the LG system and for the two HP proteins Φ and Θ. The thesis ends with a discussion and summary in sec. 4.

2 Methods

2.1 The Ising Model

The Ising model is an extensively studied model of phase transitions. It models magnetism using nearest-neighbour interacting spins in a lattice. The Hamiltonian of the Ising model can be written as

$$H = -J \sum_{\langle ij \rangle} \sigma_i \sigma_j - \mu_{\mathrm{m}} h \sum_i \sigma_i, \tag{2.1}$$

where $\sigma_i = \pm 1$ is the value of spin i, J is the interaction energy between spins, μ_{m} is the magnetic moment, and h is an external magnetic field. The first sum is over all pairs of spins σ_i and σ_j that are nearest neighbours on the lattice. It has been assumed that all spins have the same interaction energy and that all experience the same external magnetic field. Depending on the value of J, one can model different types of magnetic behaviours: $J > 0$ yields ferromagnetism whereas $J < 0$ results in antiferromagnetism. Throughout this thesis, the focus is on the ferromagnetic case. A quantity of interest in the Ising model is then the magnetisation per spin,

$$m = \frac{1}{N} \sum_i \sigma_i, \tag{2.2}$$

where N is the total number of spins (or lattice sites since they are equivalent). With this so-called order parameter, it is possible to study the phase behaviour of the Ising model.

As already stated, the Ising model exhibits phase transitions, but this only occurs in two dimensions and higher. Ising showed that there are no phase transitions for the one dimensional case [24], and Onsager proved that the two dimensional case did have transitions [25].

These conclusions were both obtained through exact solutions. For higher dimensions, there are only numerical approximations such as low/high-temperature expansions [26,27], renormalisation group theory [28], or Monte Carlo simulations [29].

In two and higher dimensions, for zero external magnetic field and low temperature, it turns out that two distinct phases coexist, characterised by magnetisations $\pm m$. The system tends to be in either of these two distinct states. As the system transitions between the states, there is a discontinuity in the magnetisation, which means that the system undergoes a first-order phase transition. Increasing the temperature, there will be a point at which the magnetisation becomes zero, beyond which only a single phase exists. The temperature at which this occurs is known as the critical temperature, T_c. The transition from two distinct phases to only one, i.e. crossing the critical temperature, will lead to a divergence in the specific heat,

$$\frac{C_V}{N} = \frac{1}{N}\frac{\mathrm{d}E}{\mathrm{d}T}, \tag{2.3}$$

which is a second-order phase transition.

2.2 The Lattice Gas Model

A simple model for gases (and liquids) is the lattice gas (LG) model [22]. It consists of particles on a lattice with nearest-neighbour interactions. Each lattice point can be occupied by at most one particle. Therefore, the energy function of the LG model can be written as

$$E = -K\sum_{\langle ij\rangle} n_i n_j, \tag{2.4}$$

where $n_i = 0, 1$ is the occupation number of lattice site i and K is the interaction energy between particles. A model property when studying the phase behaviour of the LG model is the concentration,

$$\rho = \frac{1}{N}\sum_i n_i, \tag{2.5}$$

where N is the total number of lattice sites. In the grand canonical ensemble, the LG Hamiltonian can be written as

$$H = -K\sum_{\langle ij\rangle} n_i n_j - \mu\sum_i n_i, \tag{2.6}$$

where μ is the chemical potential. In this ensemble, the concentration is no longer fixed, since both the energy and number of particles is allowed to fluctuate.

If the occupation number n_i is rewritten in terms of the Ising model's spins σ_i through $n_i = \frac{1}{2}(1+\sigma_i)$, the LG Hamiltonian (eq. (2.6)) can be written as

$$H = -\frac{K}{4}\sum_{\langle ij\rangle}\sigma_i\sigma_j - \frac{K+\mu}{2}\sum_i \sigma_i + C, \tag{2.7}$$

where C is a constant. Putting $J \mathrel{\hat{=}} K/4$ and $\mu_\mathrm{m} h \mathrel{\hat{=}} (K+\mu)/2$ results in the Ising Hamiltonian, eq. (2.1), up to a constant. This means that the thermodynamic behaviour of the Ising model is the same as that of the LG model, since the constant does not influence the behaviour. However, it is important to note that the standard Ising model at given T and h corresponds to the grand canonical ensemble in the LG model.

As the LG model is equivalent to the Ising model, there exists a critical temperature T_c below which two phases coexist, characterised by high and low concentrations, $\rho_\mathrm{H}(T)$ and $\rho_\mathrm{L}(T)$, respectively. At a given temperature $T^* < T_\mathrm{c}$, there exists a specific value of the chemical potential which results in a first-order phase transition between the two phases. This is equivalent to the first-order transitions observed in the Ising model with zero external field $h = 0$.

If the concentration is instead fixed so that the total number of particles is constant, then the canonical ensemble is obtained. This is the same as keeping the magnetisation constant in the Ising model. In such an ensemble, for some temperature $T^* < T_\mathrm{c}$ and concentration ρ such that $\rho_\mathrm{L}(T^*) < \rho < \rho_\mathrm{H}(T^*)$, the LG system is in a mixed two-phase regime, bounded by the so-called binodal curve, as shown in fig. 1. This means that there is a region in the system with a high concentration, $\rho_\mathrm{H}(T^*)$, while the rest of the system has a low concentration, $\rho_\mathrm{L}(T^*)$. Hence, the system will inhabit both phases at the same time, with a droplet of high/low concentration existing in a background of low/high concentration. Similarly, an Ising magnet also exhibits droplets under equivalent conditions, although the droplets are defined by high or low magnetisation instead of concentration.

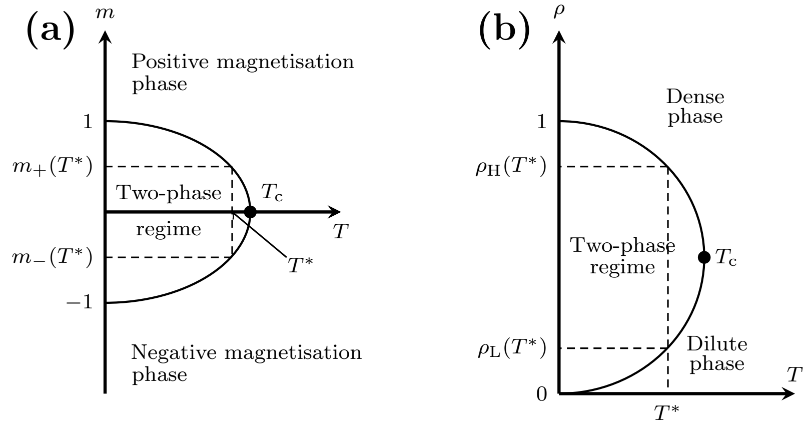


Figure 1: Schematic illustrations of the phase diagram of an Ising magnet and a lattice gas. The so-called binodal curve (solid line) is shown for both models where the maximum T-value is the critical temperature T_c. Two coexisting phases and the corresponding temperature T^* are shown. Note that T^* and T_c are not the same in the two different models. (a) The two phases of the Ising model are characterised by high and low magnetisations, $m_\pm$. (b) The two phases of the LG model are characterised by high and low concentrations, $\rho_{\mathrm{H}/\mathrm{L}}$.

2.3 The HP Model: A Minimal Explicit-Chain Protein Model

Studying real-life protein folding and droplet formation in a way that is as true-to-life as possible entails the use of all-atomic studies, where the resolution lies in the atomic world. This, however, also brings forth the problem of having a stupendous number of degrees of freedom in such model. Hence, fully atomistic models are very demanding in terms of computational power and time. It is therefore useful to simplify protein modelling with more coarse descriptions of proteins. The broader resolution is still enough to attain an insight into how proteins interact and form globular clusters. An extensively used simplified model is the minimal lattice-based hydrophobic-polar (HP) protein model [21].

The HP model consists of amino acid sequences of only two types of amino acids: hydrophobic (H) and polar (P). These sequences are represented as chains of beads which are self-avoidingly walking on a lattice. The beads interact with beads of other chains as well as beads on the same chain through nearest-neighbour interactions. Though not with nearest chain neighbours. However, it is only the H amino acid bead which is attractive; the P's do not interact with other beads. This leads to an interaction energy of the form

$$E = \begin{cases} -\epsilon, & \text{if both amino acids are hydrophobic} \\ 0, & \text{otherwise} \end{cases}, \tag{2.8}$$

where ϵ is an arbitrary positive energy scale. For each H-H interaction a factor of $-\epsilon$ is added to the total energy.

If a chain has a length of one where the single bead is an H, then the LG model is obtained with $K = \epsilon$. Therefore, the HP model can be used to simulate the LG model. Hence, an HP program can be tested by simulating one-bead H chains and comparing to existing LG and Ising results.

In this thesis, a program simulating the HP model in a three-dimensional cubic lattice with periodic boundary conditions is constructed and used. Each point in the lattice can be occupied by at most one bead and neighbouring beads on a chain are restricted to lie in neighbouring points on the lattice. For simplicity, units will be chosen so that $\epsilon = 1$. The main focus of this thesis will be two HP sequences of length 10: an alternating sequence Φ (HPHPHPHPHP), and a "blocky" sequence Θ (HHHHHPPPPP). All models will be simulated with a fixed concentration of $\rho = NN_c/V \approx 0.01$, where N is the number of chains, N_c is the number of beads on a chain, and V is the volume, i.e. the total number of lattice sites. Since the system is discrete, the concentration will not be exactly 1%.

2.4 Markov Chain Monte Carlo Basics

In a canonical system, the probability distribution of microstates s is the Boltzmann probability

$$P(s) = \frac{1}{Z} e^{-\beta E(s)}, \tag{2.9}$$

where $\beta = 1/(k_\mathrm{B}T)$, k_B is the Boltzmann constant, and the partition function $Z = \sum_s e^{-\beta E(s)}$ is a normalisation factor. Note that the expression is for discrete states, and as such, discrete energies. As the goal is to study the thermodynamical behaviour of HP chains, including single-bead H chains (LG model), the Boltzmann probabilities for all conformations need to be calculated. This is not a feasible task to perform analytically but by using Monte Carlo (MC) simulations, one can sample the correct distribution through importance sampling by generating configurations.

MC methods are algorithms that make use of random sampling. Independent samples are randomly drawn from the ensemble in order to estimate the probability of interest.

However, due to the large amount of degrees of freedom involved, this can not be done in practice. Instead, a Markov chain can be used as the basis of the MC method, a so called Markov chain Monte Carlo. A Markov chain is a sequence of events where the probability of each new state only depends on the current state. Thus, new states are generated as "perturbations" of the current state rather than being randomly drawn directly from the ensemble pool. This procedure can be described as having a transition probability

$$W(s \to s') = P(s'|s), \tag{2.10}$$

which states the conditional probability of a new state s' given that the previous state was s.

