Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____   _____
Lara A. Patel   Date

Changes in State: Simulations of Aggregation and Ordering in Finite
Systems

By

Lara A. Patel
Doctor of Philosophy

Chemistry

_____
Dr. James T. Kindt, Ph.D.
Advisor

_____
Dr. Joel M. Bowman, Ph.D.
Committee Member

_____
Dr. Francesco Evangelista, Ph.D.
Committee Member

Accepted:

_____
Lisa A. Tedesco, Ph.D
Dean of the James T. Laney School of Graduate Studies

_____
Date

Changes in State: Simulations of Aggregation and Ordering in Finite
Systems

By

Lara A. Patel
B.A., Mount Holyoke College, 2010

Advisor: Dr. James T. Kindt, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Chemistry
2018

Abstract

Changes in State: Simulations of Aggregation and Ordering in Finite Systems

By Lara A. Patel

This dissertation presents three projects that use molecular dynamics (MD) simulations to investigate finite size effects on changes in state.

The first project aims to explain the exponential decay during the sub-microsecond phase of melting kinetics of unilamellar vesicle lipid ordering observed in ultrafast IR temperature-jump experiments. MD simulations of small unilamellar vesicles of MARTINI coarse-grained DPPC lipids model the response to an instantaneous temperature jump from 280 K to final temperatures of 290 K to 310 K. Instantaneous heating led to partial or total melting, changes in vesicle shape, and the sizes and arrangements of the remaining gel-phase domains. At temperatures that produced partial melting, the gel-phase lipid content of the vesicles followed an exponential decay, consistent in form and timescale with experiment. The changing rate of melting results from the curvature stress arising from the expansion of the bilayer area upon melting competing with the confinement effect of a fixed internal volume.

The subsequent projects employ a global fitting analysis method that obtains free energies of association from equilibrated cluster size frequency distributions of unbiased constant-temperature MD simulations. Simulated systems are typically too small for the law of mass action to accurately describe the aggregation statistics. This method relies on iteratively determining a set of cluster free energies that, using appropriately weighted sums over all possible partitions of N monomers, best reproduce the cluster size distributions. To showcase the method, a united-atom model of methyl tert-butyl ether is simulated in the vapor phase and in explicit water solution over a range of system sizes and concentrations.

Bypassing the explicit generation of partitions significantly enhances the efficiency of this method and is named the Partition-Enabled Analysis of Cluster Histograms (PEACH) method. It is used to calculate the free energy surface of NaCl aggregation in MD simulations in four solvents (pure methanol, pure water, and two methanol/water mixtures). The presence of non-ideal crowding effects and the systematic concentration-dependent indicators in the results of the PEACH model fit are addressed. Insights into the proposed two-step mechanism of crystal nucleation and its dependence on solvent and degree of supersaturation are discussed.

Changes in State: Simulations of Aggregation and Ordering in Finite
Systems

By

Lara A. Patel
B.A., Mount Holyoke College, 2010

Advisor: Dr. James T. Kindt, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Chemistry
2018

Acknowledgement

It is with a strange mixture of happiness and sadness that I write these words. This body of writing marks the end of a chapter in my life and the end of a tumultuous academic journey. Like a marathon runner that has passed the point of keeping track of the miles, I keep having to remind myself that, with this dissertation, I will have completed a Ph.D. It seems fitting to thank everyone that has made this possible in chronological order.

As a kid with a learning disability, it never entered the realm of my imaginings that I would end up a scientist with a degree as prestigious as a PhD. Evil overlord, on the other hand, was definitely in the running. The commendation for averting my initial career aspirations goes to my parents, Mahesh Patel and Trine Bølling-Patel. You have gone above and beyond as parents. It was not easy. I would not be here today if you had not helped and supported me, kept me on task when lessons were abhorrent, persisted even when we were all exhausted, and encouraged me to keep striving higher.

There are three women in academia that I would like to thank in particular because you all served as role models of women in science and mathematics, and were pivotal in my journey to this moment: Catherine Berghahn, Sheila Browne, and Maria Gomez. Cathy, you taught mathematics with an uncontainable enthusiasm that was contagious! Your class was monumental in changing how I viewed myself and academia. I came to realize that having a learning disability comes with a different way of looking at problems that can be an advantage. Sheila, when I entered your organic chemistry class as a freshman, I was almost positive that I was going to major in economics. I started my first year at Mount Holyoke College with early morning classes in Econometrics and Organic Chemistry, on the off chance that chemistry might engage me. Your class captured my attention in ways that I cannot fully explain. The material and the way that you taught clicked for me. Maria, thank you so very, very much, for your diligence to teaching and explaining physical chemistry so meticulously. Thank you again for taking me as a research student and facilitating my growth as a scientist and programmer. As evident by this dissertation, those first forays into simulations and critical thinking in research planted the seeds that have come into fruition under the guidance of the professors and mentors I have had at Emory.

I would like to begin my thanks to those at Emory University by thanking Dr. James T. Kindt for all of his support, help, and teaching. It has been an honor to work with you for all these years. While I know that I will never be done learning, I am going to leave the program and your research group with a plethora of new skills and a better understanding of where my strengths and weaknesses lie. Our discussions have opened my eyes to angles I had not thought of, and the insight you have had on our projects have often boggled my mind. I became enamored with the research project that you first gave me when I went off and derived the the internal dihedral angles of a truncated icosahedron. I spent all of thanksgiving working on that derivation and figuring out how I would code the assembly of my initial vesicle structure while my family gave me space to pursue my crazed idea. I remember how chuffed I was that

I had figured out a way to get from edge lengths to dihedral angles, but I was not certain that I had not made a mistake. I gave you my calculations and asked if you would be so kind as to take a look at them. My impression when I came back later was that you were impressed by my dogged pursuit of the derivation but rather than checking all my calculations, had looked up the the dihedral angle. In my defense, I had been calling the shape a soccer ball. I learned two very important lessons that day: 1) always check to see if someone has already invented the wheel and 2) confer with your boss and colleagues as they often have good idea of where to start looking for said wheel.

I would also like to thank my committee members, Dr. Joel Bowman and Dr. Francesco Evangelista, for the time they have put aside to be on my committee and their astute observations and questions about my research. Your questions have guided me in taking a step back from the research and to remember that the accessibility of the research relies on introducing it from a common ground. In particular, Joel, your questions at my annual assessment last year is the reason that there is a full section about equilibrium constants and the law of mass action in this dissertation and my continued thoughts on finite size effects.

I would like to thank Olivia Beckwith and Robert Schneider for their contributions to the collaboration between the Kindt research group and Ken Ono's group in the mathematics department at Emory. Their insights into generating functions and the aggregation equilibrium cluster size distribution trends have been instrumental in improving the methods discussed in this dissertation. I would like to thank Xiaokun Zhang for collaborating with Dr. James Kindt and myself. I count myself as lucky to have been part of the process.

A huge thanks goes out to the member of the Kindt group: Keon Reid, Lewen Yang, Xiaokun Zhang, and Ziwei Guo. In the long pursuit of a Ph.D., the company we keep can be the thing that keeps us going when the research gets rough or incomprehensible. You have all factored into making an excellent research group that takes the time to show kindness and compassion to one another, and to discuss interesting and finicky research problems. I am so very honored to have been able to work along side all of you.

I would like to thank my best friend and long time house mate, Elizabeth Main. You have put up with me grumbling over frustrating impasses in my research, and my strange hours when I found a solution to a problem and would not stop until I had it all down in code or writing. You have been there to tell me that it is time to go out and live a little. Thank you for keeping me sane and being such a good friend.

Lastly I would like to thank my brother, Peter Patel, and his lovely wife and my wonderful new sister, Andrea Patel. Peter, you are the very best brother I could ever have had. You probably know even better than I do what sacrifices you made along the way when we were younger. I would just like you to know that I love and appreciate all of it.

Table of Contents

List of Figures

List of Tables

Chapter 1   Introduction

In this dissertation, three research projects are presented: 1) the molecular dynamics (MD) of the melting phase transitions for small frozen DPPC vesicles following an instantaneous increase in temperature, 2) the calculation of cluster free energies for single component aggregation of methyl-tert-butyl-ether (MTBE) clusters, and 3) the calculation of the cluster free energies for two component aggregation of sodium chloride (NaCl) clusters in solvent environments where we tune the NaCl solubility. At a first appraisal, the first project seems rather different in that it deals with a phase transition in the condensed phase and non-equilibrium dynamics while the two other projects involve equilibrium simulations of aggregation dynamics. In this introduction however, we aim to highlight the similarities between the three projects.

In the typical description of phase transitions between two systems of very large sizes and negligible coupling, thermodynamic equilibrium between the two bulk phases occurs when the chemical potentials and pressures of the two systems are equal. The Gibbs phase rule[1] determines that the coexistence of two phases occurs (at most) for a single temperature for a 1-component system at constant pressure; or (at most) a single concentration for a solution at constant pressure and temperature or vapor at constant temperature. The assumption here is that the properties of the two phases are unchanged when molecules exchange between the two phases. The breakdown of this assumption is vital to our findings in the vesicle melting and small-N nucleation cases and is exhibited in size dependent interfacial effects and confinement effects.

It is observed in Chapter 3 that even under the DPPC lipid melting phase transition temperature, small vesicles partially melt upon increasing the system tempera-

---

[1]The number of independent intensive thermodynamic variables $f = 2 + r - v$, where r is the number of components and v is the number of coexisting phases.[1]

ture. This is indicative of a range of equilibria between the gel and fluid phases over a temperature range due to the introduction of curvature, domain size dependent interfacial effects, and confinement effects. The bulk free energy when the system is below the phase transition temperature, opposes melting and in an infinitely large system, would ensure a single phase transition temperature. However, due to the small size of the vesicles and the rigidity of the gel phases, the frozen vesicles are faceted. The lipids melt in the regions where separate gel domains meet in order to relieve curvature stress. The initial, very fast, melting in the vesicles after a temperature jump is driven by the curvature stress and that driving force weakens as the melting progresses. The size dependent interfacial free energy of the finite gel domains on the vesicles also promotes melting of the vesicle. This driving force increases as a gel domain size gets smaller. The interplay between the opposing forces of the bulk and interfacial free energies is what leads to the notion of a critical nucleus size for gel phase domains melting on a vesicle bilayer, under which a gel domain is unstable and should melt fully.

The arresting or slowing of the phase transitions is the result of a confinement effect imposed by the competition between a fixed internal volume[2] and an increasing vesicle surface area as lipids transition from a low surface area gel phase to the higher surface area fluid phase. So while the temperature jump drives the phase transition forward, the confined and fixed internal volume acts as an increasing barrier to further melting. Since the fluid phase is relatively flexible, deformations of the vesicle bilayers alleviates what could otherwise result in a compression of the fluid phase. The kinetics of the phase transitions slows down as the opposing stresses on the interface between the two phases reach a stalemate.

Aggregation involves the collision and then association of components of a sub-

---

[2]The internal volume of the vesicles can be considered fixed because diffusion of water across the bilayer is on the time scale of 10's of milliseconds[2] and so much slower than the initial melting of the gel domains.

stance to form a larger cluster. For aggregation in a bulk solution, the association and dissociation of small clusters precedes the formation of a critical nucleus cluster size which permits the growth of a stable aggregate. In the two step nucleation mechanism experienced by crystals such as the NaCl cluster in Chapter 6, there is a second nucleation step in the phase transition from an amorphous cluster of ions to an ordered crystal structure. In the limit of a large system with a single large aggregate, an equilibrium between the aggregate and the solution (NPT) or vapor (NVT) is only reached at the saturation concentration c_{sat}.

The rare event nature of nucleating an aggregate in solution and this propensity towards a single relatively fixed size aggregate makes it challenging to sample the initial cluster sizes in a large simulation. The equilibrium cluster sizes are perturbed by employing finite size effects (in the form of the system sizes and concentrations) to alter the distributions of populated states (cluster sizes) in the simulations. Small system sizes means that the effective free monomer concentration that the aggregates experience is dependent on the clusters present, and thus a range of concentrations and their associated equilibria are sampled. Across a series of simulations we can thus achieve continuous sampling of cluster sizes. This continuous sampling is then used to globally fit a sequence of partition functions which can be used to calculate the free energy surface of nucleation using the partition enabled analysis of cluster histograms (PEACH) method (Chapter 4).

These projects also explore perturbations of an environment in the context of tuning the extent of a transition. In the case of the vesicle melting, increasing the temperature increases the driving force behind the melting phase transition. The confinement effect caused by the fixed encapsulated volume results in a range of phase coexistence temperature for the DPPC lipids in a vesicle as compared to the single melting temperature of flat bilayer with an fixed interface length. In the case of the nucleation of NaCl clusters, solvent environments were found to influence the nu-

cleation equilibria. Decreasing the solubility by increasing the fraction of methanol in the solvent, is found to increase the driving force behind the aggregation. More interesting perhaps is the observation that it also affects the transition from an amorphous cluster to an ordered cluster. This can be interpreted as a destabilization of the amorphous cluster due to a less attractive solvent environment.

We begin this thesis with a general overview of molecular dynamics (MD) and some of the simulation tools that facilitate performing MD simulations in NPT and NVT ensembles in Chapter 2. We present the results of temperature jump simulations of vesicle undergoing melting phase transitions in Chapter 3. In Chapter 4 we introduce and derive the partition enabled analysis of cluster histograms (PEACH) method which can be used to globally fit the free energy of nucleation. This method is then applied to nucleation in a single component system of MTBE amorphous clusters in Chapter 5 and then a two component system of NaCl amorphous and crystalline clusters in Chapter 6.

Chapter 2 Molecular dynamics

Molecular dynamics is a simulation technique that involves the integration of Newton's equations of motion to compute the equilibrium and/or dynamic properties of a classical many-bodied system. The dynamics of the atoms must conform to classical mechanics such that the force $\mathbf{F}_i$ experienced by an atom is determined by the partial derivative of the chemical potential $U(\mathbf{r}_1, \mathbf{r}_2, \cdots \mathbf{r}_N)$ with respect to the atomic coordinates, $\mathbf{r}_i$.

$$\mathbf{F}_i = -\frac{\partial U}{\partial \mathbf{r}_i} \qquad (2.1)$$

$$\frac{\partial^2 \mathbf{r}_i}{\partial t^2} = \frac{\partial \mathbf{v}_i}{\partial t} = \frac{\mathbf{F}_i}{m_i} \qquad (2.2)$$

For each time step in a molecular dynamics trajectory, the forces (Eqn. 2.1) are evaluated for each atom, summing the forces between all the non-bonded atom pairs and the those due to the bonded interactions. Then the configuration is updated by moving the atoms according to the forces working on them (Eqn. 2.2). There are a number of different integration algorithms that can be used, but the following projects consistently use the Verlet leap-frog integrator[3]:

$$\mathbf{v}(t + \Delta t/2) = \mathbf{v}(t - \Delta t/2) + \frac{\Delta t}{m}\mathbf{F}(t) \qquad (2.3)$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t \mathbf{v}(t + \Delta t/2) \qquad (2.4)$$

2.1 Temperature and pressure coupling

Ensembles for NPT or NVT simulations both require a coupling method that will maintain the number of atoms in the simulation, and temperature of a system. In molecular dynamics simulations, the number of atoms is maintained readily, since there is no addition or removal of particles. The volume of a system is also readily maintained when needed by not allowing the system dimensions to scale.

It is however not as simple to set the total system temperature to a single value (as one might the system volume in an NVT simulation). Temperature is an average property, reflecting a distribution of kinetic energies. A bulk temperature T, can be defined as a function of the average kinetic energy $\langle K \rangle$:

$$\langle K \rangle = \left\langle \frac{1}{2}mv_\alpha^2 \right\rangle = \frac{1}{2}k_B T \tag{2.5}$$

where α represents a single degree of freedom. The instantaneous temperature at time t can be calculated as proportional to the average of the kinetic energy over all N atoms in the system normalized with respect to the degrees of freedom N_f:

$$T(t) = \sum_{i=1}^{N} \frac{m_i v_i^2(t)}{k_B N_f}, \tag{2.6}$$

Both of the temperature coupling methods that are presented subsequently operate by scaling the simulation velocities at given intervals throughout a trajectory. Capturing the ensemble average in a simulation requires a temperature coupling method that ideally samples the ensemble distributions of kinetic energies.

2.1.1 Berendsen thermostat

The Berendsen thermostat[4] operates on the model of a system that is weakly coupled to a external temperature bath with a temperature T_0. The method corrects temperature deviations from T_0 by:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} \tag{2.7}$$

resulting in an exponential decay in the temperature deviation.[5] The temperature adjustment is performed by scaling kinetic energy of the system (or velocities of each particle) by a time dependent factor λ every n_{TC} steps such that:

$$\Delta K = (\lambda - 1)^2 K \tag{2.8}$$

The scaling factor λ is calculated as:

$$\lambda = \left[1 + \frac{n_{TC}\Delta t}{\tau_T} \left(\frac{T_0}{T(t - \Delta t/2)} - 1 \right) \right]^{1/2} \tag{2.9}$$

where τ_T, the temperature coupling time constant, is related to the time constant τ of the temperature coupling by:

$$\tau = 2C_V\tau_T/N_f k_B \tag{2.10}$$

where C_V is the total heat capacity, k_B is Boltzmann's constant, and N_f is the total number of degrees of freedom.

2.1.2 Velocity rescaling thermostat

The Berendsen thermostat does not reproduce the kinetic energy ensemble distribution accurately. The velocity rescaling thermostat[6] remedies this by ensuring that the reference temperature (or kinetic energy via the relationship in Eqn. 2.5) that the system is coupled to, samples the canonical distribution of kinetic energies. The kinetic energy K is modified according to:

$$\frac{dK}{dt} = \frac{(K_0 - K)}{\tau_T} + 2\frac{dW}{dt\sqrt{\tau_T}}\sqrt{\frac{KK_0}{N_f}} \tag{2.11}$$

where K_0 is the average reference kinetic energy, N_f is the number of degrees of freedom, and dW is a Wiener noise.

The first part of the temperature correction in Eqn. 2.11 is identical to that of the Berendsen thermostat (Eqn. 2.7). The second half of Eqn. 2.11 is a stochastic term that ensures proper sampling of the kinetic energy distribution.

2.1.3 Berendsen barostat

One of the more common ways to measure the pressure in a simulations is based on the virial equation for the pressure:

$$P = \rho k_B T + \frac{1}{dV}\left\langle \sum_{i<j} \mathbf{F}\left(\mathbf{r}_{ij}\right) \cdot \mathbf{r}_{ij} \right\rangle \tag{2.12}$$

where ρ is the density, d is the dimensionality of the system, V is the volume of the system, and $\mathbf{F}\left(\mathbf{r}_{ij}\right)$ is the force between atoms i and j at a distance $\mathbf{r}_{ij}$.[7] In the NPT

ensemble that is typically used for simulations of solutions, this requires that some form of pressure coupling is employed. The Berendsen barostat[4] uses a first order kinetic relaxation of the pressure to the reference pressure P_0 according to:

$$\frac{d\mathbf{P}}{dt} = \frac{\mathbf{P}_0 - \mathbf{P}}{\tau_p} \tag{2.13}$$

This relaxation is accomplished by rescaling the coordinate and simulation dimensions every n_{PC} simulation steps by the scaling matrix with entries of:

$$\mu_{ij} = \delta_{ij} - \frac{n_{PC}\Delta t}{3\tau_p}\beta_{ij}\left(P_{0,ij} - P_{ij}\left(t\right)\right) \tag{2.14}$$

where β is the isothermal compressibility and the actual scaling matrix is[5]:

$$\mu' = \begin{pmatrix} \mu_{xx} & \mu_{xy} + \mu_{yx} & \mu_{xz} + \mu_{zx} \\ 0 & \mu_{yy} & \mu_{yz} + \mu_{zy} \\ 0 & 0 & \mu_{zz} \end{pmatrix} \tag{2.15}$$

In the subsequent studies, we use isotropic and semi-isotropic pressure coupling. For isotropic pressure coupling, the scaling matrix elements μ_{ij} where $i = j$ are all equal and the off-diagonal elements where $i \neq j$ are zero. Semi-isotropic pressure coupling is used in slab simulations. Assuming that the slab lies in the xy-plane, the off-diagonal components are zero while the the element for the dimension parallel to the slab $\mu_\parallel$ is the average of μ_{xx} and μ_{yy} while the perpendicular direction $\mu_{zz} = \mu_\perp$ is independent of the others.

2.2   Potentials and coarse graining

Some level of smoothing over finer details is used in all classical potentials. In atomic models, electrons and nuclei of atoms are not explicitly accounted for. Electrostatic interactions are captured using a Coulomb potential by approximating the charge distribution through point charges located at atom cites:

$$U_{\text{Coulomb}}\left(r_{ij}\right) = f\frac{q_i q_j}{r_{ij}} \tag{2.16}$$

where $f = (4\pi\epsilon_0)^{-1}$, r_{ij} is the distance between particles i and j, and q is the partial charge on the particles. Other non-bonded interactions are captured by a Lennard-Jones potential:

$$U_{\text{LJ}}\left(r_{ij}\right) = 4\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) \tag{2.17}$$

where r_{ij} is the distance between particles i and j, ϵ_{ij} is the depth of the interaction well, and σ_{ij} is the distance between the particles at which the potential function is zero. The minimum of the well is located at $r_m = 2^{1/6}\sigma_{ij}$.

The selection of appropriate potential parameters (q, ϵ, σ) is typically done while attempting to reproduce experimental observables in the simulations. Models such as the MARTINI DPPC model and the TraPPE MTBE model are coarse grained potentials. This means that rather than having a set of potential parameters for each atom in the molecule, atoms are grouped together into a single simulation particle with a single set of potential parameters. For example representing a methyl group (CH_3) with a single 'bead'. While there are degrees of freedom that are lost by coarse graining a potential, it significantly reduces the computational load and speeds up dynamics. Processes that would be otherwise inaccessible given current hardware, such as the melting of a vesicle (Ch. 3), become feasible and facilitate reaching a qualitative understanding of such processes.

Chapter 3   Coarse-grained molecular simulations of the melting kinetics of small unilamellar vesicles

Unilamellar vesicles are closed bilayers encapsulating an internal volume of solvent.[1] The bilayer consists of amphiphilic lipids that self-assemble to sequester hydrophobic tail groups from polar solvent. With vesicles spanning size scales from the nm to μm, there are three subcategories: small unilamellar vesicles (SUV) with diameters less than 100 nm, large unilamellar vesicles (LUV) 100 nm to 1 μm in diameter, and giant unilamellar vesicles (GUV) with diameters exceeding 1 μm.[9] Vesicles are convenient simple model systems for biomembranes, and can also be used as nanoencapsulation agents for drug delivery and related functions.

The phase behavior of lipids in vesicles has been the subject of a number of experimental and computational studies, and has been exploited in the use of liposomes as thermosensitive drug delivery agents,[10, 11] that change their permeability at the phase transition temperature T_m.[12] Here our primary goal is to model the dynamics of phase change when a vesicle undergoes partial or total melting in response to a jump in temperature, a phenomenon observed in recent experiments by Nagarajan et al.[13] In that study, an ultrafast IR laser pulse initiated a temperature jump in a solution of LUV's of deuterated dipalmitoyl phosphatidylcholine (DPPC) near the main chain transition temperature (T_m), and the resulting change in phase was followed by time-resolved IR spectroscopy. A motivation for that study was to develop a method to induce a rapid change in the environment of lipid-associated proteins, so that the dynamics of their response can then be followed through time-dependent probe spectroscopies just as is common in T-jump experiments on soluble protein folding.[14]

---

[1]This chapter is a reiteration of results that were published in ref. [8].

Understanding the evolving structure of the vesicle following a rapid temperature jump is important to fully interpret protein responses in future experiments.

The melting process studied by Nagarajan et al.[13] was fit to two distinct stages. A fast, single-exponential submicrosecond component of the transition was interpreted as an initial partial melting from preexisting defect sites. With atomistic molecular dynamics (MD) simulation data of melting bent lamellar sheets as supporting evidence, pre-melted ridges present at the contact between facets of rigid gel domains were proposed as the sites of these defects. Faceted structures have been observed in frozen vesicles imaged by electron microscopy[15] and simulated using coarse-grained[16–19] and lattice models[20, 21]. In the second stage, melting slowed down and entered a stretched-exponential phase, which was attributed to complex coupled rearrangements of domains and the influx of water.

The goal of this research is to gain a qualitative description of the melting phase transition of a vesicle and to establish if the kinetics qualitatively match experiment. To that end, we analyze the melting dynamics of simulated frozen vesicles of 13,165 and 31,021 DPPC lipids with diameters of 33 nm and 50 nm. We have found timescales for melting of single vesicles that are in good qualitative agreement with the bulk experimental values. In the current simulations we attempt to address the processes that take place during the sub-microsecond stage and the origin of the slowing-down in melting; the timescale for the second stage remains out of reach of our simulations, so we can only offer speculations.

The MARTINI 2.0 coarse-grained potential[22, 23] was used for DPPC and for the accompanying coarse-grained water. There are some discrepancies between the phase behavior of this coarse-grained DPPC and the experimental DPPC. The experimental main phase transition temperature of DPPC is 314 K[24] (or 310 K for perdeuterated DPPC[25–27]), while estimates for the MARTINI model's transition temperature are lower, at 295 ± 5 K[22] or 302 ± 1 K[28]. A more important, qualitative difference is

that the simulation model exhibits a direct transition from an untilted gel phase (L_β) to the fluid phase (L_α), while true DPPC undergoes a pretransition from a tilted gel phase ($L_{\beta'}$) to the ripple phase ($P_{\beta'}$) and then to the fluid phase at T_m. The absence of tail tilt and the ripple phase in the simulation model simplifies the behavior of the phase transition.

The MARTINI version of DPPC has in common with the experimental system that it undergoes a considerable increase in area per lipid upon melting, from 0.48 nm^2 to 0.63 nm^2 in experiment [29–34] and from 0.47 to 0.64 nm^2 in the simulation model[22]. Over the initial microseconds of melting following the temperature jump, the area of the vesicle will expand. Given that the water content of LUV's responds to perturbations on a timescale of 10's of milliseconds[2] the solvent content and internal volume remains effectively constant over the initial stages of melting, meaning that melting will induce changes in local curvature and overall vesicle shape. Nagarajan et al.[13] rationalized the initial single-exponential melting dynamics by assuming that surface pressure builds up linearly with the degree of melting. In this study, the availability of actual structures will afford a more nuanced view of how the shape changes produced by melting feed back into the melting rate through surface curvature and interfacial line tension effects.

3.1 Methods

Molecular dynamics simulations are performed using GROMACS 4.6 [35–38] and the MARTINI 2.0 coarse-grained (CG) force field[22, 23] for DPPC lipids, water and anti-freeze (included as 10% of all solvent particles for vesicle simulations). CG anti-freeze particles effectively lower the water freezing temperature from ∼290 K to below 250 K.[23]

A time increment of 25 fs is used for all simulations. Temperature and pressure are kept constant using velocity-rescaling temperature coupling[6] with a time constant of

$\tau_T = 1.0$ ps and Berendsen pressure coupling[4] with a time constant $\tau_P = 2.5$ ps and a compressibility of $\kappa = 5 \times 10^{-5}$ bar^{-1}. Pressure coupling is fully isotropic for vesicle simulations and semi-isotropic (allowing the bilayer area to equilibrate independently of the normal dimension) for patch simulations. Graphical images of molecules were prepared using VMD.[39]

3.1.1   Bilayer patch simulations

A bilayer patch of 512 MARTINI DPPC lipids and 3656 CG water particles are equilibrated at 273 K and 1 bar for 600 ns. A similar simulation at 323 K and 1 bar is equilibrated for 200 ns to provide a base line for the fluid phase lipids. The average surface area per lipid is 0.46 nm^2 for the gel phase in agreement with simulation[22] and experimental values[32–34] and 0.64 nm^2 for the fluid phase, also in agreement with simulation[22] and experiment[29–31, 34].

3.1.2   Vesicle construction and solvation

Previous simulation studies of forming a gel phase vesicle by freezing encountered delays associated with the release of internal solvent through pores and with pore closure.[16] Direct construction of a gel phase vesicle via assembly of bilayer slabs into a truncated icosahedron was used here instead.

The advantage of a truncated icosahedron starting structure for vesicle simulations is the speed with which a stable vesicle structure in the gel phase is achieved. Truncated icosahedrons consist of 12 pentagonal faces and 20 hexagonal faces and closely approximate a spherical shape, while encouraging faceting similar to that observed in other vesicle freezing simulations.[16, 17] Starting structures were prepared in three ways: with the pentagonal faces excluded to create 12 pores that serve as defect sinks, with all of the faces present and displaced from the origin sufficiently to prevent any bad contacts between lipids, and lastly with the faces displaced closer to the origin so that the bilayers overlap initially. If a pair of lipids are within 0.45 nm

of each other, one of the lipids is removed. The first two methods lead to pores that take a long time to close (in excess of 1 μs). The last structure leads to a stable and closed structure within 1 μs that is used as a starting structure for all subsequent temperature-jump simulations.

To prepare this structure, larger gel phase bilayer slabs are constructed by removing solvent from the final bilayer patch structure at 273 K and replicating the membrane in the xy-plane. The large slab is cut into pentagonal and hexagonal slabs by lipid removal. Assembly of these pentagonal and hexagonal slabs into a truncated icosahedron is detailed in Appendix B.1.

Vesicles are solvated with CG solvent containing 10% antifreeze at an initial density of 895 kg/m^3, obtained by equilibration at 1 bar pressure and 300 K. Solvent particles within 0.6 nm of lipid beads are removed.

A summary of the final truncated icosahedron structure constants is given in Table 3.1.

Table 3.1: Summary of the total number of lipids in the initial truncated icosahedrons before removing bad contact lipids ($N_{\mathrm{Lipids,i}}$), the final number of lipids after lipids with bad contacts are removed from the structure ($N_{\mathrm{Lipids,f}}$), the radial displacement of the pentagon and hexagon slabs (r_{slab}), the edge length of the slabs (l_{edge}), the number of water particles (N_{W}), and the number of anti-freeze particles (N_{AF}).

$N_{\mathrm{Lipids,i}}$	$N_{\mathrm{Lipids,f}}$	$r_{\mathrm{slab}}/\mathrm{nm}$	$l_{\mathrm{edge}}/\mathrm{nm}$	N_{W}	N_{AF}
19,060	13,165	16.637	8.00	979,144	111,053
39,876	31,021	24.957	11.50	3,062,267	347,324

3.1.3 Vesicle equilibration

The smaller vesicle (13,165 lipids) truncated icosahedron starting structure is run for 100 ns at 260 K and a pressure of 1 bar to ensure that the vesicle does not melt completely. The vesicle is then run for 1.1 μs at 280 K, below the reported main phase transition temperature of 295 ± 5 K[23]. The transition from the gel phase to fluid phase tapers off after the first few nanoseconds as is shown in Section 3.2 but the merging of gel domains persists even after having reached a stable fraction of lipids

in the gel phase.

The larger vesicle (31,021 lipids) truncated icosahedron structure is run for 1 μs at 280 K and 1 bar.

3.1.4 Temperature jump simulations

The equilibrated vesicle structure is subjected to temperature jumps with simulations for the smaller vesicle at 290 K, 295 K, 300 K, and 310 K and for the larger vesicle at 290 K, 295 K, 297 K, and 300K. These temperatures are situated around the lamellar phase transition temperature of 295 ± 5 K to allow for simulations that undergo partial melting. These simulations are run for 500 ns at a pressure of 1 bar.

The configuration from the end of the 500 ns trajectory of the small vesicle simulated at 295 K is subjected to a refreezing simulation at 280 K and run for 1 μs at a pressure of 1 bar.

Simulations for the smaller vesicle were performed on the XSEDE computational resource Trestles. Using five 32 processor nodes, the simulation completed 160 ns/day. Simulations of the larger vesicle were performed on Comet with five 24 processor nodes and completed 60 ns/day.

3.1.5 Order parameter

During analysis of simulation results, each lipid is classified as being in the fluid or gel phase based on an order parameter. The curved surface of a vesicle necessitates an order parameter defined independently of the lipid plane.

An initial categorization of each lipid as having a gel-like or fluid-like configuration is made by calculating the angle θ_T between its two tails. This angle is defined as the

angle between vectors from C1 to C4 sites in each tail, as illustrated in Fig. 3.1.

$$\overline{v}_{T1} = \overline{v}_{C4A} - \overline{v}_{C1A} \tag{3.1}$$

$$\overline{v}_{T2} = \overline{v}_{C4B} - \overline{v}_{C1B} \tag{3.2}$$

$$\theta_T = \sin^{-1}\left(\frac{\|\overline{v}_{T1} \times \overline{v}_{T2}\|}{\|\overline{v}_{T1}\| \cdot \|\overline{v}_{T2}\|}\right) \tag{3.3}$$
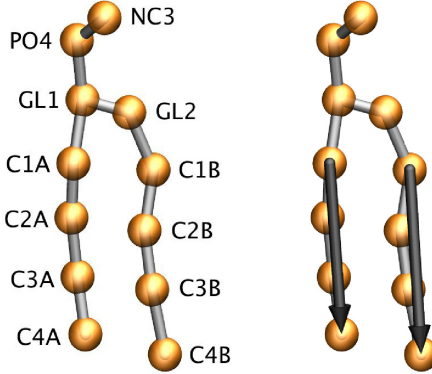
Figure 3.1: A single MARTINI 2.0 DPPC lipid with each coarse grain bead labeled and the tail vectors highlighted (black arrows).

Due to tight tail-tail packing in the ordered gel phase we anticipate that the two tails on a single gel phase lipid samples a limited range of conformations within a relatively small range of angles θ_T. The cutoff angle θ_{cut} is 15° for gel phase lipids and was chosen based on the angle distribution of gel phase bilayer patch at 273 K (Fig. 3.2, blue curve) where a strong peak exists between 0° and 20°. By comparison the angle distribution of a fluid phase bilayer patch at 323 K is significantly broader (Fig. 3.2, orange curve).

This choice of order parameter does not however account for local environment. Fluid phase lipids, while sampling a wider range of conformations and thus tail angles, still sample conformations with $\theta_T \leq \theta_{cut}$, as evident in the overlap of the two distributions in Fig. 3.2. Using only the internal order parameter to differentiate between the gel and fluid phase leads to an overestimation of the number of gel phase lipids. Lipids are therefore re-classified as fluid or gel according to their local environment: if
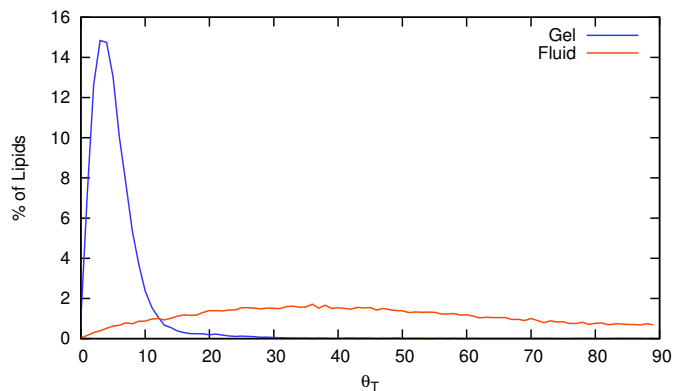
Figure 3.2: A plot of tail angle histogram for the bilayer patch of 512 DPPC lipids at 273 K (gel) and 323 K (fluid) averaged over the last 100 trajectory frames.

3 or more of a lipid's 6 nearest neighbors have $\theta_T < 15°$, it is classified as belonging to the gel phase. The implementation of this classification is demonstrated in Fig. 3.3. Interfacial lipids are then defined as gel phase lipids that have one or more fluid phase neighbors.
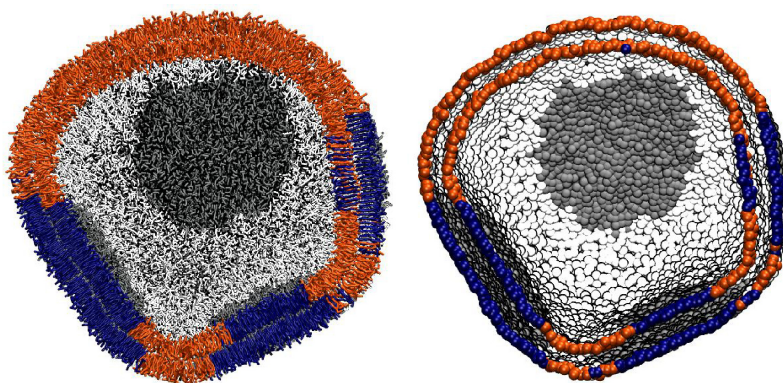
Figure 3.3: A snapshot of two cross sections of the same 13,165 lipid vesicle after melting at 290K, the left showing the full lipid structure and the right with a single bead per lipid. Each lipid is colored according to the local environment order parameter, orange being the fluid phase and blue the gel phase. Solvent although present in the simulations, is not shown in snapshots for clarity.

3.1.6 Relative shape anisotropy

The change in vesicle symmetry is quantitatively measured by the relative shape anisotropy, κ. The eigenvalues of the gyration tensor,

$$S_{m,n} = \frac{1}{N}\sum_{}^{N} r_m r_n \tag{3.4}$$

$$\mathbf{S} = \begin{bmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{yx} & S_{yy} & S_{yz} \\ S_{zx} & S_{zy} & S_{zz} \end{bmatrix}, \tag{3.5}$$

are ordered such that $S_1 \geq S_2 \geq S_3$ and κ is a function of those eigenvalues.[40, 41]

$$\kappa = 1 - 3\frac{S_1 S_2 + S_1 S_3 + S_2 S_3}{\left(S_1 + S_2 + S_3\right)^2} \tag{3.6}$$

When $\kappa = 0$ the vesicle is spherically symmetric and $\kappa = 1$ if the vesicle forms a line.

3.2 Results and Discussion

3.2.1 Vesicle equilibration

Starting from the truncated icosahedron configurations, melting occurs at the vertices and edges of domains for both SUVs within the first couple of nanoseconds. The construction of the vesicle results in there being more free volume at the edges and vertices of the vesicle for melting to begin. The percentage of gel phase can be seen dropping from 100% to 60% followed by a recovery in the gel phase (Fig. 3.4). The vesicles reach a steady state with 59.4% of the lipids in the gel phase for the 13,165 lipid vesicle and 69.8% for the 31,021 lipid vesicle. The average diameters of the SUVs are 33.47 nm and 50.28 nm.

Although the truncated icosahedron starts with 32 distinct gel domains, both of the SUVs undergo a significant consolidation of those smaller domains. This is evident in the snapshots of the vesicles in Fig. 3.5 and 3.6. Consolidation of smaller domains results in a decrease in interfacial energy between the gel and fluid phase by decreasing the length of the interfaces.
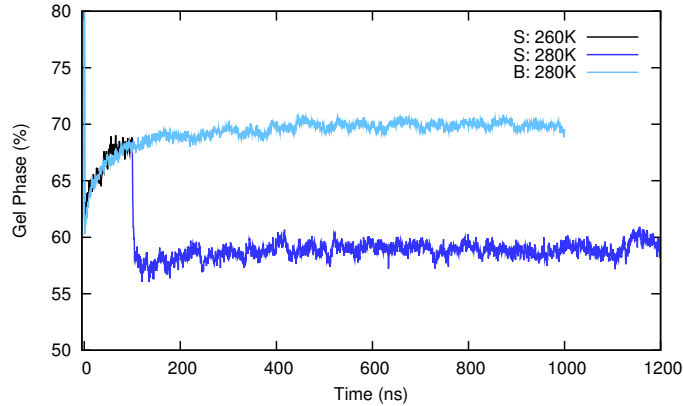
Figure 3.4: The percentage of gel phase lipids in the vesicle as it is equilibrated at 260 K (black) and then 280 K (blue) for the 13,165 lipid vesicle and 280K (light blue) for the 31,021 lipid vesicle.
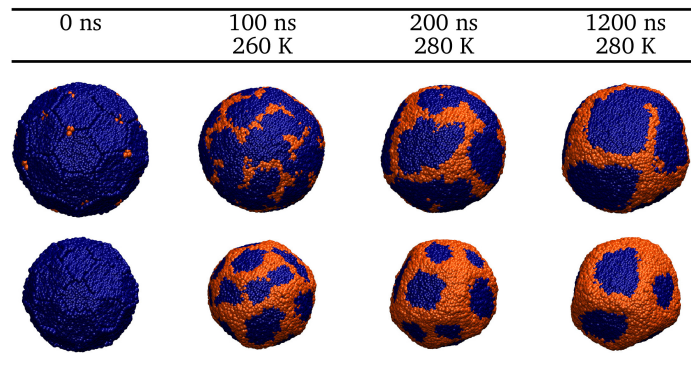
Figure 3.5: The 13165 lipid vesicle (d = 33 nm) equilibration with the outer leaflet displayed in the upper row and the inner leaflet below with the gel (blue) and fluid (orange) phases shown. In the 1.2 μs trajectory, the temperature is increased from 260 K to 280 K 100 ns into the trajectory once it is ascertained that the vesicle does not melt completely.