In order for eq. (2.10) to converge to the desired probability distribution, the Markov chain has to fulfill two conditions [30]:

- It is ergodic, i.e. any state s' can be reached from any other state s.
- The desired distribution $P(s)$ is stationary, that is $\sum_s W(s \to s')P(s) = P(s')$.

The second condition states that the desired distribution $P(s)$ is an invariant distribution of the chain, i.e. that $P(s)$ is an eigenvector to the transition probability $W(s \to s')$ with eigenvalue 1. This can be fulfilled in several ways, but an often useful property which satisfies the second condition is detailed balance. Detailed balance is fulfilled when the transition probability $W(s \to s')$ satisfies

$$W(s \to s')P(s) = W(s' \to s)P(s'). \tag{2.11}$$

This equation states that, in the stationary state $P(s)$, transitions from s to s' occur with the same frequency as transitions from s' to s.

A simple and general method for achieving detailed balance, and satisfy ergodicity, is the Metropolis-Hastings algorithm [31, 32]. There, the transition probability $W(s \to s')$ is split into two separate probabilities: a proposal probability $U(s \to s')$ and an acceptance probability $A(s \to s')$,

$$W(s \to s') = U(s \to s')A(s \to s'), \quad s \neq s'. \tag{2.12}$$

Equation (2.11) can then be rewritten as

$$\frac{A(s \to s')}{A(s' \to s)} = \frac{U(s' \to s)}{U(s \to s')}\frac{P(s')}{P(s)}. \tag{2.13}$$

There are several choices of acceptance ratios that satisfy eq. (2.13) [32]. The Metropolis choice,

$$A(s \to s') = \min\left[1, \frac{U(s' \to s)}{U(s \to s')}\frac{P(s')}{P(s)}\right], \tag{2.14}$$

will be used going forward. In the case where the proposal probabilities are symmetric, $U(s \to s') = U(s' \to s)$, and the Boltzmann distribution is used, eq. (2.14) simplifies to

$$A(s \to s') = \min\left(1, e^{-\beta\Delta E}\right), \tag{2.15}$$

where $\Delta E = E(s') - E(s)$.

2.5 Move Set

There are several types of moves or updates which can be used to simulate the HP model. The ones used in this thesis are

- One-bead updates,
- Pivot updates,
- Rigid-body updates of single chains and clusters of chains.

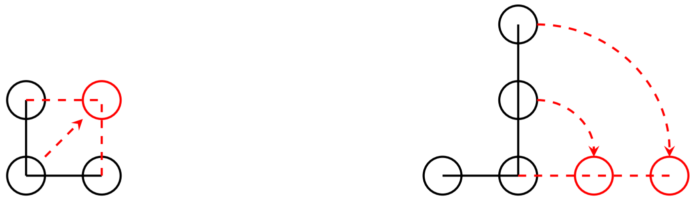
All updates are counted as moves no matter the outcome. The bulk of the updates are made up of one-bead moves, while cluster updates are the least common ones. The occurrence of pivot updates is dependent on the length of the chain: the longer the chain is the less likely it is to pivot. The rigid-body update only handles translations; rotations are handled by the other two types of updates.

2.5.1 One-Bead Update

A one-bead update works by trying to move a random single bead on a random chain. It operates differently depending on whether it is an end-bead or an internal bead. A bead on the end has four possible positions to move to, depending on if all of them are empty. An internal bead only has one possible position to move to, and only one type of move to get there: the kink-jump, as illustrated in fig. 2(a). Though, any move has to satisfy the acceptance probability as described in eq. (2.15).

2.5.2 Pivot Update

The pivot update operates by pivoting a random chain around a random bead on the chain. Which side of the chain that is kept still is randomly chosen in a uniform manner. The other side is then rotated or reflected in a random plane, as shown in fig. 2(b). Similarly to the one-bead update, the move has to satisfy the acceptance probability, eq. (2.15).



Kink-jump.

Pivot.

Figure 2: Illustrations of a kink-jump (a) and a pivot update (b).

2.5.3 Rigid-Body Updates

When chains are moving around they can form clusters. Sometimes, several clusters can form far from each other. In such cases, if there are only single-chain moves, then these clusters might never coalesce; they are too far apart for a single chain to stretch between them. In order to prevent isolation of clusters, cluster updates can be utilised.

A cluster update would need to build a cluster and then move it in some way. This can, in the most naive way, be to take a cluster (all the chains that share at least one H-H interaction), then try to move it and accept the move with the acceptance probability, eq. (2.15). However, this would violate detailed balance. If a cluster is joined with another cluster, then it is impossible to go from this new cluster back to two separate clusters using this naive method. In order to conserve detailed balance, there must not be any energetic changes.

A cluster update which conserves detailed balance while allowing the merging of two, or more, clusters is the Swendsen-Wang cluster algorithm [33, 34]. It is a fairly simple algorithm which dynamically builds a cluster according to the following steps:

1. Choose a random chain i as the seed of a cluster.
2. Each chain j in contact with i is added to the cluster with probability $1 - e^{\beta E_{i,j}}$, where $E_{i,j}$ is the total interaction energy between chain i and j.
3. Continue adding chains in accordance with step 2 for every new chain added to the cluster.
4. End when no more chains can be added.

This cluster is then trial-moved as a rigid body, and if it does not satisfy self-avoidance the move is rejected; otherwise, it is accepted. The built-in acceptance check relieves the external, post-move check. It also enables for single-chain cluster moves, i.e. a single chain moving.

For simulations of the LG model (chains of length one), there is only one type of update: single rigid-body translation. It works by randomly selecting a particle and then trying to translate it a random, short distance in a random direction. The reason for the restriction to this type of update for the LG model is twofold. Firstly, the LG simulations can reach the desired energy ranges with this update, and it is faster than including a cluster move as well. Secondly, only so-called multicanonical simulations are performed for the LG model, and to avoid too many special cases, cluster updates are omitted.

2.6 Multicanonical Simulation

An effective way to avoid the suppression of intermediate states, which is prone in canonical simulations, is to make use of so-called multicanonical simulations [35]. The principle behind such simulations is to search an extended energy range in one simulation run instead of combining several canonical simulations. What multicanonical entails in practice is a non-Boltzmann sampling. The multicanonical distribution P_{mc} is a reweighted Boltzmann distribution, $P_{\mathrm{c}}(E)$:

$$P_{\mathrm{mc}}(E) \propto W(E) P_{\mathrm{c}}(E), \tag{2.16}$$

where $W(E)$ is a reweighting factor. The canonical distribution of a phase separation is expected to be bimodal: two peaks corresponding to the two different states. For the multicanonical distribution, $W(E)$ is chosen so that $P_{\mathrm{mc}}(E)$ becomes "flat" in this bimodal structure of $P_{\mathrm{c}}(E)$. Figure 3 shows the difference of the two distributions for the LG model with 160 particles. The orange line shows the double-peak, i.e. bimodal, behaviour of the canonical distribution, whereas the blue line is the flattened multicanonical distribution. The rare intermediate states in the canonical ensemble are more likely in the multicanonical distribution, therefore, the energy landscape can be searched more effectively in the multicanonical distribution.

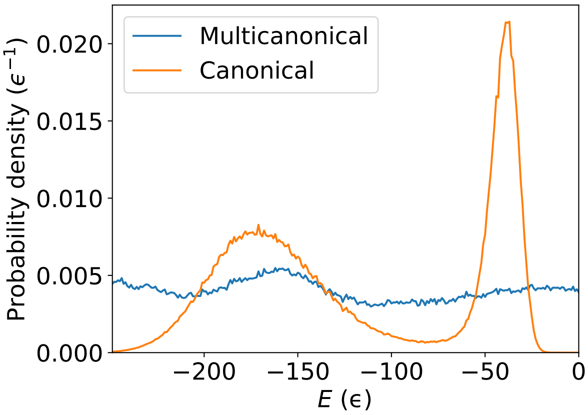


Figure 3: The jackknife estimated multicanonical distribution (blue) of the 3D LG model with 160 particles and $\rho = 1\%$, as well as the reweighted canonical distribution (orange) at the estimated transition temperature $T_{\mathrm{b}}^{(N)}$.

Choosing the reweighting factor is not trivial. There are several ways of constructing it [35]. One approach is to estimate the density of states $\Omega(E)$. The canonical distribution can be thought of as the density of states reweighted by the Boltzmann distribution,

$$P_{\mathrm{c}}(E) \propto \Omega(E)e^{-\beta E} \tag{2.17}$$

Therefore, for the multicanonical distribution to be "flat", or rather constant, its weight should be the inverse of the density of states,

$$P_{\mathrm{mc}}(E) \propto \Omega(E)\Omega^{-1}(E) \tag{2.18}$$

However, for complicated systems, $\Omega(E)$ is not a priori known; it must be estimated.

The Wang and Landau algorithm is a way to produce an estimate of the density of state, usually called $g(E)$ [36]. The algorithm starts with an initial guess, say $g(E) = 1$, then proposes a new state s' in the same way as the Metropolis algorithm and accepts it according to

$$A(s \to s') = \min\left\{1, \frac{g\left[E(s)\right]}{g\left[E(s')\right]}\right\}. \tag{2.19}$$

After the proposed state has been accepted or rejected, the histogram for the current state, $H\left[E(s)\right]$, is incremented, and the value of $g\left[E(s)\right]$ is updated in the following manner,

$$g\left[E(s)\right] \to fg\left[E(s)\right], \tag{2.20}$$

where f is a multiplicative factor. When the histogram of the different states $H(E)$ is sufficiently flat, f is updated in some decreasing manner (for instance, $f \to \sqrt{f}$) and the histogram is reset. The "walk" across the energy range is then repeated until f reaches a small value.

The density of states can be difficult to estimate, especially for complex systems, which can result in slow simulations. Therefore, speeding up the estimation of $\Omega(E)$ is desirable. One method of speeding up the estimation is to cut the energy range. While the entire energy range of a system may be very large, the interesting part, for instance the range at which droplet formation and dissolution occurs, can be much narrower. Therefore, instead of wasting computation time searching the entire range, the energy range can be cut at some energy $E_{\text{cut}} > E_{\text{min}}$ (or $E_{\text{cut}} < E_{\text{max}}$) to shorten the search. For the system in fig. 3, the entire possible energy range is $[-349, 0]$, but is shortened by roughly 30% to $[-250, 0]$. The states with $E < -240$ are heavily suppressed in the canonical distribution and as such, no important samples are lost. As an example of the speed-up, fig. 4(a) shows a multicanonical run and a part of a canonical run for $N = 320$. The canonical simulation does exhibit transitions between states with and without a droplet, but they are rare. The multicanonical simulation instead exhibits a much faster exploration of the energy landscape and is thus more efficient. To demonstrate that the low energy states are actually droplets, the evolution of the largest droplet in the canonical simulation is plotted in fig. 4(b). As can be seen, it matches perfectly with the drops in energy.

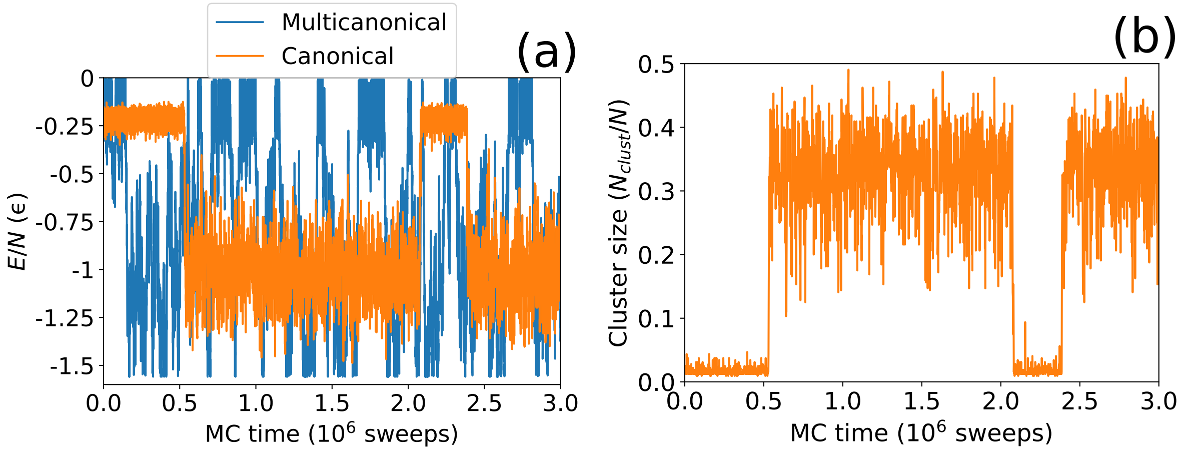


Figure 4: A comparison between the multicanonical (blue) MC evolution of the energy density E/N and the canonical evolution (orange), as well as the evolution of the largest canonical droplet for $N = 320$ at $T \approx T_{\text{b}}^{(N)}$. (a) The low and high energies correspond to states with and without a droplet, respectively. While the canonical simulation "jumps" between the two states, the jumps are rare. On the other hand, the multicanonical simulation scans the energy range much more efficiently. The energy range is cut, but encompasses the energies of interest, as seen from the canonical run. (b) The size of the largest droplet does, indeed, match perfectly with low and high energies.

2.7 Simulation Details

The program which has been constructed for the task of simulating droplet formation consists of one large C++ header that handles the entire simulation and a smaller `cpp` file which includes the header. The program generates a random starting state and then equilibrates the system so that its original state has been "forgotten". This equilibration is done for 1/1000 of the total length of the data collecting. The system reaches an equilibrium state fairly quickly and the thermalisation time is therefore quite short. The time measurement used is *sweeps*: 1 sweep = NN_c MC moves, where N is the number of chains and N_c is the length of the chain. The simulation lengths vary depending on the model and system size, and can be found in table 1.

Table 1: The simulation length in number of sweeps for different system sizes N for the LG model and the two sequences Φ and Θ. The dashes indicate that those sizes were not simulated.

System \ N	10, 20, 40	80, 160	320	640	1280, 2048, 2560	> 2560
LG	10^6	$2 \cdot 10^6$	$3 \cdot 10^6$	$4 \cdot 10^6$	$5 \cdot 10^6$	10^7
Φ	10^7	10^7	$2 \cdot 10^7$	$2 \cdot 10^7$	—	—
Θ	10^7	10^7	10^7	10^7	—	—

For the multicanonical simulations of the LG model, the simulation length is shorter than for the full sequences. This is mainly due to the nature of multicanonical simulations eliminating rare events. Simulating a system size of $N \sim 4000$ takes roughly 1 day on a modern desktop computer; the simulation time scales linearly with system size in the LG model. Simulating the two full length sequences Φ and Θ takes a longer time than the corresponding system size in the LG model. This is due to two reasons. Firstly, the cluster update (which operates in a recursive manner) is the slowest type of update. Secondly, due to the canonical ensemble, the energy landscape between the dilute one-phase and the mixed two-phase is repressed, especially for larger system sizes. Therefore, for $N < 320$, the simulation length is 10^7 sweeps, while it is $2 \cdot 10^7$ sweeps for $N \geq 320$ for sequence Φ. For sequence Θ, 10^7 sweeps is enough for all system sizes.

There are 16 individual trajectories, i.e. evolutions of the energy, generated for each temperature for the Φ sequence and for every system size for the LG model. However, for sequence Φ, multi-histogram reweighting (explained in section 2.8) was used for the largest system size (640 chains). For sequence Θ, on the other hand, there are eight trajectories generated. All temperatures in which the canonical simulations are performed are displayed in table 2. When multi-histogram techniques are used, there are several temperatures for one system size. All temperatures are chosen to be close to the maximum of the specific heat through an iterative procedure.

Table 2: The simulation temperatures in units of ϵ/k_B, for all the system sizes N for which sequence Φ and Θ were simulated in. All temperatures are chosen to be close to the maximum of the specific heat through an iterative procedure.

System \ N	10	20	40	80	160	320	640
Φ	0.466	0.4786	0.491	0.505	0.519	0.5306	0.5375, 0.5387, 0.5395
Θ	0.64	0.69	0.72	0.736	0.736	0.735	0.744

2.8 Data Analysis: Histogram Methods and Jackknife

For greater detail regarding the methods presented in this section, see the book "Computer Simulations of Surfaces and Interfaces" by Dünweg et al., specifically the chapter

"Histograms and All That" by Janke [37]. However, the text below should be sufficiently clear so that readers may themselves utilise the specific methods.

Running simulations of droplet formation can take a long time, especially when the number of chains grows large and the individual chains themselves become long. Due to this, it is beneficial to make use of some shortcuts. One such shortcut is called single-histogram reweighting, first introduced by Ferrenberg and Swendsen [38]. Without too much detail, one can describe it as follows. Assume a simulation has been made at an inverse temperature of β_0. Then the energy histogram $P_{\beta_0}(E) \propto \Omega(E)e^{-\beta_0 E}$ is known. In principle, one can then derive the energy histogram at any inverse temperature β:

$$P_\beta(E) \propto \Omega(E)e^{-\beta E} = \Omega(E)e^{-\beta_0 E}e^{-(\beta-\beta_0)E} \propto P_{\beta_0}(E)e^{-(\beta-\beta_0)E}. \tag{2.21}$$

By reweighting expectation values of functions $f(E)$ with a factor of $e^{-(\beta-\beta_0)E}$, the normalisation factor of $P_\beta(E)$ does not matter:

$$\langle f(E) \rangle (\beta) = \frac{\sum_E f(E)P_\beta(E)}{\sum_E P_\beta(E)} = \frac{\sum_E f(E)P_{\beta_0}(E)e^{-(\beta-\beta_0)E}}{\sum_E P_{\beta_0}(E)e^{-(\beta-\beta_0)E}}. \tag{2.22}$$

This all works, in principle; however, due to the statistical errors that are prone to arise from finite simulations, the accuracy of the reweighted expectation values is limited. Therefore, the probability can not be accurately reweighed to any temperature, especially temperatures far from the critical temperature T_c. The limitation on the accuracy means that errors in the estimation of parameters are likely, and as such, they must be taken into consideration. To this end, the jackknife procedure is applied. Note, that jackknife only increases accuracy close to β. To increase the accuracy range, methods beyond single-histogram reweighting can be utilised.

The jackknife method is a way of removing bias from the estimation of a parameter, O, given several independent observations [39]. Shortly explained, the jackknife procedure works as follows. Assume that N observations have been made, and from these an estimate of O_0 is produced. Then, a reduced estimate, O_i, is done for all N observations by removing observation i, and estimating O_i using the remaining $N-1$ observations. The jackknife estimate, $\hat{O}$, is then

$$\hat{O} = NO_0 - \frac{N-1}{N}\sum_{i=1}^{N} O_i. \tag{2.23}$$

Single-histogram methods are useful, but if the observable is sensitive to β, then extrapolating from a single temperature might not be enough. In those instances, a different shortcut can be useful: multi-histogram reweighting [40]. It can be described in the following way. Consider that n simulations have been performed, each at a different β_i, $i \in \{1, 2, \ldots, n\}$. Reweight all the different energy histograms P_{β_i} to a common reference inverse temperature β_0. Then, compute the error weighted averages of all the n histograms and merge them. Finally, reweight the joined histogram to any other β. For this technique to work in practice, the different β_i's need to be spaced so that the resulting histograms overlap sufficiently.

2.9 Finite-Size Scaling Theory

The systems described in sections 2.1 and 2.2 were assumed to be infinite; however, this is not the case in real life. As such, phase behaviour in finitely sized systems, especially the formation of droplets from phase separation, is a topic which has been studied over the years. Particularly, how systems scale with size, so-called finite-size scaling, has been studied and tested on various systems [17–19, 23, 41]. By performing finite-size scaling analysis, one can determine if certain systems phase separate or not. Certain key results of these studies will be demonstrated in this section (though not as rigorously) as they will be used to analyse the results of this thesis. The conventions and structure of the paper by Nilsson and Irbäck will be used [20].

Take a large but finite d-dimensional system in the canonical ensemble containing N particles in a volume V at temperature $T^* < T_\mathrm{c}$, schematically illustrated in fig. 5. Unlike in the infinite LG model described earlier (section 2.2), there is a region beneath the infinite system's binodal curve $T_\mathrm{b}(\rho)$ but above the finite system's binodal curve $T_\mathrm{b}^{(N)}(\rho)$, see fig. 5. In the lower part of the concentration, this region inhabits a supersaturated state. At some concentration $\rho_\mathrm{L}^{(N)}(T^*) > \rho_\mathrm{L}(T^*)$, the system will transition from the supersaturated state to a mixed two-phase state. As mentioned in section 2.2, this mixed state consists of a single large droplet in a dilute background. Now, consider the concentration $\rho = N/V$ to be in the supersaturated region, $\rho < \rho_\mathrm{L}^{(N)}(T^*)$. If a particle excess δN is added to the system, it can either be absorbed into the supersaturated background as a density fluctuation with a free-energy cost which scales as a Gaussian, $(\delta N)^2/N$, or a finite fraction of the excess can form a dense droplet with a free-energy cost which scales as the surface of the droplet, $(\delta N)^{(d-1)/d}$. When these two costs become similar, then droplet formation kicks in, and the linear size of the droplet R will scale as [17–19]

$$R \sim (\delta N)^{\frac{1}{d}} \sim N^{\frac{1}{d+1}}. \tag{2.24}$$

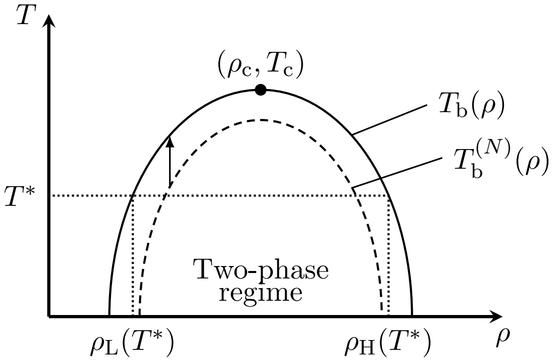


Figure 5: Schematic illustration of a finite-size phase diagram (dashed) and a infinite-size (solid) phase diagram of the LG model (compare to fig. 1(b)). Given a density ρ and a temperature T^* inside the binodal curve $T_\mathrm{b}(\rho)$, an infinitely sized system will phase separate from a single bulk phase into a mixed-two phase with high, $\rho_\mathrm{H}(T^*)$, and low, $\rho_\mathrm{L}(T^*)$, densities. On the left side of the binodal curve, the mixed phase will consist of a dense droplet in a dilute background. In a finitely sized system, the transition temperature $T_\mathrm{b}^{(N)}(\rho)$ is shifted. According to finite-size scaling predictions, $T_\mathrm{b}^{(N)}(\rho)$ converges to $T_\mathrm{b}(\rho)$ in accordance with eq. (2.25) (arrow).