In Fig. 3.7 the fraction of interfacial lipids with respect to the gel phase lipids is plotted, demonstrating that despite the majority of the melting and refreezing occurring within the first 200 ns of simulation, the decrease in interface takes longer.

Decreasing the interfacial fraction is observed to happen either by domain merging and/or by some domains melting away while others growing as seen in Ostwald ripening (Section 3.2.5). Merging to decrease interfacial energy is coupled to an increase in the local curvature because fewer facets dictates sharper angles between facets. As discussed previously,[20, 21] this curvature is localized along the fluid phase regions
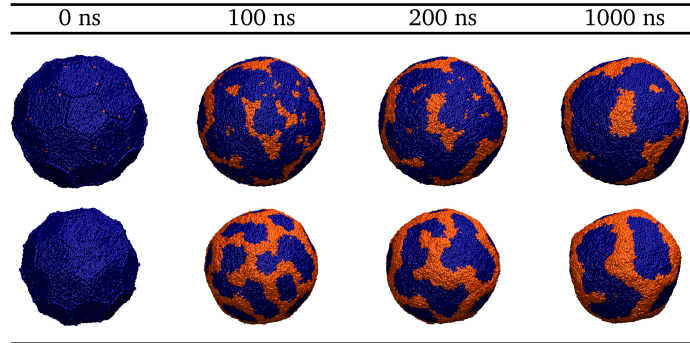
Figure 3.6: The 31,021 lipid vesicle (d = 50 nm) equilibration at 280 K with the outer leaflet displayed in the upper row and the inner leaflet below with the gel (blue) and fluid (orange) phases shown.

because the fluid phase has the lower bending modulus. Faceting is more pronounced in the smaller vesicle but present in both.
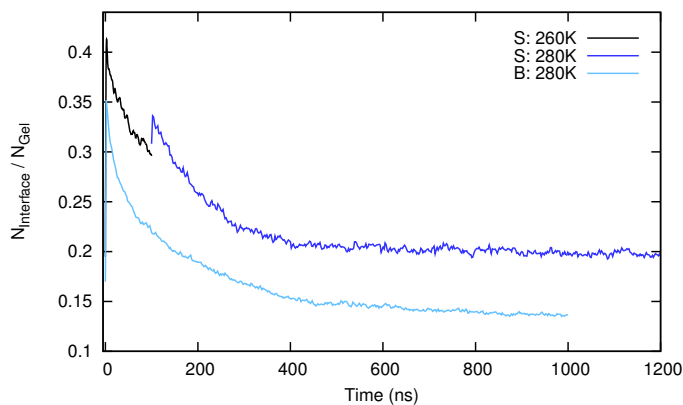
Figure 3.7: The fraction of interfacial lipids to gel phase lipids over the course of the equilibration runs.

There is no lipid flip-flopping observed in any of the simulations, equilibration or otherwise. The smaller SUV has 41.8% of the lipids in the inner leaflet while the larger SUV has 44.7%. A disparity between the population of the inner and outer leaflets is expected due to the radii of the vesicles being so close in magnitude to the lamellar thickness[22] (4.0 ±0.1 nm).

The phase composition is also asymmetric across the bilayer. In the smaller SUV 44.9% of the inner leaflet is in the gel phase as opposed to 69.8% in the outer leaflet. The larger vesicle exhibits 62.6% and 75.6% gel phase in the inner and outer leaflets

respectively.

This disparity could be promoted by curvature environments in the inner and outer leaflets differing in magnitude and direction. The inner leaflet lipid tails are permitted a larger volume and thus are more prone to melting. The radii of the vesicles are on a similar magnitude to the thickness of the bilayer and as such the radii of curvature could be different enough to have an impact on the degree of melting. The difference in phase compositions could also be an artifact, introduced by the initial construction conditions, that under populates the inner leaflet and thus promotes conversion into the higher surface area phase to compensate. The phase transition being faster than lipid flip flopping, this would be a way of relaxing any inherent surface tension or pressure within each leaflet.[42]

3.2.2 Vesicle melting rates: Comparison with experiment

Partial melting is observed in the temperature-jump simulations for the small SUV (33 nm) at 290 K and 295 K and for the larger SUV (50 nm) at 290 K, 295 K, and 297 K. Averages over the final 100 ns of the temperature jump simulations are given in Table 3.2. Plots of the percentage of lipids in the gel phase for both vesicle sizes are shown in Fig. 3.8 with single exponential fits in black. The fit constants to the single exponential form:

$$f\left(t\right) = A \exp\left(-t/\tau\right) + C \tag{3.7}$$

are given in Table 3.3. On the 100 ns timescale they exhibit approximately single exponential melting kinetics in qualitative agreement with the early stages of melting measured in IR temperature-jump experiments.[13]

The time constants (τ) are not expected to match experimental times exactly due to the effect of coarse-graining on the dynamics. Multiplying times from MARTINI model simulations by a factor of 4 to approximate experimental times has been suggested;[22] with this scaling factor, the range of timescales for melting in simulation

Table 3.2: Averages and standard deviations over the last 100 ns (200 data points) of the melting simulations for the total, inner, and outer percentages of gel lipids (x_{Gel}, $x_{Gel,i}$, $x_{Gel,o}$), the percentage of lipids in the inner leaflet (x_{Inner}), and the number of gel domains for the inner (N_i) and outer (N_o) leaflets. The percentages of lipids in the inner leaflet are 41.8% and 44.7% respectively for the smaller and larger vesicle.

13,165 Lipids (33 nm)

T/K	x_{Gel}/%	$x_{Gel,i}$/%	$x_{Gel,o}$/%	N_i	N_o
280	59.4 ± 0.8	44.9 ± 0.9	69.8 ± 0.8	6	8
290	46.2 ± 0.5	31.5 ± 0.7	56.7 ± 0.8	5	5
295	30.5 ± 0.5	17.3 ± 0.5	40.0 ± 0.6	2	3
300	0.7 ± 0.1	0.2 ± 0.1	1.2 ± 0.2	0	0
310	0.5 ± 0.1	0.1 ± 0.1	0.7 ± 0.1	0	0
280	54.3 ± 0.4	38.7 ± 0.6	65.4 ± 0.5	2	3

31,021 Lipids (50 nm)

T/K	x_{Gel}/%	$x_{Gel,i}$/%	$x_{Gel,o}$/%	N_i	N_o
280	69.8 ± 0.3	62.6 ± 0.5	75.6 ± 0.4	8	5
290	61.3 ± 0.3	52.8 ± 0.5	68.2 ± 0.4	8	5
295	51.7 ± 0.3	44.0 ± 3.1	57.7 ± 0.6	8	7
297	41.4 ± 0.5	35.0 ± 0.4	46.5 ± 0.6	5	5
300	0.7 ± 0.1	0.3 ± 0.1	1.1 ± 0.2	0	0

Table 3.3: Fitting constants for the single exponential kinetics (Eqn. 3.7) exhibited by partially melted vesicles including the time constant τ in simulation time units. Scale by 4 to compare to experimental time constants.

N_{Lipids}	T/K	τ/ns	A	C
13,165	290	91.82	7.65	46.66
	295	131.28	20.81	29.84
31,021	290	85.35	3.15	61.65
	295	133.22	9.91	51.48
	297	165.96	18.69	40.72
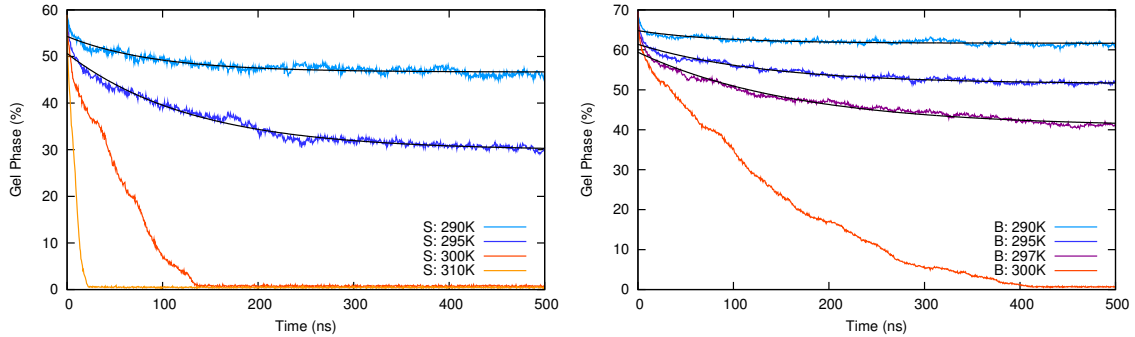
Figure 3.8: Percentage of gel phase lipids for a 33 nm SUV (left) melted at 290 K (cobalt blue), 295 K (blue), 300 K (orange), and 310 K (yellow) at 1 bar of pressure for 500 ns. Percentage of gel phase lipids for a 50 nm SUV (right) melted at 290 K (cobalt blue), 295 K (blue), 297 K (magenta), and 300 K (orange). Single exponential fits to the partially melted vesicles are shown in black.

(340 - 664 ns) matches that in experiment[13] (244 - 310 ns) to within a factor of 1.3 - 2.1.

The experiments, in measuring changes in the gel phase content averaged over an ensemble of vesicles, do not distinguish between gradual melting occurring over all the vesicles and rapid melting events distributed over a range of delay times. The qualitative similarity of the experimental ensemble average to simulated melting dynamics trends for individual vesicles strongly supports the former interpretation of the experimental data.

The initial fast melting (< 20 ns) does not fit well to the single exponential functional form indicating a different melting mechanism upon initial heating. A lack of sufficient time resolution may explain why this initial phase was not seen in experiment.

Experiments using T-jumps to final temperatures where the equilibrium state is fully melted did not yield a signal in ref. [13], presumably because the recovery time between laser shots was too short to permit nucleation of new gel phase domains.

3.2.3 Vesicle melting: Structural changes

All of the vesicles exhibit some degree of shape change upon melting, evinced by the final structures of the 33 nm vesicle in Fig. 3.9 or the 50 nm vesicle in Fig. 3.10. The number of domains (Table 3.2) in the case of the vesicle that remain relatively spherically symmetric are either static (the 50 nm vesicle at 290 K) or change only slightly (the 33 nm vesicle at 290 K and the 50 nm vesicle at 295 K).
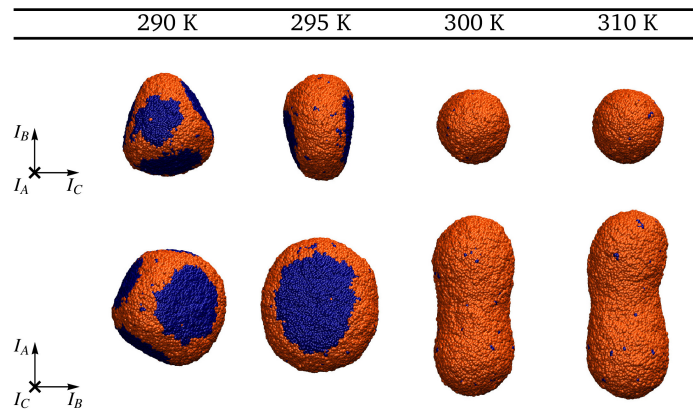
Figure 3.9: Images of the 33 nm SUV after 500 ns temperature-jump simulations from 280 K to 290 K, 295 K, 300 K, and 310 K. The gel phase (blue) and fluid phase (orange) are shown. The two rows show the same vesicles viewed down the I_A (top) and the I_C (bottom) principle moment of inertia axes. Snapshots of these vesicles melting progression can be seen in Appendix Fig. B.4.
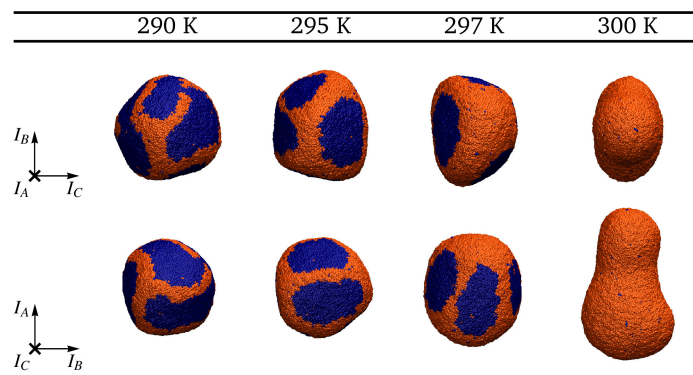
Figure 3.10: Images of the 50 nm SUV after 500 ns temperature-jump simulations from 280 K to 290 K, 295 K, 297 K, and 300 K. The gel phase (blue) and fluid phase (orange) are shown. The two rows show the same vesicles viewed down the I_A (top) and the I_C (bottom) principle moment of inertia axes. Snapshots of these vesicles melting progression can be seen in Appendix Fig. B.5.

The more drastic shape changes are coupled to a significant decrease in the number of domains. The partially melted 33 nm SUV at 295 K has asymmetric symmetry that appears to be converging on an oblate 'disk' shape, marred by a single gel domain in the outer leaflet with no matching inner domain. This domain would likely melt given a longer simulation. The approximately discoidal structure with two flat gel domains on opposite faces connected by a curved fluid-phase rim is appealingly symmetric and may represent a free energy minimum, or at least a long-lived intermediate state, for a partially melted small vesicle. A similar structure, with a single pair of gel domains present on opposite sides of a flattened (but less round) structure appears transiently during the complete melting of the 50 nm vesicle at 300 K (see Appendix Fig. B.6). The 50 nm SUV at 297 K appears to be asymmetric as well.

Fig. 3.11A and 3.11B plot the relative shape anisotropy (κ), measuring the vesicles' symmetry as they undergo partial and full melting respectively. The small SUV at 295 K undergoes the most drastic shape change of the partially melted vesicles and appears to reach a steady state. The large vesicle at 297 K also shows a large shape change but it is clear from the plot that it has not reached a steady state.

Fig. 3.8 and Fig. 3.11A show that changes in shape and the percentage of gel phase lipids occur simultaneously for the partially melted vesicles. After an initial ($\sim$20 ns) increase, the ratio of interfacial lipids to gel phase lipids steadily decreases throughout the trajectories (Fig. 3.11C). Shape changes are accompanied by the melting and merging of domains, thereby decreasing the length of gel/fluid interfaces (Fig.3.11E). These shape changes include changes in curvature and thus curvature energy in the simulations.

Full melting is observed in the smaller SUV (33 nm) at 300 K and 310 K (Fig. 3.9) and for the larger SUV (50 nm) at 300 K (Fig. 3.10) and produces structures with approximately prolate symmetry. Prolate or pear-shaped structures similar to those seen for the 33 nm vesicle and 50 nm vesicle respectively, have been reported as the
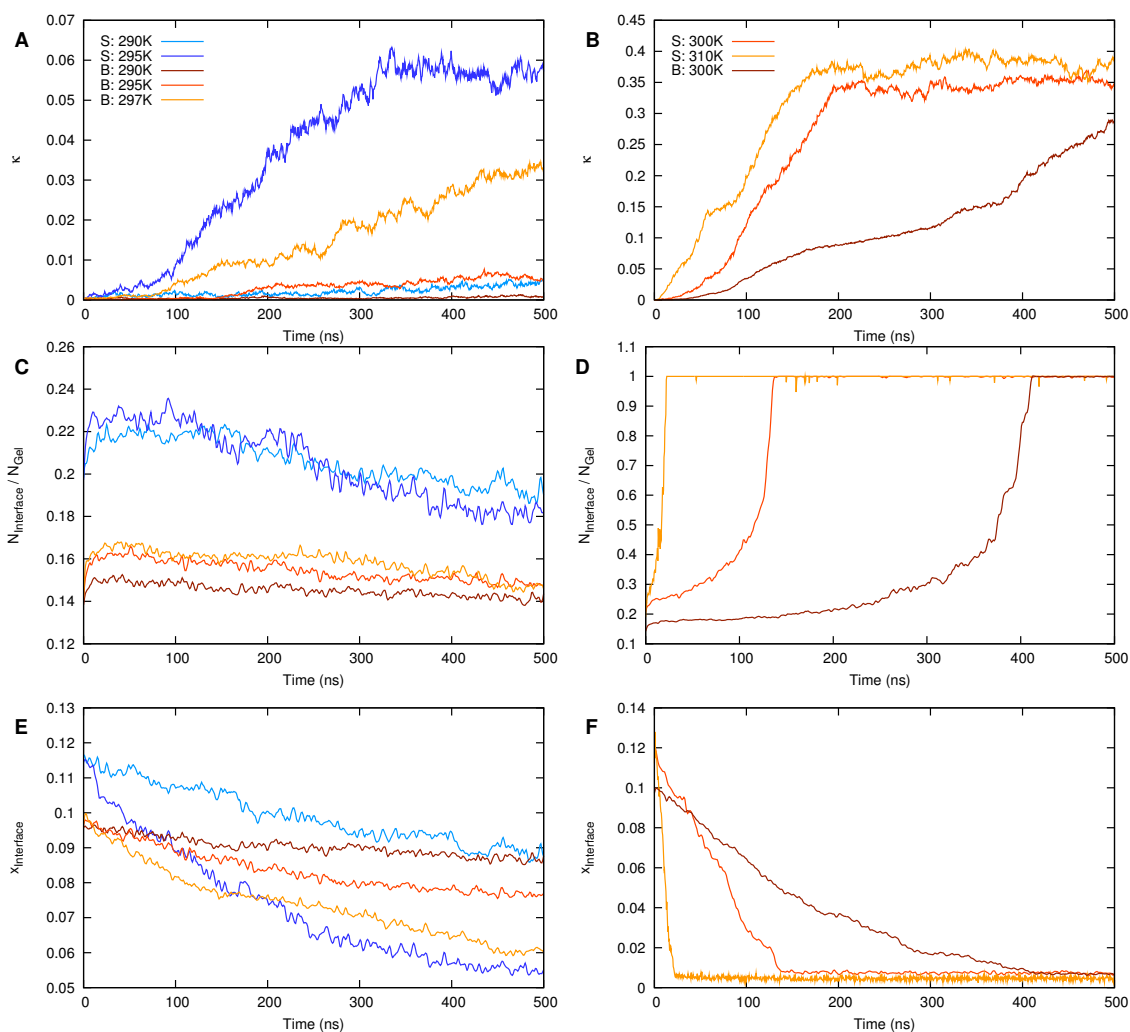
Figure 3.11: Plot of the relative shape anisotropy κ for the partially melted vesicles (A) and the fully melted vesicles (B). (More detailed moment of inertial plots can be seen in Appendix Fig. B.10.) Plot of the fraction of gel phase lipids at a gel-fluid interface for the partially melted vesicles (C) and the fully melted vesicles (D). Plot of the absolute fraction of lipids at the gel-fluid interfaces for the partially melted vesicles (E) and fully melted vesicles (F).

elastically stable structures for moderately deflated vesicles;[43] a more quantitative comparison to those predictions is given in the next section.

3.2.4 Vesicle melting rates: Interpretation

To understand the connection between the changing rate of melting and the evolving vesicle structure we consider the mesoscale contributions to the structure's free energy.

The interplay between curvature energy, line tension, and domain shape in vesicles containing both ordered and fluid domains has been investigated through experiment[44] and theory[45] in the case of GUV's where long ribbon-like domains predominate.

Sknepnek et al.[20] have described faceted vesicle structures arising from the presence of a mixture of soft and stiff bending components with some interfacial free energy penalty governing their phase separation. The bending stress of the system is supported entirely by the soft component. The structures of partially melted vesicles seen here can be understood in the same framework, except that the fraction of the different mixing components is not fixed but determined by the equilibrium between gel and fluid phases.

The degree of melting in the vesicle is determined at equilibrium by minimization of a total free energy,

$$F_{tot} = F_0 + F_{\text{interf}} + F_{\text{curv}} \tag{3.8}$$

composed of the bending energy of the bilayer (F_{curv}) and the interface between fluid and gel phases (F_{interf}) in addition to the intrinsic free energy balance (F_0) between gel and fluid phases. The state of the equilibrated vesicle before the temperature jump can be assumed to have reached a local minimum in this overall free energy with respect to the degree of melting at the edge of each gel domain:

$$\frac{dF}{d\Delta N} = 0 \tag{3.9}$$

where $\Delta N = N_f - N_g$ is the difference between the number of fluid and gel phase lipids. Throughout this discussion, we will assume the constraints of constant total number of lipids, constant internal volume and constant number of lipids in each leaflet, the latter two imposed by the slow rates of permeability and flip-flop.

For a system brought out of equilibrium through a temperature jump, it is reasonable to assume that melting rate is determined both by the length of the gel/fluid interface and the magnitude of this driving force. The total interfacial length does not change enough (Fig. 3.11E) to account for the near total slowing down observed (Fig. 3.8), so we look to each free energy component to identify how its derivative with respect to phase composition will evolve during melting. The first, intrinsic or bulk term in the full free energy can be approximated as:

$$F_0 \approx N_f \Delta H \left(1 - \frac{T}{T_m} \right) \tag{3.10}$$

using the fully frozen, planar bilayer as a reference. The derivative of F_0 with respect to the number of fluid lipids is therefore a constant at a given temperature; as $T_m = 302$ K [28], the bulk tendency opposes melting at all temperatures where incomplete melting was observed.

The interfacial contribution to the free energy also promotes melting here, because for the faceted geometry, melting decreases the area of each domain and thus the length of its boundary. We can assume that F_{interf} is given by the product of a line tension Λ and the length of the interface, which for a single circular domain scales as $(N_{\text{domain}} \, a_g)^{1/2}$. The change in F_{interf} with respect to melting in a single bilayer leaflet is:

$$\frac{dF_{\text{interf}}}{d\Delta N} = -\frac{\Lambda}{4} \left(\frac{\pi \, a_g}{N_{\text{domain}}} \right)^{1/2} \tag{3.11}$$

That is, the driving force towards melting due to the line tension increases as melting proceeds. It cannot therefore be responsible for the slowing-down of melting (unless the number of domains changes).

Table 3.4: Critical nucleus sizes (N^*) for a leaflet of a bilayer based on Eqn. 3.12 for temperatures below the bulk phase transition temperature (302 K) and a corresponding estimate of the diameter (d^*) of a circular domain consisting of that many lipids. For the MARTINI DPPC model with a bilayer line tension Λ of 10.0 ± 1.5 pN,[28, 47] a conversion enthalpy ΔH of 25.8 ± 1.5 kJ/mol per molecule,[28] and a bulk phase transition temperature T_m of 302 K[28]. The approximate diameter ranges for domains in the inner (d_i) and outer (d_o) leaflets of the two vesicle sizes are reported based on structures at the end of the trajectories.

T/K	280	290	295	297	300
N^*	4	12	37	72	449
d^*/nm	1.5	2.7	4.6	6.4	16.2
33 nm Vesicle					
d_i/nm	9-16	6-20	12-18	-	-
d_o/nm	15-24	14-23	14-33	-	-
50 nm Vesicle					
d_i/nm	10 - 50	6 - 42	6 - 35	13 - 34	-
d_o/nm	16 - 97	15 - 86	17 - 41	19 - 40	-

Equating the tendency of the line tension to shrink the gel phase domain (Eqn. 3.11) with the bulk tendency to add lipids to it (Eqn. 3.10) yields a temperature-dependent critical size, below which the (monolayer) domain is unstable:[46]

$$N^* = \pi \, a_g \left(\frac{\Lambda/2}{\Delta H \left(1 - T/T_m \right)} \right)^2 \tag{3.12}$$

Using the most recently obtained values for the transition temperature,[28] enthalpy of melting,[28] and line tension,[28, 47] the critical diameters of gel phase domains at the temperatures of interest have been calculated and are given in Table 3.4. In cases where partial melting is seen, the final domain sizes begin and remain above the critical domain size calculated at that temperature. So, while line tension is undoubtedly an influence on the melting rates, it cannot account for the abrupt decrease in melting rate in the first 20 ns or the more graduate slowing down over 100's of ns seen in experiment and simulation.

Thermal expansion of the internal solvent is a possible source of initial rapid partial melting, but the degree of expansion over the temperature jumps is limited to $1.1 - 1.8\%$ as the linear fit in the solvent volume (nm^3/bead) as a function of

temperature shows: $v_{\mathrm{sol}} = 0.09234 + 0.00014T$. Approximating the vesicle as spherical, the change in the surface area of the vesicle ΔA_{melt} due to solvent expansion, the change in the surface area per lipid transitioning from the gel phase to the fluid phase Δa_{melt}, and the number of lipids N_{Lip} in the vesicle can be used to calculate the change in the fraction of lipids in the gel phase:

$$
\begin{aligned}
\Delta Gel &= \frac{-2\left(\Delta A_{\mathrm{melt}}/\Delta a_{\mathrm{melt}}\right)}{N_{\mathrm{Lip}}} \\
&= \frac{-2(36\pi)^{1/3}\left(\left(n_{\mathrm{sol}}v_{\mathrm{sol},H}\right)^{2/3} - \left(n_{\mathrm{sol}}v_{\mathrm{sol},C}\right)^{2/3}\right)}{N_{\mathrm{Lip}}\left(a_f - a_g\right)},
\end{aligned}
\tag{3.13}
$$

where n_{sol} is the number of coarse grained solvent particles encapsulated by the vesicle, $v_{\mathrm{sol},C}$ and $v_{\mathrm{sol},H}$ are the volume per solvent bead for the temperatures before and after the T-jump, and the areas per lipid for the gel and fluid phase are respectively a_f and a_g. The solvent expansion as such would result in a decrease of $1.5 - 2.3\%$ for the small vesicle and $1.7 - 2.9\%$ by comparison to that seen in Fig. 3.8. Thus solvent expansion alone cannot account for the 20 ns of rapid initial melting.

The coupling of degree of melting to curvature is difficult to treat formally except under the idealized case of a straight ridge of constant curvature (i.e., a section of a cylinder) that joins two flat gel phase domains that maintain a fixed relative angle θ. The radius of curvature of a ridge of length L containing N_f lipids is $N_f a_f (L\theta)^{-1}$.

$$
F_{\mathrm{curv}} = \frac{\kappa}{2}\frac{N_f a_f}{r^2} = \kappa\frac{L^2\theta^2}{2N_f a_f}
\tag{3.14}
$$

At a fixed angle between facets and ridge length, increasing N_f lowers F_{curv}:

$$
\frac{dF_{\mathrm{curv}}}{dN_f} = -\frac{\kappa}{2a_f}\left(\frac{L\theta}{N_f}\right)^2
\tag{3.15}
$$

As the number of fluid-phase lipids N_f increases, the favorable contribution of curvature stress towards bending becomes weaker. This is therefore the only free energy contribution that can account for the decrease in melting rate observed in the simulations after the initial ~ 20 ns, over which the 2-D shapes of the domains and

the 3-D shapes of the vesicles do not qualitatively change (see Figs. B.4 and B.5) and the mean domain size decreases (as indicated by the increase in ratio of interfacial to gel phase lipids, Fig. 3.11C). For the larger vesicle heated to 290 K, this accounts for nearly the entire extent of the observed melting. In all other cases where partial melting is seen the melting proceeds in concert with shape changes over the remainder of the 500 ns trajectories, as evident in the final structures in Figs. 3.9 and 3.10, and in comparing time dependences of the relative shape anisotropy (Fig. 3.11A) and the degree of melting (Fig. 3.8).

In general, once the fluid zones separating gel phase domains are wide enough, the domains can adopt more compact structures (less constrained by the edges of the overall polyhedral structure) and can migrate to form a different arrangement of domains and/or fuse with others. Furthermore, complete melting of an isolated domain was seen in some cases which is expected once melting brings the domain size below the critical nucleus size.

In the case of the 300 K vesicle, the inner leaflet domain sizes are below the critical nucleus size which could in part account for the vesicle melting so rapidly that the shape change lags behind. Values for the 50 nm vesicle (Table 3.4) are also above the critical domain sizes.

For T-jumps to temperatures above (310 K) or slightly below (300 K) the planar bulk transition temperature, the initial domain sizes are either below the critical nucleus size or near enough that a small amount of shrinkage brings them below, producing complete melting. The shapes of the vesicles continue to evolve after melting is complete, in the case of the smaller vesicles reaching a peanut- or dumb-bell shaped, approximately prolate structure that persists for hundreds of nanoseconds. (The larger 300 K vesicle is continuing to evolve at the end of the trajectory, but the changes in its moments of inertia (Appendix Fig. B.9) are consistent with an approach to a prolate structure.)

Seifert et al.[43] used two different models to construct phase diagrams for elastically stable structures for vesicles subjected to a full range of deflated volumes, the spontaneous-curvature model and the coupled bilayer model.

To make a comparison to the phase diagrams produced by Seifert et al.[43], we calculated the corresponding reduced area difference, Δa, and reduced volume, v, variables based on our simulations. These two variables are normalized with respect to a spherical, fully hydrated vesicle with the same number of lipids. The radius R_0 is related to the fully melted and hydrated vesicle such that,

$$R_0 = (A/4\pi)^{1/2} = \left(\frac{N_{\mathrm{inner}} a_f}{4\pi} \right)^{1/2} , \tag{3.16}$$

where N_{inner} is the number of inner leaflet lipids and a_f is the area per lipid in the fluid phase. The reduced volume variable is:

$$v = \frac{V}{4\pi/3 R_0^3} , \tag{3.17}$$

where V is the vesicles actual encapsulated volume. The 33 nm vesicle encapsulates approximately 87,250 solvent particles. At a temperature of 300 K, the solvent density is 7.41 beads/nm^3 and the vesicle volume is 11,648.6 nm^3. The radius of the inner leaflet of a vesicle with the same lipid population in the fluid phase and full hydration is calculated to be 16.7 nm and the reduced volume for the fully melted vesicle at 300 K is 0.59.

The area difference between the leaflets ΔA,

$$\Delta A = A^{\mathrm{outer}} - A^{\mathrm{inner}} = (f^{\mathrm{outer}} - f^{\mathrm{inner}})N_{\mathrm{Lip}} * a_f , \tag{3.18}$$

where f_{outer} and f_{inner} are the fractions of lipids in the inner and outer leaflets respectively, is related to the integrated mean curvature M and distance between the leaflets D such that $M \approx \Delta A/2D$. Thus the reduced area is:

$$\Delta a = \frac{M}{4\pi R_0} \approx \frac{\Delta A}{8\pi D R_0} \tag{3.19}$$

Table 3.5: Relative volume v and change in area Δa parameters for comparison to the phase diagrams produced by Seifert et al.[43] as they apply to the fully vesicles at 300 K.

N_{Lip}	R_0/nm	V/nm^3	v	Δa
13,165	16.7	11648.6	0.59	1.48
31,021	26.6	48783.0	0.62	1.42

The actual value of Δa is highly sensitive to the definition of D, which depends on the reference plane. The "pivotal plane" is defined such that the number of lipids in a curved leaflet is proportional to its area.[48] It has been calculated for MARTINI DMPC as 0.85 nm away from the midplane, or 0.89 nm closer to the midplane than the mean distance of the phosphate beads.[48] For MARTINI DPPC, in which the phosphate beads lie at 2.0 nm from the midplane,[22] we may therefore estimate that the pivotal plane is at 1.11 nm from the midplane, and the effective thickness of the bilayer is 2.22 nm. The reduced area difference then is $\Delta a = 1.48$ for the smaller vesicle and 1.42 for the larger. Fluid vesicles with these combinations of v and Δa are predicted to adopt a prolate dumb-bell shape according to the phase diagram calculated by Seifert et al.[43] The final structures of smaller vesicles simulated at 300 and 310 K (shown in Fig. 3.9) are consistent with this prediction. The larger vesicle at 300 K appears as a pear-shaped structure after 500 ns CG simulation (Fig. 3.10), but its shape is clearly still changing (as evident from its anisotropy parameter, shown in Fig. 3.11A) and it is reasonable to assume it will adopt a similarly symmetric prolate shape as well. Using $D = 4.0$ nm in Eqn. 3.19 yields Δa values near 0.8, which would be predicted to produce cup-like stomatocyte shapes.

For both vesicle sizes, the shape change lags behind the vesicle phase transition. We can surmise as a result that the driving force behind the melting is larger than the barrier incurred by a high curvature energy surface 'transition state'. Ultimately, the fully melted vesicles would be expected to adopt spherical shapes over timescales long enough for solvent permeation and lipid flip-flop to remove the effective constraints on v and Δa respectively.

3.2.5 Topological evolution

Changes in curvature occur in two ways - melting at pre-existing gel-fluid interfaces along the edges of the vesicle and changes in the number and arrangement of gel domains. In this section we will discuss some of the more interesting trends in the topological events of vesicles that undergo partial melting.

There are two mechanisms observed that facilitate a change in the number of domains: complete melting of a domain (inner and outer) and the fusion/fission of domains. In the first case, the inner leaflet domain is the first to melt completely. With a smaller domain size, a higher curvature and a different orientation of lipids with respect to that curvature, the melting phase transition is favored by the inner leaflet. The outer leaflet can persist for tens of nanoseconds but eventually melts.

Examples of domain fusion can be seen in Fig. B.7 (domain A). In the equilibration of the smaller 33 nm vesicle, a single outer leaflet domain merged together to bridged two inner leaflet domains. The evolution of the domain's continued fusion in the first case at 290 K shows that one of the inner domains has to melt completely before the remaining one is permitted to grow and become a single mirroring domain. Alternatively in the case of the vesicle at 295 K, the two inner domains melt completely followed by the outer leaflet domain.

Domain fission is shown in Fig. B.8 (domain B) as the other direction that a bridging domain can take. The vesicle at 290 K first undergoes fission of the outer leaflet domain. The smaller inner leaflet domain melts completely and is then followed by it's mirroring outer leaflet domain. For the vesicle at 295 K the same domain in the outer leaflet undergoes a similar process except the remaining outer leaflet domain then fuses with a different neighboring domain in the outer leaflet followed by complete melting of the smaller inner domain to form a single pair of inner and outer leaflet domains.

Based on these observations, the fusion of two domains is more easily accomplished

through the outer leaflet. Subsequently one of the inner leaflet domains has to melt to facilitate a transition to a single set of inner and outer leaflet domains. Coupling between the leaflets and the curvature environment of the inner leaflet prevent them from migrating or growing until forming contact. Similar dynamics were seen in the larger vesicle as shown in Fig. B.9 (domains C and D).

3.2.6 Refreezing

In the experiments performed by Nagarajan et al. a temperature jump (T-jump) is affected over a 10 ns dead time during which there are no IR measurements. The dynamics of the melting is then probed using time resolved infrared spectroscopy (TRIR) over 10^9 to 10^2 s.[13] It is assumed that the sample remains at a constant temperature for approximately 12 ms before it cools via diffusion of heat.[14]

Thus if the experiments involve cycles of heating and cooling, does the vesicle reach a true equilibrium structure or simply vacillate between two meta-stable states? By extension, this brings into question whether the initial structure of the vesicles with many domains used here, accurately represents the structure of a vesicle in the cool part of the heating cycle.

To address the question of reversibility in light of the changes in shape and number of domains, we refreeze the small SUV structure at 295 K to 280 K. The vesicle structure at 295 K is considered the most interesting case with an almost oblate symmetry and the most extreme shape change of the partially melted SUVs (see Fig. 3.11A). The outer leaflet has 3 gel domains while the inner leaflet has 2 gel domains, thus presenting a situation where one of the gel domains does not have a mirror in the inner leaflet.

Images of the initial and final structures of the vesicle are displayed in Fig. 3.12 from four viewpoints looking down the I_A, I_B, and I_C moment of inertia principle axes of the 280 K vesicle after 1 μs of simulation.
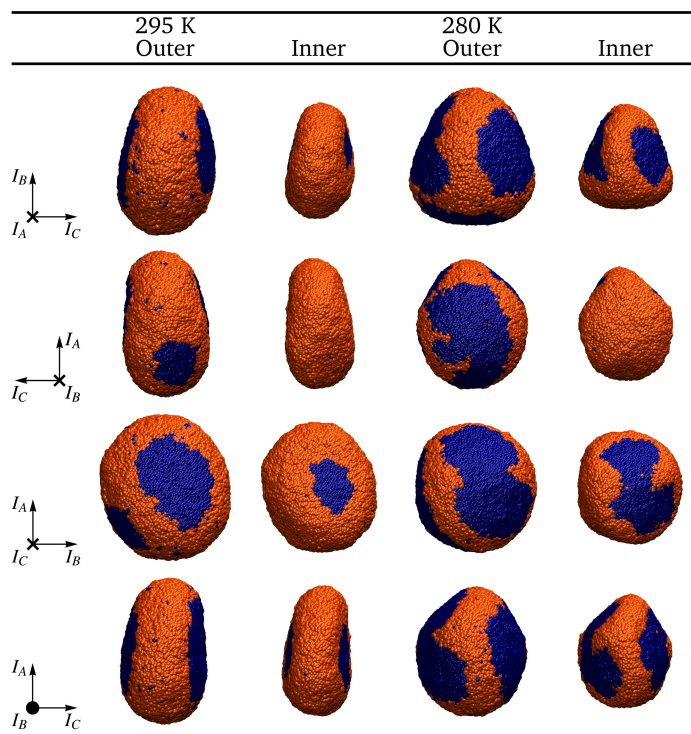
Figure 3.12: Snapshots of the 33 nm vesicle's inner and outer leaflets at 295 K (left) and at the end of a 1 μs refreezing simulation at 280 K (right).

The total number of domains in the inner and outer leaflets does not change in the 1 μs simulation. The domains all grow in size with the total percentage of lipids in the gel phase almost recovering completely at 54.3% (Table 3.2). The refreezing process is slower than the melting process as seen in Fig. 3.13 A. We propose that the rate of freezing after the initial increase in the fraction of gel phase lipids is tempered by processes governing changes in the vesicle curvature such as domain nucleation and fission.