Given eq. (2.24), the density shift $\rho_{\mathrm{L}}^{(N)}(T) - \rho_{\mathrm{L}}(T) = \delta N/V$ will scale as $\rho_{\mathrm{L}}^{(N)}(T) - \rho_{\mathrm{L}}(T) \propto N^{-1/(d+1)}$. This scaling is equivalent to

$$T_{\mathrm{b}}^{(N)}(\rho) - T_{\mathrm{b}}(\rho) \propto N^{-\frac{1}{d+1}}, \tag{2.25}$$

with ρ fixed rather than T. Equation (2.25) states that the finite-size transition temperature $T_{\mathrm{b}}^{(N)}(\rho)$ converges to the infinite-size transition temperature $T_{\mathrm{b}}(\rho)$ slowly. For a proper first-order phase transition, the finite-size temperature shift scales as N^{-1} [42].

The transition is smeared as well as shifted. For fixed ρ, the smearing, or rather the width w_T, can be expressed as the temperature interval over which $|\beta\Delta F| \lesssim 1$ [19, 23], where ΔF is the free-energy difference between the single-phase state and the mixed-phase state. Taylor expanding $\beta\Delta F$ to leading order around the transition temperature $T_{\mathrm{b}}^{(N)}$ yields

$$\beta\Delta F = (\beta\Delta F)|_{T=T_{\mathrm{b}}^{(N)}} - \left.\frac{\Delta E}{k_{\mathrm{B}}T^2}\right|_{T=T_{\mathrm{b}}^{(N)}} \left(T - T_{\mathrm{b}}^{(N)}\right), \tag{2.26}$$

since $\frac{\partial\beta F}{\partial\beta} = E$. At the transition temperature, ΔF vanishes in the limit of large system sizes as the two phases have equal probability. Assuming that there are no interactions in the low concentration phase, ΔE is dominated by the droplet energy, which depends on the droplet volume,

$$\Delta E \sim R^d \sim N^{\frac{d}{d+1}}. \tag{2.27}$$

By applying $|\beta\Delta F| \sim 1$ and eq. (2.27) to eq. (2.26), the width w_T should scale as

$$w_T \propto N^{-\frac{d}{d+1}}. \tag{2.28}$$

Another thermodynamical property that can be used to study droplet formation is the specific heat C_V/N (which was already mentioned in eq. (2.3)). At the point of transition, the specific heat has a distinct peak. From the energy fluctuations, the specific heat can be calculated using

$$\frac{C_V}{N} = \frac{\langle E^2\rangle - \langle E\rangle^2}{Nk_{\mathrm{B}}T^2}. \tag{2.29}$$

By using the specific heat, the transition temperature $T_{\mathrm{b}}^{(N)}$ and the width w_T can be computed. The position of the peak of the specific heat is $T_{\mathrm{b}}^{(N)}$ and the width is w_T. As such, the width of the peak will decrease in accordance with eq. (2.28), and the height of the peak $C_V^{\max}/N$ will grow with increasing system size. How the peak scales with N can be derived from eq. (2.29). Assume that there are two states, 1 and 2. Then, if the probability of having the energy of state 1 E_1 is x, the probability of having energy E_2 is simply $1 - x$. Equation (2.29) therefore becomes

$$\frac{C_V}{N} = \frac{1}{Nk_{\mathrm{B}}T^2}\left\{xE_1^2 + (1-x)E_2^2 - [xE_1 + (1-x)E_2]^2\right\} = \frac{1}{Nk_{\mathrm{B}}T^2}x(1-x)(\Delta E)^2, \tag{2.30}$$

where $\Delta E = E_1 - E_2$. At the peak, both states are equally probable ($x = 1/2$) leading to

$$\frac{C_V^{\max}}{N} \approx \frac{(\Delta E)^2}{4Nk_{\mathrm{B}}T^2}. \tag{2.31}$$

Combining eq. (2.31) with eq. (2.27) then yields how $C_V^{\max}/N$ scales:

$$\frac{C_V^{\max}}{N} \sim N^{\frac{d-1}{d+1}}. \tag{2.32}$$

This is slightly different to what has been suggested previously, namely a scaling of the form $C_V^{\max}/N \sim N^{d/(d+1)}$ [23]. If it assumed that droplet formation is a first order phase transition (which it is not), then this would be the scaling of the maximum specific heat $C_V^{\max}/N$ with respect to N. As the area under the peak is constant for a first-order phase transition, $C_V^{\max}/N \sim w_T^{-1}$ must hold. However, in the case for droplet formation, $\Delta E/N$ vanishes in the large-N limit, therefore, the area under the peak should also vanish. Hence, $C_V^{\max}/N$ should grow slower than $w_T^{-1} \sim N^{d/(d+1)}$ with N, as it does in eq. (2.32).

3 Results

The HP-program was first tested with LG simulations. System sizes from $N = 10$ all the way up to $N = 5120$ were studied with multicanonical simulations (see table 1 for total simulation lengths). The volume was adjusted so that $\rho = 0.01$ holds for all system sizes (or as close as possible since the system is discrete). For each size, 16 runs are performed. In each run, the entire energy span of interest ($[E_{\text{cut}}, 0]$) is traversed frequently. The results were analysed and compared to known results from theory and simulations [23].

Then, the two sequences Φ (HPHPHPHPHP) and Θ (HHHHHPPPPP) were canonically simulated at temperatures near the onset of droplet formation for a range of system sizes, from $N = 10$ to $N = 640$. As in the LG simulations, the volume was adjusted so that $\rho = 0.01$ is true for all the system sizes. For sequence Φ, 16 runs were performed for every system size, while 8 were generated for sequence Θ. For both sequences, cluster formation and dissolution occurred frequently in each run.

3.1 Multicanonical LG

The specific heat of a system which phase separates is expected to diverge. Plotting the specific heat data from runs with system sizes of 10–640 particles (chains with only a single H-bead) in fig. 6, it is clear that the specific heat diverges as the system size increases. As the number of particles grow, the peak becomes higher and slimmer. This behaviour is consistent with theory and experiments [23]. The HP-program therefore seems to properly simulate a LG model; however, by comparing the data to the finite-size scaling prediction, a stronger case for the validity of the program can be made.

Starting with the finite-shift of the transition temperature, the difference $T_{\text{b}}^{(N)} - T_{\text{b}}$ is predicted to scale as $N^{-1/4}$ (eq. (2.25)). Figure 7(a) shows the data for $T_{\text{b}}^{(N)}$ plotted against $N^{-1/4}$ with a fitted line of the form $T_{\text{b}}^{(N)} = T_{\text{b}} + aN^{-1/4}$, where T_{b} and a are parameters. This type of fit is indeed good, even for fairly small systems ($N \leq 40$). The transition temperature T_{b} predicted from this is $T_{\text{b}}k_{\text{B}}/\epsilon \approx 0.626$, which is in the vicinity of the low-temperature expansion $T_0k_{\text{B}}/\epsilon \approx 0.622$ [23, 26]. However, even the largest system ($N = 5120$) has a $T_{\text{b}}^{(N)}$ which is roughly 10% below the transition temperature

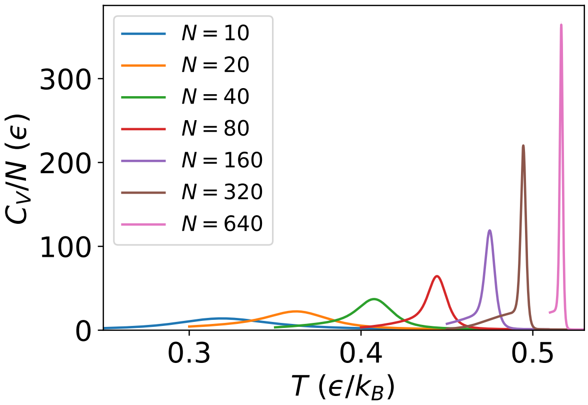


Figure 6: The specific heat C_V/N plotted as a function of temperature T of the LG model for system sizes $N = 10, 20, 40, \ldots, 640$ as computed from the reweighted multicanonical data. The statistical uncertainties are shown as shaded bands, however, they are very small and therefore not visible. The specific heat exhibits a single peak that steadily gets higher and narrower with increasing system size.

of the infinite system. This is expected given that the shift scales as $N^{-1/4}$, i.e. that the temperature converges slowly.

Moving on to the width of the transition, w_T, i.e. the width of the specific heat. It is predicted to scale as $w_\mathrm{T} \propto N^{-3/4}$ (eq. (2.28)). Therefore, w_T against N is log-log plotted in fig. 7(b). From this plot, it can be seen that for very small systems ($N \leq 20$) not much can be said about the scaling behaviour; larger systems are needed. For the small to intermediate interval ($40 \leq N \leq 320$), a different scaling behaviour than the one predicted from theory is observed, namely $w_\mathrm{T} \propto N^{-1}$. It is not until $N \geq 640$ that the correct scaling behaviour is observed. Therefore, if one only has data from the intermediate region, an incorrect scaling behaviour is found. Nevertheless, for large N, the predicted scaling is observed.

Lastly, the maximum specific heat C_V^max/N is plotted as a log-log plot against N in fig. 7(c). If $\Delta E \sim N^{3/4}$, then the expected scaling is $C_V^\mathrm{max}/N \sim N^{1/2}$ (eq. (2.32)). As can be seen though, for small and intermediate system sizes ($10 \leq N \leq 2048$), the maximum specific heat seems to scale inversely proportional to the width w_T. This is consistent with a first-order phase transition, and what has been suggested in a previous paper [23]. However, as mentioned in section 2.9, droplet formation is not a phase transition; the area under the specific heat peak should become zero as N goes to infinity. By increasing the system size to very large systems ($N \geq 2560$), the scaling shifts to $C_V^\mathrm{max}/N \sim N^{1/2}$, as predicted by finite-size scaling theory for droplet formation.

Given that the specific heat peak depends on ΔE, the scaling behaviour found for C_V^max/N should be reflected in how ΔE scales. Figure 8(a) shows a log-log plot of ΔE as a function of N for intermediate to large system sizes ($1280 - 5120$). Indeed, for intermediate system sizes, ΔE scales as $N^{7/8}$. This scaling behaviour corresponds to $C_V^\mathrm{max}/N \sim N^{3/4}$, which explains fig. 7(c). Increasing the system size, the correct scaling behaviour is found ($\Delta E \sim N^{3/4}$), as expected, though it is not very pronounced. In fig. 8(b), the probability distribution of the shifted and rescaled energy $E^* = (E + b)/N^{3/4}$, where b is a parameter

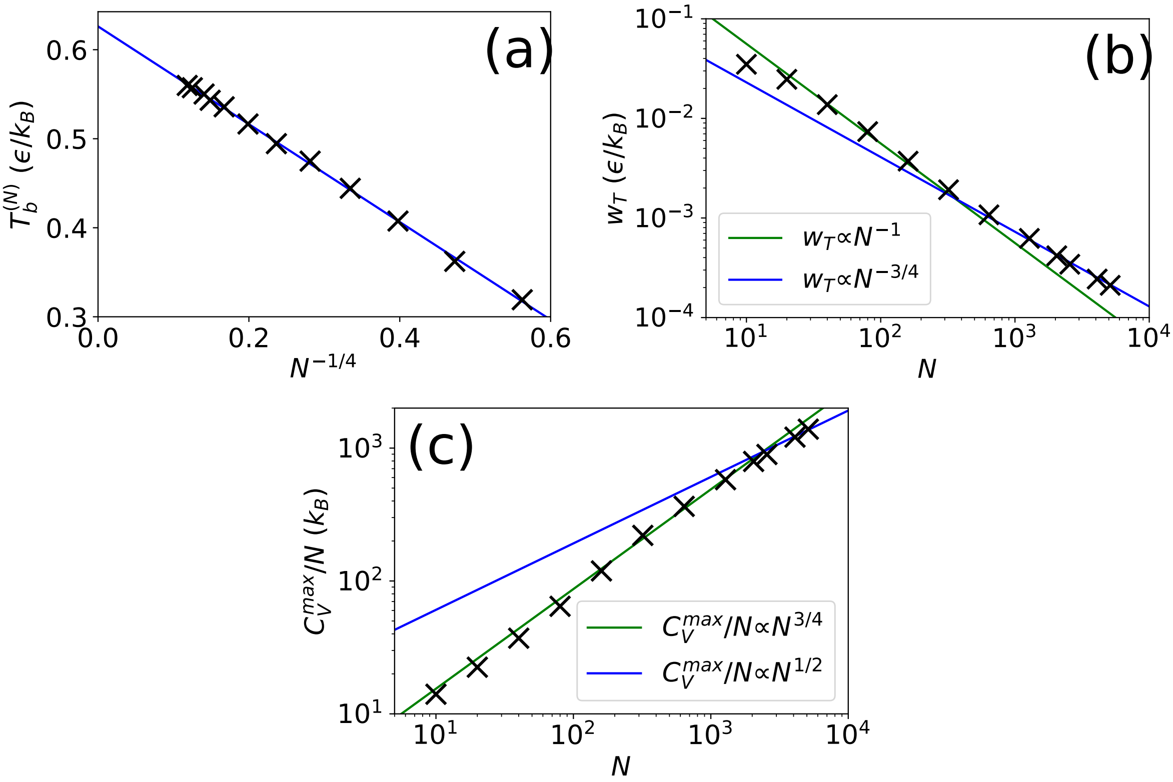


Figure 7: Finite-size scaling of the LG model from multicanonical simulations with 10–5120 particles (chains of length one). (a) The transition temperature $T_{\mathrm{b}}^{(N)}$ plotted as a function of $N^{-1/4}$. The line is a fit of the form $T_{\mathrm{b}}^{(N)} = T_{\mathrm{b}} + aN^{-1/4}$, with a and the transition temperature for infinite system size T_{b} as fit parameters. The fitted value of T_{b} is $T_{\mathrm{b}}k_{\mathrm{B}}/\epsilon = 0.626$. The theoretical value from low temperature expansions of the three dimensional Ising model is $T_{\mathrm{b}}k_{\mathrm{B}}/\epsilon \approx 0.622$ [23, 26]. (b) Log-log plot of the finite-size width of the transition w_T against N. The width is defined as the temperature interval over which $C_V > 0.8C_V^{\max}$. The lines are fits of the forms $w_T \sim N^{-1}$ (green) and $w_T \sim N^{-3/4}$ (blue). (c) Log-log plot of the maximum specific heat $C_{\mathrm{V}}^{\max}/N$ against N. The lines are fits of the forms $C_{\mathrm{V}}^{\max}/N \sim N^{3/4}$ (green), and $C_{\mathrm{V}}^{\max}/N \sim N^{1/2}$ (blue).

independent of E, is shown for the three largest systems. By shifting and rescaling the energy, it is possible to see that the probability distribution of the energy is bimodal. It is also possible to see that the gap between the two peaks stays the roughly the same (as already hinted from fig. 8(a)). Therefore, the scaling behaviour of the maximum specific heat is reflected in the scaling of the energy difference.

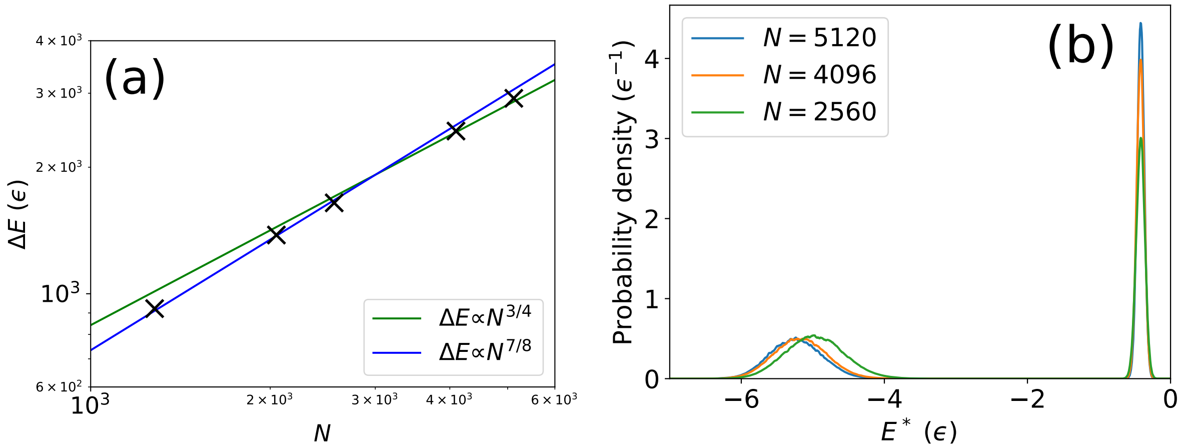


Figure 8: Energy difference between the mixed phase and the solved phase for five biggest systems, and the canonical distribution for the three biggest system sizes at $T \approx T_{\mathrm{b}}^{(N)}$. (a) The scaling behaviours of intermediate sized and large sized system match what is expected from fig. 7(c). (b) The shifted and rescaled energy $E^* = (E+b)/N^{3/4}$ (with $b = 0.15N$) shows that $\Delta E/N^{3/4}$ stays roughly the same with increasing system size.

Based on the preceding analyses, the correct finite-size scaling relations are found for the LG model, alhough large system sizes were needed to observe the correct scaling behaviours for the width and the specific heat peak. It has thus been demonstrated that the HP-program is able to simulate a system which phase separates into a single large droplet. Therefore, further investigations of the longer sequences Φ and Θ can be performed.

3.2 HP sequences

Moving on to the main focus, the two sequences Φ and Θ were analysed and compared. As the specific heat should diverge for a phase separating system (as was shown in fig. 6), the comparison between the two sequences begins with the specific heat. In fig. 9, the specific heat data from simulations of a range of different system sizes for Φ (a) and Θ (b) is shown (up to 640 chains). Sequence Φ's peak grows higher and narrower with increasing system size indicating a divergence. The same trend can be seen for sequence Θ; however, at large sizes, the peak decreases and becomes slightly multimodal. This behaviour suggests the presence of multiple droplets rather than a single large droplet. However, given that sequence Θ clearly does not display the behaviour expected of a system undergoing LLPS, no efforts were made to extend the temperature range in order to definitely observe multimodality. Note that for small systems, the specific heat peak is higher for sequence Θ than for sequence Φ.

To clarify the behaviours of the respective sequences, and to further specify whether or

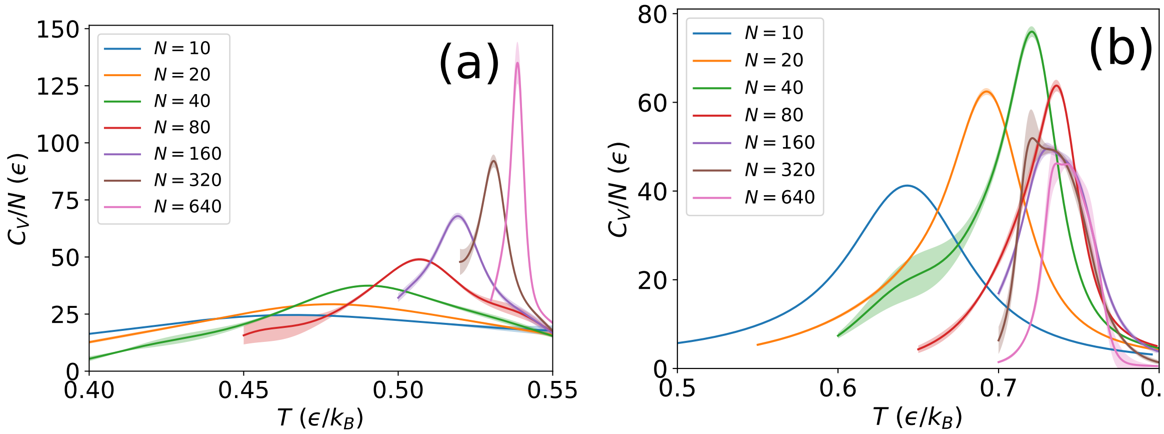


Figure 9: The specific heat C_V/N plotted as a function of temperature T of the HP model for a range of system sizes. The curves are computed using single- and multi-histogram reweighting techniques [38, 40] from canonical MC simulations at temperatures near the onset of droplet formation. Statistical uncertainties are shown as shaded bands. (a) Sequence Φ features a single peak that grows higher and thinner as the system size is increased (similar to the LG model fig. 6). (b) Sequence Θ exhibits a similar trend, but only for small systems. For larger systems, a precursor to a multimodal specific heat starts to appear, indicating the presence of multiple droplets.