The domain shown in the third row in Fig. 3.12 and on the left hand side of the fourth row images is mirrored across the inner and outer leaflets. The process of splitting the domain is facilitated in part by the domains bending outwards so that the lipids melt along the 'hinge' - a forming edge between two facets. This would suggest that given more time, the vesicle will revert to a structure with more domains. In Fig. 3.13B the relative shape anisotropy demonstrates a reversion to a
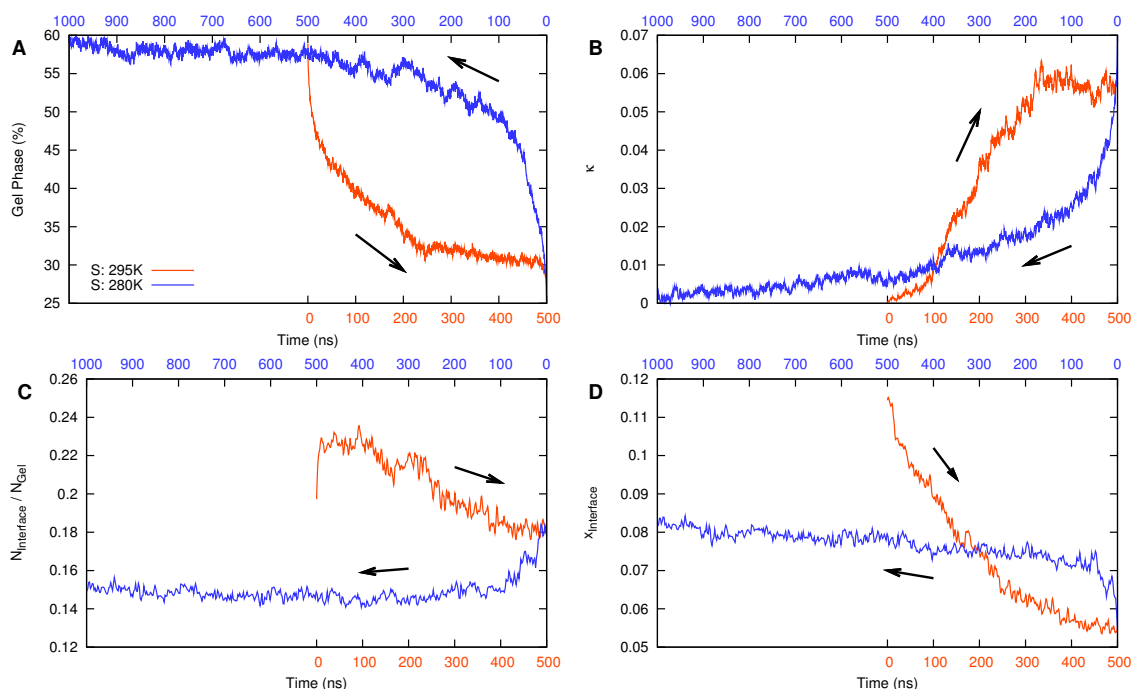
Figure 3.13: Plots of the percentage of lipids in the gel phase (A) and the relative shape anisotropy (B), the fraction of gel phase lipids at a gel-fluid interface (C) and the absolute fraction of lipids at a gel-fluid interface (D) for the vesicle melting at 295 K (orange) and subsequent refreezing at 280 K (blue).

relatively spherical symmetry that is facilitated by the bending, splitting, and growth of domains.

The outer leaflet gel domain shown in the second row of Fig. 3.12 lacks a mirroring gel domain in the inner leaflet from start to finish, although the outer domain grows quite large upon refreezing. This would confirm that the nucleation of a gel domain in the inner leaflet is a rare event. The nucleation event is hampered by the curvature of the inner leaflet. The same domain can be seen from the side on the left side of the vesicle in the third row. The size of the outer leaflet domain and rigidity of the gel phase promotes flattening of a portion of the inner leaflet. Although not seen in simulation, given enough time the flattened portion could facilitate a gel domain nucleating on the inner leaflet, thus stabilizing the gel domain in the outer leaflet.

Fig. 3.13C shows that the fraction of gel lipids at the interface initially decreases as a result of the growth of a fixed number of gel domains. The absolute fraction

of interfacial lipids (Fig. 3.13D) has a fast initial growth followed by a steady increase, indicating that the refreezing process does not reach a steady state within the simulation time.

3.2.7 Implications for dynamics beyond 500 ns

Although our simulations capture only the initial stages of melting following an ultrafast temperature jump, based on our observations we may speculate about the later stages. We hypothesize that the system approaches a plateau level of melting where the bending, interfacial, and intrinsic effects of further melting balance out, and that this accounts for the initial single-exponential phase of melting observed experimentally. Fluctuations around this degree of melting (or fluctuations in the partitioning of the gel phase fraction among multiple domains) will continue. The permeation of solvent on a time scale of 10^{-2} s is likely too slow to account for the subsequent stretched exponential. Infrequent large fluctuations may break the stability of the arrangement of domains, provoking (for instance) the complete melting of a domain and the rearrangement of the remaining gel domains, which will adopt a new plateau level of melting. Such rare events with high activation barriers could account for the stretched-exponential melting component observed in experiment. While the permeation of solvent (and associated volume expansion) is expected to have a profound effect on the shape and phase composition of the fully equilibrated vesicle structure, with a time scale of 10^{-2} s it is apparently too slow to account for the melting in this regime, approximately half of which takes place between 10^{-6} and 10^{-5} s.[13]

3.3 Conclusions

Vesicles with diameters of 33 nm and 50 nm are equilibrated to form faceted frozen topologies in agreement with previous simulations. The inner leaflet displays

a greater degree of conformational disorder than the outer leaflet.

Vesicles that undergo partial melting at temperatures below 300 K exhibit single exponential kinetics in qualitative agreement with experiment, accompanied by shape changes driven by changes in the surface area to volume ratio and curvature stress. The internal volume of the vesicle is constant over the time scale of the phase transition due to the low permeability of the bilayer to solvent, as lipids convert from the gel phase to the fluid phase, the surface area of the vesicle increases. This initial conversion serves to decrease the local curvature of the edges on the faceted structure where the fluid phase is localized. Further melting after the first (~ 20 ns) relaxation changes curvature stress until a domain is fully melted or merges with a neighboring domain, changing the vesicle shape and thus the curvature stress free energy surface, thereby influencing the melting rates of other domains.

Further significant melting will then occur upon rare fluctuations large enough to change the number of domains (complete melting or fusion). The large and heterogeneous barriers to such events may contribute to the stretched exponential nature of the slow phase of melting observed in experiment.

Vesicles that undergo full melting have distinctly different kinetics since the driving force behind the phase transition exceeds the barrier created by a build up of stresses. The final structures of these vesicles assume the prolate symmetry of a dumbbell shaped surface, a shape that has been found as a minimum curvature energy surface in previous studies using the bilayer-coupling model.[43]

Finally the refreezing of a partially melted vesicle would indicate that the cycles of heating and cooling of vesicles using T-jumps and TRIR would not simply cause existing domains to grow and shrink but for there to be significant topological changes in the number of domains present on the surface of the vesicle. This supports the hypothesis that the phase transition temperature and thus the rate of melting in the single exponential phase of the melting kinetics is a dependent on the curvature

energy of the vesicle.

3.4   Acknowledgements

Chapter 4   PEACH Method Development

This chapter will present the background in statistical thermodynamics, and in the area of number theory known as partitions theory, that underly the methods used to predict and analyze equilibrium cluster size distributions in small systems. Connection will be made to the law of mass action, the familiar relationship that governs equilibrium size distributions in the limit where the system is large, and finite size effects on equilibrium size distributions. We will explain the motivations to move away from the explicit generation of partitions for the calculation of cluster size frequencies, the subsequent development of the partition enabled analysis of cluster histograms (PEACH) method and using the grand canonical partition function Ξ as a generating function for the canonical partition function Q. We wrap up this chapter by describing the global fitting method used to extract cluster partition functions q_i° for a standard state volume V° from a series of simulation cluster size histograms and how the free energy surface of nucleation is calculated from the partition functions and in classical nucleation theory (CNT).

4.1   What is a partition function?

A partition function is essentially a sum over all the states available to an atom, particle, or molecule. For the case where the ground state energy is defined as zero, the partition function is roughly a count of all the possible states available that have thermally accessible energies. A molecular partition function can be expressed as the sum over all possible quantum states j with energies ϵ_j,

$$q = \sum_j \exp\left[-\beta\epsilon_j\right] \tag{4.1}$$

The difference between energy levels relative to the energy levels themselves at

room temperature is orders of magnitude smaller,[1] and permits the approximation of these levels as a continuum. An exception is intramolecular vibration modes; this is one reason (along with computational efficiency) that in many classical atomic simulations, bond lengths and some bond angles are constrained to a fixed value. For the classical partition function for a single particle this is:

$$q_{\text{class}} = \frac{1}{h^3} \int \exp\left[-\beta H\left(\mathbf{p}, \mathbf{q}\right)\right] \mathbf{dp}\ \mathbf{dq} \tag{4.2}$$

where $\beta = (k_B T)^{-1}$, and $H\left(\mathbf{p}, \mathbf{q}\right)$ is the Hamiltonian function dependent on a set of momenta $\mathbf{p}$ and generalized coordinates $\mathbf{q}$. This is an integral over many dimensions where $\mathbf{p}$ and $\mathbf{q}$ run over N_f degrees of freedom. The Hamiltonian for a classical system is composed of the kinetic energy as a function of the momentum $\mathbf{p}$ and mass m of the atoms, and a potential energy as a function, $U\left(\mathbf{q}\right)$. For a monatomic gas, the Hamiltonian is expressed as follows:

$$H\left(\mathbf{p}, \mathbf{q}\right) = \frac{1}{2m} \sum_{j=1}^{N} \left(p_{xj}^2 + p_{yj}^2 + p_{zj}^2\right) + U\left(\mathbf{q}\right) \tag{4.3}$$

The total partition function Q_{class}, can thus be taken as the total count of states available to a system of atoms or molecules. Given a monatomic gas with a single atomic type, the total partition function can be taken as the product of the partition functions q_{class} for N atoms normalized with respect to the $N!$ number of permutations available to those atoms. Essentially we want to normalize for the indistinguishability of one atoms of said gas from other atoms of the same gas.

$$Q_{\text{class}} = \frac{q_{\text{class}}^N}{N!} = \frac{1}{N!} \left(\frac{2\pi m k_B T}{h^2}\right)^{3N/2} Z_N \tag{4.4}$$

Integration of the potential energy contribution to the Hamiltonian results in Z_N, the classical configuration integral.

$$Z_N = \int_V \exp\left[-\beta U\left(x_1, \cdots z_N\right)\right] dx_1 \cdots dz_N \tag{4.5}$$

---

[1]See chapter 5 of ref. [49] where, in deriving the translational partition function for a monatomic ideal gas, it is shown that there is approximately 30 orders of magnitude difference between the difference between the energy levels and a typical value for a translational energy level.

In the case where the atoms are non-interacting, $U(\mathbf{q}) = 0$ and the configuration integral is $Z_N = V^N$. This result is equivalent to the calculation of the translational partition function $q_{\text{trans}}(V,T)$ for a particle in a box.

$$Q_{\text{class}} = \frac{1}{N!} \left(\left(\frac{2\pi m k_B T}{h^2} \right)^{3/2} V \right)^N = \frac{q_{\text{trans}}(V,T)^N}{N!} \tag{4.6}$$

If we extend beyond an ideal monatomic gas to molecules or to atoms/molecules that interact with one another to form clusters, the partition functions become more complex, requiring the partition function to include an internal component. The internal part of the partition functions has to account for all the different configurations that the molecules and/or cluster can assume while also accounting for the indistinguishability of atoms within the molecule or atoms/molecules in an aggregated cluster.

$$
\begin{aligned}
q(V,T) &= q_{\text{class}}(V,T) \; q_{\text{internal}}(T) \\
&= q_{\text{trans}}(V,T) \; q_{\text{rot}}(T) \; q_{\text{vib}}(T) \; q_{\text{elec}}(T) \; q_{\text{nucl}}(T)
\end{aligned}
\tag{4.7}
$$

For a single molecule in a quantum mechanical context, accounting for the permutations of atoms within the molecule occurs in the rotational partition function $q_{\text{rot}}(T)$ where a symmetry number σ normalizes for permutations in rotational symmetry in the bonded atoms of the molecule.[2]

The method we will be outlining in this chapter does not require the formal evaluation of integrals or hamiltonians to solve for the partition function. It is however essential to establish what states we define our partition functions to encompass.

We can define a partition function for any grouping of atoms that is distinguishable, whether that is on an atomic, molecular or aggregate scale. The level at which we make this separation is somewhat arbitrary[3] but as we discuss the partition functions

---

[2]See chapter 8 of ref. [49] for a discussion of the partition functions for polyatomic molecules.

[3]Caution should be used when defining distinguishable aggregates. Not accounting for the distinguishability of the components forming a cluster can lead to very poor fitting of partition functions. For example, using a single component approach for clusters consisting of two components such as NaCl leads to very poor results in reproducing cluster frequency distributions.

for clusters, we assume that the partition function for an i-mer cluster, q_i, includes all of the possible unique conformations that a cluster of i components can assume. While conformations with different internal coordinates are considered to be unique, interchanging two identical atoms within a single conformation is a permutation we assume the partition function does not double count. It is assumed that the partition function of an i-mer includes the normalization for internal permutations of like atoms or molecules.

4.1.1 Volume scaling of the partition function

An important concept to reinforce is that the partition functions scale linearly relative to the system volume. Recall that the partition function can be separated into a translational and internal portion wherein the translational portion is a function of both volume and temperature.

$$q\left(V, T\right) = q_{\text{trans}}\left(V, T\right) q_{\text{internal}}\left(T\right) \tag{4.8}$$

This linear dependence means that partition functions for the same molecule or cluster but two different system volumes, V_1 and V_2 are related by,

$$\frac{q\left(V_1, T\right)}{V_1} = \frac{q\left(V_2, T\right)}{V_2}. \tag{4.9}$$

It is this volume scaling that enables us to perform global fitting over simulated systems with different solute concentrations and system sizes. We can define a standard state volume V° (which will be the inverse of a standard state concentration c°) that can be used to relate each of the simulated systems with a volume V to the standard state partition function q°. In section 4.4, the standard state partition functions are progressively fit using the following function to convert from a standard state partition function to the partition functions for the simulated volume.

$$q\left(V, T\right) = q^\circ\left(V^\circ, T\right) V/V^\circ. \tag{4.10}$$

4.2 The law of mass action

The law of mass action is a relationship that is taught to students in general chemistry classes as a way to calculate the equilibrium constant for a reaction based on the reaction stoichiometry and the reaction component concentrations. For example in a reaction,

$$\nu_A A + \nu_B B \rightleftharpoons \nu_C C + \nu_D D \tag{4.11}$$

the equilibrium constant is defined as the following by the law of mass action:

$$K_{\text{LoMA}}(T) = \frac{(\rho_C V^\circ)^{\nu_C} (\rho_D V^\circ)^{\nu_D}}{(\rho_A V^\circ)^{\nu_A} (\rho_B V^\circ)^{\nu_B}} = \frac{q_C^{\circ\,\nu_C} q_D^{\circ\,\nu_D}}{q_A^{\circ\,\nu_A} q_B^{\circ\,\nu_B}} \tag{4.12}$$

where ν is the stoichiometric coefficient, ρ is the concentration (N/V), q° is the partition function for a standard state volume V°. Normalizing with respect to volume serves to make the partition functions only a function of temperature.

There are two assumptions built into this equation:

1. The equilibrium constant, regardless of how we calculate it, is a function of temperature only for ideal systems. When non-ideal effects such as crowding and interactions between components beyond the momentary interaction of reacting or aggregating (as is the case in our studies) occur, the equilibrium constants begin to be concentration dependent. This can influence the quality of the PEACH global fitting as will be shown for aggregation of NaCl clusters in water in section 6.3.3.

2. The law of mass action only works well for systems whose concentrations are negligibly perturbed by fluctuations in the extent of the forward or backwards reaction, which is only possible in the limit of large system sizes where the reactant and product populations can be treated as continuous variables.

This second aspect can be seen in the derivation of the law of mass action from

equilibrium conditions.[49][4] If we start from the equation for the change in Helmholtz free energy at constant temperature and volume and define a variable for the extent of the reaction such that $dN_j = \nu_j d\lambda$,

$$dA|_{T,V} = \sum_j \mu_j dN_j = \left(\sum_j \mu_j \nu_j \right) d\lambda. \tag{4.13}$$

At equilibrium, the change in free energy relative to the progression in the reaction $(\partial A / \partial \lambda)_{T,V} = 0$:

$$0 = \nu_C \mu_C + \nu_D \mu_D - \nu_A \mu_A - \nu_B \mu_B \tag{4.14}$$

If the chemical species are independent and distinguishable, the total partition functions is defined as the product of the partition functions of the components:

$$Q\left(\{N_j\}, V, T\right) = \prod_j Q(N_j, V, T) = \prod_j \frac{q_j^{N_j}}{N_j!} \tag{4.15}$$

and the chemical potential for a given species can be defined as the partial derivative of the Helmholtz free energy with respect to the population of that species. In taking the partial derivative with respect to N_j we are making the assumption that the number of j in the system is large enough that we can treat N_j as a continuous variable rather than a discrete one.

$$\mu_j = -k_B T \left(\frac{\partial \ln Q}{N_j} \right)_{\partial N_{i,i \neq j}, V, T} = -k_B T \ln \frac{q_j\left(V, T\right)}{N_j}. \tag{4.16}$$

Stirling's approximation states that $N! \approx N \ln N - N$ in the limit of large N, and is what allows us to derive the second expression for the chemical potential. Combining Eqn. 4.16 and 4.14, and using Eqn. 4.10 to convert the partition functions to a standard state partition function, allows us to get the equilibrium constant in Eqn. 4.12. This matters in the context of simulations since the system sizes are typically finite.

So what does this mean in the context of applying the law of mass action to simulated systems, where computational expense imposes limitations on system sizes

---

[4]This derivation is a reiteration of the one given in Ch. 9 of ref. [49]

and thus N_j is in the realm of being a discrete variable? We focus on this question in the context of using the law of mass action to calculate the free energy surface of aggregation where it is important to continuously sample aggregate sizes up to some aggregate size limit.

4.2.1 The limitations of the law of mass action

We use a free energy profile for the aggregation of surfactants into micelles derived by Maibaum et al.[50], to illustrate the system size dependence of the law of mass action equilibrium association constants. Micelle aggregation is ideal for this demonstration because micelles have a finite optimal cluster size. This allows us to be at the saturation concentration or critical micelle concentration ρ_{cmc} (Eqn. 4.18) and form the maximum micelle (cluster) size; all while still allowing for a variety of other cluster sizes or multiples of the maximum cluster size to coexist in the simulation.

The free energy (Eqn. 4.17) is a function of the temperature T (293.15 K), the oil-water surface tension γ_{ow} (15.0×10^{-3} N m^{-1}), the free energy difference between a hydrophobic tail within the micelle core and solvated in water $\Delta\mu$ (11 kJ mol^{-1}), the surfactant length δ (2 nm), and the girth of the surfactant a (0.5 nm). The values for these constants were chosen to achieve a free energy minimum at a micelle size of 40 (Fig. 4.1).

$$\frac{\Delta G}{k_B T} = -\frac{\Delta\mu}{k_B T} n + \frac{(36\pi)^{1/3} \gamma_{\text{ow}} a^2}{k_B T} \left(\frac{\delta}{a}\right)^{2/3} n^{2/3} + \left(\frac{3}{4\pi}\right)^{2/3} \frac{96}{49} \left(\frac{a}{\delta}\right)^{4/3} n^{5/3} \quad (4.17)$$

$$\rho_{\text{cmc}} = \frac{1}{a^3} \exp\left[\left(\frac{5832}{49}\right)^{1/3} \left(\frac{\gamma_{\text{ow}} a^2}{k_B T}\right)^{2/3} - \frac{\Delta\mu}{k_B T}\right] \quad (4.18)$$

The free energy is used to generate the partition functions for a system with a surfactant concentration equal to the critical micelle concentration for the free energy surface for increasing system sizes N_{sys} and system volume V. For each system size, the total partition functions were calculated using iterative derivatives of the grand canonical partition function for that system volume and size (using the formulation
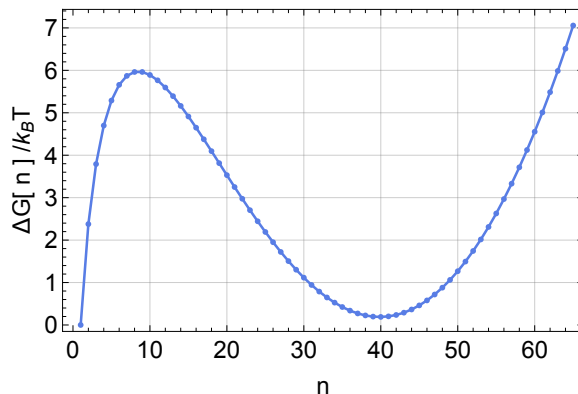
Figure 4.1: Free energy of association (Eqn. 4.18) relative to the monomer free energy for micelle aggreagation.

in section 4.3.4). As is demonstrated in Appendix A.2.1, the relative divergence between the actual equilibrium constant and the law of mass action equilibrium constants (Eqn. A.14) can be calculated solely using ratios of total partition functions $Q(N, V, T)$. The results of this analysis is shown in Fig. 4.2.
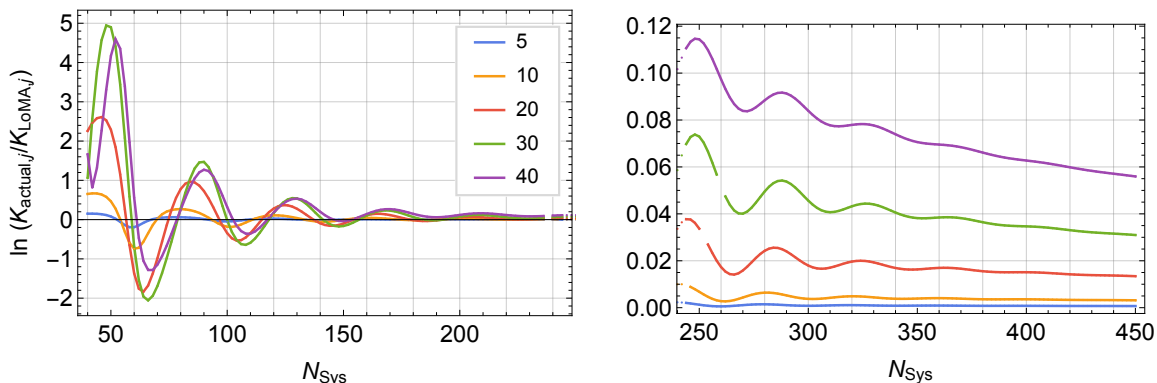
Figure 4.2: The ratio (Eqn. A.14) between the exact and law of mass action equilibrium association constants for surfactants in clusters of size 5, 10, 20, 30, and 40 as the total number of surfactants in the systems (N_{Sys}) increase.

The oscillation periods in the deviation between the equilibrium constants are approximately equal to the optimal cluster size of 35-37 surfactants, as determined by the most frequent large micelles seen in the cluster size distributions in Fig. 4.3 for the largest N_{sys}. The oscillations and magnitudes of the divergence between the equilibrium association constants in Fig. 4.2 appear to die away as the system size increases. For a system size that can accommodate a maximum of 7 micelles ($\sim$260

surfactants), the error in the law of mass action equilibrium constant for a 40-mer is an underestimation by $\sim 10\%$.
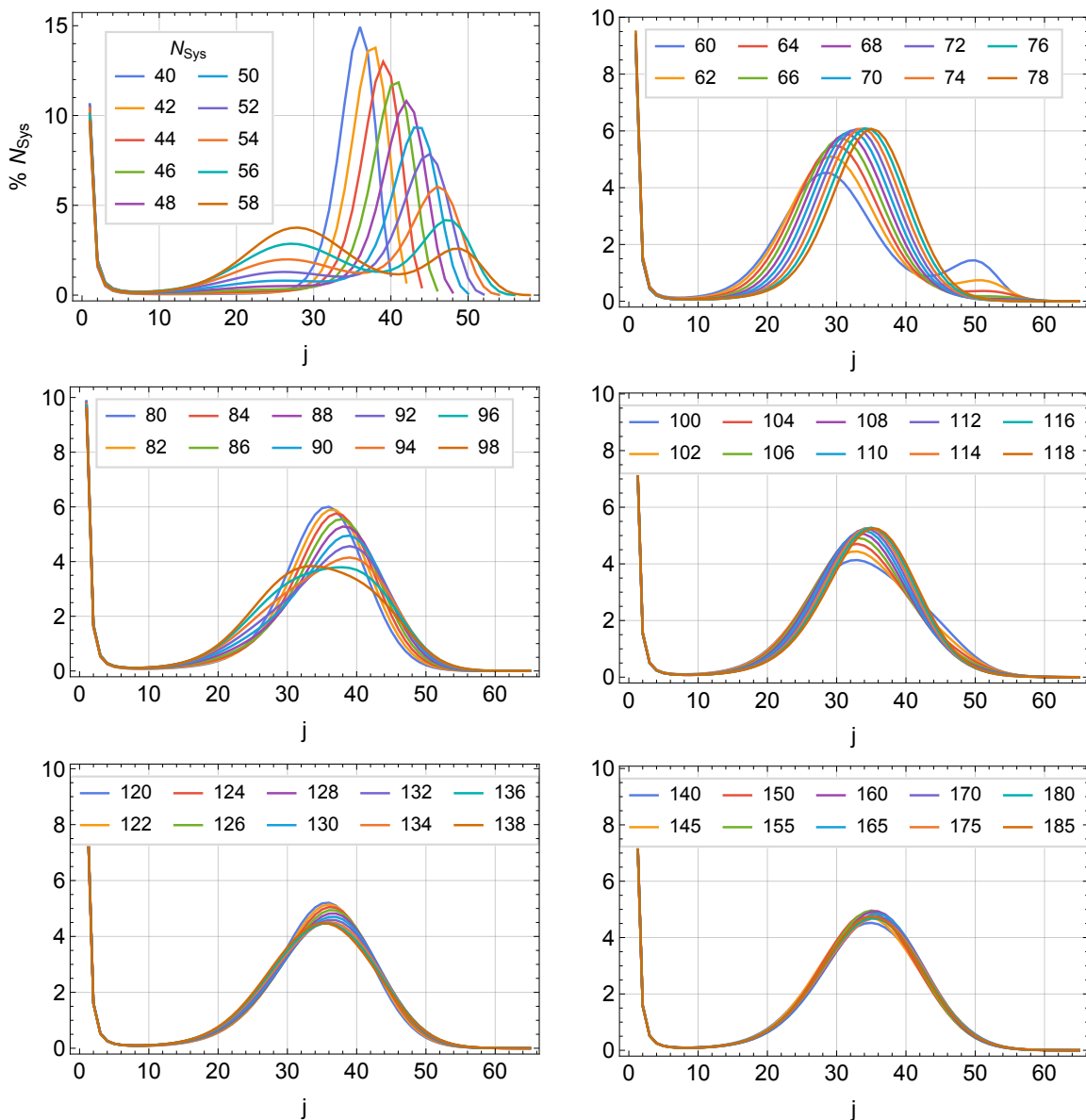
Figure 4.3: Micelle size distributions for systems of size N_{sys} based on the free energy surface in Fig. 4.1 and a surfactant concentration equal to the critical micelle concentration given by Eqn. 4.18.

There are very pronounced finite size effects reflected by the micelle size distributions in Fig. 4.3 for small systems ($N_{\text{sys}} < 100$) that lead to a bimodal micelle size distribution. This comes about as a result of not having enough surfactants in the

system to form the optimal cluster size in the presence of an ideal monomer concentration. The two peaks signify two less than optimal cluster sizes that are energetically competitive for the systems.

Johnston et al. [51] claim that for simulations of micelles to produce relatively good average micelle size frequencies even for the largest micelles, the system size must contain enough material to form several micelles. They proposed that simulations capable of forming at least 5 of the large micelles were likely to produce good statistics for micelle size distributions. While it is not possible to draw direct comparisons between their simulations and the results of the theoretical free energy used here, there appears to be relatively good agreement between the two estimations for how large the surfactant sizes have to be in order to escape finite size effects.

But what happens when one simulates a nucleating system where there is an energetic barrier to the formation of a cluster of the critical nucleus size and then, given a sufficient supply of free monomers, the aggregate continues to grow until the simulation size limits its growth?

4.2.2   Finite size effects: What happens to the free energy surface in a simulation?

Let us begin by considering nucleation of a crystal in an experimental context where we have a saturated solvent; no further solute can be solvated in the solution ($c_{\mathrm{sys}} = c_{\mathrm{sat}}$). It has been posited that the conditions for crystal growth are set by localized density fluctuations of solute in the solution where clusters form and break apart in the high density solute regions. The cluster's growth is furthered via nucleation, a rare event where a cluster grows sufficiently large to overcome the nucleation barrier, forming a relatively stable crystal at the critical nucleus size. There may or may not be a phase transition to an ordered structure as the crystal grows.

Crystal growth in an experimental context is driven by evaporation of the solvent. We can think of the solute concentration being minutely raised above the satura-

tion concentration by evaporation. Deposition of the solute onto the crystal nucleus depletes the free solute (monomers) in the solvent, returning the free solute concentration to the saturation concentration. This mechanism of crystal growth maintains an equilibrium and thus an equal chemical potential between the free monomers in solution and those associated with the crystal.

Simulations in the canonical (NVT) and isothermal-isobaric (NPT) ensembles have a fixed population of solutes and solvents, and thus encounter finite size effects. In the dilute limit ($c_{\mathrm{sys}} \leq c_{\mathrm{sat}}$), only small clusters ($i << N_{\mathrm{sys}}$) form with statistically relevant frequency since the formation of the clusters deplete the free monomers solution to below the saturation concentration, impeding further growth. Chen and Pappu[52] achieved consistent results calculating the equilibrium constants for the aggregation of small ionic cluster (size ≤ 6) in relatively large and dilute unbiased molecular dynamics simulations using the law of mass action.

To observe larger cluster sizes, simulations must be performed at concentrations significantly higher than the saturation threshold.[53] The the formation of a cluster of solutes depletes the number of available free monomers, decreasing the free solute concentration. If we start a simulation at the saturation concentration c_{sat}, the formation of a cluster of any size effectively lowers the solvent concentration below c_{sat} and thus destabilizes larger crystals. If the total concentration is not well above c_{sat}, this depletion effect impedes the formation of cluster nuclei.[54] Salvalaglio et al. termed this a consequence of a confinement effect, like a fixed system size, where the stability of clusters does not depend on just the supersaturation but also the system size.[55]

Fig. 4.2.2 shows a very simplified effective free energy (blue lines) that exemplifies how the depletion effect alters the true free energy surface (orange lines) for nucleation of MTBE clusters in a finite system of 30 molecules. The model assumes that the system is populated by one cluster of size i, and $N_{\mathrm{sys}}-i$ free monomers.[5] The effective

---

[5]This is very simplified since it artificially imposes a limit of one partition for each cluster size i. Even in a finite system, the full scope of partitions $N - i$ should be accounted for.
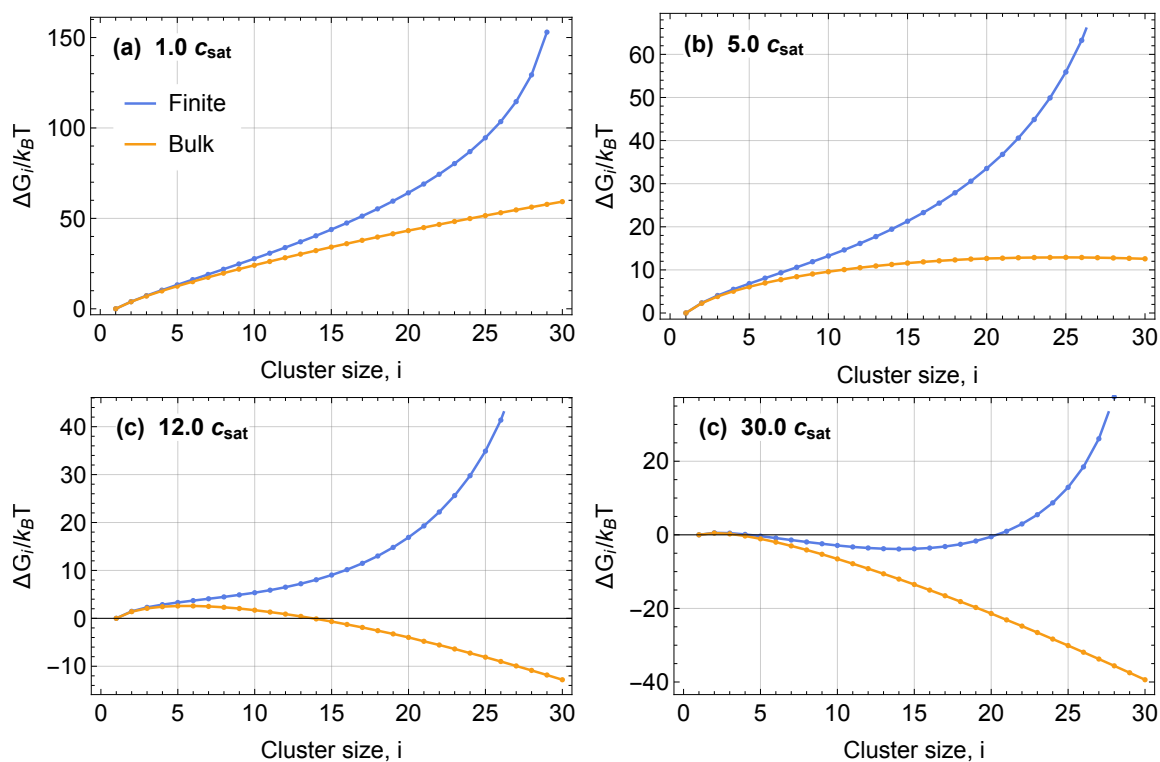
Figure 4.4: The effective free energy surfaces for MTBE (Chapter 5) aggregation at scaled monomer concentrations of the saturation concentration c_{sat} (14.52 mM based on the classical nucleation theory (CNT) fit (Table 5.2)). The blue lines indicate the effective free energy surface for a system with a total of 30 MTBE at the given concentrations assuming the presence of one large cluster of size i and the rest of the monomers. This gives an approximate indication of the depletion effect of forming a cluster on the free energy surface. The corresponding free energy of a system large enough to not experience any notable depletion effect is shown in yellow.

monomer concentration is thus a function of the cluster size i, $c_{1,i} = (N_{\text{sys}} - i)\,V_{\text{sys}}$. The effective free energy for an i-mer is calculated as the true free energy for the i-mer given the adjusted monomer concentration $c_{1,i}$ and a total system size N_{sys}. A similar model has been explored by Wedekind et al..[54]

While the model used in this case likely over estimates the degree of supersaturation required to encourage the formation of larger clusters, it does demonstrate the qualitative behavior of a system under confinement. In systems that undergo a nucleation type mechanism of growth, once the critical nucleus forms, this propels the system into rapid cluster size growth. The formation of clusters depletes the concentration of free monomers in solution to around the saturation threshold. The equilibrium state consists of some cluster size fluctuation around one large cluster in the system in the presence of free monomers and small clusters. This makes it impossible to establish equilibrium constants for the transient intermediate cluster sizes from high concentration and large unbiased simulation.

However the finite size effects can be used to tune cluster size sampling for a range of cluster sizes. Across a number of simulations at different concentration (well above c_{sat}) and/or system sizes, the influence of finite size effects allows continuous coverage of all the cluster sizes around and past the critical nucleus size. At these smaller system sizes, the formation and breaking of larger clusters significantly alters the concentration of other cluster sizes in the system and the law of mass action no longer serves as a good model to calculate the equilibrium association constants. In the subsequent sections we will derive the PEACH (Partition enabled analysis of cluster histograms) method functional forms, and explain the global fitting methods that allow the calculation of equilibrium constants and thus free energy surfaces for a range of cluster sizes using simulation data from relatively small system simulations.

4.3 PEACH: Partition enabled analysis of cluster histograms

It is important before proceeding to have a thorough understanding of what we mean by a partition. In mathematics, an integer partition of N denotes a set of positive integers whose sum equals N. In its physical manifestation, a partition is simply one way of grouping N objects and is independent of the order in which those parts are grouped. The partition count $p(N)$ is the number of distinct partitions that can be formed from N identical objects independent of ordering.

Fig. 4.3 shows the example for the partitions of five objects which in the context of the subsequent sections will be individual atoms or molecules. Note that partition count in this case is $p(5) = 7$.
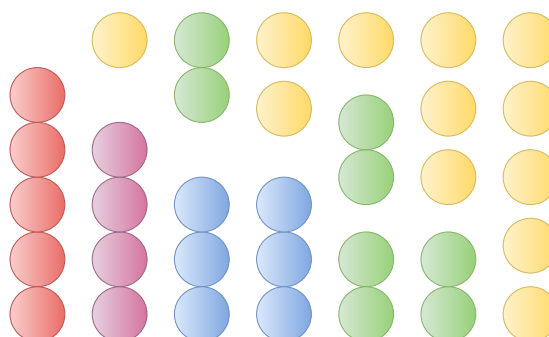
Figure 4.5: Each column in this figure shows one distinct way 5 objects can be grouped and thus is a partition. Clusters of different sizes are given different colors. Note that permutations on these groupings are not included.

For a large number of parts N the number of partitions $p(N)$ can be approximated using the Hardy-Ramanujan-Rademacher asymptotic formula,[56]

$$p\left(N\right) \approx \frac{1}{4N\sqrt{3}}e^{\pi\sqrt{2N/3}}. \tag{4.19}$$

For a set of objects with two distinguishable types, where the number of parts 1 and the number of parts 2 are equal ($N_1 = N_2 = N$), the number of partitions grow even

more rapidly,

$$p\left(N, N\right) \approx \frac{a}{n^{55/36}} \exp\left[bN^{2/3} + cN^{1/3}\right] \qquad (4.20)$$

where a, b, and c are constants.[57, 58]

We begin by outlining how explicit enumeration of partitions can be used to calculate thermodynamic averages in section 4.3.1 and can be used to fit the free energies of a system. The exponential growth in the number of partitions rapidly makes the calculation of ensemble averages though the generation of individual partitions very computationally expensive however, and this motivated the development of the partition-enabled analysis of cluster histograms (PEACH)[59] methods laid out in section 4.3.2-4.3.4. The derivations herein are for a single component system. The derivations and equations for two-component (or bipartite) systems are given in ref. [59].

4.3.1 Canonical ensemble partition function

It is most logical to start an explanation of how partition functions can be used to generate average quantities in the canonical ensemble where the number of molecules N, the volume V and the temperature T are kept constant. These are the same conditions maintained in the MD simulations we have performed for MTBE vapor phase cluster formation.