not they phase separate, the behaviour and evolution of the systems can be observed. In fig. 10, snapshots of representative configurations of sequence Φ (a) and sequence Θ (b) for $N = 640$ are shown. A single large droplet is observed for sequence Φ with a dilute background containing small clusters, indicative of a LLPS. For sequence Θ, however, multiple droplets can be seen, all of which are smaller than the single droplet formed by sequence Φ. Studying how the systems' evolve with time (runtimes shown in fig. 11), there is a clear difference between the two sequences. In fig. 11(a), the evolution of the energy density for the two sequences is shown. Sequence Θ appears to be in only one phase, with rapid "jumps" in energy, possibly indicating a rapid formation and dissolution of small droplets. Sequence Φ, on the other hand, forms a single large droplet which survives for a longer period of time. Also, the energy in sequence Θ's system does not become as low as in the system with sequence Φ. This also demonstrates that no sequence Θ droplets become as big as the sequence Φ droplets. Indeed, studying the evolution of the largest droplet of each sequence confirms this. These evolutions are shown in fig. 11(b), and it is clear that sequence Θ does not produce any large droplets and consequently it does not undergo LLPS. Sequence Φ, on the other hand, does form a single large droplet which lasts on average for $\gtrsim 10^6$ sweeps. However, comparing sequence Φ to the LG (see fig. 4), the "jumps" in energy and droplet size are not as sudden nor as big.

Finite-size scaling theory is based on the probability distribution of the energy being bimodal. Therefore, fig. 12 shows the probability distribution of E/N for sequence Φ for 320 and 640 chains at a temperature near the onset of droplet formation. It clearly shows that for $N = 320$ there is no bimodal structure, and that for $N = 640$ there is only a slight bimodality. Therefore, the correct finite-size scaling behaviours might not be observed.

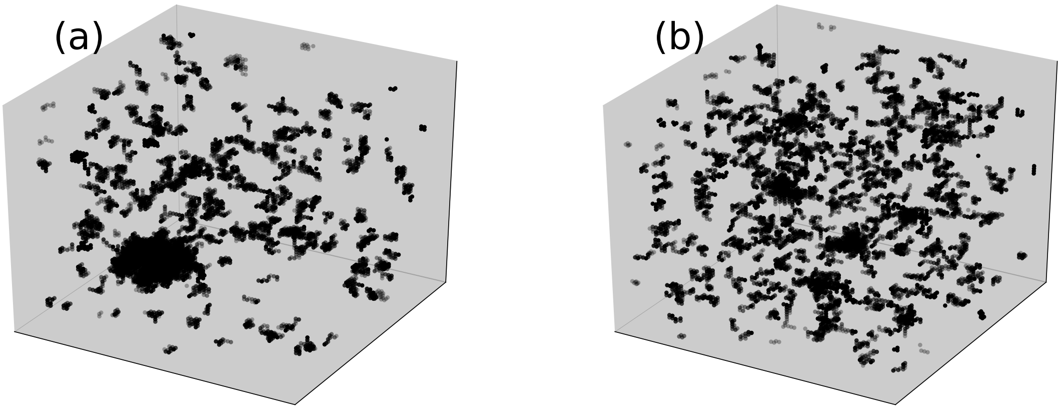


Figure 10: Snapshots of configurations displaying droplets representative of the different sequences at system size $N = 640$ and temperatures $T \approx T_{\rm b}^{(N)}$. Each bead is represented by a dot. (a) A single large droplet can be seen for sequence Φ. (b) A few small droplets are observed for sequence Θ.

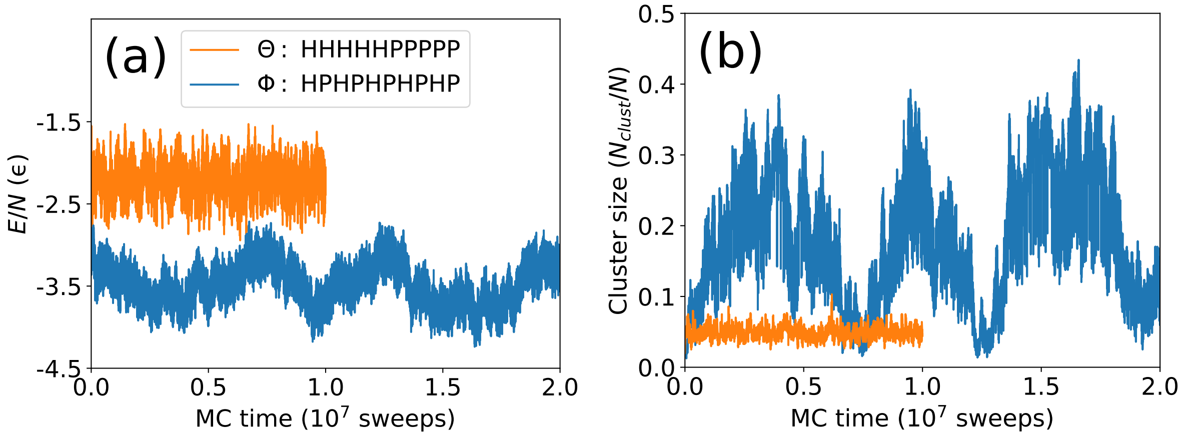


Figure 11: MC evolution of the energy density E/N for sequence Φ and Θ as well as the evolution of the largest droplet for $N = 640$ and $T \approx T_{\rm b}^{(N)}$. High and low energies correspond to states without and with a droplet, respectively. The two sequences were simulated for different number of sweeps and are therefore of different lengths (table 1). (a) Sequence Θ rapidly shifts in energy, whereas sequence Φ has more pronounced energy states. (b) A large droplet forms for sequence Φ which varies greatly in size. Sequence Θ does not grow any large droplets, which further demonstrates that the sequence does not undergo LLPS.

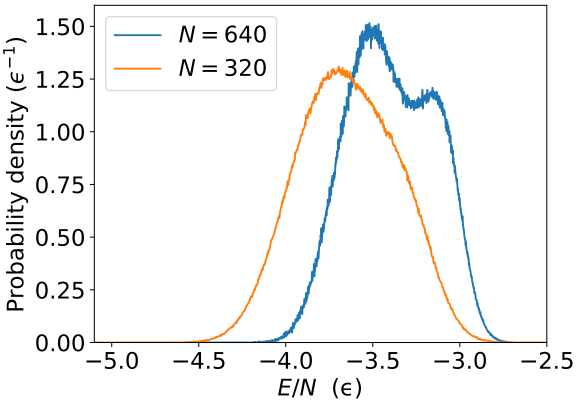


Figure 12: The probability distribution of the energy per number of chains for $N = 320$ and $N = 640$ at $T = T_{\mathrm{b}}^{(N)}$. There is no bimodality for 320 chains and a slight bimodality for 640 chains.

The results so far suggest that sequence Φ may undergo LLPS, and that sequence Θ clearly does not. Therefore, no further investigation regarding LLPS was performed on sequence Θ. On the other hand, sequence Φ was further analysed using finite-size scaling analysis.

The finite-shift of the transition temperature $T_{\mathrm{b}}^{(N)} - T_{\mathrm{b}}$ should scale as $N^{-1/4}$, if a system undergoes LLPS (eq. (2.25)). Therefore, fig. 13(a) shows the data for $T_{\mathrm{b}}^{(N)}$ plotted against $N^{-1/4}$. A fit of the form $T_{\mathrm{b}}^{(N)} = T_{\mathrm{b}} + aN^{-1/4}$, with T_{b} and a as parameters, shows that eq. (2.25) does describe intermediate- to large-N data ($80 \leq N \leq 640$). The predicted transition temperature from this fit is $T_{\mathrm{b}} k_{\mathrm{B}} / \epsilon \approx 0.538$. As already stated, $T_{\mathrm{b}}^{(N)}$ should converge slowly. Indeed, for the largest system, the shifted transition temperature is roughly 8% below the fitted value of T_{b}.

In fig. 13(b), the data for the width of the transition w_{T} is plotted against N in a log-log plot. The width is predicted to scale as $N^{-3/4}$. For small systems ($N \leq 20$), not much can be said about the scaling behaviour. Increasing N, a scaling of $w_{\mathrm{T}} \propto N^{-1}$ is observed, which is incorrect. This scaling is likely due to the present system sizes not being sufficiently large. In fact, the same scaling behaviour ($w_{\mathrm{T}} \propto N^{-1}$) was observed for the intermediate system sizes in the LG model. No correct scaling of $w_{\mathrm{T}}^{-3/4}$ was observed.

Finally, the maximum specific heat $C_V^{\max}/N$ is predicted to scale as $N^{1/2}$ (eq. (2.32)). In fig. 13(c), $C_V^{\max}/N$ is plotted against N in a log-log plot. Again, small systems do not follow the predicted scaling behaviour, as would be expected since the predictions assume asymptotically large N. The data for the four largest systems, however, fit very well with the theoretical predictions.

Despite the fact that bimodality is not observed until the largest system size ($N \geq 640$), several scaling behaviours are correctly identified. Unlike for the LG model, the predicted specific heat peak scaling is found for comparably small systems. The shift of the transition temperature is found to match perfectly with the predicted scaling behaviour. However, the correct scaling behaviour of the width of the transition is not observed. Instead, the same intermediate scaling behaviour from the LG is observed. Hence, it is likely that the

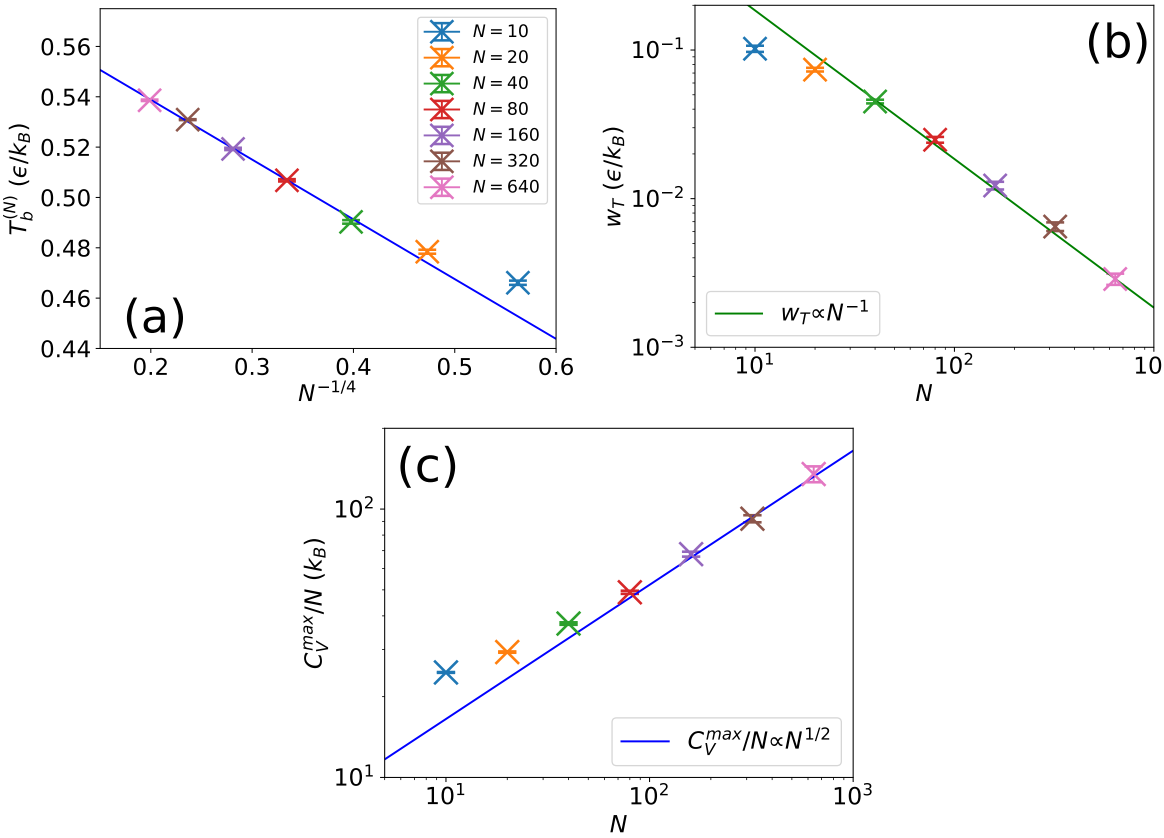


Figure 13: Finite-size scaling data of sequence Φ from simulations with 10-640 chains. (a) The transition temperature $T_\mathrm{b}^{(N)}$ plotted as a function of $N^{-1/4}$. The line is a fit of the form $T_\mathrm{b}^{(N)} = T_\mathrm{b} + aN^{-1/4}$, with a and the transition temperature for infinite system size T_b as fit parameters. The fitted value of T_b is $T_\mathrm{b}k_\mathrm{B}/\epsilon = 0.585$ (b) Log-log plot of the finite-size width of the transition w_T against N. The width is defined as the temperature interval over which $C_\mathrm{V} > 0.8C_\mathrm{V}^\mathrm{max}$. The line is a fit of the form $w_T \sim N^{-1}$. (c) Log-log plot of the maximum specific heat, $C_\mathrm{V}^\mathrm{max}/N$, against N. The line is a fit of the form $C_\mathrm{V}^\mathrm{max}/N \sim N^{1/2}$.

system sizes studied are not large enough for the correct scaling behaviour to be observed for w_T. Nevertheless, based on the results, there are strong indications that sequence Φ unlike sequence Θ, indeed undergoes LLPS.

3.2.1 Droplet characteristics

As has been shown in the comparison of the specific heat data of sequence Φ and Θ, the two sequences have different phase behaviour, even though they have the same length and constituent parts. To understand why this is, a more thorough study of some of the basic structural properties of the droplets was performed. The data studied was gathered from simulations with the largest system size ($N = 640$) and at temperatures near the onset of droplet formation.

As shown previously, different types of droplets, or rather clusters, are formed depending on the sequence. Therefore, the size, or mass, of the droplets formed is an important characteristic of the phase behaviour. Figure 10 demonstrated that sequence Φ can form more massive droplets than sequence Θ. However, to determine whether this is always the case or not, a study of the cluster mass distribution for both sequences was performed. The distribution is plotted in fig. 14 and shows that the most common droplet formed by sequence Φ contains about 170 chains. Sequence Θ, on the other hand, typically forms droplets with around 30 chains in them. Intermediate-sized droplets are statistically suppressed for both sequences, but not by a great deal. This might be due to the lack of a bimodal energy distribution. Nevertheless, a single large droplet is observed for sequence Φ, which is expected if phase separation occurs [17–19].

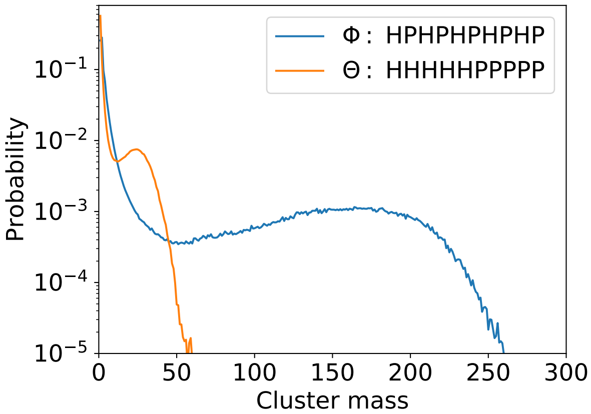


Figure 14: Mass fraction of clusters with N_{clust} chains achieved with the largest system size (640 chains) at a temperature near the onset of droplet formation. Sequence Φ forms droplets of roughly 170 chains, and intermediate-mass clusters are statistically suppressed, though not by much. Sequence Θ instead forms droplets with roughly 30 chains, i.e. in the intermediate-mass range.

The internal structure of the droplets is also important. Therefore, the average bead density around the center of mass of large droplets was studied. The density at a distance r is defined as the concentration of beads in the volume $V(r + l) - V(r)$, where l is a constant. A large droplet is a cluster containing a number of chains greater than a threshold (70 for sequence Φ and 20 for sequence Θ). The resulting density profiles for respective sequence are shown in fig. 15. From the density profile for sequence Φ (a), it can be seen that the density is essentially constant for both small and large r_{cm} so it can be assumed that these regions are representative for the interior of droplets and the dilute surrounding, respectively. Based on this assumption, the droplet density is over 100 times greater than the background density, where the total bead density is roughly $3.7 \cdot 10^{-3}$. The droplets are, like the background, homogeneous in H and P. On the other hand, sequence Θ's droplets are denser and consist of almost only H beads. Given that the sequence is split into one H half and one P half, a droplet with a mainly H core can only hold a finite amount of such chains; it very quickly becomes crowded. Therefore, sequence Θ's droplets can only grow so large, despite an increase in system size.

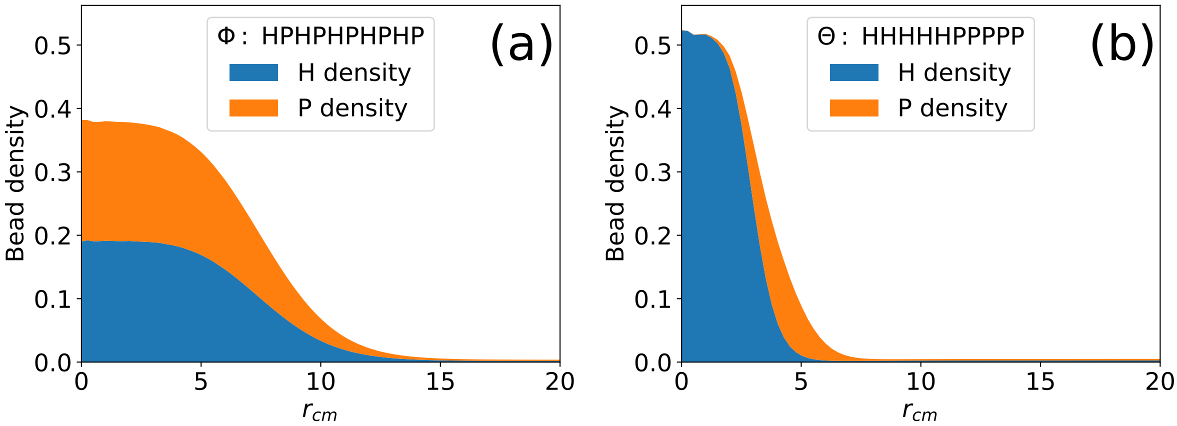


Figure 15: Bead density profiles calculated as a function of the distance r_{cm} from the center of mass of large clusters, obtained using $N = 640$ and a temperature near the onset of droplet formation. The density at distance r_{cm} is defined as the concentration of beads in the volume $V(r+l) - V(r)$, where $l = 0.25$. A large droplet is a cluster with a number of chains greater than a threshold (70 and 20 for sequence Φ and sequence Θ, respectively). The total density is split into H and P densities.

It is clear that sequence Φ phase separates into a droplet with a dilute background, but how liquid-like the droplets are remain unknown. By constraining the model to a lattice, a droplet might become solid-like due to the restriction of movement, which would lead to a poor model for LLPS. It can be expected that a liquid-like droplet would exchange its constituents with the surrounding. It should also have a dynamic interior with chains moving around even if they are in the center of the droplet. To clarify this, an analysis of the rigidity of the droplets formed by sequence Φ was performed. For this purpose, configuration data from every 1000^{th} sweep is used, which is much shorter than the average droplet lifetime of $\gtrsim 10^6$ sweeps. As before, only large droplets ($N_{\mathrm{clust}} > 70$) were used. Two dynamical characteristics are studied: firstly, how rapid the exchange between the droplet and its surrounding is, and secondly, how rigid the core of a droplet is. To analyse the exchange with the surrounding, the chain contents of two consecutive snapshots with a droplet are compared. From such an analysis, it is found that roughly 17% of the initial chains in the droplet are lost by the next snapshot. This may seem like a small amount, but considering that during a droplet's lifetime, it loses, or replaces, 17% of its constituents 1000 times or more. This demonstrates that the droplet has a rather dynamic behaviour; it rapidly exchanges with its surrounding compared to its lifetime.

To study how rigid or dynamic the core of a droplet is, internal chain pairs were analysed. This was done by first finding connected pairs of chains which each interact with at least another 5 chains. This threshold ensures that only interior chains were studied and was determined after manual inspection. Next, the fate of each pair was analysed in the next consecutive snapshot containing a droplet, similar to the exchange analysis. From such an analysis, it appears that the interior of the droplet is also dynamic. On average, 27% of the pairs remain connected, while 62% of the pairs have separated due to internal dynamics. Therefore, about 11% of internal chain pairs are broken due to at least one of the chains leaving the droplet. This demonstrates that the droplets formed from a lattice constrained model are dynamic in terms of their exchange with the surroundings and with their internal configurations.

4 Discussion & Conclusions

Biomolecular condensates form through a LLPS process, which is often driven by IDPs. However, the forces driving the LLPS of IDPs remain incompletely understood. Molecular simulation has the potential to provide insight into the sequence-dependent LLPS of IDPs, but deals with relatively small system, due to computational limitations. In this thesis, finite-size scaling theory has been applied to analyse protein droplet formation in the lattice-based HP model.