We begin with the familiar canonical partition function for N indistinguishable, non-interacting and thus non-clustering molecules with a partition function $q\left(V, T\right)$ in Eqn. 4.21. This is sometimes referred to as the ideal gas assumption and holds true in the limit of dilute systems. Admittedly, in order for clusters to form, a system must be saturated or super-saturated, thus violating this assumption at the very least when a cluster forms or breaks apart. The extension to a set of clusters is however similar.

$$Q(N, V, T) = \frac{q\left(V, T\right)^N}{N!} \qquad (4.21)$$

If we make the assumption that the clusters of different sizes are independent and distinguishable species, we can express the total partition function for a given partitioning of N monomers, $\{n_i\}$, as the product of the individual partition functions:[49]

$$Q\left(\{n_i\}, V, T\right) = \prod_{i=1}^{N} Q_i\left(n_i, V, T\right) = \prod_{i=1}^{N} \frac{q_i^{n_i}}{n_i!} \tag{4.22}$$

where the partition is constrained so that total number of monomers N is constant:

$$N = \sum_{i=1}^{N} i \, n_i \tag{4.23}$$

Summing Eqn. 4.22 over all possible partitions of N monomers, we get the canonical ensemble partition function:

$$Q\left(N, V, T\right) = \sum_{k=1}^{p(N)} \prod_{i=1}^{N} \frac{q_i^{n_{i,k}}}{n_{i,k}!} \tag{4.24}$$

To test a set of partition functions $\{q_i\}$ against simulation results we take the product of a measurable quantity, for example the number of a given cluster size s ($n_{s,k}$), in a given partition k multiplied by the probability weighting for that partition summed over all the possible $p(N)$ partitions for N molecules. It is worth noting that in this situation, the calculating of the average cluster size frequency would require the explicit generation of each possible partition. As shown in Eqn. 4.19, the number of partitions grows exponentially. While Eqn. 4.25 holds true for any system size, the generation of partitions becomes computationally intractable for larger systems and even for relatively small bipartite systems (Eqn. 4.20).

$$\langle n_s \rangle = Q(N, V, T)^{-1} \sum_{k=1}^{p(N)} \left(n_{s,k} \prod_{i=1}^{N} \frac{q_i^{n_{i,k}}}{n_{i,k}!} \right) \tag{4.25}$$

4.3.2 Grand canonical ensemble partition function

As in other problems in statistical mechanics (Bose-Einstein condensation, for instance[49]), some qualities are easier to calculate in the grand canonical ensemble than in the canonical ensemble. We begin by defining the grand canonical partition

function Ξ as a function of the absolute activity $\lambda = e^{\mu/k_B T}$ and the canonical partition functions Q with a total number of particles N, a volume V, and temperature T.

$$\Xi(\lambda, V, T) = \sum_{N=0}^{\infty} Q(N, V, T) \lambda^N \tag{4.26}$$

For an a system of non-interacting, indistinguishable particles that do not form clusters, the canonical partition function is the rather familiar Eqn. 4.21[49] and substitution into Eqn. 4.26 garners a sum over all system sizes,

$$\Xi(\lambda, V, T) = \sum_{N=0}^{\infty} \frac{(q_1 \lambda^1)^N}{N!} = \exp\left[q_1 \lambda^1\right]. \tag{4.27}$$

The extension of this function to a system which can form clusters is established by taking the product of a similar sum for each cluster size. The sum in Eqn. 4.28 accounts for all the possible variations of the number of i-mers that are possible in a configuration independently of the total number of particles in said configuration. Addition being equivalent to an 'or' statement in probability, it makes sense that the number of an i-mer in any configuration is exclusive in the sense that a configuration cannot simultaneously have no i-mers and have n i-mers. The product ensures that all possible combinations of cluster sizes are accounted for.

$$\Xi(\lambda, V, T) = \prod_{i=1}^{\infty} \left[\sum_{m_i=0} \frac{(q_i \lambda^i)^{m_i}}{m_i!}\right] \tag{4.28}$$

The connection back to Eqn. 4.26 can be established by expanding the series up to the $m = 3$ terms and grouping according to the power of lambda:

$$
\begin{aligned}
\Xi &= \prod_{i=1}^{\infty} \left[\frac{(q_i \lambda^i)^0}{0!} + \frac{(q_i \lambda^i)^1}{1!} + \frac{(q_i \lambda^i)^2}{2!} + \cdots\right] \\
&= \left(1 + \frac{q_1 \lambda}{1!} + \frac{(q_1 \lambda)^2}{2!} + \cdots\right)\left(1 + \frac{q_2 \lambda^2}{1!} + \cdots\right)\left(1 + \frac{q_3 \lambda^3}{1!} + \cdots\right)\cdots \\
&= \underbrace{1}_{Q(0,V,T)} \lambda^0 + \underbrace{\left(\frac{q_1}{1!}\right)}_{Q(1,V,T)} \lambda^1 + \underbrace{\left(\frac{q_1^2}{2!} + \frac{q_2}{1!}\right)}_{Q(2,V,T)} \lambda^2 + \underbrace{\left(\frac{q_1^3}{3!} + \frac{q_1 q_2}{2! \cdot 1!} + \frac{q_3}{1!}\right)}_{Q(3,V,T)} \lambda^3 + \cdots
\end{aligned}
$$

where $Q(N, V, T)$ is of the form seen in Eqn. 4.24, and map onto the groupings of partition functions q_i for the integer partitions of 0, 1, 2, and 3 respectively.

The functional form of the grand canonical partition function in Eqn. 4.28 can again be tweaked by recognizing that the sum in that expression is equal to the Maclaurin series expansion of an exponential where $x = q_i \lambda^i$, the probability weighting per i-mer cluster:

$$\exp[x] = \frac{x^0}{0!} + \frac{x^1}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^k}{k!} + \cdots = \sum_{m=0}^{\infty} \frac{x^m}{m!}. \tag{4.29}$$

Thus, the third functional form for the grand canonical partition function is,

$$\boxed{\Xi(\lambda, V, T) = \exp\left[\sum_{i=1}^{\infty} q_i \lambda^i\right].} \tag{4.30}$$

These three expressions for the grand canonical ensemble partition function are essential to the methods that will be unveiled in the subsequent sections and how they can be implemented in the calculation of the average cluster size distributions results in significantly faster algorithms.

4.3.3 $\langle n_s \rangle_N$ from $\langle n_s \rangle_\lambda$

It should be a relatively familiar concept that one can calculated the average of a measurable variable by taking the sum over all products of the value of said variable and the probability meeting the conditions where that variable holds said value. Thus the average frequency of a cluster of s molecules $\langle n_s \rangle_\lambda$ in the grand canonical ensemble under the conditions of a fixed chemical potential and thus a fixed absolute activity λ this corresponds to:

$$\langle n_s \rangle_\lambda = \sum_{N=0}^{\infty} \langle n_s \rangle_N P(N, \lambda) \tag{4.31}$$

where $\langle n_s \rangle_N$ is the s-mer average cluster size frequency in the canonical ensemble with a system size of N molecules and $P(N, \lambda)$ is the probability of that system size N in the grand canonical ensemble given an absolute activity of λ. This can therefore

be rewritten as:

$$\langle n_s \rangle_\lambda \;\; = \;\; \sum_{N=0}^{\infty} \langle n_s \rangle_N \, \frac{Q\left(N,V,T\right)\lambda^N}{\Xi} \tag{4.32}$$

Suppose now that we take a moment to think about which partitions contribute to this weighted average. To clarify the matter, it helps to substitute the equation for $\langle n_s \rangle_N$ which was given in Eqn. 4.25.

$$\langle n_s \rangle_\lambda \Xi \;\; = \;\; \sum_{N=0}^{\infty} \langle n_s \rangle_N \, Q\left(N,V,T\right)\lambda^N \tag{4.33}$$

$$= \;\; \sum_{N=0}^{\infty} \lambda^N \sum_{k=1}^{p(N)} \left(n_{s,k} \prod_{i=1}^{N} \frac{q_i^{n_{i,k}}}{n_{i,k}!} \right) \tag{4.34}$$

$$= \;\; \left(\sum_{n_s=1}^{\infty} \frac{n_s \left(q_s \lambda^s\right)^{n_s}}{n_s!} \right) \left(\prod_{i=1, i \neq s}^{\infty} \left[\sum_{m=0}^{\infty} \frac{\left(q_i \lambda^i\right)^m}{m!} \right] \right)$$

$$= \;\; \left(\frac{1\left(q_s\lambda_s\right)^1}{1!} + \frac{2\left(q_s\lambda_s\right)^2}{2!} + \frac{3\left(q_s\lambda_s\right)^3}{3!} + \cdots \right) \left(\prod_{i=1, i \neq s}^{\infty} \left[\sum_{m=0}^{\infty} \frac{\left(q_i \lambda^i\right)^m}{m!} \right] \right)$$

$$= \;\; q_s \lambda^s \prod_{i=1}^{\infty} \left[\sum_{m=0}^{\infty} \frac{\left(q_i \lambda^i\right)^m}{m!} \right]$$

$$= \;\; q_s \lambda^s \Xi$$

Partitions where the cluster size s does not show up ($n_{s,k} = 0$) do not actually contribute to $\langle n_s \rangle_\lambda \Xi$ and thus this product is equivalent to multiplying the grand canonical partition function (Eqn. 4.28) by the probability weighting of a single s-mer, $q_s \lambda^s$.

$$\boxed{\langle n_s \rangle_\lambda \Xi = q_s \lambda^s \Xi} \tag{4.35}$$

4.3.4 Bypassing a sum over partitions in the calculation of $\langle n_s \rangle_N$

In section 4.3.1 we demonstrated that $\langle n_s \rangle_N$ can be calculated using a sum over all partitions of N (Eqn. 4.25).[6] In early versions of the algorithms we used to cal-

---

[6]This section is a results of a collaboration between Professor James Kindt's group in the chemistry department and Professor Ken Ono's group in the mathematics department at Emory. In particular, this is the result of work done by Olivia Beckwith, James Kindt, Lara Patel, Robert Schneider, and Xiaokun Zhang.

culate $\langle n_s \rangle_N$ we had to generate all the different partitions of N for each iteration of fitting. This was problematic however since the number of partitions for a system grows exponentially with respect to the total number of integer parts. To be specific, for large N the number of partitions $p(N)$ can be approximated using the Hardy-Ramanujan-Rademacher asymptotic formula,[56]

$$p(N) \approx \frac{1}{4N\sqrt{3}} e^{\pi\sqrt{2N/3}}. \tag{4.36}$$

This means that calculating $\langle n_s \rangle_N$ becomes too time consuming using Eqn. 4.25 for large values of N. For a bipartite system the problem becomes ever more intractable. These methods and functional forms were derived primarily with the goal of applying them to bipartite systems. It is easier however to start by discuss the concepts in the context of a single molecule type aggregation equilibrium.

Before we proceed any further it is helpful to recognize that the grand canonical ensemble partition function Ξ, for which we have already given several functional forms already (Eqn. 4.26, 4.28, 4.30), can be expressed as a polynomial expansion in orders of λ where c_N is the coefficient for a system with N molecules.

$$\Xi(\lambda, V, T) = \sum_{N=0}^{\infty} c_N \lambda^N = \sum_{N=0}^{\infty} Q(N, V, T) \lambda^N \tag{4.37}$$

Thus for a system with N particles, the c_N coefficient is equal to the canonical ensemble partition function:

$$c_N = Q(N, V, T) \tag{4.38}$$

It is this trick that will make it possible to derive an equation for the average cluster frequency in the canonical ensemble $\langle n_s \rangle_N$ in terms of two total canonical partition functions Q and the partition function of the s-mer q_s. The same polynomial representation can be applied to $\langle n_s \rangle_\lambda \Xi$ in orders of λ with coefficients denoted as g_N:

$$\langle n_s \rangle_\lambda \Xi = q_s \lambda^s \Xi = \sum_{N=0}^{\infty} g_N \lambda^N \tag{4.39}$$

Let us begin with the representation for Ξ from Eqn. 4.26:

$$
\begin{aligned}
q_s \lambda^s \Xi &= q_s \lambda^s \sum Q(N, V, T) \lambda^N \\
&= q_s \sum Q(N - s, V, T) \lambda^N
\end{aligned}
\tag{4.40}
$$

Then using Eqn. 4.35, 4.33, and 4.39 the following becomes evident:

$$
\sum_{N=0}^{\infty} g_N \lambda^N = q_s \sum Q(N - s, V, T) \lambda^N = \sum \langle n_s \rangle_N Q(N, V, T) \lambda^N
\tag{4.41}
$$

This means that if we look at the g_N coefficient of the polynomial expansion,

$$
g_N = q_s Q(N - s, V, T) = \langle n_s \rangle_N Q(N, V, T)
\tag{4.42}
$$

we are able to get an equation for $\langle n_s \rangle_N$ of the canonical ensemble in terms of two total partition functions Q and the s-mer partition function q_s:

$$
\boxed{\langle n_s \rangle_N = \frac{q_s\, Q(N - s, V, T)}{Q(N, V, T)}}
\tag{4.43}
$$

You, as the reader, may be asking yourself how exactly this functional form is an advantage over that given in Eqn. 4.25 since the function we have thus far given for calculating the total canonical ensemble partition function (Eqn. 4.24) still involves a sum over all the ways to partition the total number of molecules in the system N. It turns out that the way by which we can bypass the need to generate all the possible partitions of N is by recognizing that it is possible to get $Q(N, V, T)$ from the grand canonical partition function by taking N derivatives with respect to the absolute activity λ:

$$
\frac{\partial^N}{\partial \lambda} \Xi[\lambda, V, T] = \sum_{N'=0}^{\infty} \frac{(N + N')!}{N!} Q(N + N', V, T) \lambda^{N'}
\tag{4.44}
$$

Since we are solving for Q, which is independent of λ, we are free to choose any convenient value of λ and will set $\lambda = 0$ to simplify this expression:

$$
Q(N, V, T) = \frac{1}{N!} \left(\frac{\partial^N}{\partial \lambda} \Xi[\lambda, V, T] \right)_{\lambda=0}.
\tag{4.45}
$$

Thus is can be seen that the grand canonical ensemble is a generating function for the canonical ensemble functions.

The reason that this is so convenient is because of the grand canonical ensemble partition function given in Eqn. 4.30. The exponential functional form is particularly advantageous in that it is preserved over the iterations of taking the derivatives and since we take the liberty of setting $\lambda = 0$ at the end of each derivative to get the 0th order coefficient, the exponential portion of the equation equals 1 and can be effectively ignored. On top of this, it is possible to collect a library of $Q(N, V, T)$ from $N = 1$ up to $N = N_{\text{tot}}$ for a simulation in a single iterative loop and then subsequently calculated all $\langle n_s \rangle_N$ using Eqn. 4.43.

To give an example of how this would work, let us make make a substitution in notation for the grand canonical ensemble partition function such that,

$$
\begin{aligned}
\Xi[\lambda, V, T] &= f_2(\lambda) \exp[f_1(\lambda)] \\
f_1(\lambda) &= \sum_{i=1}^{N_{\text{tot}}} q_i \lambda^i \\
f_2(\lambda) &= 1
\end{aligned}
$$

It is possible to truncate the sum in f_1 at the total number of molecules in a given simulation as we are not really interested in the value of the grand canonical partition function but rather the constituent canonical partition functions.

In each successive derivative, coefficients with the same power of λ are grouped and the 0th order coefficient is saved as a canonical ensemble partition function. For example, the first derivative would give the following where $f_3(\lambda)$ is the polynomial function after grouping like terms of λ:

$$
\frac{\partial^1}{\partial \lambda} \Xi = \underbrace{\left(\frac{\partial f_2(\lambda)}{\partial \lambda} + f_2(\lambda) \frac{\partial f_1(\lambda)}{\partial \lambda} \right)}_{f_3(\lambda)} \exp[f_1(\lambda)] \tag{4.46}
$$

$$
Q(N = 1, V, T) = \frac{1}{1!} \left(\frac{\partial^1}{\partial \lambda} \Xi \right)_{\lambda=0}
$$

Second derivative would then be:

$$\frac{\partial^2}{\partial\lambda}\Xi = \underbrace{\left(\frac{\partial f_3(\lambda)}{\partial\lambda} + f_3(\lambda)\frac{\partial f_1(\lambda)}{\partial\lambda}\right)}_{f_4(\lambda)}\exp\left[f_1(\lambda)\right] \tag{4.47}$$

$$Q(N = 2, V, T) = \frac{1}{2!}\left(\frac{\partial^2}{\partial\lambda}\Xi\right)_{\lambda=0}$$

This would continue all the way up to the total number of molecules in a given system N_{tot} such that the last derivative would be:

$$\frac{\partial^{N_{\text{tot}}}}{\partial\lambda}\Xi = \underbrace{\left(\frac{\partial f_{N_{\text{tot}}+1}(\lambda)}{\partial\lambda} + f_{N_{\text{tot}}+1}(\lambda)\frac{\partial f_1(\lambda)}{\partial\lambda}\right)}_{f_{N_{\text{tot}}+2}(\lambda)}\exp\left[f_1(\lambda)\right] \tag{4.48}$$

$$Q(N = N_{\text{tot}}, V, T) = \frac{1}{N_{\text{tot}}!}\left(\frac{\partial^{N_{\text{tot}}}}{\partial\lambda}\Xi\right)_{\lambda=0}$$

The average cluster size frequencies for a given system size and concentration can be calculated using Eqn. 4.43. This is significantly more efficient than the original algorithm where partitions had to be generated.[60]

4.4   Global fitting

To obtain globally optimal fitting of the equilibrium association constant parameters K_i across the full set of either vapor phase or aqueous phase cluster size distributions, where some cluster sizes i appear frequently in some simulation conditions and rarely in other, it is useful to quantify the statistical certainty of each cluster size frequency data point $\langle n_{i,j}\rangle_{\text{sim}}$ in the distribution. To do so, the post-equilibration portions of the trajectory is split into 25-50 ns segments, the cluster size distribution is calculated for each, and the standard error of the mean $\sigma\left(\langle n_{i,j}\rangle_{\text{sim}}\right)$ over the segments is calculated for each cluster size i in trajectory j.

The relative statistical certainty $w_{i,j}^{\text{rel}}$ of that data point is calculated as:

$$w_{i,j}^{\text{rel}} = \frac{\langle n_{i,j}\rangle_{\text{sim}}}{\sigma\left(\langle n_{i,j}\rangle_{\text{sim}}\right)} \tag{4.49}$$
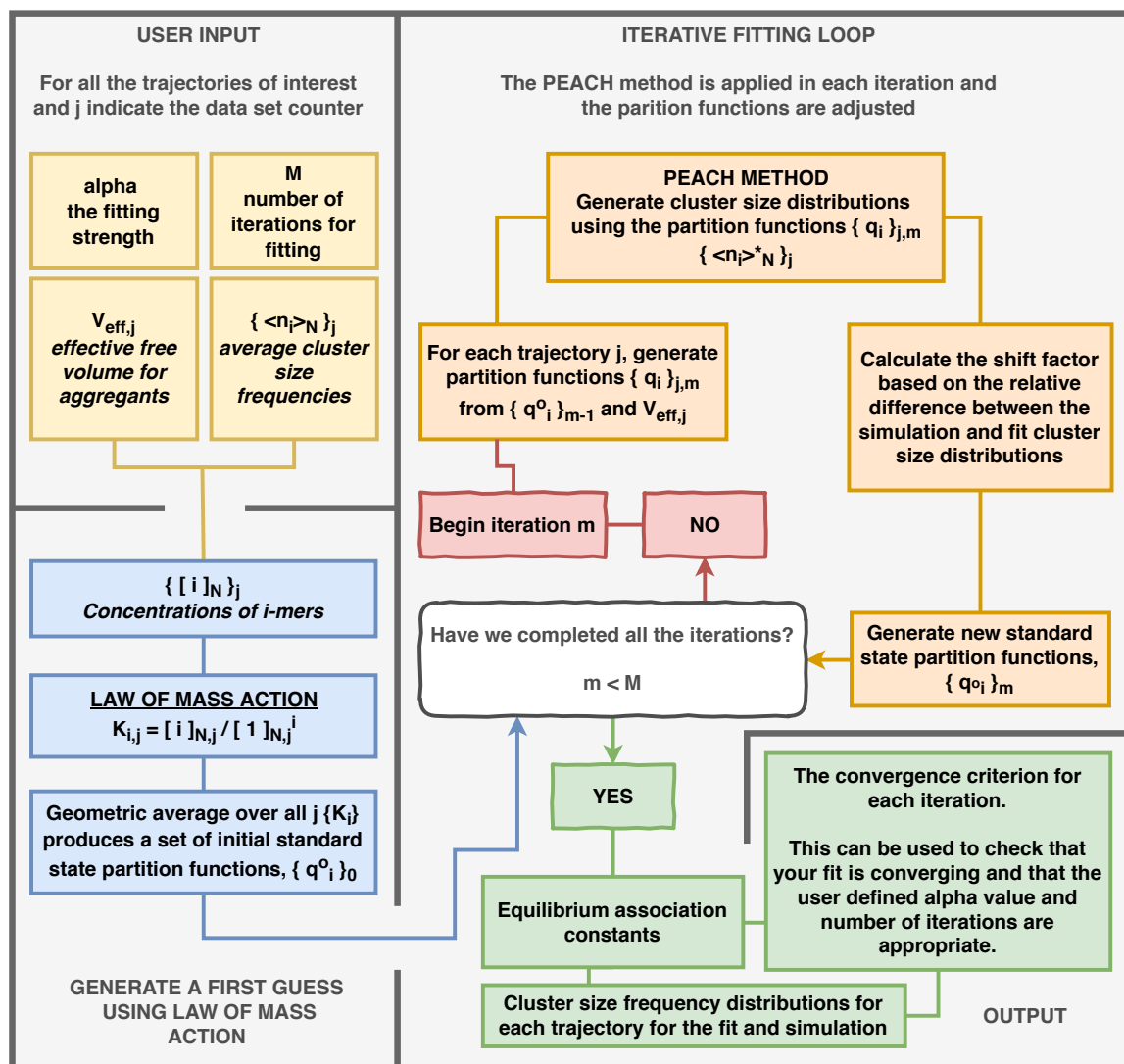
Figure 4.6: Graphical representation of implementing the PEACH method in an iterative global fitting algorithm.

To control the degree to which the fit should aim to reproduce a data point for the i-mer from trajectory j relative to all other trajectories where the i-mer is present, a normalized relative weighting factor, $w_{i,j}$ is defined:

$$w_{i,j} = w_{i,j}^{\text{rel}} \left(\sum_{j=1}^{N_{\text{traj}}} w_{i,j}^{\text{rel}} \right)^{-1} \tag{4.50}$$

As in previous work,[61] initial guesses for the values of K_i are based on the apparent equilibrium constants $K_{i,j}'^{\text{sim}}$, as calculated using the law of mass action. (We designate "apparent" equilibrium constants with a prime to indicate that they are not in fact constant upon varying N or V_{eff} for small systems.)

$$K_{i,j}'^{\text{sim}} = \frac{c_{i,j}}{(c_{1,j})^i} = \frac{\langle n_{i,j} \rangle / V_{\text{eff},j}}{(\langle n_{1,j} \rangle / V_{\text{eff},j})^i} \tag{4.51}$$

The effective volume is assumed to be the free volume in the system or the volume occupied by solvent.[7] As one value of the equilibrium association constant must be given for each i, and Eqn. 4.51 will give different values for different trajectory conditions, the initial guess for the true (bulk) equilibrium constants $K_i^{\text{fit},0}$ is taken as the weighted geometric mean of $K_{i,j}'^{\text{sim}}$ from all trajectories j that feature clusters of size i:

$$K_i^{\text{fit},0} = \exp \left(\sum_{j=1}^{J_i} w_{i,j} \log \left(K_{i,j}'^{\text{sim}} \right) \right). \tag{4.52}$$

We define an energy scale such that $q_1^\circ = 1$, such that we may substitute K_i for q_i° in Eqn. 4.10 so that $q_{i,j} = K_i^{\text{fit}} V_{\text{eff},j} / V^\circ$.

The next step is to calculate the calculate the cluster size distribution $\langle n_{i,j} \rangle_{\text{fit}}$ for each trajectory j given a total number of solutes N_j and effective volume $V_{\text{eff},j}$. There are two ways to go about this. In the analysis of MTBE aggregation (Chapter 5) the cluster size distribution is calculated following the procedure of generating and

---

[7]For example, in simulations of MTBE (Chapter 5), the effective volume is calculated as $V_{\text{eff}} = V - N v_{MTBE}$, with a molecular volume of MTBE taken as $v_{MTBE} = 0.191$ nm^3 from bulk liquid simulations . For the simulations of NaCl (Chapter 6), it was calculated as $V_{\text{eff}} = N_{\text{Sol}} v_{\text{Sol}}$ where v_{Sol} was the average volume per solvent from a solvent only simulation.

averaging over the integer partitions of N_j as described in ref. [61]:

$$\langle n_{i,j} \rangle_{\text{fit}} = Q_{\text{tot}}^{-1} \sum_{k=1}^{p(N_j)} \left(n_{i,k} \prod_{i'=1}^{N_j} \frac{(q_{i,j})^{n_{i',k}}}{n_{i',k}!} \right) \tag{4.53}$$

$$Q_{\text{tot}} (N_j, V_j, T) = \sum_{k=1}^{p(N_j)} \prod_{i'=1}^{N_j} \frac{(q_{i,j})^{n_{i',k}}}{n_{i',k}!} \tag{4.54}$$

For larger or bipartite systems such as NaCl aggregation (Chapter 6), the method employed for the calculation of $\langle n_{i,j} \rangle_{\text{fit}}$ proceeds as outlined in section 4.3.4 through iterative derivatives of the exponential form of the grand canonical partition function. It should be noted that equations 4.55-4.57 are for a single component system.

$$\langle n_{i,j} \rangle_{\text{fit}} = \frac{q_{i,j} \, Q \left(N_j - i, V_{\text{eff},j}, T \right)}{Q \left(N_j, V_{\text{eff},j}, T \right)} \tag{4.55}$$

$$Q \left(N_j - i, V_{\text{eff},j}, T \right) = \frac{1}{(N_j - i)!} \left(\frac{\partial^{(N_j - i)}}{\partial \lambda} \Xi \left(\lambda, V_{\text{eff},j}, T \right) \right)_{\lambda=0} \tag{4.56}$$

$$\Xi \left(\lambda, V_{\text{eff},j}, T \right) = \exp \left(\sum_{i=1}^{N_j} q_{i,j} \, \lambda^i \right) \tag{4.57}$$

To calculate the next iteration for K_i^{fit}, we first calculate the apparent equilibrium constants $K_{i,j}'^{\text{fit}}$ from the results for each trajectory:

$$K_{i,j}'^{\text{fit}} = (V_{\text{eff},j})^{i-1} \frac{\langle n_{i,j} \rangle_{\text{fit}}}{\left(\langle n_{1,j} \rangle_{\text{fit}} \right)^i} \tag{4.58}$$

An adjustment factor g_i is calculated to improve the fit in the equilibrium association constants:

$$K_i^{\text{fit}} \leftarrow g_i K_i^{\text{fit}} \tag{4.59}$$

The the adjustment factor's strength is controlled by a user defined variable α that has a value less than 1. The aim should be to use the largest value of α for the data set being fit that guarantees stability in the fitting to ensure the fastest fitting. The adjustment factor can be calculated in two ways:

$$g_i = \exp \left(\alpha \sum_{j=1}^{J_i} w_{i,j} \ln \left(\frac{K_{i,j}'^{\text{sim}}}{K_{i,j}'^{\text{fit}}} \right) \right) \tag{4.60}$$

$$g_i' = \exp\left(\alpha \sum_{j=1}^{J_i} w_{i,j} \ln\left(\frac{\langle n_{i,j}\rangle_{\text{sim}}}{\langle n_{i,j}\rangle_{\text{fit}}}\right)\right) \tag{4.61}$$

The calculation of $\langle n_{i,j}\rangle$ is then repeated based on the updated set of equilibrium constants. The cycle of Equations 4.53-4.61 are repeated, updating K_i^{fit} at each cycle, until the user defined number of iterations are completed. Given enough iterations, the cluster size distributions calculated by Eqn. 4.53 should match the simulated distributions for well sampled data points.

In the case of fitting to MTBE aggregation data (Ch. 5), Eqn. 4.60 was found to be the most reliable. However for the bipartite systems such as the NaCl aggregation data (Ch. 6), it was found that doing two passes of fitting worked better. In the first pass, the first guess for the equilibrium association constants was made using the law of mass action (Eqn. 4.51-4.52) and the adjustment factor in Eqn. 4.60. The resultant equilibrium association constants are then used as the first guess for the second pass of fitting where the adjustment factor was calculated as in Eqn. 4.61 which requires a better first guess to ensure stable convergence in the fit.

4.4.1 Evaluating the quality of the fit

As a quantitative measure of the agreement between the fitted results and the simulation data, a convergence criterion C_i is calculated for each cluster size i as a weighted geometric mean of the relative squared difference between the average cluster frequency from the fit and that of the simulations over the J_i simulations in which that cluster size is present.

$$C_i = \exp\left(\sum_{j=1}^{J_i} w_{i,j} \ \log\left(\frac{\langle n_{i,j}\rangle_{\text{fit}} - \langle n_{i,j}\rangle_{\text{sim}}}{\langle n_{i,j}\rangle_{\text{sim}}}\right)^2\right) \tag{4.62}$$

The arithmetic average of these convergence criteria is then defined as the overarching convergence criterion C_{tot}:

$$C_{\text{tot}} = N_{\text{max}}^{-1} \sum_{i=1}^{N_{\text{max}}} C_i \tag{4.63}$$
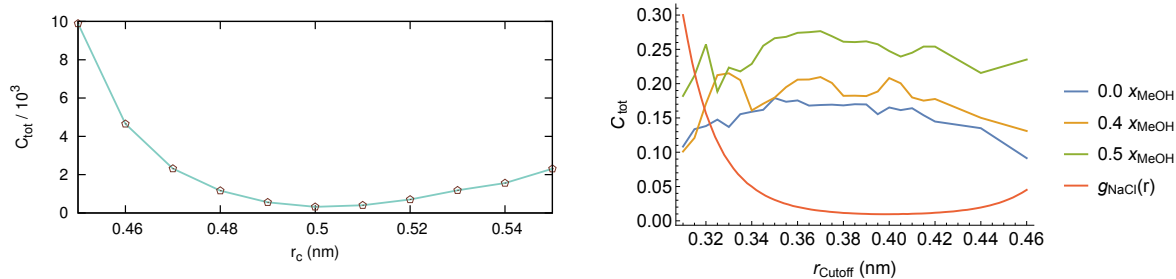
Figure 4.7: Plot of the final convergence criterion C_{tot} for different cluster definition r_c distances used in the global fitting to aggregation data for vapor phase MTBE (left) and solution phase NaCl (right) solvated in pure SPC/E water and the MeOH mixtures. The plot on the right also shows the scaled radial distribution function (red line) for the Na^+ and Cl^- in the region between the first and second peaks.

where N_{max} is the maximum cluster size observed in the simulations.

The convergence criterion serves two functions. The primary one indicates whether the choice in the fit scalar α produces stable fitting, and whether the user has performed a sufficient number of fitting iterations for the criteria to converge. In the limit where non-ideal effects (such as interactions between clusters) are small, the convergence criterion can also be used to identify an optimal cluster definition.

In both of the aggregation studies presented here (Ch. 5-6), the clusters are defined by a series of neighbor associations where molecules or ions within a cutoff r_c are considered part of the same cluster. Thus an optimal choice in the cutoff r_c is one that minimizes C_{tot}. If there are not enough data sets in the fit that overlap or strong non-ideal effects are present (as is shown in the NaCl aggregation), the convergence criterion fails to show a reliable minimum. Both cases are presented in Fig. 4.7.

In section 5.4, it is shown that the selection of r_c that minimizes C_{tot} corresponds to the cutoff that ensures that the second virial coefficient for the monomer $b_{1,1}$ is almost equal to the volume per monomer v_1 used to calculate the effective volume ($V_{eff} = V - N v_1$).

4.5 Free energy profile

At constant temperature and pressure, the change in Gibbs free energy can be expressed as a sum over the i-mer chemical potentials μ_i and the change in the population of i-mers dn_i.[49]

$$dG|_{p,T} = \sum_i \mu_i \, dn_i \tag{4.64}$$

For a reaction of i monomers aggregating into a single i-mer at constant temperature and pressure, the change in free energy can thus be expressed as,

$$\Delta G_i = \mu_i - i\mu_1. \tag{4.65}$$

We assume that in the dilute systems, aggregates exhibit negligible interactions in between aggregation events. A change in the concentration is therefore related to the standard state by the change in entropy, and the chemical potential for a given concentration $(c = N/V)$ is given by:

$$\mu = \mu^\circ + k_B T \ln \left(\frac{c}{c^\circ} \right) \tag{4.66}$$

where μ° and c° are the standard state chemical potential and concentration respectively.[49] If we assume that the concentration of i-mer is equal to the concentration of monomers and substitute Eqn. 4.66 into Eqn. 4.65, the difference in Gibbs free energy is,

$$\Delta G_i = \Delta G_i^\circ + k_B T \left(1 - i \right) \ln \left(\frac{c}{c^\circ} \right) \tag{4.67}$$

The standard state change in Gibbs free energy for a reaction can be expressed in terms of the equilibrium association constant $K_{\text{eq},i}$ for the aggregation of i molecules:

$$\Delta G_i^\circ = -k_B T \ln K_{\text{eq},i} \tag{4.68}$$

Assuming that the concentrations of the i-mer and monomer are equal, and using the final globally fitted equilibrium constant $K_i'^{\text{fit}}$ in place of $K_{\text{eq},i}$, Eqn. 4.67 becomes:

$$\frac{\Delta G_i}{k_B T} = -\ln K_i'^{\text{fit}} - (i - 1) \ln \left(\frac{c}{c^\circ} \right) \tag{4.69}$$

Free energy profile for a bipartite system

We start by defining the free energy it in terms of two monomer concentrations, $c_{1,0}$ and $c_{0,1}$, and the equilibrium association constant $K_{i,j}$ for a cluster composed of i and j monomers of the two components in the system (for example Na^+ and Cl^-).

$$\frac{\Delta G_{i,j}}{k_B T} = -\ln K_{i,j} - (i-1) \ln \left(\frac{c_{1,0}}{c^\circ} \right) - (j-1) \ln \left(\frac{c_{0,1}}{c^\circ} \right) \tag{4.70}$$

The problem here is that the free energy from Eqn. 4.70 is not given relative to the monomer free energies (it does not equal zero for the monomers). If we subtract the free energy of a monomer we get a free energy surface $\Delta G'_{i,j}$ for clusters:

$$\begin{aligned} \frac{\Delta G'_{i,j}}{k_B T} &= \frac{\Delta G_{i,j}}{k_B T} - \frac{\Delta G_{1,0}}{k_B T} \\ &= -\ln K_{i,j} - (i-1) \ln \left(\frac{c_{1,0}}{c^\circ} \right) - (j) \ln \left(\frac{c_{0,1}}{c^\circ} \right) \end{aligned} \tag{4.71}$$

For the production of our free energy surfaces in the main paper, we assume that the monomer concentrations are equal, $c_{1,0} = c_{0,1} = c_1$ and thus:

$$\frac{\Delta G'_{i,j}}{k_B T} = -\ln K_{i,j} - (i+j-1) \ln \frac{c_1}{c^\circ} \tag{4.72}$$

4.5.1 Classical nucleation theory (CNT)

CNT typically defines the free energy of association, ΔG_i, of a cluster of size i as a function of the cluster size and the concentration of free monomers c_1. In an experimental context, the cluster size is proportional to the dimensions such as cluster radii but can be expressed in terms of the number i of molecules it contains. If the aggregate shape is not size-dependent, the area of the interface is proportional to the cluster size $i^{2/3}$ with a constant of proportionality A_{geom} that is shape dependent.[62] Treating the chemical potential of an MTBE molecule in the bulk liquid as $\mu = 0$, the chemical potential of an i-mer cluster at the standard state concentration is then:

$$\mu^\circ = \gamma A_{geom} i^{2/3} + C \tag{4.73}$$

with C a constant. The change in free energy associated with adding a monomer to an i-mer is:

$$\Delta G_+ = \mu_{i+1} - \mu_i - \mu_1 \tag{4.74}$$
$$= \gamma A_{\text{geom}} \left((i+1)^{2/3} - i^{2/3} - 1\right) - C - k_B T \ln \left(\frac{c}{c^\circ}\right).$$

In the limit of large i, the addition of a free monomer produces a negligible ($\propto i^{-1/3}$) change to the surface area, and this quantity depends only on c_1. We can define the free monomer concentration where ΔG_+ equals zero in the bulk limit as $c_1 = c_{\text{sat}}$, meaning the concentration where the free monomers have reached saturation in the vapor or solution phases at equilibrium with the bulk liquid. This yields:

$$C = -\gamma A_{\text{geom}} - k_B T \ln \left(c_{\text{sat}}/c^\circ\right) \tag{4.75}$$

Combining equations 4.65, 4.66, 4.73, and 4.75, and assuming that $c_1 = c_i = c$ then gives:[63]

$$\Delta G_i = \gamma A_{\text{geom}} \left(i^{2/3} - 1\right) - (i-1)k_B T \ln \left(c/c_{\text{sat}}\right) \tag{4.76}$$

Amorphous clusters are fluid and free to assume a spherical shape, minimizing the surface area and thus reducing the penalty of the interface. For a relatively spherical cluster, $A_{\text{geom}} = (36\pi v_1^2)^{1/3}$ where v_1 is the monomer volume.[8]

---

[8]For the MTBE simulations in Ch. 5, $v_1 = 0.191$ nm^3 in bulk phase liquid simulations at 273 K.

Chapter 5 Cluster free energies from simple simulations of small numbers of aggregants: Nucleation of liquid MTBE from vapor and aqueous phases

The reversible formation of aggregates, whether from the gas phase or from solution, is a phenomenon important to a wide range of biological, technological, and environmental processes.[1] The determination of the free energy of aggregation, as a function of the number of monomers per aggregate, is a goal that has been pursued via numerous creative techniques in molecular simulation. One area where experimental characterization of aggregates is particularly difficult, and simulation has been correspondingly helpful, is in the field of nucleation. The formation of a new condensed phase from either a supersaturated vapor or a supersaturated solution proceeds via the formation and growth of finite aggregates that are unstable and transient. Building on years of development of Monte Carlo (MC) methods to treat nucleation from the vapor phase,[64, 65] the aggregation volume (AV) bias Monte Carlo[66] has emerged as an effective approach to evaluate the thermodynamics of aggregation of both single-component and mixed molecular clusters.[67] This method relies on biased Monte Carlo exchange moves that efficiently transport monomers between aggregates and randomly selected positions within the mostly empty space of the vapor phase system. An umbrella sampling (US) bias potential can then be used to ensure good statistical sampling over the desired range of aggregate sizes, with monomers exchanging between a single aggregate and a virtual reservoir of monomers at a well-defined chemical potential.[68] Although innovations continue in this area,[69] the AV-US method and other approaches that require insertion moves[70, 71] pose challenges in simulations

---

[1]This chapter is a reiteration of results published in ref. [60].

of aggregates in solution modeled with explicit solvent, because the solvent gets in the way of the insertion move attempts.

Recently Parrinello and co-workers have used well-tempered (WT) metadynamics[72] to obtain aggregation free energies along the nucleation pathway for urea crystallizing from aqueous solution.[73] In that work, the WT metadynamics algorithm accelerates the sampling of aggregates of varying size. This is a powerful method, but the control over sampling aggregate sizes relies on the a priori definition of collective variables and their implementation within the actual simulations, which takes some customization on a system-by-system basis. The effects of simulating a closed (constant-N) system are also considered in that work using an approach introduced by Wedekind et al.[54] and used in studies of nucleation of 2-component systems[74] to account for the depletion of monomers in the solution phase that accompanies growth of an aggregate. The main difference between the current method and the approach of ref. [54] is that the present analysis explicitly accounts for all clusters present in the system rather than focusing on the largest one.

Here, we approach the thermodynamics of cluster growth using an embarrassingly simple (by contrast to the methods mentioned above) approach: through conventional, constant NVT (in the vapor case) or NpT (in the solution case) simulations with N rather small. In previous reports[61, 75] we have described how the free energy of cluster formation can be related to the equilibrium distribution of cluster sizes at small constant N without applying the law of mass action, which is unreliable in small systems.

To review how this is achieved, we start by treating a cluster of i molecules within the quasi-chemical framework as a distinct chemical entity, with its characteristic single-particle partition function q_i that reflects the internal interactions within within the cluster (and the cluster's interface with the solvent, if any). This partition function includes a translational component that is proportional to volume, and so it is useful

to define a standard partition function q° corresponding to a cluster that is free to move within a standard volume V°. As in the conventional derivation of the equilibrium expression from the canonical partition function of a mixture of reactants and products,[49] we may set the probability of finding a set of clusters in the system such that n_i clusters of size i are present (where $i = 1$ indicates a free monomer) as:

$$P\left(n_1, n_2, \cdots n_N\right) = \frac{1}{Q\left(N, V, T\right)} \prod_{i=1}^{N} \frac{(q_i^\circ)^{n_i} \left(V/V^\circ\right)^{n_i}}{n_i!} \tag{5.1}$$

with the sum of the product $i \times n_i$ constrained to equal N. In the limit of large N, the average composition can be found by maximizing this function with respect to each n_i, treating the numbers of clusters of each type as a continuous variable; this will produce equilibrium expressions consistent with the law of mass action. At small N, however, fluctuations away from the most probable composition may be large and the discrete integer nature of the numbers of clusters the system cannot be ignored.