To ensure that the program developed and used is capable of simulating phase separation, it was first tested on the simple LG model, in which droplet formation through phase separation is known to occur. The simulation data was in good agreement with results from previous simulations by Zierenberg and Janke [23]. Furthermore, the results matched perfectly with predictions from finite-size scaling theory for droplet formation. Hence, the program can be used to study phase separation in the LG system. However, it is worth noting that in order to observe the behaviour predicted by finite-size scaling theory the system must be sufficiently large ($\gtrsim$ 2560 particles).

As a main objective in this work, two HP sequences were studied in terms of their finite-size scaling properties. The blocky sequence Θ was found not to undergo LLPS, but instead formed aggregates reminiscent of micelles: hydrophobic cores with a polar shell. By contrast, finite-size scaling analysis provides strong evidence that sequence Φ indeed undergoes LLPS. However, for the width of the transition, the predicted scaling behaviour was not observed, most likely due to insufficient system size. In fact, it is only for the largest system size (640 chains) that a bimodal energy distribution is observed. The predicted scaling behaviour of the maximum specific heat and the transition temperature is observed already for the present system sizes. It is interesting to note that a bimodal energy distribution is observed already for much smaller systems in the related continuous model studied by Nilsson and Irbäck [20].

The above analysis does not show whether or not the droplets formed by sequence Φ are liquid-like in character. This question was addressed through additional measurements. It was found that the droplets have a rapid exchange with the surroundings, relative to the droplets' lifespans, and a dynamic interior, with chains relocating within droplets at a rapid pace. Thus, although the use of a lattice restricts the mobility of the chains, the droplets formed by sequence Φ appear to behave more like a liquid than like a solid.

If one wants to determine the underlying phase diagram, the finite-size shift of the transition temperature, $T_{\mathrm{b}}^{(N)} - T_{\mathrm{b}}$, needs to be taken into account. It is worth noting that the convergence of the transition temperature to its value for infinite system size is slow, $T_{\mathrm{b}}^{(N)} - T_{\mathrm{b}} \sim N^{-1/4}$. In fact, in our simulations of sequence Φ, $T_{\mathrm{b}}^{(N)}$ is still 8% lower than T_{b} for the largest system (640 chains). By using finite-size scaling theory, the shift of the transition temperature can be estimated and accounted for.

Acknowledgements

Anders Irbäck has been a terrific supervisor, with his expertise in the field and his great communication skills, he has helped me a ton and I am very grateful of the opportunity to have had him as a supervisor. The CBBP floor has helped me feel at home for the three semester that I have spent with this project, and I am grateful for all the people that I have gotten to know there, and for all the floorball and badminton we have played.

I am extremely thankful for all the people that I have met during my studies who have enriched my life in many different and interesting ways. In particular, my proofreaders (and housemates) Emil Boman and Mattias Sjö for their fantastic spirits and the many shared spirits. I would also, from the bottom of my heart, like to thank Cornelia Säll for her support and for making the last 12 months a fantastic experiences outside of the university (and for allowing me to rant). Lastly, Emil Andersson, with whom I have spent the last 18 months sharing an office, thank you for the fantastic, wonderful and weird discussions, rubber ducking, and for making me laugh.

References

[1] C. P. Brangwynne, C. R. Eckmann, D. S. Courson, A. Rybarska, C. Hoege, J. Gharakhani, F. Jülicher, and A. A. Hyman, “Germline P granules are liquid droplets that localize by controlled dissolution/condensation,” Science **324**, 1729 (2009).

[2] S. F. Banani, H. O. Lee, A. A. Hyman, and M. K. Rosen, “Biomolecular condensates: organizers of cellular biochemistry,” Nat. Rev. Mol. Cell Biol. **18**, 285 (2017).

[3] T. J. Nott, E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowietz, T. D. Craggs, D. P. Bazett-Jones, T. Pawson, J. D. Forman-Kay, and A. J. Baldwin, “Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles,” Mol. Cell **57**, 936 (2015).

[4] A. Molliex, J. Temirov, J. Lee, M. Coughlin, A. P. Kanagaraj, H. J. Kim, T. Mittag, and J. P. Taylor, “Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization,” Cell **163**, 123 (2015).

[5] K. A. Burke, A. M. Janke, C. K. Rhine, and N. L. Fawzi, “Residue-by-residue view of in vitro FUS granules that bind the C-terminal domain of RNA polymerase II,” Mol. Cell **60**, 231 (2015).

[6] M. L. Huggins, “Solutions of long chain compounds,” J. Chem. Phys. **9**, 440 (1941).

[7] P. J. Flory, “Thermodynamics of high polymer solutions,” J. Chem. Phys. **10**, 51 (1942).

[8] S. Qin and H.-X- Zhou, “Fast method for computing chemical potentials and liquid-liquid phase equilibria of macromolecular solutions,” J. Phys. Chem. B **120**, 8164 (2016).

[9] T. S. Harmon, A. S. Holehouse, M. K. Rosen, and R. V. Pappu, “Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins,” eLife **6**, e30294 (2017).

[10] G. L. Dignon, W. Zhen, Y. C. Kim, R. B. Best, and J. Mittal, “Sequence determinants of protein phase behavior from a coarse-grained model,” PLoS Comput. Biol. **14**, e1005941 (2018).

[11] G. L. Dignon, W. Zhen, R. B. Best, Y. C. Kim, and J. Mittal, “Relation between single-molecule properties and phase behavior of intrinsically disordered proteins,” Proc. Natl. Acad. Sci. USA **115**, 9929 (2018).

[12] S. Das, A. Eisen, Y.-H. Lin, and H. S. Chan, “A lattice model of charge-pattern-dependent polyampholyte phase separation,” J. Phys. Chem. B **122**, 5418 (2018).

[13] S. Das, A. N. Amin, Y.-H. Lin, and H. S. Chan, “Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters,” Phys. Chem. Chem. Phys. **20**, 28558 (2018).

[14] T. S. Harmon, A. S. HoleHouse, and R. V. Pappu, "Differential solvation of intrinsically disordered linker drives the formation of spatially organized droplets in ternary systems of linear multivalent proteins," New J. Phys. **20**, 045002 (2018).

[15] J. McCarty, K. T, Delaney, S. P. O. Danielsen, G. H. Fredrickson, and J.-E. Shea, "Complete phase diagram for liquid-liquid phase separation of intrinsically disordered proteins," J. Phys. Chem. Lett. **10**, 1644 (2019).

[16] Y. Lin, J. McCarty, J. N. Rauch, K. T. Delaney, K. S. Kosik, G. H. Fredrickson, J.-E. Shea, and S. Han, "Narrow equilibrium window for complex coacervation of tau and RNA under cellular conditions," eLife **8**, e42571 (2019).

[17] K. Binder and M. H. Kalos, ""Critical clusters" in a supersaturated vapor: theory and Monte Carlo simulation," J. Stat. Phys. **22**, 363 (1980).

[18] M. Biskup, L. Chayes, and R. Kotecky, "On the formation/dissolution of equilibrium droplets," Europhys. Lett. **60**, 21 (2002).

[19] K. Binder, "Theory of the evaporation/condensation transition of equilibrium droplets in finite volumes," Physics A **319**, 99 (2003).

[20] D. Nilsson and A. Irbäck, "Finite-size scaling analysis of protein droplet formation," submitted manuscript (2019).

[21] K. F. Lau and K. A. Dill, "A lattice statistical mechanics model of the conformational and sequence spaces of proteins," Macromolecules **22**, 3986 (1989).

[22] D. Chandler, "Introduction to modern statistical mechanics," 1st ed. Oxford University Press (1987).

[23] J. Zierenberg and W. Janke, "Exploring different regimes in finite-size scaling of the droplet condensation-evaporation transition," Phys. Rev. E. **92**, 012134 (2015).

[24] E. Ising, "Beitrag zur Theorie des Ferromagnetismus", Z. Phys. **31**, 253 (1925).

[25] L. Onsager, "Crystal statistics. I. A two-dimensional model with an order-disorder transition", Phys. Rev. **65**, 117 (1944).

[26] G. Bhanot, M. Creutz, and J. Lacki, "Low-temperature expansion for the Ising model," Phys. Rev. Lett. **69**, 1841 (1992).

[27] P. Butera and M. Pernici, "High-temperature expansions of the higher susceptibilities for the Ising model in general dimension d", Phys. Rev. E **86**, 011139 (2012).

[28] A. Falicov, A. Nihat Berker, and S. R. McKay, "Renormalization-group theory of the random-field Ising model in three dimensions", Phys. Rev. B **51**, 8266 (1995).

[29] A. M. Ferrenberg, J. Xu, and D. P. Landau, "Pushing the limits of Monte Carlo simulations for the three-dimensional Ising model", Phys. Rev. E **97**, 043301 (2018).

[30] C. P. Robert and G. Casella, "Monte Carlo statistical methods," 2nd ed. New York: Springer (2004).

[31] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," J. Chem. Phys. **21**, 1087 (1953).

[32] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," Biometrika **57**, 97 (1970).

[33] R. H. Swendsen and J. S. Wang, "Nonuniversal critical dynamics in Monte Carlo simulations," Phys. Rev. Lett. **58**, 86 (1987).

[34] D. Frenkel and B. Smit, "Understanding molecular simulation: From algorithms to appplications," Computational science series (Elsevier Science, 2001).

[35] W. Janke, "Multicanonical Monte Carlo simulations," Physica **A254**, 164 (1998).

[36] F. Wang and D. P. Landau, "Efficient, multiple-range random walk algorithm to calculate the density of states," Phys. Rev. Lett. **86**, 2050 (2001).

[37] B. Dünweg, D. P. Landau, and A. I. Milchev, "Computer simulations of surfaces and interfaces," 1st ed. Netherlands: Springer (2003).

[38] A .M. Ferrenberg and R. H. Swendsen, "New Monte Carlo technique for studying phase transitions," Phys. Rev. Lett. **61**, 2635 (1988); *ibid.* **63**, 1658 (1989) (Erratum).

[39] R. G. Miller, "A trustworthy jackknife," Ann. Math. Statist. **35**, 1594 (1964).

[40] A .M. Ferrenberg and R. H. Swendsen, "Optimized Monte Carlo data analysis," Phys. Rev. Lett. **63**, 1195 (1989).

[41] M. Schrader, P. Virnau, and K. Binder, "Simulation of vapor-liquid coexistence in finite volumes: a method to compute the surface free energy of droplets," Phys. Rev. E **79**, 061104 (2009).

[42] C. Borgs and R. Kotecký, "Finite-size effects at asymmetric first-order phase transitions," Phys. Rev. Lett. **68**, 1734 (1992).