Instead, an average cluster size distribution can be obtained from the partition functions by calculating the weights for all possible combinations of cluster sizes (which map on to the integer partitions of N, which can be generated by standard algorithms[76]) and finding a probability-weighted average number $\langle n_i \rangle$ of each size i of cluster. This "small-N" mapping of the equilibrium distribution to a set of q_i° is rigorous under the assumption that interactions between clusters can be neglected. An iterative fitting procedure may then be used to find the best set of q_i° to fit an equilibrium distribution of $\langle n_i \rangle$ obtained from simulation. The standard partition function q_i° is related to the equilibrium constant K_i for formation of the cluster from i monomers and to the standard free energy of aggregation ΔG_i° as:

$$K_i = \frac{q_i^\circ}{(q_1^\circ)^i} = e^{\Delta G_i^\circ / k_B T} \tag{5.2}$$

It is common in the literature on nucleation to define a nucleation free energy for a specific free monomer concentration c as:[68, 77, 78]

$$\Delta G_i = -k_B T \left[\ln K_i + (i-1) \ln \left(\frac{c}{c^\circ} \right) \right] \tag{5.3}$$

In the current study, a series of vapor- and aqueous-phase simulations of methyl tert-butyl ether (MTBE) in varying numbers and at varying concentrations have been per-formed. Each simulation yields some distribution of aggregate sizes; simulation conditions were chosen so that there is some overlap among size ranges in different simulations. A globally optimized fit of the size-dependent free energies to these distributions has been performed using a multi-set adaptation of the method in ref. [75]. The results are compared with the predictions of classical nucleation theory based on surface tensions obtained from simulations of planar slab interfaces. The choice of MTBE was motivated by its modest solubility in water, which allowed the observation of distinct clusters that fluctuated in size to an extent that allowed converged equilibrium distributions to be seen over moderate system sizes and trajectory lengths. MTBE is in fact soluble enough that leaks of gasoline containing MTBE as an additive have resulted in groundwater contamination,[79] leading to a shift towards alternative oxygenating additives.[80]

The remainder of this report will describe the details of the vapor- and aqueous-phase simulations of MTBE, review the fundamentals of the "small-N" statistical analysis and describe the extensions made to permit a global fit over several data sets, and present the results interpreted within the framework of classical nucleation theory.

5.1   Methods

5.1.1   General MD simulation methods

All simulations are performed using the GROMACS 4.6 software package.[35, 37, 38] Newton's equations of motion are evaluated with the leap-frog algorithm integrator with a time-step of 2 fs.

MTBE is simulated using the united-atom TraPPE-UA potential which was parameterized to reproduce phase behavior.[81] The TIP4P water model[82] is used for

all aqueous-phase MTBE simulations. TIP4P exhibited the best performance in a study by Lundsgaard et al.[83] on the partition coefficients of TraPPE-UA squalene in water/ethanol mixtures with several water models.

The temperature is maintained at a reference temperature of 273 K ($\tau_T = 1.0$ ps) with the velocity rescaling thermostat.[6] NpT simulations (for aqueous systems) implement the Berendsen barostat[84] with a reference pressure of 1 bar ($\tau_p = 5.0$ ps and compressibility 4.5×10^{-5} bar^{-1}).

The Lorentz-Berthelot combining rule for LJ interactions between different atomic species is used for TIP4P water and the TraPPE-UA MTBE molecules:

$$\sigma_{ij} = \frac{1}{2}\left(\sigma_{ii} + \sigma_{jj}\right), \quad \epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}} \tag{5.4}$$

Electrostatic interactions are calculated by the Particle Mesh Ewald method.[85] The short-range neighbor list and Coulombic cut-off radius are 1.9 nm. Van der Waals inter-actions implement a switch function at 1.2 nm with a cut-off radius of 1.4 nm.

5.1.2 Simulations of equilibrium cluster aggregation

Vapor-phase NVT cluster aggregation simulations of MTBE were run for 1 μs at densities (ρ_{MTBE}) of 284 and 510 mM. Simulations were initiated from a random distribution of MTBE. At each concentration, a set of populations (N_{MTBE}) from 20 to 60 MTBE molecules increasing in increments of 5 molecules are simulated. The choice in increment size was made to ensure sufficient overlap in cluster size distributions for continuous sampling of clusters sizes of 1 to 60 MTBE. The 510 mM simulation set was extended up to a system size of 95 molecules as it shows sufficient cluster size distribution overlap. The full set of cluster size distribution plots for the vapor-phase simulations are shown in Appendix Fig. C.1. MTBE coordinates were saved at intervals of 2.0 ps.

Aqueous-phase NpT cluster aggregation simulations were run for 500 ns at total MTBE concentrations of 761, 1019, 1275 and 1537 mM. A summary of solution

phase configurations is given in Table 5.1 with the number of MTBE (N_{MTBE}) and the number of TIP4P water molecules (N_{TIP4P}). MTBE coordinates were saved at intervals of 0.2 ps.

Table 5.1: Solution phase MTBE cluster aggregation simulation parameters

[MTBE] /mM	761	1019	1275	1537
N_{MTBE}	30	40	50	60
N_{TIP4P}	2025	1957	1891	1822

Cluster size distributions are calculated using a modified version of the GROMACS utility *g_clustsize*, which was adapted to give results at higher precision. An initial analysis of each trajectory is performed to determine the equilibration period, which was taken to be the time required for the system to initially reach its average maximum cluster size. The equilibration of the vapor phase systems typically took no more than 1 ns while the aqueous phase systems took 5-10 ns. The cluster size distributions for analysis were calculated with the first 20 ns period excluded, giving the mean number of clusters of size i, $\langle n_i \rangle$.

5.1.3   Calculation of surface tension

Simulations of a slab of liquid MTBE in contact with either its vapor phase or with liquid water were performed to obtain an independent measure of the surface tensions of these interfaces under the simulation conditions. A slab was placed within the simulation box forming two interfaces in the xy-plane. Both simulations were run for 20 ns at 273 K.

The aqueous-phase simulation consists of 384 MTBE and 2822 TIP4P water molecules. Semi-isotropic pressure coupling at a pressure of 1 bar with a coupling time constant of 5.0 ps was maintained using the Berendsen barostat.[84] The x and y dimensions were fixed at 5.2 nm while the z-dimension was permitted to vary with a compressibility of 4.5×10^{-5} bar^{-1}.

The MTBE portion of the final configuration of the aqueous-phase simulation was

then used as the initial configuration of the vapor-phase simulation. This was run with a constant volume in a simulation box with dimensions of $5.2 \times 5.2 \times 5.8$ nm.

The average MTBE slab thickness is 2.9 and 2.7 nm for the vapor and aqueous phases respectively.

Surface tension, γ, is calculated using the g_energy GROMACS utility based on the following equation:

$$\gamma(t) = \frac{L_z}{n} \left(P_{zz} - \frac{P_{xx} + P_{yy}}{2} \right) \tag{5.5}$$

The interface is assumed to be in the xy-plane of the simulation. The simulation box dimension perpendicular to the interface is L_z. The variables P_{xx}, P_{yy}, and P_{zz} are the pressure along the x, y, and z dimensions. The number of interfaces, n, is 2 in these simulations.

5.2 Estimation of error in the global fit

We account for the sensitivity of the fit to the statistical uncertainty in the input data by generating $\sim$100 new sets of cluster size distributions introducing random fluctuations consistent with the estimated standard errors in $\langle n_{i,j} \rangle$ for each trajectory j, and performing a global fit for each. The resulting standard deviation in the Gibbs free energy profiles over these 100 fits are shown as the shaded regions in Fig. 5.5 for the vapor phase and Fig. 5.7 for the solution phase.

Each new cluster size frequency $\langle n_{i,j} \rangle_{\text{new}}$ is generated assuming a Gaussian distribution in the cluster size frequency for a cluster size i in the jth simulation. Accounting for 95% of the Gaussian distribution in $\langle n_{i,j} \rangle$ centered at $\langle n_{i,j} \rangle_{\text{sim}}$, it is possible to generate the new cluster frequency and an acceptance probability:

$$\langle n_{i,j} \rangle_{\text{new}} = \langle n_{i,j} \rangle_{\text{sim}} + 4.0 \left(R_1 - 0.5 \right) \sigma \left(\langle n_{i,j} \rangle_{\text{sim}} \right) \tag{5.6}$$

$$P_{\text{acc}} = \frac{1}{\sigma \left(\langle n_{i,j} \rangle_{\text{sim}} \right) \sqrt{2\pi}} \exp \left(\frac{\left(\langle n_{i,j} \rangle_{\text{new}} - \langle n_{i,j} \rangle_{\text{sim}} \right)^2}{2\sigma \left(\langle n_{i,j} \rangle_{\text{sim}} \right)^2} \right) \tag{5.7}$$

The new cluster frequency is accepted if $P_{acc} > R_2$ where R_1 and R_2 are random numbers generated between 0 and 1.

5.3 Results

5.3.1 Cluster definition

The free energy of a cluster depends on how the cluster is defined, which has been a concern to researchers from early days of simulations on clusters.[86, 87] Too restrictive a definition skews the cluster size distribution towards small cluster sizes while too loose a definition skews it towards large cluster sizes. The concept of an "optimal" cluster definition has been considered recently using the framework of variational transition state theory,[88] with the surface dividing the state of an i-mer cluster from that of an $(i-1)$-mer and a free monomer defined so as to minimize the flux between them, favoring persistent rather than transient groupings of molecules. In the present work, the global analysis of cluster size distributions allows the comparison of different cluster definitions from the same set of trajectories, and the optimization of a cluster criterion based on consistency in fitting size distributions under different system size and concentration conditions. This enables us to evaluate cluster definitions using purely structural data with relatively little computational expense.

Here a cluster is defined as a group of MTBE molecules that are linked via a sequence of neighbor associations, as first implemented by Stillinger.[89] Neighbors are molecules within a specified radial distance r_c of one another. We compared using the distances between molecular centers of mass (COM) to using distances between united atom sites of the MTBE molecules for the cluster definitions. While both are effective, the second definition makes more chemical sense as it measures the closest contact points between the molecules. The positions of the first peaks (or shoulders) in the site-site radial distribution functions (Fig. 5.1) indicate that the distance r_c used in the site-site cluster definition should be between 0.40 and 0.60 nm.
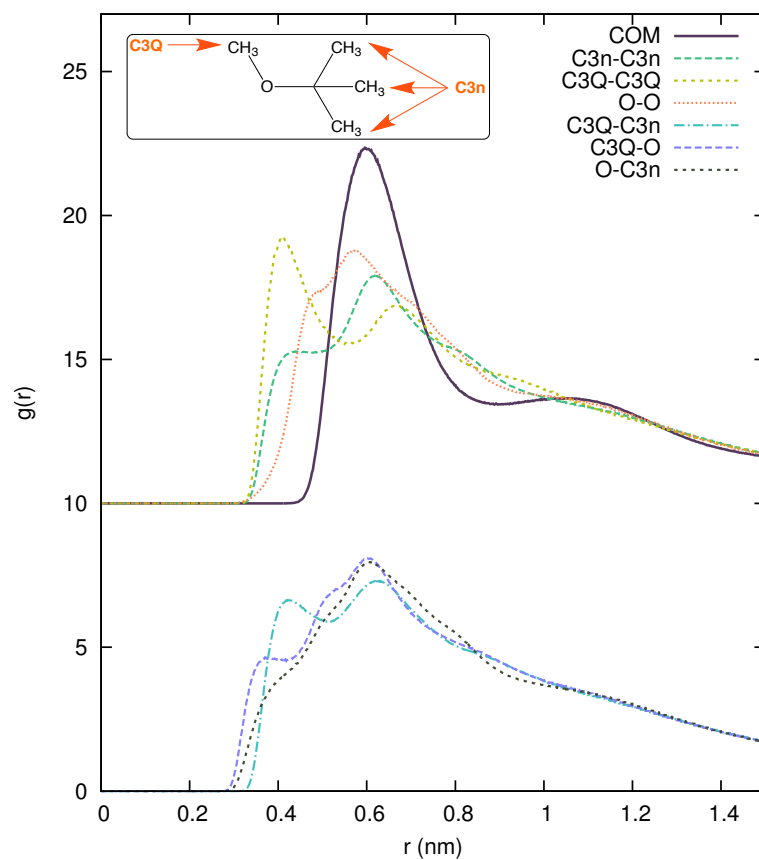
Figure 5.1: Plot of the radial distribution functions (RDF) for the 510 mM vapor phase MTBE simulation containing 30 molecules. For clarity, the first four distributions are offset by 10 and radial distributions involving the tertiary carbon are excluded. The bold purple line indicates the RDF for the molecular centers of mass.
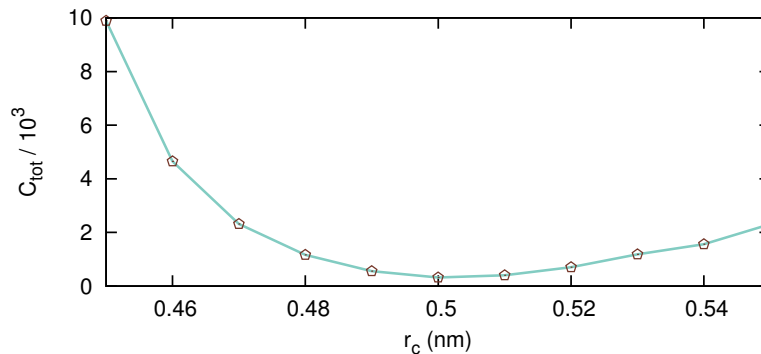
Figure 5.2: Plot of the convergence criterion for different cluster definition r_c distances used in the global fitting to all 25 vapor phase simulation cluster size distributions.

An optimal cut-off distance for the cluster definition is established as the cut-off that minimizes the convergence criterion (Eqn. 4.63) of the global fit to the vapor phase simulations. Fig. 5.2 demonstrates that the convergence criterion minimum is located at 0.50 nm. Decreasing the cut-off by 0.02 nm leads to doubling the convergence criterion.

Inclusion of simulation data from both concentrations was necessary to constrain the cluster distance cut-off r_c and achieve a consistent minimum in the convergence criterion. The inclusion of both concentrations ensured that there were two independent cluster distributions for each system size (N_{MTBE}). By comparison, when only a single concentration data set was used, the minimum in the convergence criterion shifted depending on the number of data sets used in the global fitting procedure. Plots showing the comparison between using a data set from a single concentration versus using two concentrations are given in the Appendix Fig. C.1 and C.2. Due to the smaller number of solution phase simulations it is not possible to reliably establish the optimal cluster definition distance criteria. The vapor phase cluster distance criterion is used for the solution phase analysis as well.
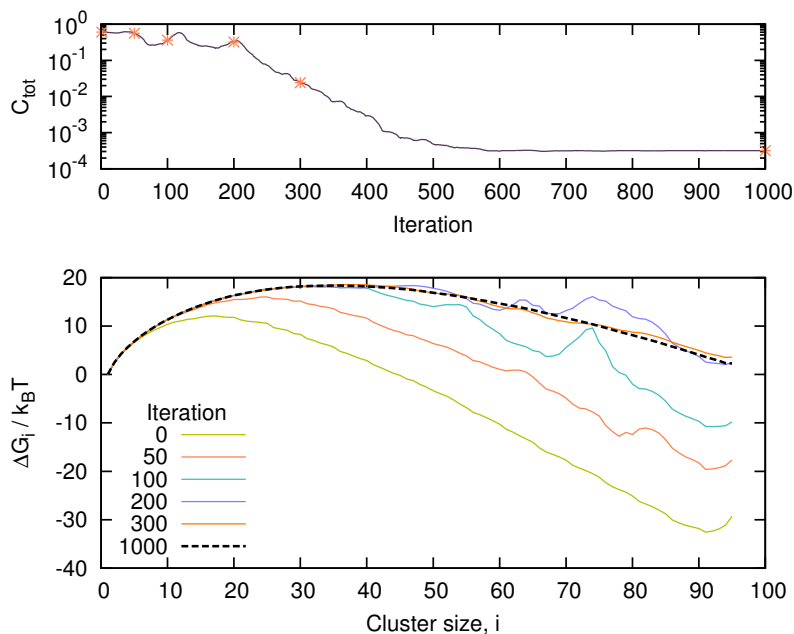
Figure 5.3: A plot of the convergence criterion in Eqn. 4.63 for fitting to the vapor phase simulations (top) with * marking the iterations displayed in the subsequent plot and the ΔG of association at intervals over the 1000 fitting iterations (bottom) as calculated for a monomer concentration of 60 mM.

5.3.2 Global fitting

The fitting exponent α (Eqn. 4.60) used for the global fitting to the 25 vapor phase simulations is 0.01 and requires approximately 600 iterations to converge. The progression in globally fitting partition functions to the cluster distributions can be seen in Fig. 5.3 for the vapor phase simulations. The first plot shows the convergence criterion for the fitting with asterisks (*) marking the iterations chosen to display the Gibbs free energy profiles (Eqn. 4.69) plotted in the second plot. Iteration 0 is the initial guess based on Eqn. 4.52.

The quality of the global fit for vapor phase MTBE is exemplified in Fig. 5.4 where a selection of cluster size distributions from simulation and the global fitting routine are displayed in tandem. In regions of the distributions where the standard error of the cluster population is small, the fit has good agreement with the simulated distribution. Enlarged figures for the distributions for both vapor phase simulation
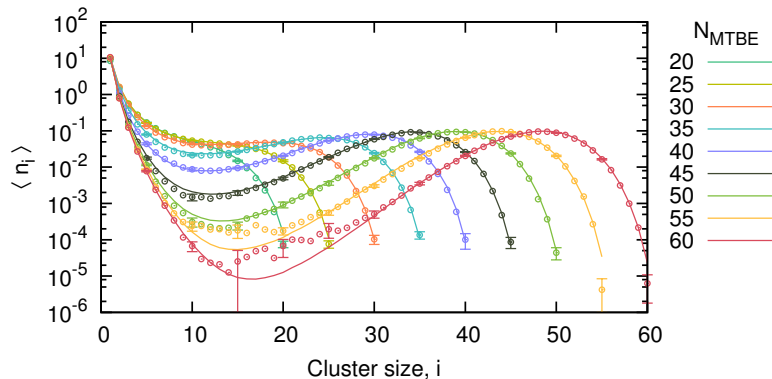
Figure 5.4: Plot for the 284 mM MTBE vapor phase of the cluster size distributions based on the MD simulation data (points) and the cluster distributions produced by the global fitting routine (lines). Error bars are given for every fifth cluster size as the standard error calculated by block averaging over nineteen 50 ns segments of the trajectory.

concentrations of 284 mM and 510 mM are given in Appendix Fig. C.3.

The equilibrium association constants that are the parameters used in the global fit to simulation data are conveniently expressed as free energies of association using Eqn. 4.69 assuming a free monomer concentration of 60 mM for the vapor phase, and are shown in Fig. 5.5 (other choices for the monomer concentration would produce a linear i-dependent shift in the curve). As discussed in previous work,[61, 75] finite-size effects shift the equilibrium cluster size distributions (shown in Fig. 5.4) away from what would be expected for a set of cluster free energies in the limit of large N. For cluster free energies curves that involve a nucleation barrier, this produces a flattening of the distribution that allows a broad range of cluster sizes to appear with reasonably high frequency, improving the sampling efficiency without requiring any explicit bias to be applied during the simulation. Fig. 5.5 also shows the least squares linear regression fit of the CNT functional form (Eqn. 4.76) to the final Gibbs free energy profile.

The standard deviation in the free energy curve due to uncertainties in the input cluster size distributions have been shown in the shaded region of the top plot in Fig. 5.5. The standard deviations are shown separately due to their small magnitude.
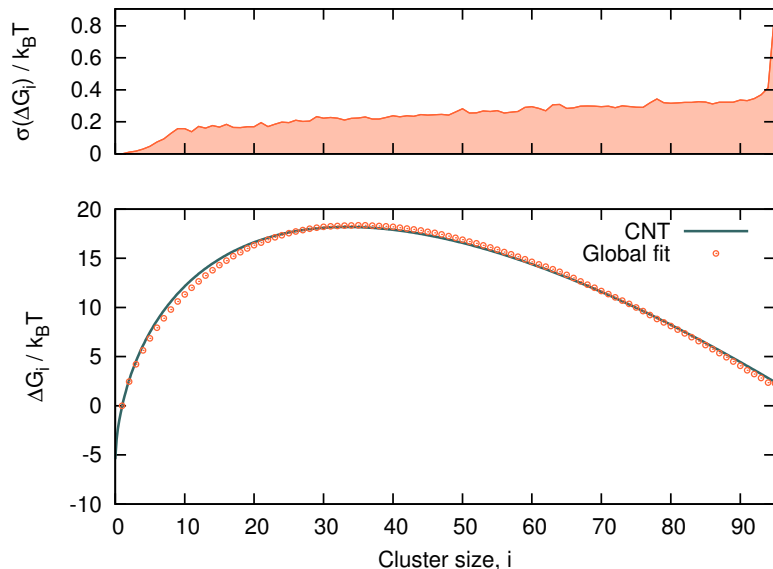
Figure 5.5: Plot (top) of the standard deviation ($+2\sigma$) in the Gibbs free energy of association due to uncertainties in the simulated cluster distributions. Plot (bottom) of the CNT functional form (line) fit to the Gibbs free energy of association for monomer concentration of 60 mM based on the equilibrium constants of the vapor phase (points) from the global fit to the simulated cluster distributions.

The fitting exponent α used for the global fitting to the 4 aqueous phase simulations is 0.11 and requires approximately 300 iterations to converge. A plot of the progression in the globally fitted partition functions to the cluster distributions of the aqueous phase simulations is exhibited in Fig. C.5. The smaller number of data sets in the fit is shown to reduce the number of iterations required to reach a converged fit. The quality of the global fit is demonstrated in Fig. 5.6 by the application of the equilibrium association constants in reproducing the cluster size distributions for each of the aqueous phase simulations. Like the vapor phase simulation results (Fig. 5.4) it is evident that cluster sizes with large standard errors in simulation have worse consistency with the global fit.

The fit of the CNT functional form (Eqn. 4.76) to the Gibbs free energy curve (Eqn. 4.69) assuming a monomer concentration of 183 mM is displayed in Fig. 5.7.

The standard deviation in the free energy curve for the solution phase simulations are large enough to be shown in the same plot as the shaded region surrounding the
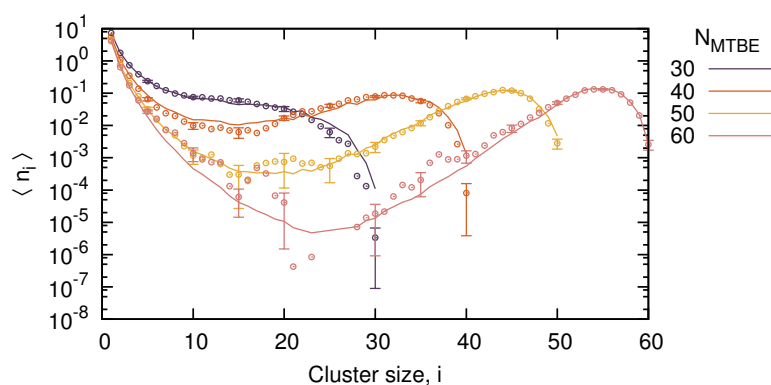
Figure 5.6: Plot for the aqueous phase of the cluster size distributions based on the experimental data (points) with error bars based on the standard error for every five cluster sizes and the cluster distributions produced by the global fitting routine (line). Error bars are given for every fifth cluster size as the standard error calculated by block averaging over nineteen 25 ns segments of the trajectory.
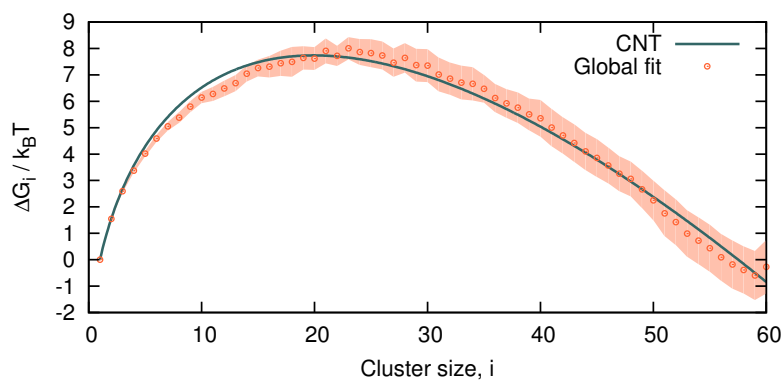
Figure 5.7: Plot of the CNT functional form (line) fit to the Gibbs free energy of association based on the equilibrium association constants (points) for the aqueous phase MTBE simulations and a monomer concentration of 183 mM solvated in TIP4P water. The shaded region corresponds to the standard deviation ($\pm 2\sigma$) in ΔG_i due to uncertainties in the simulated cluster distributions.

global fit results (points).

The CNT fitting parameters for the vapor and aqueous phases are given in Table 2, with c the monomer concentration used to generate the free energy curve, $\ln\left(c/c_{\mathrm{sat}}\right)$ accounting for the entropic difference between the saturation concentration and given monomer concentration, and γ_{CNT} the surface tension based on the CNT fit.

The uncertainties in $\ln\left(c/c_{\mathrm{sat}}\right)$ and γ_{CNT} are approximated as the perturbation in either variable from its optimal value (based on the least squares fitting) required to double the standard deviation of the CNT fit with respect to the global fit. For example, to estimate the uncertainty in γ_{CNT}, least square regression optimizations of $\ln\left(c/c_{\mathrm{sat}}\right)$ are performed holding γ_{CNT} fixed over a range of values near its globally optimal value. The standard deviation of the fit is calculated as a function of the deviation of γ_{CNT} from its global optimum value, and the uncertainty in γ_{CNT} is defined as the change in γ_{CNT} that doubles the standard deviation at the global minimum.

The surface tension calculated from the MTBE slab simulations μ_{sim} (see section 5.1) and the experimental surface tensions γ_{exp} are given for comparison. The saturation concentration is given based on (1) the extrapolation from the CNT fit $c_{\mathrm{CNT}}^{\mathrm{sat}}$, (2) the concentration of free monomers in the slab simulations $c_{\mathrm{sim-slab}}^{\mathrm{sat}}$, (3) an estimate based on extrapolation from the Clausius-Clapeyron plot of Fig. 11 in ref. [81] at 273 K, derived from simulations with the TraPPE forcefield, and (4) experimental saturation concentrations $c_{\mathrm{exp}}^{\mathrm{sat}}$ for the vapor[90] and aqueous[91] phases.

5.4   Discussion

The success of the global fitting is exemplified in Fig. 5.4 and Fig. 5.6 where the reproduction of the cluster size distributions show good consistency with the simulated distributions. The efficacy of the data weighting is demonstrated by the close agreement (contingent on sufficient sampling and thus smaller standard errors in

Table 5.2: Fitting parameters to CNT functional form.

Phase	Vapor	Aqueous
c/mM	60	183
$\ln\left(c/c_{\mathrm{sat}}\right)$	1.24 ± 0.11	1.12 ± 0.15
γ_{CNT}/mNm^{-1}	16.06 ± 1.06	10.72 ± 1.27
γ_{sim}/mNm^{-1}	17.68 ± 0.13	12.60 ± 0.33
γ_{exp}/mNm^{-1}	23.6 [a]	7.8 [b]
$c^{\mathrm{sat}}_{\mathrm{CNT}}$/mM	14.52 [c]	59.45 [c]
$c^{\mathrm{sat}}_{\mathrm{sim-slab}}$/mM	13.41 ± 0.98 [e]	62.64 ± 10.91 [e]
$c^{\mathrm{sat}}_{\mathrm{sim-CC}}$/mM	7.05 [d]	-
$c^{\mathrm{sat}}_{\mathrm{exp}}$/mM	4.59 [f]	975.36 [g]

[a] Ref. [92] at 273 K. [b] Ref. [93] at 278 K. [c] Calculated from fitting to $y = \ln\left(c_{\mathrm{sat}}/c_1\right)$ for CNT fits at different monomer concentrations. [d] Estimated by extrapolation from the Clausius-Clapeyron plot of Fig. 11 in ref. [81] at 273 K. [e] Saturation concentration approximated based on the concentration of free monomers in the slab simulations. [f] Based on Wagner vapor pressure equation parameters given in ref. [90] at 273 K. [g] Extrapolation based on the van't Hoff thermodynamic parameters from ref. [91] at 273 K.

the reported populations) between the cluster size frequencies reproduced from single set of K_i from the global fit, to cluster size frequencies from multiple simulations. As discussed, optimization of the fit with respect to the size cut-off r_c gives an objective measure for the cluster definitions. The good agreement of the best-fit distributions with simulation data points under different concentration and system size conditions indicate that any non-ideal effects here are accounted for using a simple approximation for the free volume V_{eff}.

The reproduced cluster size distributions based on the global fit appear less noisy than individual trajectories data sets due to the overlap in data sets. The larger number of data sets, the degree of overlap in the cluster size distributions, and the better quality of the data in regions of overlap in the vapor phase simulations results in smoother distributions by comparison to the solution phase.

The optimization technique converges reliably as is demonstrated in Fig. 5.3 for the vapor phase. The best parameters used to define the fitting operation (α and

number of iterations) depend to the number of data sets being fitted over. For a larger number of data sets, the α value must be smaller and number of iterations larger to maintain stability (data not shown).

The CNT expression (Eqn. 4.76) fits the free energy profile based on the globally fitted K_i unexpectedly well even down to cluster sizes with only a few monomers, although clusters with size $i < 20$ exhibit the largest deviation from the CNT functional form for the vapor phase simulations. The values for the surface tension γ_{CNT} obtained in fits of both vapor and aqueous MTBE cluster free energies of association are lower than those (γ_{sim}) calculated from simulations of planar slab interfaces. This discrepancy persisted when clusters with $i < 20$ were excluded from the fit (data not shown). Conversely, using the surface tension derived from planar slab simulations to predict size-dependent free energies would yield excessively large barriers to cluster growth. Similar deviations have been seen in previous simulation studies of clusters in gas phase. Chen et al. found that the CNT-based predictions of vapor-liquid nucleation barriers of alkanes using a planar surface tension were consistently too high.[78] A lower apparent surface tension for small clusters, relative to the planar value, was also found in studies of a coarse-grained water model by Factorovich et al.[70, 71], and characterized in terms of a positive Tolman length. Deviations in the other direction, involving a higher apparent surface tension for clusters than for planar interfaces, have been observed in simulation studies of water cluster nucleation.[94, 95] The saturation concentrations c_{sat} in the vapor phase or solution phase, as calculated from the cluster free energies or from the slab calculations, agreed to within the error estimates. Given that these error bars were rather large, however, it is difficult draw conclusions.

The concentration in the vapor phase simulated at equilibrium with a slab of pure liquid can be compared directly with experiment and with the results presented by the developers of the TraPPE force-field. The vapor pressure obtained in the

current simulations is nearly twice the value obtained from the Clausius-Clapeyron parameters reported by Stubbs et al.[81] We believe that the origin of this discrepancy is our neglect of long-range Lennard-Jones (LR-LJ) interactions. To support this claim we calculate the ratio of the vapor pressure including the LR-LJ interactions $P_{\text{vap}}^{\text{Long}}$ with respect to the vapor pressure excluding LR-LJ interactions $P_{\text{vap}}^{\text{Short}}$ as:

$$\frac{P_{\text{vap}}^{\text{Long}}}{P_{\text{vap}}^{\text{Short}}} = \exp\left(\frac{-\Delta G_{\text{vap}}^{\text{Long}} + \Delta G_{\text{vap}}^{\text{Short}}}{k_B T}\right) = \exp\left(\frac{\mu_{\text{LR}}}{k_B T}\right) \tag{5.8}$$

where k_B is the Boltzmann constant and T is the temperature. The LR-LJ interactions for bulk-phase MTBE with a density ρ, summed over the set of all pairs of sites i, j in the MTBE molecule with LJ well depth of ϵ_{ij} and LJ distance of σ_{ij} can be calculated as:[96]

$$\mu_{\text{LR}} = \frac{16}{9}\pi\rho\sum^{\{i,j\}}\epsilon_{ij}\sigma_{ij}^3\left(\left(\frac{r_{\text{vdw}}}{\sigma_{ij}}\right)^{-9} - 3\left(\frac{r_{\text{vdw}}}{\sigma_{ij}}\right)^{-3}\right) \tag{5.9}$$

where r_{vdw} is the Van der Waals interaction cut-off distance of 1.4 nm. For a temperature of 273 K, the vapor pressure ratio (Eqn. 5.8) is 0.56 and very close to the ratio in saturation concentrations $c_{\text{sim}-\text{CC}}^{\text{sat}}/c_{\text{CNT}}^{\text{sat}} = 0.49$. It may also be possible that the under prediction of the surface tension (with respect to experiment, as it was not reported in ref. [81] from simulation) is related to the neglect of LR-LJ interactions, whose effect on surface tension is rather more complicated to estimate.[97]

The greatest discrepancy between experiment and theory here is seen in the solubility of MTBE in water, where simulation yields a solubility that is too low by a factor of 16. The solubility arises from the competition of MTBE-MTBE interactions with the MTBE-water interactions. Given that the TraPPE force-field over-predicts the vapor pressure slightly, the current discrepancy is unlikely to be caused by excessively strong MTBE-MTBE interactions. Neglect of long-range interactions is unlikely to play a major part, as the differences between these terms for MTBE in neat liquid and in aqueous solution are likely to be smaller than between MTBE and in vapor, where the effect was only a factor of 2. The water-MTBE interactions might be expected

to also be reasonably good due to the good (within a factor of 2) agreement with experiment reported for the partitioning of cis-3-hexenol between squalane and water or water/ethanol mixtures.[83] For an explanation as to why MTBE does not follow cis-3-hexenol one might blame the methyl site - water interactions, as MTBE contains four instead of just one. Indeed, Ashbaugh et al.[98] have calculated TIP4P/2005 hydration free energies with the TraPPE parameters for several alkanes, and found that the value for neopentane (also with four methyl groups) exceeds the experimental value by about 3.4 kJ/mol at 273 K while n-butane exceeds it only by 1.2 kJ/mol. Remarkably small adjustments to the water oxygen/CH_n site LJ parameters could bring simulation results in line with experiment.[98]

The model used in our analysis of cluster free energies assumes that the presence of a cluster of one size does not influence the stability of a different cluster, and that the effective volume available to all clusters for a given simulation is the difference between the total box volume and the sum of the monomer volumes $V_{\text{eff}} = V - Nv_{\text{MTBE}}$. (There is an important distinction between this "non-interacting cluster" assumption and the ideal gas or ideal-dilute solution assumptions, which neglect the possibility that the clusters themselves form. Conditions that lead to clusters are clearly non-ideal from the perspective of monomer-monomer interactions, but may or may not satisfy the non-interacting cluster assumption.)

For simulations of MTBE cluster formation in the vapor phase, the distance cut-off criterion for defining the clusters was selected (as shown in Fig. 5.2) to give a set of cluster size distributions that optimized the global fit to the model under the "non-interacting" cluster assumption. The quality of the fits across different concentration and volume conditions would seem to suggest that the assumption is valid, because its breakdown would imply that the relative weights of different partitions would not scale consistently with the model under different volume conditions. Another possibility is that violations of the assumption affect cluster free energies in a consistent way

across these different conditions. It is useful to examine the conditions under which real, interacting cluster statistics would agree with the non-interacting cluster model. To incorporate non-ideal effects at the level of individual partitions of N monomers, within the second-virial approximation we can write:[99]

$$P\left(n_1, n_2, \cdots n_N\right) = Q^{-1}\left(\prod_{i=1}^{N}\frac{(q_i^\circ)^{n_i}\left(V/V^\circ\right)^{n_i}}{n_i!}\right) \times \left(1 - \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{n_i\left(n_j - \delta_{ij}\right)b_{ij}}{V}\right) \tag{5.10}$$

where δ_{ij} is the Kronecker delta function and where b_{ij} is the second virial coefficient that relates to the potential of mean force F between distinct clusters of size i and j at a distance r between their centers-of-mass:

$$b_{ij} = \int_0^\infty 2\pi r^2\left(1 - \exp\left[-\beta F_{i,j}\left(r\right)\right]\right)dr \tag{5.11}$$

Here the force is assumed to behave as a hard wall at any surface where the clusters merge to become a single cluster. If the potential of mean force is zero beyond that surface, then b_{ij} represents one-half the mutual excluded volume of a cluster of size i and one of size j.

In the analysis, the influence of composition-dependent non-ideal effects arising from the second virial terms on the probability weighting of each partition is represented only by a factor of $(1 - Nv_{MTBE}/V)$ for each cluster, which comes from the use of V_{eff} in place of V. This will be a good approximation provided that:

$$1 - \sum_{i=1}^{N}n_i\frac{Nv_{MTBE}}{V} \approx \left(1 - \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{n_i\left(n_j - \delta_{ij}\right)b_{ij}}{V}\right) \tag{5.12}$$

on average over the distribution of partitions. The most typical partitions in these nucleating systems contain a single N'-mer cluster and the remaining $n_1 = N - N'$ molecules distributed as free monomers. For these partitions, Eqn. 5.12 simplifies to:

$$\left(n_1 + 1\right)Nv_{MTBE} \approx 2\,n_1\,b_{1,N'} + \left(n_1^2 - n_1\right)b_{1,1} \tag{5.13}$$

which holds approximately true when $b_{1,N'} \approx 1/2N'v_{MTBE}$ and $b_{1,1} \approx v_{MTBE}$ for all

N'. The error in the probability weighting from using the effective volume approximation then is,

$$P_{\text{model}}/P \sim \left(1 - (N + n_1) \, v_{\text{MTBE}}/V\right) \tag{5.14}$$

Considering first the condition on $b_{1,N'}$, we see that it is likely to be satisfied, as the excluded volume between a small monomer and a large cluster ($= 2b_{1,N'}$) should approximately equal the volume of the large cluster ($N'v_{\text{MTBE}}$). The implications of the second condition are less obvious. For like hard-sphere particles with no attractions, the excluded volume is a sphere whose radius is the sum of the radii of the two spheres, giving an excluded volume that is 8 times the actual volume. The condition $b_{1,1} = v_{MTBE}$ can only be met then if the monomer-monomer excluded volume is balanced in part by a medium-to-long ranged attraction between monomers that are classified as "free".

To check whether this condition is in fact satisfied for the cluster criterion used, we used data from a simulation of 20 MTBE molecules in the vapor at 0.049 M, where 87% of the molecules on average remained as free monomers and the largest cluster observed was a hexamer, to estimate $b_{1,1}$ using Eqn. 5.11. The radial distribution function $g(r)$ based on center-of-mass distances r_{COM} for free monomers alone was calculated under site-site cut-off distances of 0.5 nm (the apparently optimal value) as well as for 0.48 and 0.52 nm. These are shown, with the full $g(r)$ function that includes monomers within clusters, in Fig. 5.4. (If a center-of-mass cut-off had been used, the $g(r)$ functions that exclude monomers within clusters would equal zero at distances below the cut-off and coincide with the center-of-mass $g(r)$ for distances greater than the cut-off.) The function $g(r)$ in the dilute limit provides a good approximation[100] to the Boltzmann factor in Eqn. 5.11. The value of the integral is sensitive to small deviations of $g(r)$ from 1 at large r, so we evaluated the integral in $b_{1,1}$ with an upper limit of $r = 1.75$ nm. As evident from Fig. C.5, the cut-off distance of 0.5 nm gives $b_{1,1} = 0.16$ nm^3, near the theoretical optimum of $b_{1,1} = v_{\text{MTBE}} = 0.191$
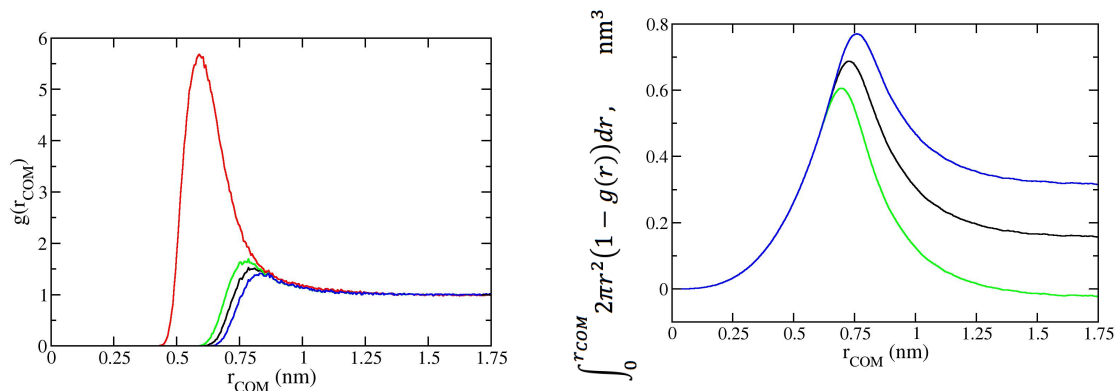
Figure 5.8: Left panel: radial distribution function of MTBE center-of-mass positions from simulation of 20 molecules in vapor phase at 0.049 M. Red curve: all molecules; black curve: only molecules classified as free monomers for a site-site cutoff distance of 0.50 nm; green curve: free monomers only using a 0.48 nm cutoff; blue curve: free monomers only using a 0.52 nm cutoff. Right panel: b_2 integral evaluated for the free monomer radial distribution functions shown on the left, with x-axis giving the upper limit to the integral.

nm^3, while nearby cut-offs of 0.48 nm or 0.52 nm give much worse agreement (-0.02 nm^3 and 0.32 nm^3 respectively).

It appears, therefore, that by optimizing the cluster criterion to give the most consistent fit (across different simulation conditions) between simulated and model cluster size distributions, we are finding the position of the dividing surface between monomer and dimer that comes closest to the condition $b_{1,1} = v_{\mathrm{MTBE}}$. When this is true, the exchange of an MTBE from a large cluster into the surrounding gas has a minimal effect on non-ideal contributions to the free energy: the net unfavorable effect on the non-ideal interaction free energy of the free monomers from adding one to their number (itself coming from a balance between excluded volume and attractive terms) is balanced by the favorable effect of reducing the volume occupied by the cluster. Some cut-off definition that satisfies the condition $b_{1,1} = v_{\mathrm{monomer}}$ should exist in most cases where attractions between monomers are strong enough to nucleate a cluster.

We can estimate the error in free energy (in units of $k_B T$) determined by the fitting procedure to a given cluster size i as the logarithm of the proportional weighting

error of a partition containing that cluster. If the optimal condition is met, then per Eqn. 5.14 there remains an error of order Nv_{MTBE}/V, the volume fraction of the box occupied by all monomers, which is consistent with the magnitude in errors that we see between simulation data and fits in the well-sampled conditions. The error introduced by non-ideal effects in the difference in free energies of an i-mer and an $(i+1)$-mer is then on the order of v_{MTBE}/V in units of $k_B T$, which suggests that the error in c_{sat} from non-ideal effects is similarly small for the vapor-phase simulations. (For the solution phase, the data obtained were insufficient to give a clearly optimal choice of cut-off, and much greater errors due to non-ideal effects are possible there.) If non-ideal effects were ignored completely (that is, if the total volume V and not the effective volume V_{eff} were used), this error would be increased by a factor of the number of monomers in the cluster, with catastrophic consequences for fitting in all but the most dilute systems. Non-ideal effects are far from negligible in this system, but their effect on cluster free energies can be controlled with a simple definition of an effective free volume and an adjustment of the cluster definition to tune the second virial coefficient of free monomers.

5.5   Conclusions

A global fitting analysis of cluster size distributions obtained from standard unbiased, equilibrated constant-temperature molecular dynamics trajectories has been used to obtain cluster free energies that are consistent over a range of simulation conditions. The analysis allows the definition of what constitutes a cluster to be determined after the simulations have been completed and provides a guide to optimizing the cluster definition based on achieving best consistency in fitting trajectories with different volumes and numbers of particles. The ability to achieve consistency across different concentrations furthermore provides confirmation that cluster-cluster interactions have little effect on aggregation statistics, a condition achieved in turn

by the optimized cluster definition.

Although the thermodynamics of MTBE clusters formations studied here present no special surprises, these calculations have demonstrated that cluster free energies can be extracted from the simplest, most modest simulations of equilibrium aggregation behavior. In principle these same trajectories could be analyzed to extract rate constants for monomer addition to and dissociation from clusters, providing enough information to make predictions of nucleation rates.[67] The ability to extract cluster free energies from unbiased simulations should find also applications in phenomena other than nucleation, for instance in the modeling of micelles, where unbiased simulations are part of standard protocol and finite-N effects are a common worry.[51]

This approach is optimally suited for systems with relatively weak association energies, as the rates of association and dissociation under the temperature and concentration conditions modeled need to be high enough to observe many events over the course of the simulation. For aqueous solutions, monomer solubilities or critical micelle concentrations below $\sim$2 mM would require inconveniently large systems ($>$100,000 solvent molecules) to obtain a spread of cluster sizes in a single trajectory. On the other hand, for clusters that are only observed at high monomer concentrations, non-ideal effects (monomer-monomer or cluster-cluster interactions) will complicate the analysis but could in principle be incorporated.

5.6 Acknowledgments

Chapter 6 Simulations of NaCl aggregation from solution: Solvent determines topography of free energy landscape

6.1 Introduction

Nucleation is a ubiquitous process by which molecules, atoms, or ions in a solution, a melt or a gas phase locally adopt a density, order, or composition that initiates the formation of a new phase.[1] Its influence can be seen across disciplines ranging from the formation of fullerenes from a hot gas of carbon dimers,[101] the impact of water ice-crystal nucleation on climate change or intracellular freezing, the formation of the amyloid fibers associated with Alzheimer's disease, crystal formation in kidney stone disease and gout, and clathrate hydrate aggregation in natural gas pipelines.[102] Nucleation of NaCl crystals may seem comparatively mundane, but in recent years it has been found that the nucleation mechanism is rather nuanced. The simplicity of the system, and the availability of experimental data on nucleation rates and solubility, have made it an appealing system to simulate.

Crystal nucleation begins on a size scale of nm that is not currently accessible using experimental methods and therefore lends itself to being understood through atomistic simulations. Great strides have been made in understanding NaCl crystallization since molecular dynamics (MD) simulations were first attempted in the 1990s by Ohtaki and Fukushima.[103] MD has been used to capture the dynamics of nucleation in supersaturated systems with many ions in a number of studies;[104–109] a drawback of this approach is that the nucleation event typically occurs only once per trajectory. Trajectory path sampling (TPS), an unbiased improved sam-

---

[1]This chapter present results that are under review to be published in the Journal of Computational Chemistry.

pling technique, has also been used with success to examine the nucleation pathway of NaCl.[110–113] The rare event nature of nucleation has also encouraged the development of improved sampling methods include biased MD simulations techniques like umbrella sampling,[114–116] and metadynamics.[53, 72, 73, 117, 118] A drawback of these methods is that the user is required to dictate the final state and/or the pathway to some extent.[119]

Classical nucleation theory (CNT),[120] while able to describe the growth of the critical nucleus in advanced stages, fails to adequately describe the initial stages of nucleation most often because it is very difficult to adequately describe the interfacial energy of small clusters. Furthermore, there is a growing body of research indicating that NaCl nucleation (like a number of other cases of crystal formation from solution[121]) proceeds via a two-step nucleation mechanism consistent with Ostwald's rule of stages.[122]

The two-step nucleation mechanism involves the formation of amorphous clusters or liquid droplets that then transition into an ordered NaCl crystal nucleus. Unbiased MD simulations of NaCl crystallization from solution performed by Chakraborty and Patey indicate that nucleation is preceded by density fluctuations that lead to regions of high concentrations of Na^+ and Cl^- ions but in an amorphous arrangement;[104, 105] nucleation events were observed only where the local concentration exceeded the average concentration. Spatial ordering of the ions into the crystalline arrangement was then the rate-determining step.[105] Lanaro and Patey performed direct MD simulations revealing that 90% of the clusters that failed to nucleate had lifetimes shorter than 2 ns.[106] Clusters with higher crystallinity had longer survival times for all cluster sizes ($s \leq 6$) and it was concluded that cluster crystallinity/order has significant influence over the probability of nucleating.

Giberti et al.[53] performed metadynamics biased MD simulations on NaCl nucleation and also observed the formation of relatively unstable small amorphous clusters

that would redissolve in 1-2 ns in unbiased MD simulations. Unlike the direct simulation techniques, they saw two longer lived ordered structures (taking 25-30 ns to re-dissolve): the expected rocksalt structure and a wurtzite structure. Geometric optimization of the wurtzite structure showed that it corresponds to a metastable minimum of the bulk NaCl potential energy landscape. They proposed that the wurtzite polymorphism would be more stable for small NaCl clusters (radii< 5.7 nm) due to the lower surface tension at a wurtzite/water interface. Zimmerman et al.[123] use the analysis of the growth of nanocrystalline seeds at high concentration to determine an apparent surface tension (which can be related to a cluster free energy via CNT) as well as to estimate nucleation dynamics overall.

A full characterization of the thermodynamic landscape that influences NaCl crystal nucleation from solution requires calculation of the free energy of formation of ordered and/or amorphous NaCl clusters from aqueous solution. Chen and Pappu used the law of mass action on relatively large, unbiased simulations of ionic cluster formation in the dilute limit to extract equilibrium association constants for small clusters ($s \leq 6$) with consistent results across a range of simulated salt concentrations.[52, 73] Using the law of mass action to calculate the equilibrium association constant, and thus the free energy surface, works rather well for bulk solutions. It requires a system where the formation of a cluster does not significantly alter the concentration of the other clusters in the system. Extracting equilibrium association constants for larger clusters is more challenging in large systems. At the dilute limit, only small clusters form with statistically relevant frequency. The solute concentration must be near or above the saturation concentration to form larger clusters. In systems that undergo a nucleation type mechanism of growth, once the critical nucleus forms, this propels the system into rapid cluster size growth. The equilibrium state consists of some cluster size fluctuation around one large cluster in the system in the presence of free monomers and small clusters. This makes it impossible to establish equilibrium

constants for intermediate cluster sizes as they are transient.

To observe larger cluster sizes in a small system, simulations must be performed at concentrations significantly higher than the saturation threshold.[53] In simulations of a small system, the formation of clusters depletes the concentration of free monomers in solution to around the saturation threshold. If the total concentration is not well above the saturation concentration, this depletion effect can impede the formation of cluster nuclei.[54] Salvalaglio et al. termed this a consequence of a confinement effect, like a fixed system size, where the stability of clusters does not depend on just the supersaturation but also the system size.[55] Across a number of simulations at different concentration and/or system sizes, this allows continuous coverage of all the cluster sizes around and past the critical nucleus size. At these smaller system sizes, the formation and breaking of larger clusters significantly alters the concentration of other cluster sizes in the system and the law of mass action no longer serves as a good model to calculate the equilibrium association constants. Our previous work on methyl t-butyl ether (MTBE) cluster formation shows that the deviation from the real association constants increase as cluster size increases.[60] The recently developed Partition Enabled Analysis of Cluster Histograms (PEACH) method avoids this problem as it accounts for the discrete nature of state partitions explicitly.[59] It allows extraction of free energies of formation from cluster distributions obtained through unbiased simulations of small systems, even when a large fraction of the monomers are involved in cluster formation, and has been applied to the free energy of assembly of anionic surfactant micelles.[59]

In this paper, the PEACH method is used to calculate the free energy surface for the nucleation of NaCl clusters solvated in water, methanol, and a mixture of water and methanol from atomistic MD simulations. Differences among the free energy landscapes inferred from these simulations indicate that the cluster size at which ordered clusters become more stable than amorphous clusters is sensitive to

solvent environment. This factor, and the accompanying effects on the rate of the ordering transition, suggest how solubility influences whether nucleation proceeds via a two-step process. Furthermore, results from PEACH analysis of the aqueous system highlight the importance of non-ideal effects in promoting cluster formation at high concentrations.

6.2 Methods

All atomistic molecular dynamics simulations were performed using GROMACS[35–38] (version 4.6.5). Molecular images from the trajectories were made using VMD.[124] Simulations were performed in the NPT ensemble for four different system compositions of NaCl solvated in: pure SPC/E water,[125] pure OPLS-AA (all atom) methanol,[126] and a mixture of the two solvents with a mole fraction, x_{MeOH}, of methanol to total solvent of 0.4 and 0.5. The Na$^+$ and Cl$^-$ ions were simulated using the OPLS force field parameters[39, 127] given in Table 1; these parameters were chosen for consistency with the work of Chakraborty and Patey showing two-step nucleation.[104, 105]

Table 6.1: Ion force field parameters.

Ion	σ/nm	$\epsilon/\mathrm{kJ\ mol^{-1}}$
Na$^+$	3.330×10^{-1}	1.160×10^{-2}
Cl$^-$	4.417×10^{-1}	4.928×10^{-1}

Simulation pressure was maintained at 1 atm with $\tau_P = 0.2$ ps and a compressibility of 4.5×10^{-5} bar^{-1} using an isotropic Berendsen barostat.[4] The temperature was maintained at 300 K (pure water) or 323 K (all other solvent environments) by velocity rescaling with $\tau_T = 0.1$ ps.[6] Non-bonded interactions between different species were handled using the $\epsilon_{ij} = (\epsilon_{ii}\epsilon_{jj})^{1/2}$ and $\sigma_{ij} = 0.5\,(\sigma_{ii} + \sigma_{jj})$ combination rules. Lennard-Jones interactions were truncated at 1.0 nm for aqueous systems and 1.2 nm for methanol-containing systems. Coulomb interactions were handled using

Particle Mesh Ewald (PME)[128] summations with a real space cutoff of 1.0 nm.

Simulations of all systems containing methanol are performed with a time step of 1 fs. Trajectory configurations were saved every 500 steps for the ions, and every 5000 steps for the whole system. Simulations of NaCl clustering in pure methanol were run for 1000 ns for the simulations with 1 to 5 N_{NaCl} and for 500 ns otherwise. Composition details are given in Table 6.2.

Table 6.2: Simulation details for NaCl aggregation in pure methanol ($x_{\mathrm{MeOH}} = 1.0$) at 323K and 1 bar.

N_{NaCl}	[NaCl]/m	N_{Sol}	[NaCl]/m	N_{Sol}
1	0.05	672	0.03	988
2	0.09	670	0.06	986
3	0.14	668	0.10	984
4	0.19	666	0.13	982
5	0.24	664	0.16	980
6	0.28	662	0.19	978
7	0.33	660	0.22	976
8	0.38	658	0.26	974
9	0.43	656	0.29	972

The simulations of NaCl clustering in solvent mixtures of SPCE/E water and OPLS-AA methanol were run for 200 ns for concentrations of 0.42-1.69 molal (m). Details about the composition are given in Table 6.3.

Table 6.3: Simulation details for NaCl aggregation in solvent mixtures consisting of methanol and SPC/E water at 323 K and 1 bar with 20 ion pairs.

N_{NaCl}	[NaCl]/m	[NaCl]/m
500	1.69	1.60
700	1.21	1.14
800	1.06	1.00
900	0.94	0.88
1000	0.85	0.80
1500	0.56	0.53
1900	-	0.42
x_{MeOH}	0.4	0.5

Simulations of NaCl solvated in pure water were performed for systems containing 20, 30, and 45 NaCl pairs. The simulation composition details are given in Table 6.4

for each concentration and system size. These MD simulations were performed with a time step of 2 fs. Snapshots of the trajectory are printed every 500 steps for the ions and every 50,000 steps for the whole configuration. These simulations were run for 500 ns except for the simulations of 30 NaCl pairs at concentrations of 2.78-3.89 m, which were run for 2000 ns. Only amorphous clusters were observed in the 0.56-3.89 m simulations, which were used for the multiset fitting.

Table 6.4: Simulation details for NaCl aggregation in SPC/E water at 300 K and 1 bar. Only simulations with concentrations of 0.56-3.89 m were used in the PEACH global fit analysis.

$[\text{NaCl}]/\text{m}$	N_{Sol}	N_{Sol}	N_{Sol}	t_c/ns
7.21	-	231	346	170
6.66	-	250	375	-
6.10	-	273	409	320
5.55	-	300	450	10
5.00	-	333	500	-
4.44	-	375	562	610
3.88	286	429	643	-
3.33	333	500	750	-
2.78	400	600	900	-
2.22	500	750	-	-
1.67	667	1000	-	-
1.11	1000	1500	-	-
0.56	2000	3000	-	-
N_{NaCl}	20	30	45	

The higher concentration simulations (4.44-7.21 m) were performed for systems with 30 and 45 NaCl pairs to see if ordered crystals would appear. The trajectories for the 30 and 45 NaCl simulations were respectively 2000 ns and 500 ns long. A large crystal with a rocksalt structure was formed in the 45 NaCl simulation at concentrations of 4.44 and 5.55-7.21 m. No large ordered crystals were formed in the simulations with 30 NaCl pairs or in the simulation at 5.00 m with 45 NaCl pairs. The approximate time to crystalize, t_c, is given in Table 6.4 for systems with 45 NaCl pairs.

To ascertain whether the simulations of NaCl clusters in water are limited to the

formation of relatively small amorphous clusters at the lower set of concentrations (Table 6.4) by kinetic or thermodynamic factors, a fully formed crystal with a rock-salt structure was taken from the simulation of 45 NaCl ion pairs at 5.55 m and "seeded" in solvent environments mimicking those used in the aggregation simulations (Table 6.5). Each seed was constrained for a 1 ns period to allow for solvent equilibration around the crystal and then the constraints were removed to see if the crystal structure persists over a 100 ns trajectory. The first set of simulations were performed at the same concentration (3.88 m) while decreasing the total number of ion pairs in the system. The second set of simulations maintained a total of 45 ion pairs while decreasing the concentration.

Table 6.5: Simulation details for testing the resilience of an NaCl crystal with a rock-salt configuration under solvation conditions equivalent to those used for the aggregation simulations (Table 6.4).

N_{NaCl}	$[\text{NaCl}]/\text{m}$	N_{Sol}
35	3.88	500
40	3.88	571
45	3.88	643
45	3.33	750
45	2.78	900
45	2.22	1125

For the application of the PEACH method, it is assumed that the cluster size frequency distribution data of a simulation is gathered over a set of equilibrium configurations. Simulations of the cluster formation started from configurations with ions randomly dispersed in solution. We consider the equilibrium condition to be met when the maximum cluster size in the trajectory reaches the average maximum cluster size in the simulation and then proceeds to fluctuate around that value. For the simulations in SPC/E water this process took 5-10 ns, for the mixed solvent simulations it took 5-15 ns, and for the pure MeOH solvent simulations the equilibration took 5-10 ns with the exception of the 0.22 m simulation containing 7 N_{NaCl} where it took ~40 ns. In the averaging of the cluster size frequency distributions over each

trajectory, we excluded the first 20 ns of simulation time. The trajectories were split into 20 ns blocks to generate the standard error in the cluster size distributions.

6.2.1 Cluster definition

We use a cluster definition that was first implemented by Stillinger.[89] A cluster consists of a group of ions that are linked by a sequence of neighbor associations. Ions of a different charge from the observation ion that are within a cutoff distance r_c are defined as neighbors. The choice of cutoff distance $r_c = 0.34$ nm was made based on a distance intended to capture the first peak of the Na-Cl radial distribution function (Fig. 6.1).
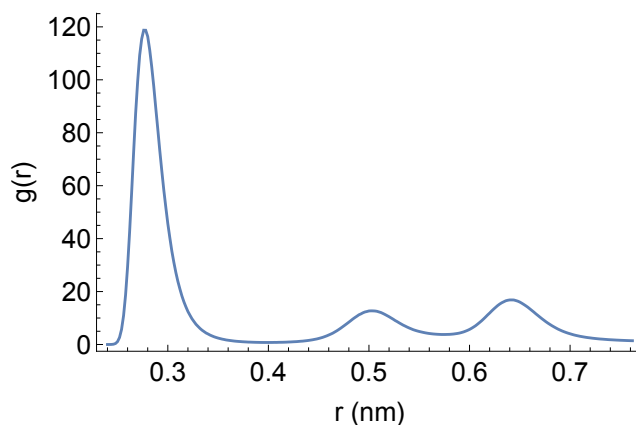
Figure 6.1: Radial distribution function $g(r)$ for Na and Cl ions of the simulation of 20 NaCl ion pairs at 1.60 m in the solvent mixture with $x_{\mathrm{MeOH}} = 0.5$.

There was no consistent correlation between the the quality of fit and the cluster definition (see Appendix D.1 Fig. D.1) and thus it was not possible to identify a cutoff that minimized the error of the global fit as was done in ref. [60].

6.2.2 Order parameter definition

To identify ordered rock-salt structures the second and third order invariant Steinhardt order parameters[129] within clusters were calculated. There are a number of papers that have used Q_l, the second order invariant Steinhardt order parameter.[106, 130–132] While it is less expensive to compute than the third order invariant, Mickel

et al. found that that the value of Q_l is not independent of the number of nearest neighbors used to calculate its value,[133] limiting its usefulness for very small clusters. The magnitude of the third order invariant Steinhardt order parameter $\widehat{W_l}$ is however supposed to be unaffected by changing the definition of the nearest neighbor and $\left|\widehat{W_l}\right|$ provides a direct index of the symmetry for a particular cluster.[129] This does not mean that changing the number of nearest neighbors does not change the value of $\widehat{W_l}$ for a very small cluster.

To calculate the third order Steinhardt invariant $\hat{W}_l$, we must first define $Q_{lm}(\vec{r})$, where l and m are quantum numbers and $l \leq m \leq -l$. $Q_{lm}(\vec{r})$ is the value of a spherical harmonics function for a bond with a vector $\vec{r}$ between two atoms that satisfies $|\vec{r}| < r_c$. This definition of $\vec{r}$ differs slightly from the original publication[129] but produces the same results.

$$Q_{lm}(\vec{r}) = Y_{lm}(\theta(\vec{r}), \phi(\vec{r})) \tag{6.1}$$

For each ion, $Q_{lm}(\vec{r})$ is averaged over the nearest neighbors within the cutoff distance r_c and we define the nearest neighbor bonds as only existing between oppositely charged ions.

$$\bar{Q}_{lm} = \langle Q_{lm}(\vec{r}) \rangle \tag{6.2}$$

From the second order Steinhardt invariant $\bar{Q}_{lm}$, the third order Steinhardt order parameters is calculated as,
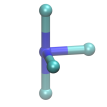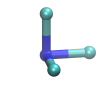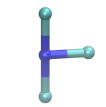
$$\hat{W}_l = \left(\sum_{m=-l}^{l} |\bar{Q}_{lm}|^2\right)^{-2/3} \sum_{\substack{m_1,m_2,m_3 \\ m_1+m_2+m_3=0}} \begin{bmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{bmatrix} \times \bar{Q}_{lm_1}\bar{Q}_{lm_2}\bar{Q}_{lm_3} \tag{6.3}$$

where the coefficients in the matrix are Wigner $3j$ symbols.

Since we are interested in evaluating the structure of small clusters, we evaluate the third order invariants for each atom and its direct neighbors individually to ascertain whether it meets the criteria of a rock-salt structure. Table 6.6 shows the Steinhardt third order invariant for a central ion coordinated with $3-6$ ions of a different charge,

assuming a rock-salt structure. Depending on which of the six bonds are excluded in a rock-salt conformation, the $\hat{W}_l$ values are liable to change.

Table 6.6: Third order Steinhardt invariants for $l = 4, 6, 8$ for the typical conformation of a coordinated ion in the rock-salt for a range of possible coordination numbers.

Coord. #	6	5	4	3	4*	3*
$\hat{W}_4$	0.1593	0.1529	0.1534	0.1593	0.1250	0.1275
$\hat{W}_6$	0.0132	-0.0067	-0.0341	0.0132	-0.0072	-0.0433
$\hat{W}_8$	0.0585	0.0599	0.0608	0.0585	0.0638	0.0659

$\widehat{W}_l$ values for $l = 4$ are sufficient in distinguishing between the rock salt ordered states and the amorphous states. We show in Fig. 6.2 that in a simulation where the ordered state forms regularly (1.60 m) versus a simulation where the ordered state is comparatively infrequent (1.14 m), the presence of an ordered state shifts $\widehat{W}_4$ to values greater than 0.14. We define ordered atoms as having $\widehat{W}_4 > 0.145$ (indicated with a grey line in Fig. 6.2).
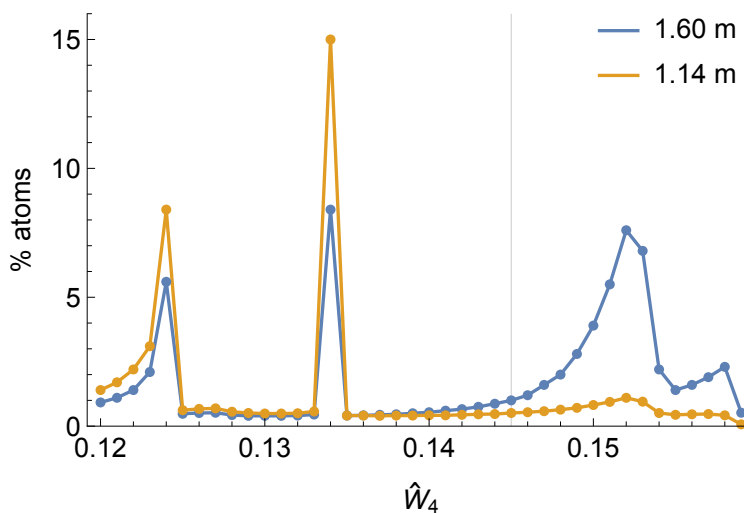
Figure 6.2: The figure on the right shows the results from the mixed solvent simulation with $x_{\text{MeOH}} = 0.5$ and 20 NaCl pairs solvated at 1.60 and 1.14 m. The simulation at 1.60 m transitions between a large ordered cluster and amorphous clusters.

6.2.3 Implementation of the PEACH method

The partition-enabled analysis of cluster histograms (PEACH) method uses simulation cluster size distributions from a set of simulations over a range of concentration and/or system sizes to generate equilibrium association constants $K_{i,j}$ for each cluster with i and j representing the numbers of Na$^+$ and Cl$^-$ ions in the cluster.[59] Using the exact relationship between the single-cluster partition functions $q_{i,j}$ $(= q_{i,j}^{\circ} V/V^{\circ}$, with q° the partition function of a single cluster in a standard volume $V^{\circ} = 1/c^{\circ})$ and the canonical partition functions $Q(N_A, N_B, V, T)$ for a system with a total of N_A and N_B parts of each molecule type, the average cluster frequency is,

$$\langle n_{i,j} \rangle_{N_A, N_B} = q_{i,j} \times \frac{Q(N_A - i, N_B - j, V, T)}{Q(N_A, N_B, V, T)}. \tag{6.4}$$

The PEACH method thus generates a set of average cluster frequencies that are then used in tandem with the raw simulation cluster size frequencies to adjust the standard state partition functions $q_{i,j}^{\circ}$ to better reproduce the simulated average cluster size frequencies. If the cluster definitions and observed statistics are consistent with an equilibrium distribution of non-interacting clusters, the PEACH method converges to an optimal fit in the partition functions which can then be used to generate a final set of equilibrium associations constants, $K_{i,j}$. The free energy of aggregation of each cluster is defined for some free monomer concentration c as,

$$\Delta G_{i,j} = -k_B T \ln \left(K_{i,j} (c/c^{\circ})^{i+j-1} \right) \tag{6.5}$$

In an actual system there may be different concentrations of the two free ions (see Appendix section 4.5), but Eqn. 6.5 is most convenient for representing the association constants. Using the same assumption of equal concentrations of free anions and cations, we can assume that at equilibrium the relative populations of clusters with the same total number of ions $n = i + j$, will be proportional to $K_{i,j}$. The average and standard deviation of cluster charge $z = i - j$ can then be calculated as a function of n.

6.3 Results

Results from simulations of NaCl cluster formation from methanol, mixed methanol/water, and pure water solutions will be presented and followed by a comparative discussion of how solvent affects the free energy landscape of cluster formation.

6.3.1 Pure methanol

NaCl cluster formation from methanol solution proves to be the simplest cluster formation system, although the most computationally challenging for the PEACH method. Due to the low solubility and high ion-ion affinity, the dissociation rate of individual ions or ion pairs is slow on the simulation timescale (resulting in slow convergence of cluster size distributions) and the average numbers of free ions or ion pairs are also low (resulting in narrow equilibrium cluster size distributions). The infrequent monomer exchange does not allow cluster size distributions with sufficient overlap between the cluster size distribution of different simulations for an effective global fit. Fig. 6.3 demonstrates that the ions in most of the simulations spend more than 60% of the simulation time in a single maximum cluster size for the system. This propensity is slightly less pronounced for the dilute set of simulations of 900 molecules and ions, but not enough to improve the overlap of cluster size distributions.

The surface resulting from the global fitting method is shown in Fig. 6.4 for a monomer concentration of 0.0009 NaCl/nm^3 (0.002 m) for the Na$^+$ and Cl$^-$ ions. The free energies of clusters with extra negative charge are lower than those with extra positive charge, presumably from the greater desolvation energy of the (smaller) Na$^+$ relative to Cl$^-$.

In contrast to the relatively infrequent changes in cluster size, we observe rapid fluctuation in cluster structures over the lifetimes of individual clusters, including oscillations between ordered and disordered structures. Fig. 6.5 shows the fraction of clusters that contain at least 50% of their ions in rock-salt environments as measured
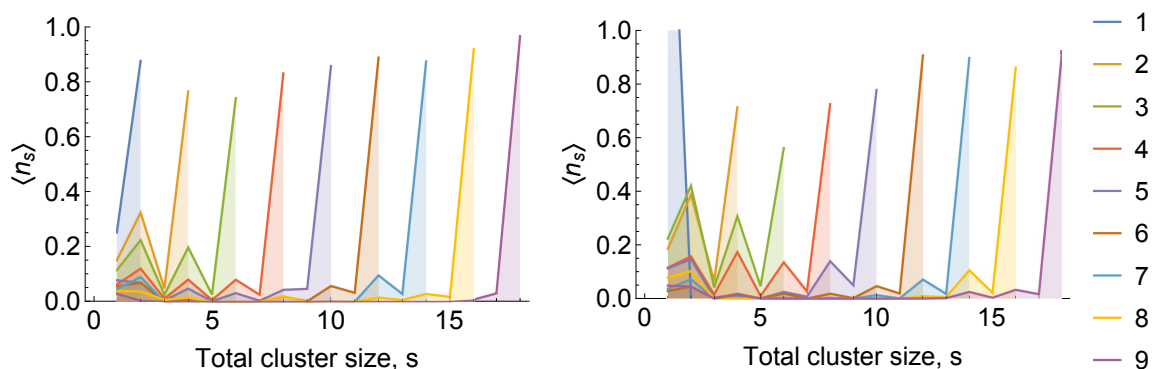
Figure 6.3: Cluster size distributions in terms of cluster number frequency for the simulations of Na and Cl ions solvated in methanol with a total of 674 (top) and 900 (bottom) molecules and ions. The legend on the right side of each plot indicate the number of NaCl ion pairs are present in each simulation. Cluster constituents are not distinguished in this plot for clarity.

by the $\widehat{W_4}$ order parameter (see Methods). It is noteworthy that this a measurable fraction even for the smallest cluster that can form a complete NaCl unit, i.e. Na_4Cl_4, and that the degree of order grows non-monotonically with cluster size. "Magic numbers" with higher degree of order are observed, as expected, for cluster sizes of 8 (a single "box"), 12 (pair of "boxes" sharing a face), and 18 (cluster of 4 "boxes" arranged in a square, see Fig. 6.6). The degree of order vs. size is nearly superimposable for two independent simulations in different volumes, indicating that sampling of different structures is well-converged over the simulation runs.
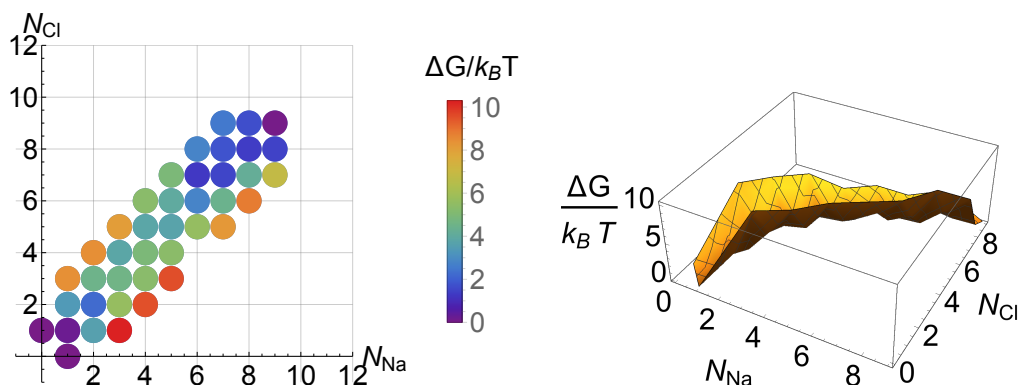
Figure 6.4: Free energy surface for the formation of NaCl cluster solvated in methanol as a 2-D (left) and 3-D surface (right) for a monomer concentration of 0.0009 NaCl/nm^3 (0.0021 m).
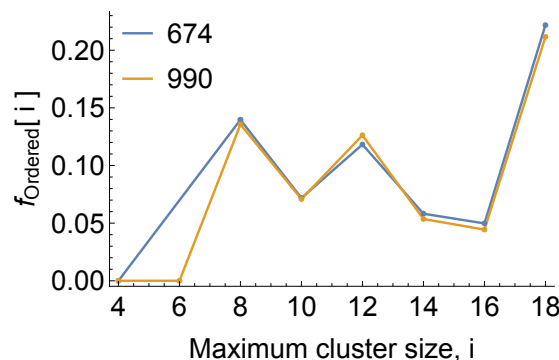
Figure 6.5: Order parameter trends for the maximum cluster size present in the NaCl cluster formation simulation solvated in methanol with systems containing 674 and 990 molecules. A cluster is considered ordered if there are 50% or more atoms in the cluster with $\widehat{W}_4 \geq 0.145$.

One particularly interesting mode of fluctuation of the Na_9Cl_9 structure is captured in snapshots shown in Fig. 6.6. Two well-ordered rocksalt structures, which differ only in which internal Na-Cl bonds are formed, interconvert through an intermediate state consisting of 3 stacked hexagons. (It should be noted that there are no solvent molecules within the cylindrical cavity.) The transition is complete in under 100 ps.
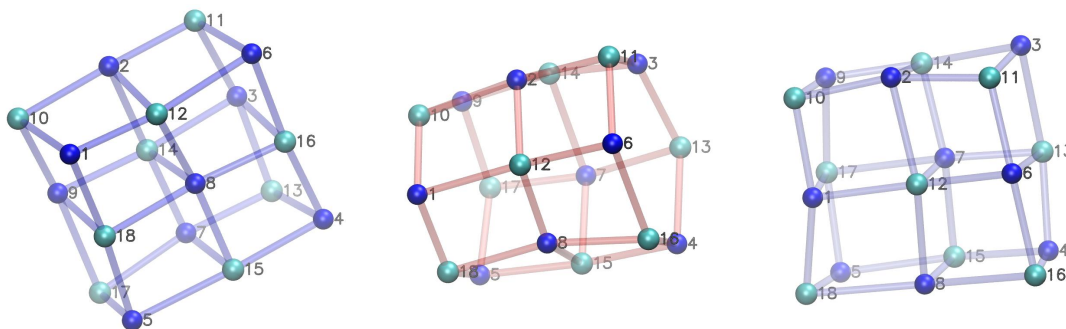
Figure 6.6: Snapshots of a 9 Na and 9 Cl cluster fluctuating between two cubic ordered states through a stacked hexagonal configuration transition state when the internal Na-Cl bonds break. The snapshots correspond to 499.647, 499.691, and 499.731 ns in the trajectory. A short movie of this transition is included in the supplemental material.
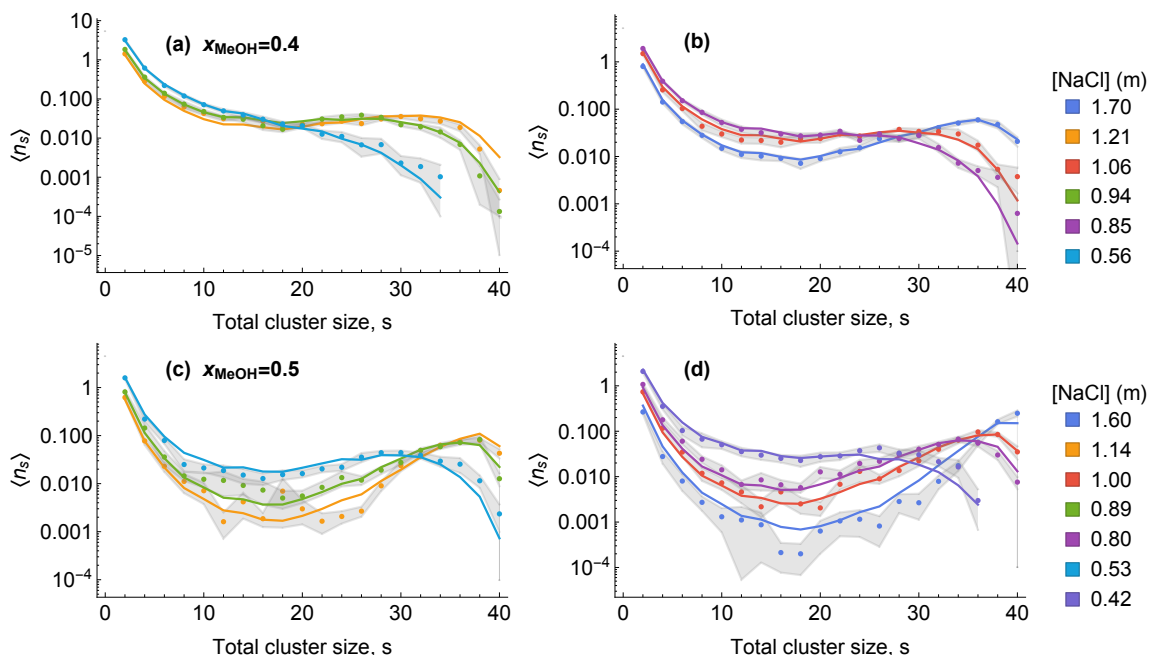
Figure 6.7: Plots of the cluster size frequency distributions based on the simulated raw results from NaCl in methanol/water mixtures (points) with the standard error (grey shaded regions) and the result of the global fit from the PEACH analysis (lines). The distributions were split into two plots for visibility with data for simulation with $x_{MeOH} = 0.4$ given in plots (a) and (b), and $x_{MeOH} = 0.5$ given in plots (c) and (d).

6.3.2 Mixed solvent simulation results

The mixed MeOH/H$_2$O solvent simulations (with solvent mole fractions $x_{MeOH} = 0.4$ and 0.5) are characterized by frequent monomer exchange at concentrations where large clusters can form and then break apart. Simulations of 20 NaCl over a range of concentrations sample neutral clusters of all sizes as well as clusters with charge asymmetries of up to ± 4.

The PEACH analysis produced a good fit over a range of concentrations for $x_{MeOH} = 0.4$, but was less successful in the $x_{MeOH} = 0.5$ system, where the fit underpredicted the prevalence of large clusters at high concentrations. Fig. 6.7 shows the observed and fitted levels of neutral clusters only; full plots showing all species are given in the Appendix Fig. D.6 and D.7.

We propose that the less successful fitting in the case of the $x_{MeOH} = 0.5$ system is

likely due to the onset of crystalline ordering in large clusters. We find that the size at which clusters display ordering is larger in these mixed solvents than in pure methanol. Whereas over 10% of very small (Na_4Cl_4 or Na_6Cl_6) clusters aggregating from pure methanol adopt an ordered rocksalt structure, in the mixed solvents this fraction was under 1%. (Fig. 6.8). The proportion of clusters showing significant ordering climbs steeply for sizes over 30 total ions. In principle the existence of multiple forms (e.g. ordered and disordered) of a cluster of a certain size s does not invalidate the framework used here to analyze cluster distributions; the partition function $q_{i,j}$ associated that size cluster will incorporate contributions from all varieties. The kinetics of the order/disorder transition may, however, introduce a sampling problem. Fig. 6.8 (center plots) shows the maximum cluster size and the fraction of atoms in the maximum cluster size that meet the order criteria, x_{Ordered}, over the course of the trajectories. The latter plot has the appearance of a two-state system, but with considerable "noise" overlaid. Transitions between an ordered and amorphous state occur 10 times over a 200 ns trajectory for the 1.60 m simulation, which is insufficient to establish the equilibrium populations of the two states. In the lower concentration simulations, 1.14 and 1.00 m, the transitions are even less frequent. Since most of the trajectories have minimal sampling of ordered states for the larger cluster sizes, the one simulation where the ordered state is present more than 30% of the time for clusters sizes larger than 35 is poorly reproduced by a global fit that favors the amorphous state.

In contrast to the simulations in pure methanol, here the sampling issue is not so much the formation of a cluster of a given size but the transition between conformations for that cluster size. The rapid oscillations in the fraction of ions in ordered environments, in both the ordered and amorphous states, reflects the dynamic restructuring of surface ions (in the ordered case) and the rapid appearance and disappearance of locally ordered environments within the amorphous clusters.
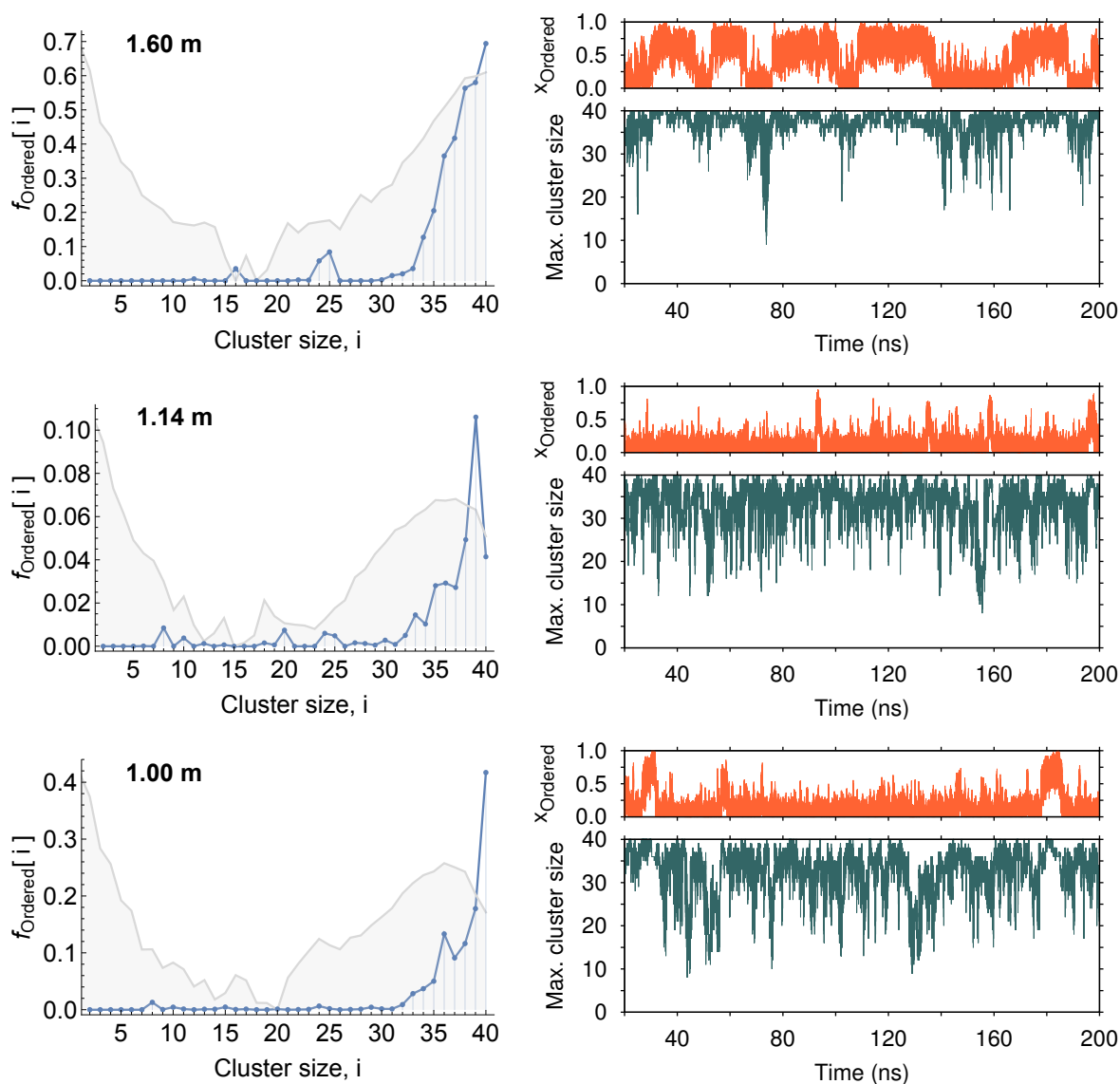
Figure 6.8: Order parameter trends for the three highest concentration simulations of NaCl solvated in the 0.5 x_{MeOH} solvent mixture. The figures on the left show the fraction of clusters of size i where more than 50% of the atoms meeting our order criteria (blue line and points) $f_{\mathrm{Ordered}}[i]$, and the 1-D cluster size histogram shaded in light grey in the background to indicate the relative cluster size frequencies for that simulation (irrespective of the cluster composition). The plots on the right show the maximum cluster size (teal green line) and x_{Ordered}, the fraction of atoms in that cluster that meet the order criteria, versus the simulation time.
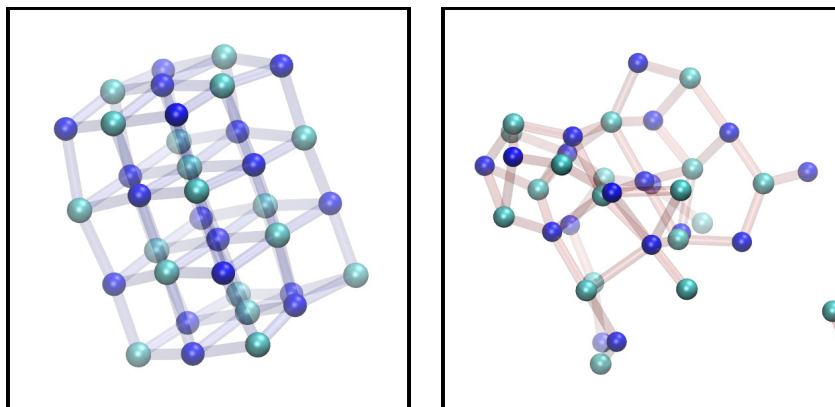
Figure 6.9: The snapshots show the ordered (blue bonds) and disordered (red bonds) structures of a 36-mer from the simulation of NaCl solvated in the 0.5 x_{MeOH} solvent mixture at a solute concentration of 1.60 m.

The transitions within this small, fluctuating cluster are presumably closely related to the nucleation of ordered crystals from within larger amorphous clusters observed by Chakraborty and Patey.[104]

Poor sampling may also explain why the degree of ordering vs. cluster size was not consistent over three concentrations in the $x_{\mathrm{MeOH}} = 0.5$ mixed solvent (Fig. 6.8), in contrast to the consistent degree of order observed in clusters forming in pure methanol under different simulation conditions (Fig. 6.5). Two other explanations related to systematic effects are also plausible. The first relates to the mixed solvent. In a small constant-composition system with a mixed solvent, if the number of one solvent type solvating some population of free ions or clusters is a significant fraction of the total number of waters in the system, the remaining solvent will be depleted of water to a degree that depends on system size and will provide a different solvation environment for the other clusters. It is also possible that there are non-ideal effects relating to the interaction between the cluster and surrounding "free" ions or other clusters. This possibility can be ruled out as a sole cause (although it might be a contributing factor) since the discrepancy is observed even for the 40-mer cluster, which will always be the only ion cluster in the simulation box.

With the caveat that the fitting may not properly capture the equilibrium between
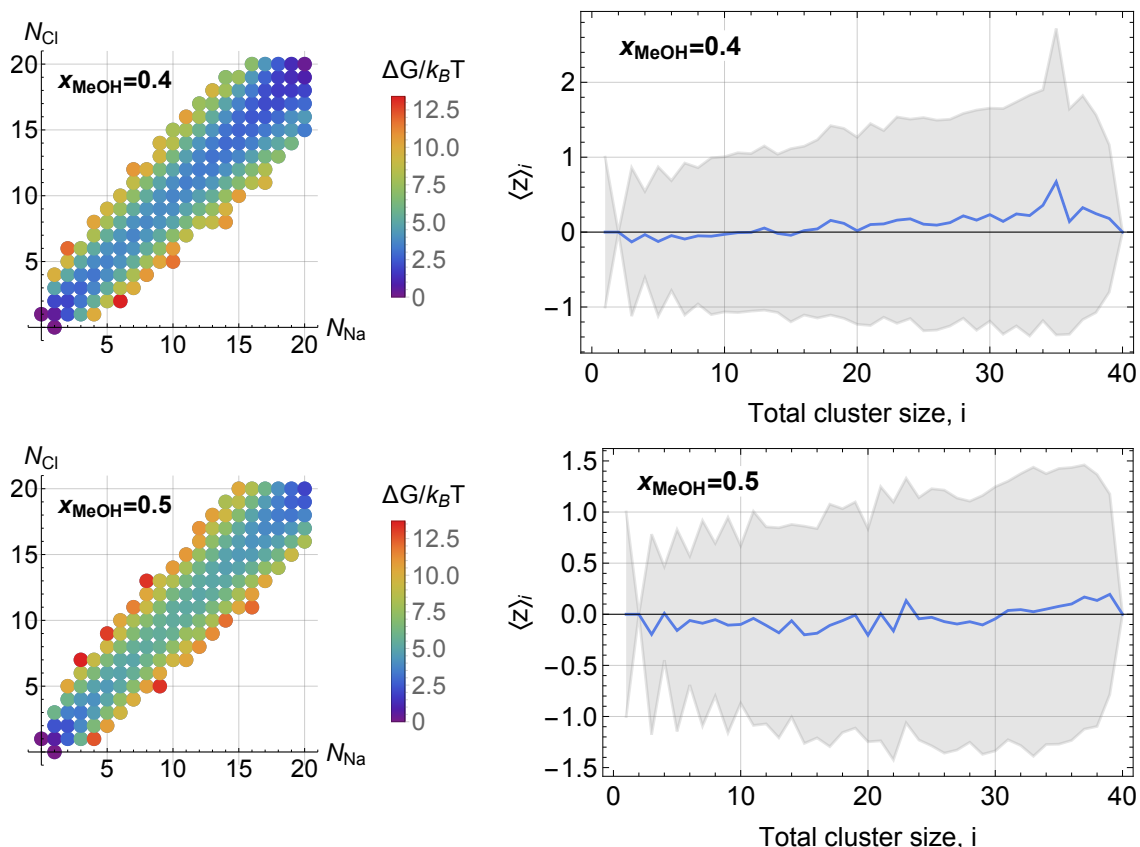
Figure 6.10: Gibbs free energy of association for NaCl cluster formation in mixed solvent with a monomer concentration of 0.085 NaCl/nm^3 (0.17 m) for the $x_{\mathrm{MeOH}} = 0.4$ solvent mixture and 0.050 NaCl/nm^3 (0.10 m) for the $x_{\mathrm{MeOH}} = 0.5$ solvent mixture. The right column of plots shows the average cluster charge $\langle z \rangle_i$ for cross sections of the free energy surface where the total cluster size i is the same. The region shaded in grey indicates the weighted standard deviation in the cluster charge.

ordered and disordered clusters at large sizes, it is interesting to examine the free energy surfaces themselves, shown in Fig. 6.10. In both cases, free energies rise and fall again moving from small to large cluster sizes, qualitatively consistent with the nucleation of a new phase of amorphous NaCl, which at some point (at or near the upper size limit here) will become unstable with respect to the formation of ordered nanocrystallites. The free energy curves are roughly symmetrical with respect to sodium and chloride content. Considering charge as the second dimension of the growth of NaCl clusters, with fluctuations away from the neutral necessary if the growth is made by single ions, we plot the mean and rmsd charge as a function
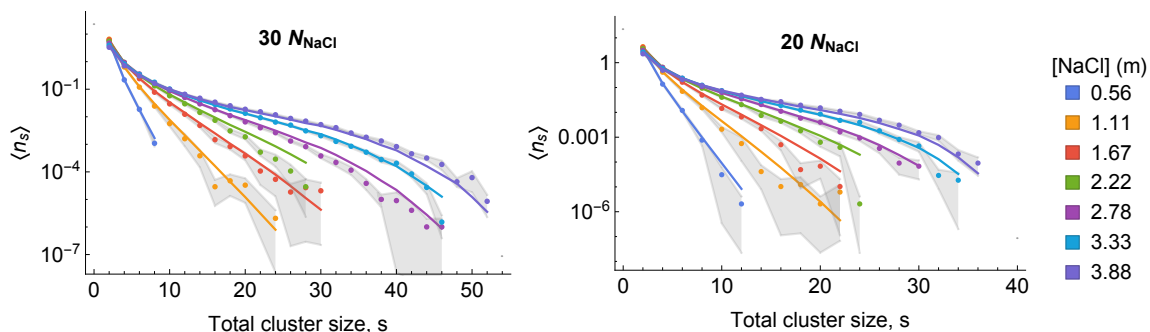
Figure 6.11: Plot of the cluster size frequency distributions for clusters solvated in SPC/E water ($x_{\text{MeOH}} = 0$) based on the simulated raw results (points) with the standard error (grey shaded regions) and the result of the global fit from the PEACH analysis (lines). The two plots display the data for simulation containing 30 and 20 NaCl pairs; a plot showing results from 45 NaCl pairs is given in Appendix Fig. D.3 and full 2D surface plots are given in Appendix Fig. D.3-D.5.

of cluster size (Fig. 6.10, right), assuming the concentrations of free ions of either charge are identical. From the data we infer that there is no definitive cluster charge preference for clusters solvated in either of the mixed solvent environments in this size range.

6.3.3 Pure SPC/E water solvent simulation results

NaCl cluster formation in pure SPC/E water requires significantly higher salt concentrations than the other two solvent environments to promote the formation of large and/or ordered clusters. Large amorphous clusters that contain strings of alternating anions and cations, could form cycles across periodic boundaries; such structures are artificially stabilized by the boundary conditions, so trajectories that produced such structures were not used in the PEACH analysis. First, we will discuss the set of 7 concentrations from 0.56 m to 3.88 m with numbers of NaCl pairs 20, 30, and 45, in which exclusively amorphous clusters were observed.

As we saw for the mixed solvent simulations, the agreement in magnitude between the cluster size distributions from simulation and those reproduced using the global fit equilibrium association constants is at first glance rather good, shown in Fig. 6.11 for neutral clusters. Closer inspection, however, shows discrepancies between the fitted

and observed distributions that are significant both for their magnitude relative to the standard error in the simulation and because they are systematic in their dependence on total NaCl concentration. At concentrations below 3.33 m, the fit consistently predicts higher levels of large cluster formation than observed in simulation, while at the single concentration above 3.33 m, the fit predicts lower levels for large clusters. So, even after finite-N effects are handled using the PEACH analysis, the equilibrium constants describing aggregation into large clusters (> 8 ions) appear to be growing with increased concentration. Unlike the mixed-solvent simulations where sampling of the transitions between the ordered and amorphous states was a complication, no persistent ordered structures were observed in the simulations used for the global fitting. The inability to achieve a satisfactory global fit across concentrations is therefore most likely due to non-ideal effects associated with cluster-cluster interactions (where "cluster" encompasses cluster of size 1, i.e. free ions.)

An obvious origin for non-ideal effects is the long-ranged Coulomb interaction among clusters and free ions. The qualitative trend predicted by the Debye-Hückel screening formalism is that increased ionic strength will reduce the activity of charged species but not of neutral species,[134] and so would reduce the apparent equilibrium constants for the formation of neutral products. The trend observed in the simulations, that high concentrations are more favorable to clustering than an ideal model would suggest, runs contrary to this prediction. Even though Debye-Hückel theory is not expected to be quantitatively accurate for the concentrations of concern here, its failure to predict the qualitative trend away from ideal behavior suggests that electrostatics are not the main source of non-ideality. (Electrostatic interactions within a cluster, and between cluster and solvent, are accounted for in the cluster formation free energy and do not contribute to non-ideality.)

A better explanation for the observed systematic error is a non-ideal crowding effect. Crowding in solution can be viewed from two (equivalent) perspectives: as

an effect of excluded volume or of solute effects on solvent chemical potential.[135] Free ions and clusters are associated with tightly bound waters of hydration, some of which are released upon aggregation. In dilute solution, the water chemical potential depends only very weakly on concentration. Free solvent becomes scarcer at high solute concentrations, lowering the chemical potential of water and making the release of a given number of waters an increasingly favorable contribution to the free energy of cluster formation. The effect is a greater association constant for clustering at high concentration, as seen in the data.

From the excluded volume perspective, the release of solvent upon aggregation increases the effective volume available to other ions and clusters. The numbers of closely associated solvent per ion (estimated here as 5.4 for Na^+ and 7.2 for Cl^-, similar to values in the literature;[136–138] see Appendix section D.3.1 for details) are comparable to than the total number of solvents per ion pair (14) in the 3.88 m system, it is not surprising that solvent scarcity will influence aggregation in a non-ideal way.

Our previous report of MTBE cluster formation discussed how non-ideal contributions to cluster free energies from excluded volume effects might be finessed with an appropriate choice of cluster cut-off,[60] through which a change in free volume associated with monomer association or dissociation is balanced by changes in attractive interactions between monomers. Varying cut-offs did not, however, improve the systematic non-ideal effects revealed in the present NaCl systems. This case might be different because the volumes associated with the free ion solvent shells are too large for this balancing to work.

If the systematic errors in the fit are indeed the result of non-ideality, the corresponding free energy surface can be seen as an effective free energy surface averaged over the environments represented in the simulations. With this caveat, the free energy surface obtained from global fit to three sets of simulations with 20, 30 and 45
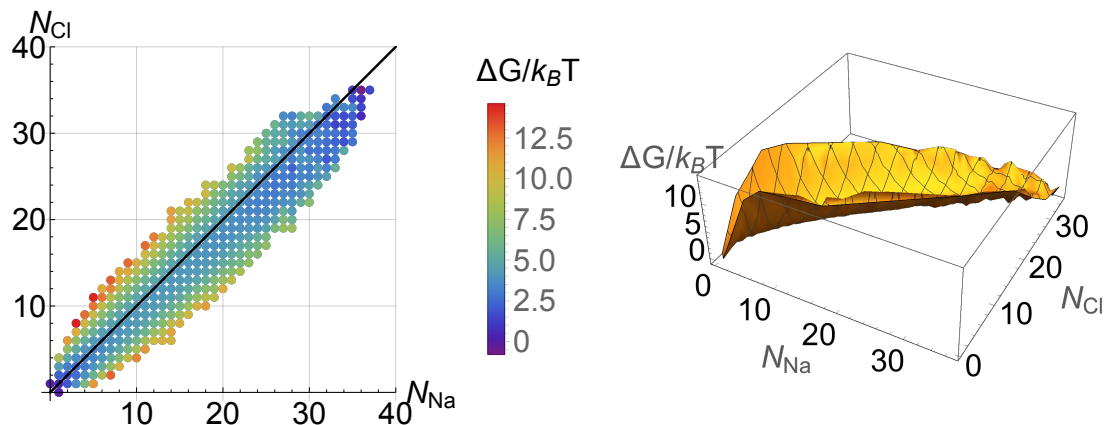
Figure 6.12: Gibbs free energy surface for NaCl cluster aggregation solvated in SPC/E water with a monomer concentration of 0.58 NaCl/nm3 (0.97 m). The top plot shows a contour plot of the free energy resulting from the PEACH method being applied to simulations containing 20, 30, and 45 NaCl pairs at concentrations of 0.56-3.88 m and the bottom plot shows a 3-D representation of the same free energy surface.

NaCl at concentrations of 0.56-3.88 m is presented in Fig. 6.12.

Charge distributions as a function of cluster size, derived from this plot are represented in Fig. 6.13. In contrast to the cluster formation in methanol-containing solvents, this free energy surface favors the incorporation of positive ions into the cluster, especially for larger cluster sizes, and also predicts a wider distribution of charges for any given cluster size. The observed trend that the standard deviation in charge increases with cluster size is not surprising, as a larger cluster allows a greater distance between excess charges and so reduces Coulomb repulsion. An explanation for the preference for formation of positively charged clusters might come from the stability of the cluster and/or the stability of the free ion. Noting that this preference is much weaker or absent in the mixed solvent systems, solvation effects must be playing a role; differences in how the ions pack in the clusters themselves will not address this question, unless these differences are solvent-dependent. Na+ may lose less of its favorable hydration enthalpy upon adding to the surface of a neutral cluster than Cl^-. It is unlikely that the release of waters of hydration is a driving force for the cluster charge preference since Cl^- ions are more hydrated than Na^+ ions but it
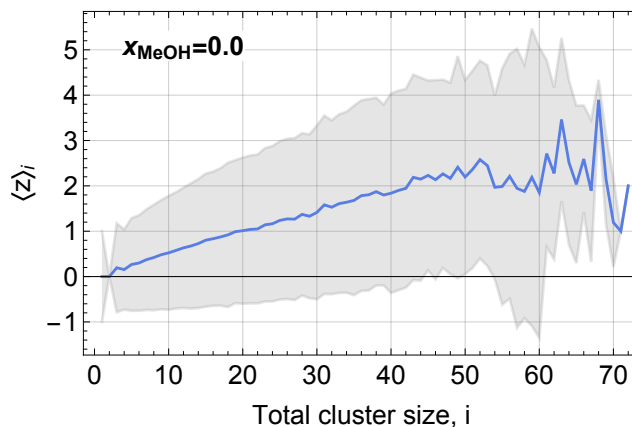
Figure 6.13:   Plot of the average cluster charge $\langle z \rangle_i$ for cross sections of the free energy surface where the total cluster size i is the same. The region shaded in grey indicates the weighted standard deviation in the cluster charge.

may be the case that sodium desolvates more completely than chloride.

The cluster charge trend has been touched on in several other simulations.[106, 139, 140] In developing a polarizable potential for water and ions, Soniat et al. reported a very clear preference for the formation of positively charged clusters with an increasing average charge as the cluster size increased.[140] Lanaro and Patey reported a preference for positively charged Na^+ to be in a more ordered environment than the Cl^- ions for small clusters, a trend that decreased with increasing cluster size.6 In a study by Oh et al. on the nucleation of water and methanol droplets around a Lenard Jones cations and anions of the same size, it was found that the formation of a SPC/E water droplet around the anion was slightly more favorable.[141] While there are size effects that must be taken into account, a lower solvation energy for the anions could lead to the trend in favoring positively charged clusters.

6.3.4   Ordered clusters in water

Aqueous systems of 30 NaCl pairs did not form ordered clusters over 500 ns of simulation even for concentrations up to 7.2 m. Larger systems (45 NaCl pairs) displayed irreversible formation of an ordered cluster at concentrations as low as 4.44 m, with widely varying onset times (between 10 ns and 610 ns, see t_c values in Table
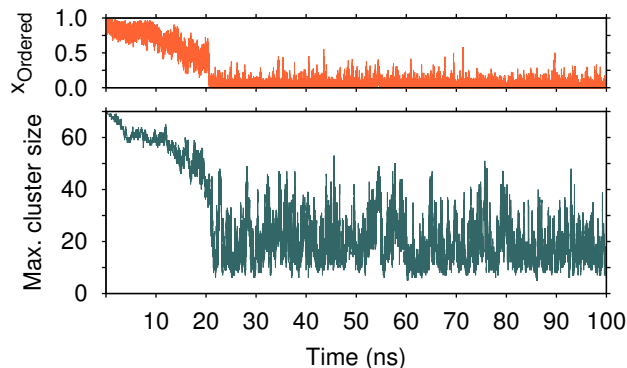
Figure 6.14: Plots of the maximum cluster size in the trajectory (green) and the fraction of ions in that cluster that would be considered ordered (orange), from simulation of 35 NaCl nanocrystallite seeded into a box of 500 SPC/E water.

6.4). In two runs, at 5.00 and 6.66 m, ordering was not observed over 500 ns of simulation. When placed into solvent boxes, ordered clusters of 45 NaCl pairs were stable over 100 ns at concentrations as low as 2.22 m. The concentration of free ions in solution in these cases appears to stabilize at about 0.5 m, which suggests the true solubility of this model NaCl in SPC/E water is far lower than reported previously.[105] An ordered cluster of 35 NaCl pairs was seen to reach this stable plateau initially (Fig. 6.14) with 10 ions entering the solvent in the initial 4 ns and the remaining 60-ion cluster remaining stable and ordered for an additional 10 ns. Over an additional 6 ns, the cluster shrank and lost its crystalline order rapidly upon reaching a total size of about 40 ions. From this we conclude that the threshold size, above which ordered clusters are more stable than disordered clusters in aqueous solution, is between 40 and 90 ions for the current force field. The barrier between the ordered and disordered forms is apparently too large for a more precise conclusion from the current data.

This uncertainty is not surprising since sampling rare events like nucleation using unbiased MD can require extremely long simulation times to see a single nucleation event, especially at low supersaturations.[119, 142] Even in situations where it is possible to see nucleation events within the simulation time, the time required to

observe a nucleation event is concentration and system size dependent. Chakraborty and Patey compared systems with 4000 and 500 NaCl ion pairs at a concentration of 3.97 m and found that the nucleation onset times were $\leq$ 10 ns and $\approx$ 90 ns respectively.[104]

6.3.5 Classical nucleation theory and amorphous NaCl clusters

The 1-D effective cluster free energy surfaces (for even-numbered, neutral clusters) for all four solvent systems were fit to a functional form taken from classical nucleation theory (CNT),[143]

$$\Delta G_i = \gamma A_{\text{geom}} \left(i^{2/3} - 1 \right) - (i - 1) \, k_B T \ln \left(\frac{c}{c_{\text{sat}}} \right) \tag{6.6}$$

where γ is the interfacial surface tension, A_{geom} is a geometric area constant for the cluster, and c is the monomer concentration. The values are shown in Table 6.7. Cluster free energies for all four systems, with corresponding fits, are shown in Fig. 6.15 under conditions of equivalent supersaturation, specifically at a degree of supersaturation $s = 1.5$, i.e. a free monomer concentration of 1.5 c_{sat}. The c_{sat} values, although derived from a simple model and from primarily amorphous rather than crystalline structures, are conspicuously low compared with experimental NaCl solubilities for the pure solvents (6.17 m for H_2O and 0.214 m for MeOH).[144] The results are therefore to be taken for the phenomena and trends that they illuminate rather than for accurate predictions about NaCl in these solvents.

Table 6.7: Values from the fits to neutral cluster free energies using the classical nucleation theory functional form (Eqn. 6.6).

Solvent (x_{MeOH})	$\gamma A_{\text{geom}}/k_B T$	c_{sat}/m
1.0	3.63	3.7×10^{-4}
0.5	2.42	0.044
0.4	1.97	0.081
0.0	1.46	0.70
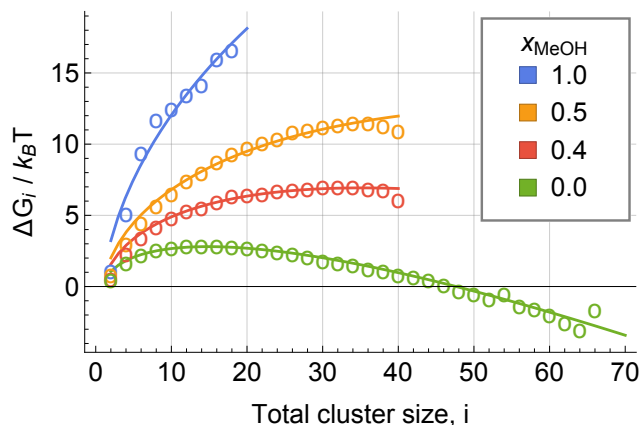
Figure 6.15: Plot of neutral clusters ΔG_i calculated using Eqn. 6.5 (circles) with the PEACH analysis results for NaCl cluster simulations under different solvent conditions, and the results of the fit to the CNT expression in Eqn. 6.6 (line). The values for the CNT fit are given in Table 6.7. Both calculations assume a monomer concentration of $c_{Na^+} = c_{Cl^-} = 1.5\,c_{\text{sat}}$.

Except for the pure methanol case and the largest cluster sizes in the solvent mixtures, these curves describe primarily or exclusively amorphous clusters. They can therefore be interpreted as depicting the nucleation barrier and critical nucleus size to form a metastable amorphous phase as the endpoint to unrestricted amorphous cluster growth. We would identify the result of this growth with the "liquid droplet" invoked in the 2-step nucleation mechanism.[121]

As discussed above, non-ideal effects complicate the interpretation of the curve and fitting parameters given for pure water, but they should be adequate for this qualitative discussion; it is physically reasonable that increasing solvent quality would lower the effective surface tension. Given that the crystalline structure becomes the more stable structure at large cluster sizes, these curves will be crossed by a more steeply descending curve that describes the free energy of crystalline clusters. Changing supersaturation will shift the position of the maximum along the amorphous phase curve (the critical nucleus size), and so can influence whether the crossover occurs before or after this critical size is reached. If the crossover occurs at a higher cluster size than the critical nucleus size for the amorphous phase, a two-step nucleation process

is much more likely. We identify the crossover size near 40 total ions for the two mixed-solvent cases and at a larger size (between 40-90 total ions) for pure water. Accordingly, at $s = 1.5$ we would expect crystal nucleation in pure water to proceed via a 2-step mechanism, since the nucleation barrier to forming a dense amorphous phase occurs at cluster sizes where the ordered state is not yet stable. In the mixed solvents at the same supersaturation, the growing amorphous cluster would have the opportunity to cross over to an ordered state near the critical nucleus size, and might bypass the formation of a larger amorphous cluster or "liquid droplet". Greater confidence in this type of prediction would require more study of the rates of growth of the amorphous cluster relative to the (size-dependent) rate of change of the amorphous cluster to a crystalline one.

The fit of Eqn. 6.6 to the free energies of NaCl clusters in pure methanol is shown only for completeness. It does not account for "magic number" effects discussed above, and contains significant contributions from both amorphous and ordered structures, so does not represent the pathway to a dense amorphous phase. The 2-step mechanism would not be a possibility under any supersaturation conditions for this system, because the ordered structure becomes stable at small cluster sizes where the barrier to crossing over from amorphous to ordered is negligible; growth and ordering can follow an smooth path across the global free energy surface.

6.4   Conclusions

Although the three solvent conditions presented challenges to the determination of a global free energy of association for NaCl cluster formation, the forms these challenges took illustrate contrasting free energy landscapes. In pure methanol, crystalline structures became stable even at low cluster sizes (10 NaCl pairs); furthermore, dissociation was slow relative to structural transformations. The implication for the effectiveness of the PEACH method is that sampling across different cluster sizes was

poor; hybrid methods that combine PEACH with enhanced sampling approaches are needed to treat cases like this. The implication for crystal nucleation is that we expect a growing cluster to cross smoothly over from amorphous to ordered structures at small cluster sizes, so that no two-step process would be observed.

In 40% or 50% aqueous methanol, ordered structures become more stable than amorphous structures once clusters grow to about 20 NaCl pairs. Transitions between ordered and disordered forms for clusters near this transitional size are significantly slower than exchange of ions with the surroundings but are fast enough to observe occurring reversibly in simulations. It is conceivable that a growing cluster, at high enough supersaturation, could grow larger than this crossover size before making the transition. To draw firm conclusions about the likelihood of a two-state mechanism, more information would be needed on how the rate of the ordering transition depends on the size of the amorphous cluster size. In 40% methanol, fast sampling across wide ranges of cluster sizes allowed for excellent fits to the data using PEACH. In 50% methanol, slow interconversion between ordered and disordered states of the largest clusters may be the cause of somewhat poorer fits.

In pure water, the crossover to favoring the ordered cluster appears to be shifted to yet larger cluster sizes, apparently in the 20-40 NaCl pair range. It is hard to be more precise about the crossover because transitions were so slow that they could only be seen under conditions where they were irreversible. The slow rate of dehydration involved in conversion from a partially hydrated amorphous cluster to a compact crystal is a likely contribution to this slowness. For this reason, it is easy to imagine conditions favoring two-step nucleation:[105, 121] an initially formed amorphous cluster grows beyond the crossover size (even into a metastable amorphous phase) without undergoing ordering, and an ordered crystal nucleates from within this high-concentration, disordered environment. The systematic deviation between simulation data and PEACH best-fit models (which rely on the assumption of non-interacting

clusters) suggests that non-ideal effects are an important factor favoring the growth of amorphous clusters at high concentrations. We propose that the release of solvent into a crowded solution environment is the source of this non-ideality and a driving force for amorphous cluster formation in this aqueous NaCl model. Further elaboration of the PEACH method to incorporate non-ideal effects explicitly is an important goal for the future.

6.5 Acknowledgments

Appendix A   Derivations

A.1   Ensemble averages

In the following section we derive a number of useful ensemble averages related the the cluster size histograms that are and could be useful. In particular we want to use the relationships between the generating function (the grand canonical partition function Ξ), the canonical partition functions Q, the partition functions for individual clusters q_i, and ensemble averages.

A.1.1   Cluster co-frequency $\langle m_j m_s \rangle_N$

Similarly to the derivation of the cluster size frequency in section 4.3.3, we begin by defining the grand canonical ensemble average in the cluster size co-frequency as a sum over the canonical ensemble cluster size co-frequencies weighted by the probability of the system having N particles and a given activity λ, $Q(N,V,T)\lambda^N/\Xi(\lambda,V,T)$ as related by a probability $P(N,\lambda)$ weighted average over all the possible system sizes N.

$$
\begin{aligned}
\langle m_j m_s \rangle_\lambda &= \sum_{N=0}^{\infty} P\left(N,\lambda\right) \langle m_j m_s \rangle_N \\
&= \sum_{N=0}^{\infty} \frac{Q\left(N,V,T\right)\lambda^N}{\Xi} \langle m_j m_s \rangle_N
\end{aligned}
\tag{A.1}
$$

Expanding the expression for the canonical ensemble average of the co-frequency $\langle m_j m_s \rangle_N$ to a sum over all partitions of N, $P(N)$, where an s-mer and j-mer are

present,

$$
\begin{aligned}
\langle m_j m_s \rangle_\lambda \, \Xi \;&=\; \sum_{N=0}^{\infty} \frac{Q\left(N,V,T\right)\lambda^N}{Q\left(N,V,T\right)} \left(\sum_{k=0}^{P(N)} m_{j,k} m_{s,k} \prod_{i=1}^{N} \frac{q^{m_{i,k}}}{m_{i,k}!} \right) \\
&=\; \left(\sum_{m_j=0}^{\infty} m_j \frac{(q_j \lambda^j)^{m_j}}{m_j!} \right) \left(\sum_{m_s=0}^{\infty} m_s \frac{(q_s \lambda^s)^{m_s}}{m_s!} \right) \left(\prod_{\substack{i=1 \\ i\neq j,s}}^{\infty} \sum_{m_i=0}^{\infty} \frac{(q_i \lambda^i)^{m_i}}{m_i!} \right) \\
&=\; q_j q_s \lambda^{j+s} \left(\sum_{m_j=1}^{\infty} \frac{(q_j \lambda^j)^{m_j-1}}{(m_j-1)!} \right) \left(\sum_{m_s=1}^{\infty} \frac{(q_s \lambda^s)^{m_s-1}}{(m_s-1)!} \right) \left(\prod_{\substack{i=1 \\ i\neq j,s}}^{\infty} \sum_{m_i=0}^{\infty} \frac{(q_i \lambda^i)^{m_i}}{m_i!} \right) \\
&=\; q_j q_s \lambda^{j+s} \left(\prod_{i=1}^{\infty} \sum_{m_i=0}^{\infty} \frac{(q_i \lambda^i)^{m_i}}{m_i!} \right)
\end{aligned}
$$

the expression is rearranged to extract a factor of $q_j q_s \lambda^{j+s}$ and return to the grand canonical function Ξ such that the following hold true:

$$
\boxed{\langle m_j m_s \rangle_\lambda \, \Xi = q_j q_s \lambda^{j+s} \Xi} \tag{A.2}
$$

Using the relationship expressed in eqn. A.1 for the left side of eqn. A.2 and expanding the expression for Ξ on the right side of eqn. 4.34 into a polynomial in powers of λ we get the following relationship:

$$
\sum_{N=0}^{\infty} Q\left(N,V,T\right) \lambda^N \langle m_j m_s \rangle_N = q_j q_s \sum_{N=0}^{\infty} Q\left(N-(s+j),V,T\right) \lambda^N \tag{A.3}
$$

By isolating the coefficients for the Nth power of λ we can get an expression for the canonical ensemble average of the co-frequency in j-mers and s-mers as a function of the partition functions q and the canonical ensemble partitions functions Q:

$$
\boxed{\langle m_j m_s \rangle_N = q_j q_s \frac{Q\left(N-j-s,V,T\right)}{Q\left(N,V,T\right)}} \tag{A.4}
$$

A.1.2 Frequency of seeing two of the same cluster size simultaneously $\left\langle m_j^2 - m_j \right\rangle_N$

The derivation of cluster size co-frequency here is a variant of the previous derivation except the assumption in this case is that the cluster sizes are the same $(s = j)$.

We begin by stating the relationship between the cluster size co-frequencies in the canonical ensemble $\langle m_j(m_j - 1)\rangle_N$ and grand canonical ensemble $\langle m_j(m_j - 1)\rangle_\lambda$.

$$
\begin{aligned}
\left\langle m_j^2 - m_j \right\rangle_\lambda &= \sum_{N=0}^{\infty} P(N, \lambda) \left\langle m_j^2 - m_j \right\rangle_N \\
&= \sum_{N=0}^{\infty} \frac{Q(N, V, T) \lambda^N}{\Xi(\lambda, V, T)} \left\langle m_j^2 - m_j \right\rangle_N
\end{aligned}
\tag{A.5}
$$

Expanding the expression for the canonical ensemble cluster size variance $\left\langle m_j^2 - m_j \right\rangle_N$ to a sum over all the partitions of N, $P(N)$, we are then able to rearrange the expression until a factor of $(q_j \lambda_j)^2$ can be extracted.

$$
\begin{aligned}
\left\langle m_j^2 - m_j \right\rangle_\lambda \Xi &= \sum_{N=0}^{\infty} \frac{Q(N, V, T) \lambda^N}{Q(N, V, T)} \left[\sum_{k=1}^{P(N)} \left(m_{j,k}^2 - m_{j,k} \right) \prod_{i=1}^{N} \frac{q_i^{m_{i,k}}}{m_{i,k}!} \right] \\
&= \left(\sum_{m_j=0}^{\infty} \left(m_j^2 - m_j \right) \frac{(q_j \lambda^j)^{m_j}}{m_j!} \right) \left(\prod_{i=1, i \neq j}^{\infty} \left[\sum_{m_i=0}^{\infty} \frac{(q_i \lambda^i)^{m_i}}{m_i!} \right] \right) \\
&= \left(q_j \lambda^j \right)^2 \left(\sum_{m_j=2}^{\infty} \frac{(q_j \lambda^j)^{m_j-2}}{(m_j-2)!} \right) \left(\prod_{i=1, i \neq j}^{\infty} \left[\sum_{m_i=0}^{\infty} \frac{(q_i \lambda^i)^{m_i}}{m_i!} \right] \right) \\
&= \left(q_j \lambda^j \right)^2 \left(\prod_{i=1}^{\infty} \left[\sum_{m_i=0}^{\infty} \frac{(q_i \lambda^i)^{m_i}}{m_i!} \right] \right)
\end{aligned}
$$

In this last expression we can see that we have returned to the grand canonical partition function, Ξ, such that the relationship between the grand canonical ensemble average of the cluster size variance can be expressed as:

$$
\boxed{\left\langle m_j^2 - m_j \right\rangle_\lambda \Xi = \left(q_j \lambda^j \right)^2 \Xi}
\tag{A.6}
$$

Substituting eqn. A.5 for the left side of the eqn. A.6 and expanding the right side of eqn. A.6 as a polynomial in powers of λ, allows us to reach an expression that relates two polynomials of λ as:

$$
\sum_{N=0}^{\infty} \left\langle m_j^2 - m_j \right\rangle_N Q(N, V, T) \lambda^N = q_j^2 \sum_{N=0}^{\infty} Q(N - 2j, V, T) \lambda^N
\tag{A.7}
$$

Upon isolating the coefficients of the polynomials with the same power, we are able to find the functional form for the canonical ensemble average of the cluster size variance

for a j-mer in terms of the partition function q_j and the canonical ensemble partition functions Q:

$$\left\langle m_j^2 - m_j \right\rangle_N = q_j^2 \frac{Q\left(N - 2j, V, T\right)}{Q\left(N, V, T\right)} \tag{A.8}$$

A.2 Defining the exact equilibrium association constant as a function of co-frequency

We begin by defining the exact equilibrium association constant in terms of the standard state partition functions q_j° for the association/dissociation of a j-mer and an s-mer. Using the volume scaling $q_j = q_j^\circ V/V^\circ$ between the standard state partition function and a partition function for a given volume V,

$$K_{\text{actual},j+s} = \frac{q_{j+s}^\circ}{q_j^\circ q_s^\circ} = \frac{\left(V/V^\circ\right)}{\left(V/V^\circ\right)^2} \frac{q_{s+j}}{q_j q_s} \tag{A.9}$$

We apply a mathematical trick of multiplying Eqn. A.9 by 1 (i.e. the total partition functions divided by themselves). Upon rearranging the equation so that we can see the equivalence to Eqn. 4.43 and Eqn. A.4,

$$K_{\text{actual},j+s} = \frac{\left(V/V^\circ\right)}{\left(V/V^\circ\right)^2} \frac{q_{s+j} Q\left(N - (s+j), V, T\right)}{Q\left(N, V, T\right)} \frac{Q\left(N, V, T\right)}{q_j q_s Q\left(N - (s+j), V, T\right)}$$

we get the following:

$$K_{\text{actual},j+s} = \frac{\left\langle m_{j+s} \right\rangle_N}{\left\langle m_j m_s \right\rangle_N} \frac{\left(V/V^\circ\right)}{\left(V/V^\circ\right)^2} \tag{A.10}$$

This naturally raises a question concerning the global fitting method (section 4.4): why do we not use this to calculate the first guess for the equilibrium association constants rather than the law of mass action? The answer is that we could use the co-frequencies to generate the initial guess for the equilibrium constants but the generation of the data we use to improve the fit (in practice the cluster size frequency $\left\langle n_s \right\rangle_N$) would still involve choosing one reaction pathway for each cluster size for a 1-1 mapping in the Eqn. 4.58 and thus the subsequent calculation of the adjustment factor (Eqn. 4.60-4.59).

A.2.1 Deriving $\ln\left[K_{\text{actual},j}/K_{\text{LoMA},j}\right]$

In the search from a correction to the law of mass action based association constant for j monomers associating into a j-mer $\left(j\,[\,1\,]\,\underset{\text{dissoc}}{\overset{\text{assoc}}{\rightleftharpoons}}\,[\,j\,]\right)$, the question arose as to whether we could exactly calculate the difference between it and the exact association constant. It is understood qualitatively that the law of mass action only works well in the limit of bulk solutions where the formation of a product does not significantly perturb the concentration of reactants or products. As the system size increases, the ability of the law of mass action to reproduce the equilibrium association constant improves.

$$K_{\text{actual},j} = \frac{q_j^\circ}{(q_1^\circ)^j} = \frac{q_j}{(q_1)^j}\left(\frac{V}{V^\circ}\right)^{j-1} \tag{A.11}$$

$$K_{\text{LoMA},j} = \frac{c_j}{(c_1)^j} = \frac{\langle n_j\rangle}{\langle n_1\rangle^j}\left(\frac{V}{V^\circ}\right)^{j-1} \tag{A.12}$$

To quantify the convergence between these two functional forms, they must both be reframed in terms of the partition functions. Using the definition for the average cluster frequency $\langle n_j\rangle$ from Eqn. 4.43, the law of mass action equilibrium association constant in Eqn. A.12 becomes:

$$K_{\text{LoMA},j} = \frac{q_j}{(q_1)^j}\left(\frac{V}{V^\circ}\right)^{j-1}\frac{Q\left(N-j,V,T\right)}{Q\left(N,V,T\right)}\left(\frac{Q\left(N,V,T\right)}{Q\left(N-1,V,T\right)}\right)^j \tag{A.13}$$

The ratio of the exact association constants (Eqn. A.11) and the law of mass action (Eqn. A.13) is:

$$\boxed{\ln\left[\frac{K_{\text{actual},j}}{K_{\text{LoMA},j}}\right] = j\ln\left[\left(\frac{Q\left(N-1,V,T\right)}{Q\left(N,V,T\right)}\right)\right] - \ln\left[\frac{Q\left(N-j,V,T\right)}{Q\left(N,V,T\right)}\right]} \tag{A.14}$$

and reaches 0 as the law of mass action converges on the true equilibrium association constant in the limit of $N\to\infty$.

Deriving $\ln\left[K_{\text{actual},j}/K_{\text{LoMA},j}\right]$ using co-frequencies

Another variant of the derivation is based on the relationship between an exact equilibrium constant $K_{\text{actual},j'}$ and the cluster co-frequencies as defined in Eqn. A.10

in section A.2. Suppose we define an aggregation pathway for the formation of a j-mer as the association of a monomer and a $(j-1)$-mer.

$$[\,1\,] + [\,j-1\,] \underset{\text{dissoc}}{\overset{\text{assoc}}{\rightleftharpoons}} [\,j\,]$$

$$K_{\text{actual},j'} = \frac{\langle n_j \rangle}{\langle n_1 n_{j-1} \rangle} \frac{(V/V^\circ)}{(V/V^\circ)^2} = \frac{q_j^\circ}{q_1^\circ\, q_{j-1}^\circ} = \frac{q_j}{q_1\, q_{j-1}} \frac{(V/V^\circ)}{(V/V^\circ)^2} \tag{A.15}$$

$$K_{\text{LoMA},j'} = \frac{\langle n_j \rangle}{\langle n_1 \rangle \langle n_{j-1} \rangle} \frac{(V/V^\circ)}{(V/V^\circ)^2} \tag{A.16}$$

We denote the equilibrium association constants with a j' since this is a different pathway to the reaction described in the previous section.

To make the connection to an equivalent expression of Eqn. A.14, one must recognize that the reaction for the equilibrium constant there is for the reaction pathway:

$$j\,[\,1\,] \underset{\text{dissoc}}{\overset{\text{assoc}}{\rightleftharpoons}} [\,j\,]$$
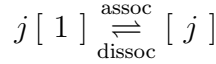
And we can get to this though a progression of reactions such that,

$$K_{\text{actual},j} = \prod_{n=0}^{n=j-2} K_{\text{actual},(j-n)'} \tag{A.17}$$

This definition results in an equilibrium association constant based on the law of mass action that is no different from the one reported in Eqn. A.12. Likewise, if we use the definition for $K_{\text{actual},j'}$ based on the partition functions q° and substitute it into Eqn. A.17 gets back to the same expression for the equilibrium constant given in Eqn. A.11. The equilibrium constant as a function of the co-frequencies can be expressed as:

$$
\begin{aligned}
K_{\text{actual},j} &= \frac{\langle n_j \rangle}{\langle n_1 n_{j-1} \rangle} \frac{\langle n_{j-1} \rangle}{\langle n_1 n_{j-2} \rangle} \frac{\langle n_{j-2} \rangle}{\langle n_1 n_{j-3} \rangle} \cdots \\
&\qquad \frac{\langle n_{j-(j-3)} \rangle}{\langle n_1 n_{j-(j-3+1)} \rangle} \frac{\langle n_{j-(j-2)} \rangle}{\langle n_1 (n_1 - 1) \rangle} \left(\frac{(V/V^\circ)}{(V/V^\circ)^2} \right)^{j-1} \tag{A.18} \\
&= \langle n_j \rangle \frac{\langle n_{j-1} \rangle}{\langle n_1 n_{j-1} \rangle} \frac{\langle n_{j-2} \rangle}{\langle n_1 n_{j-2} \rangle} \cdots \frac{\langle n_2 \rangle}{\langle n_1 n_2 \rangle} \frac{1}{\langle n_1 (n_1 - 1) \rangle} \left(\frac{(V/V^\circ)}{(V/V^\circ)^2} \right)^{j-1} \\
&= \left[\frac{\langle n_j \rangle}{\langle n_1 (n_1 - 1) \rangle} \prod_{m=2}^{j-1} \frac{\langle n_m \rangle}{\langle n_1 n_m \rangle} \right] \left(\frac{(V/V^\circ)}{(V/V^\circ)^2} \right)^{j-1} \tag{A.19}
\end{aligned}
$$

The ratio of equilibrium association constants can therefore be calculated as a function of co-frequencies:

$$
\begin{aligned}
\ln\left[\frac{K_{\text{actual},j}}{K_{\text{LoMA},j}}\right] &= \ln\left[\frac{\langle n_1\rangle^j}{\langle n_j\rangle}\frac{\langle n_j\rangle}{\langle n_1(n_1-1)\rangle}\prod_{m=2}^{j-1}\frac{\langle n_m\rangle}{\langle n_1 n_m\rangle}\right] \\
&= \ln\left[\frac{\langle n_1\rangle^j}{\langle n_1(n_1-1)\rangle}\prod_{m=2}^{j-1}\frac{\langle n_m\rangle}{\langle n_1 n_m\rangle}\right] \qquad\text{(A.20)}
\end{aligned}
$$

Appendix B Coarse-grained molecular simulations of the melting kinetics of small unilamellar vesicles

B.1 Truncated icosahedron assembly

It is possible to calculate internal dihedral angles between faces that share an edge for a truncated icosahedron through application of geometric identities and the constraints that:

1. Every vertex of a truncated icosahedron is the product of an intersection of a pentagonal face and two hexagonal faces.

2. The length of all pentagon and hexagon edges are equal.

3. The edge length l_{edge} is directly correlated to the distance, r_{face}, between the center of the truncated icosahedron and any face on the shape.

The results of this can be seen in sections B.1.1 and B.1.2.

The displacement of the lipid slabs if they were a two-dimensional surface is given by Eqn. B.1 for a given edge length l_{edge}. The slabs however occupy a three-dimensional space and their displacement must account for the thickness of the bilayer l_{thick} ($\sim$4 nm). If we assume that the distance r_{face} is defined as the distance to the center of mass of the pentagonal or hexagonal slab, a good first approximation to the appropriate slab displacement is given by Eqn. B.2. In practice, we generated a series of truncated icosahedron with different displacements, chose the displacement that provided sufficient overlap of lipids in both leaflets at edges. That configuration was then subjected to a script that identifies lipid pairs that have a minimum separation

less than 0.45 nm and removes one of the lipids.

$$r_{\text{face}}\left(l_{\text{edge}}\right) = l_{\text{edge}}\left(\frac{9}{2}\sqrt{\frac{17+6\sqrt{5}}{109}}\right) \tag{B.1}$$

$$r_{\text{slab}}\left(l_{\text{edge}}\right) = r_{\text{face}}\left(l_{\text{edge}}\right) + l_{\text{thick}}/2 \tag{B.2}$$

B.1.1 Hexagonal faces

The dihedral angle between a pentagon face and a neighboring hexagon face θ_{ph} is 37.4° (Eqn. B.3) and the dihedral angle between two hexagon faces θ_{hh} is 41.8° (Eqn. B.4).

$$\theta_{\text{ph}} = \frac{180}{\pi}\left(\pi - \sin^{-1}\left[\frac{2\sqrt{65-22\sqrt{5}}}{15-3\sqrt{5}}\right] - \sin^{-1}\left[\frac{3+\sqrt{5}}{6}\right]\right) \tag{B.3}$$

$$\theta_{\text{hh}} = \frac{180}{\pi}\left(\pi - 2\sin^{-1}\left[\sqrt{\frac{3+\sqrt{5}}{6}}\right]\right) \tag{B.4}$$

$$\theta_{\text{rot}} = \frac{2\pi}{5}\left(\frac{180}{\pi}\right) \tag{B.5}$$

Starting with a hexagonal slab, Fig. B.1 shows that a strip of the truncated icosahedron hexagons can be constructed by rotating a hexagonal slab displaced from the origin by a distance r_{slab} by the angles θ_{ph} and θ_{hh}. Rotating the whole strip by the angle θ_{rot} of 72.0° about a $\hat{\text{C}}_5$ rotational axis of the truncated icosahedron four times gives rise to the hexagon-only structure in the center of the top row of Fig. B.3.[1]

B.1.2 Pentagonal faces

The angle between two pentagons separated by the edge of two intersecting hexagons θ_{pp} is 62.2° (Eqn. B.6).

The first pentagonal slab of lipids is displaced from the origin by a distance r_{slab}. The next pentagon is generated by rotating the first by 36.0° about the radial vector that goes from the origin of the truncated icosahedron to the middle of the pentagon

---

[1]A variation of this structure with 12 intentional pores (the pentagonal faces) to serves as defect sinks was used in early calculations but the resulting structure was slow to close.
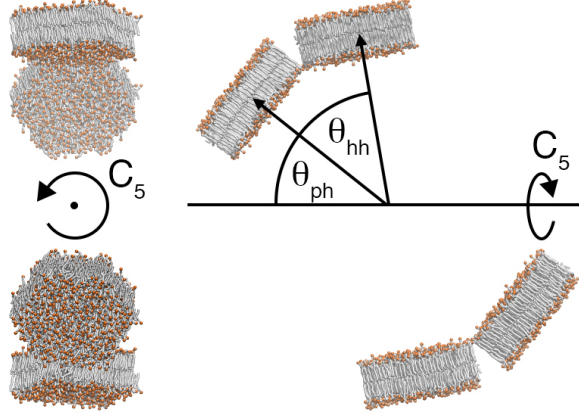
Figure B.1: The diagram above illustrates the relevant angles used in the construction of a strip of four hexagonal slabs. The angle between two hexagonal faces θ_{hh} and the angle between a hexagonal and pentagonal face θ_{ph} are marked on the right side image relative to a $\hat{C}_5$ rotational axis. Four subsequent rotations of the strip around that axis are then used to place the rest of the hexagons as shown in the left side image of Fig. B.3.

slab (the $\hat{C}_5$ rotational axis) and then rotating by θ_{pp} towards the $\hat{C}_2$ rotational axis which is marked in Fig. B.2. This pentagon is then subsequently rotated about the $\hat{C}_5$ rotational axis by θ_{rot} four more times so that the hemisphere in Fig. B.2 is the result. This structure is then rotate by 180.0° around the axis perpendicular to the $\hat{C}_2$ and $\hat{C}_5$ axes to get the pentagon-only structure seen in the first image in the first row of Fig. B.3.

$$
\begin{aligned}
\theta_{pp} &= \frac{180}{\pi} \left(\cos^{-1}\left[\frac{36419 + 5296\sqrt{5}}{52369} \right] + \right. \\
&\qquad \left. 2\cos^{-1}\left[\frac{-9\left(55 + 13\sqrt{5}\right)}{-5 + \sqrt{5}} \sqrt{\frac{5 - \sqrt{5}}{122149 + 49863\sqrt{5}}} \right] \right)
\end{aligned}
\tag{B.6}
$$

The pentagon-only truncated icosahedron (Fig. B.3, left) and hexagon-only truncated icosahedron (Fig. B.3, center) are combined (Fig. B.3, right), and a bad contact distance of 0.45 nm is implemented. Any pair of lipid within 0.45 nm of one another are flagged and one of the lipids is removed. The result of the combination and effect of removing bad contacts can be seen in the bottom row of snapshot in Fig. B.3.
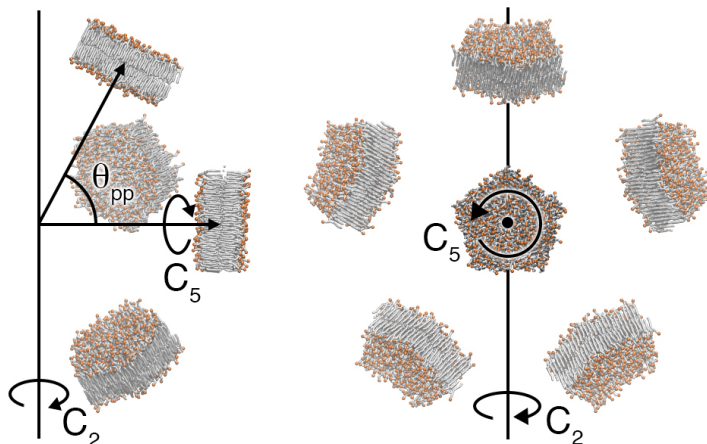
Figure B.2: The diagram illustrates the the relevant angles used in the construction of the first half of the pentagons used to construct a truncated icosahedron. This half is constructed by first rotating the middle pentagon out towards the $\hat{C}_2$ axis by θ_{pp}. This pentagon is then rotated around the central pentagon by θ_{rot}. The half of the pentagons is then rotated around the $\hat{C}_2$ by $180.0°$.

B.2 Supplementary figures

B.2.1 Topological evolution

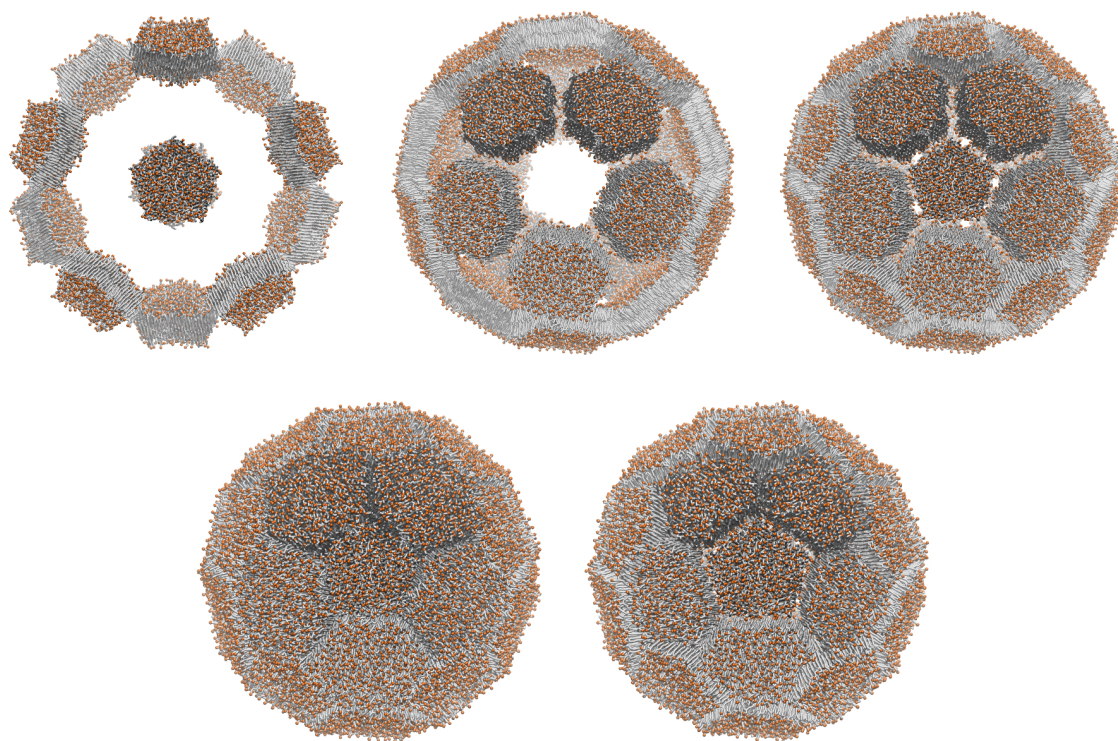Figure B.3: The first row of images are a 'blown up' schematic of the vesicle assembly. From right to left we have the pentagonal faces of a truncated icosahedron constructed from gel phase slabs of DPPC lipids, the hexagonal faces of a truncated icosahedron constructed likewise from gel phase DPPC lipids, and the result of concatenating the two sets of lipids to form the full truncated icosahedron. The second row shows what happens when a shorter displacement is used. The first vesicle is one with bad contacts between lipids and the second is the same vesicle after lipids within 0.45 nm of another lipid have been removed.

	0.0 ns	10.0 ns	30.0 ns	100.0 ns	250.0 ns	500.0 ns
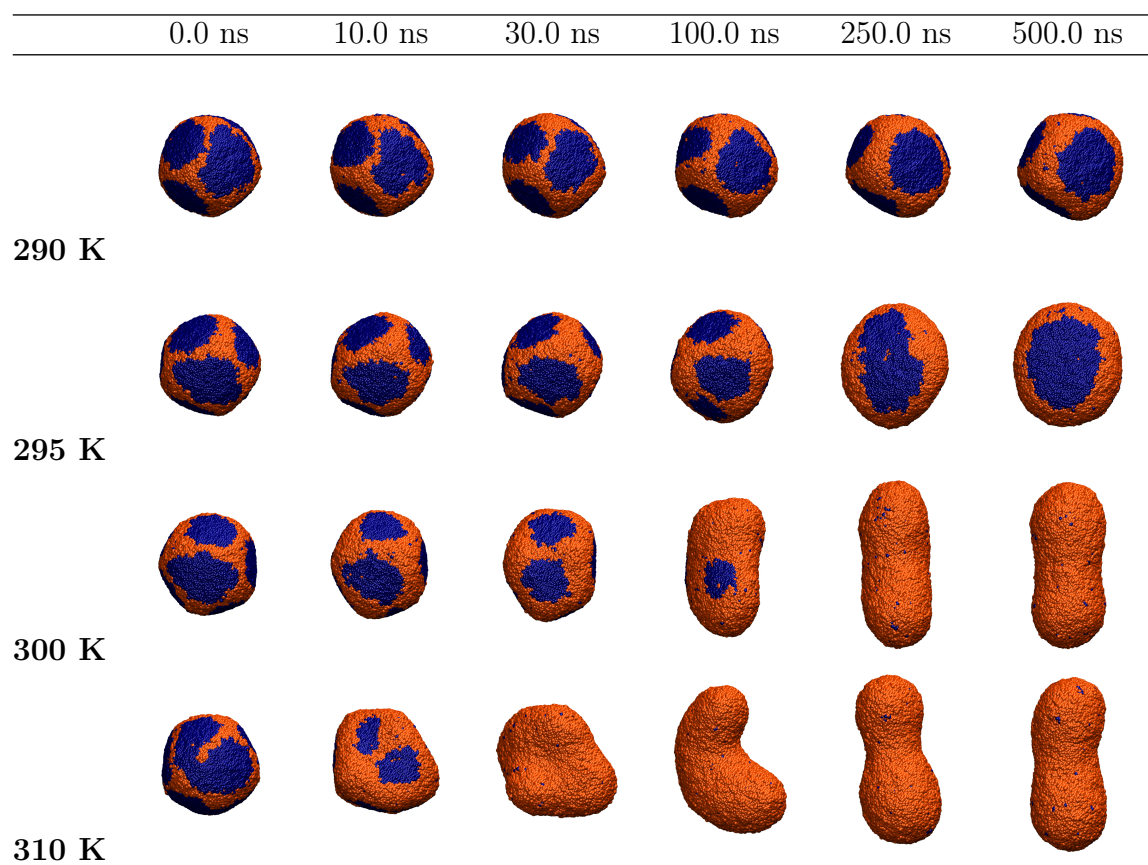
290 K

295 K

300 K

310 K

Figure B.4: Melting progression for the 33 nm vesicle melting at 290 K, 295 K, 300 K and 310 K.

	0.0 ns	10.0 ns	30.0 ns	100.0 ns	250.0 ns	500.0 ns
290 K						
295 K						
297 K						
300 K						
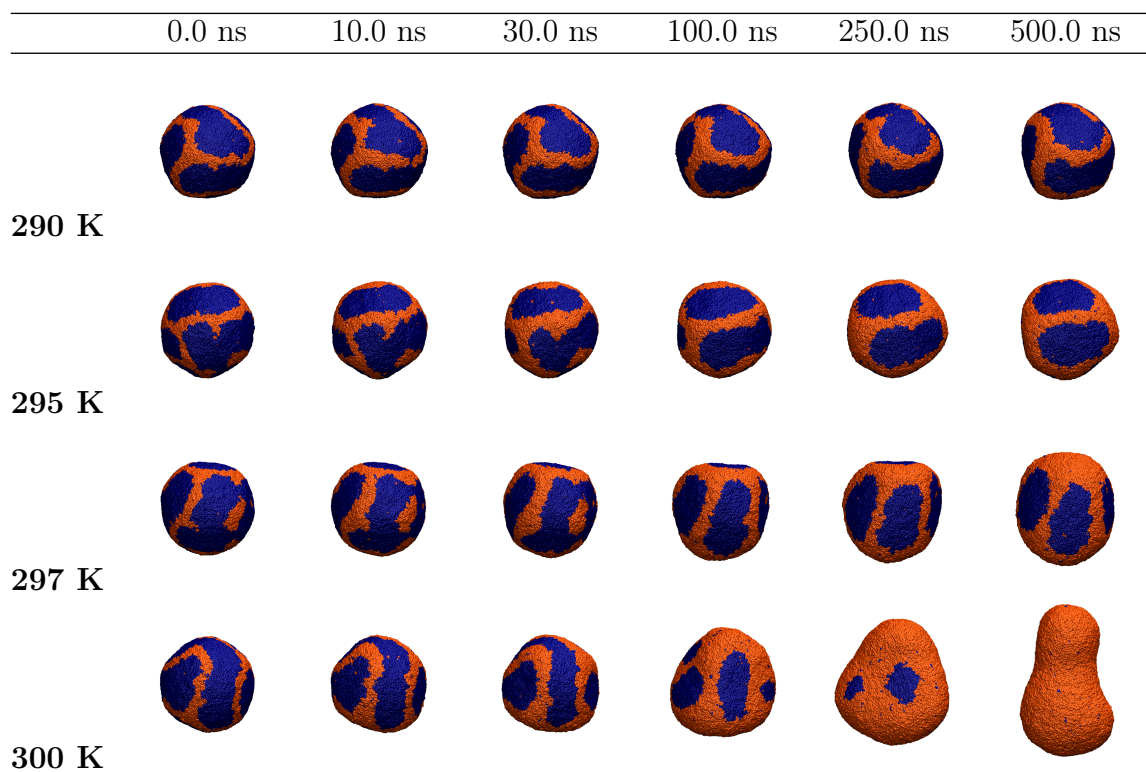
Figure B.5: Melting progression for the 50 nm vesicle at 290 K, 295 K, 297 K, and 300 K.

θ_{y}	0.0 ns	50.0 ns	140.0 ns	268.5 ns	359.0 ns	415.0 ns	500.0 ns
0°							
90°							
180°							

Figure B.6: Shape of the 50 nm melting at 300 K over a 500 ns trajectory. The rows show the vesicle rotated about the vertical axis of the vesicle by 90°.

Bridging domain A melting at 290 K.
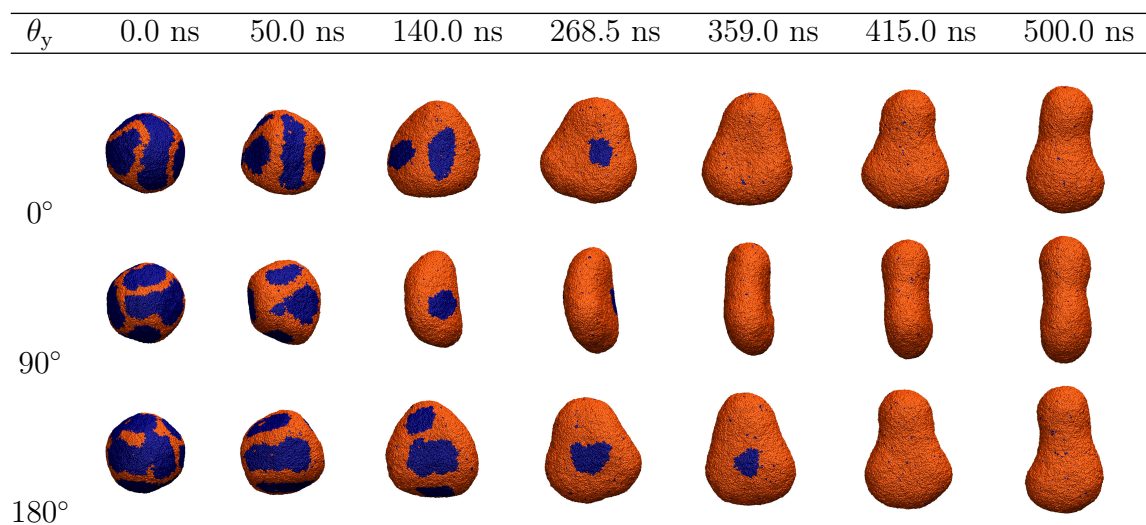
0 ns	175.0 ns	188.0 ns	500.0 ns
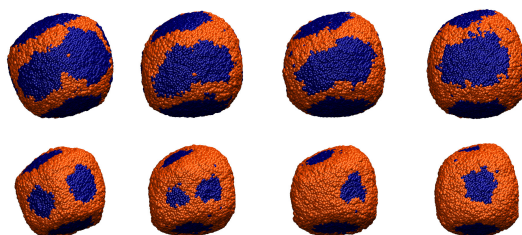
Bridging domain A melting at 295 K.

0 ns	16.5 ns	73.5 ns	500.0 ns

Figure B.7: Snapshots of one of the bridging domains melting in the inner and outer leaflets of vesicles with diameters of 33 nm.

Bridging domain B melting at 290 K.

0 ns	171.0 ns	413.0 ns	500.0 ns

Bridging domain B melting at 295 K.

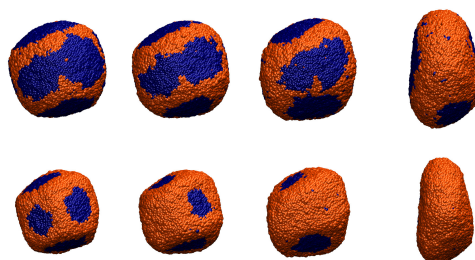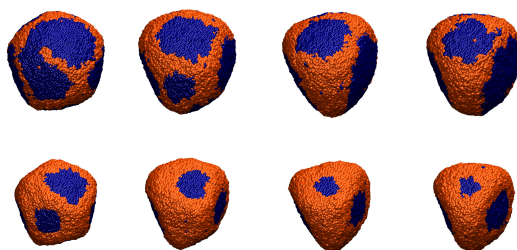| 0 ns	46.0 ns	138.0 ns	500.0 ns

Figure B.8: Snapshots of a second bridging domains melting in the inner and outer leaflets of vesicles with diameters of 33 nm.

Bridging domain C melting at 295 K.

| 0 ns | 73.0 ns | 180.0 ns | 500.0 ns |

Bridge domain C melting at 297 K.

| 0 ns | 39.5 ns | 96.5 ns | 500.0 ns |

Bridging domain D melting at 297 K.

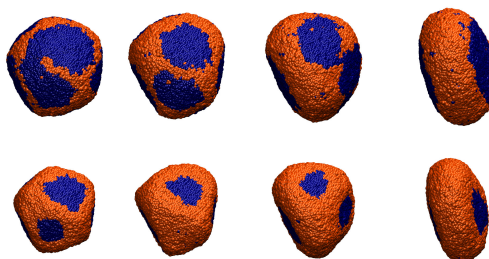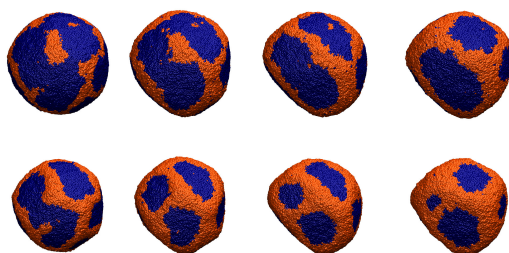| 0 ns | 118.5 ns | 287.5 ns | 500.0 ns |
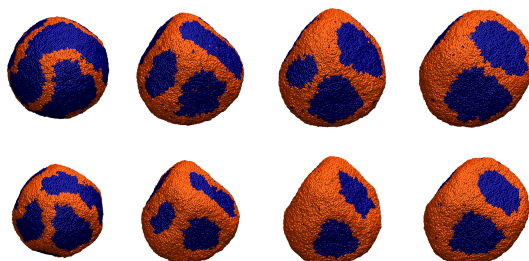
Figure B.9: Snapshots of bridging domains melting in the inner and outer leaflets of vesicles with diameters of 50 nm.
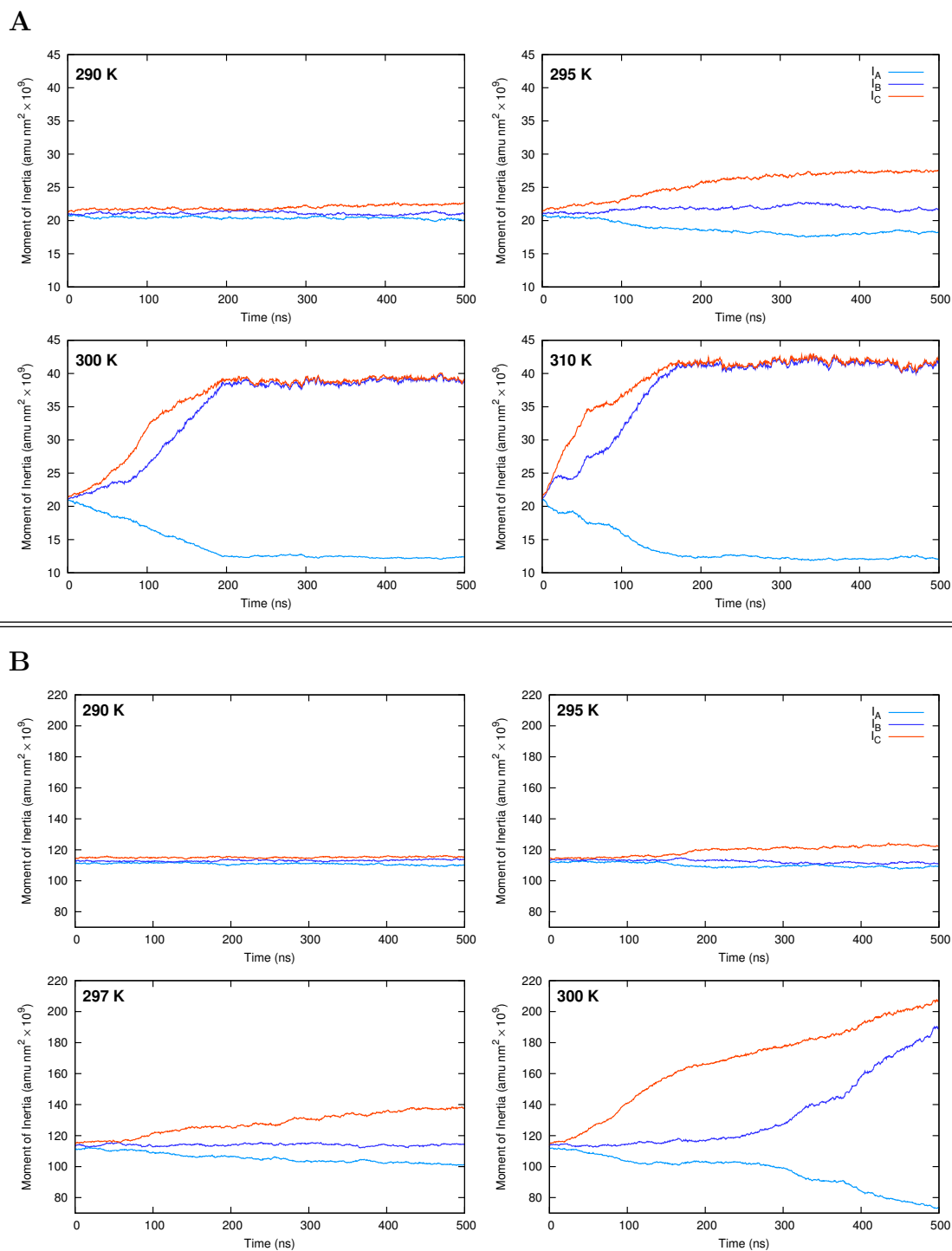
Figure B.10: Plots of the primary moments of inertial for the 33 nm vesicle (A) and the 50 nm vesicle (B) over the 500 ns trajectory of melting. The moments of inertial are arranged such that $I_A \leq I_B \leq I_C$. Partial melting occurs for the 33 nm vesicle at 290 and 295 K with full melting to a prolate symmetry at 300 and 310 K. Partial melting for the 50 nm vesicle occurs at 290, 295, and 297 K and complete melting at 300 K to end with a pear shape.

Appendix C Cluster free energies from simple simulations of small numbers of aggregants: Nucleation of liquid MTBE from vapor and aqueous phases

C.1 Establishing an optimal cluster definition

In section 5.3.1, we mentioned that the optimization of the Stillinger cluster definition cutoff distance between atoms of the molecules in the same cluster required:

1. Two independent sets of data at two different concentrations.

2. Simulations in both independent data set with the same total number of MTBE molecules.

The evidence for this requirement is given in Fig. C.1 and C.2.

The final convergence criterion C_{tot} for cutoff distance r_c used for the Stillinger cluster definition is plotted for a given number of data sets, $N_{DataSets}$, for simulations that had a total number of MTBE molecules up to $N_{\max}$.

In Fig. C.1 we show the progression in the C_{tot} plots when increasing the number of data sets (N_{DataSets}) and thus $N_{\max}$ for the simulations at a concentration of 510 mM or 284 mM in the vapor phase. It is immediately evident that using the data from a single sequence of simulations where the cluster size distributions overlap but the system sizes are not repeated, results in inconsistent minima in C_{tot}. By contrast, the inclusion of both data sets (Fig. C.2) results in a consistent minimum in C_{tot} at 0.5 nm with only two exceptions where the minimum rests at 0.51 nm. We are therefore able to conclude that at least two independent sets of simulations covering the same range of cluster sizes at different concentrations are needed to use the quality of the fit to determine an optimum cluster definition.

Figure C.1: The convergence criterion C_{tot} (over a range of inter-atomic cutoff distances r_c used to define member of a cluster) from the global fitting procedure using cluster size distributions from simulations at only a concentration of 510 mM (left) and from simulations at only a concentration of 284 mM (right) in the vapor phase.
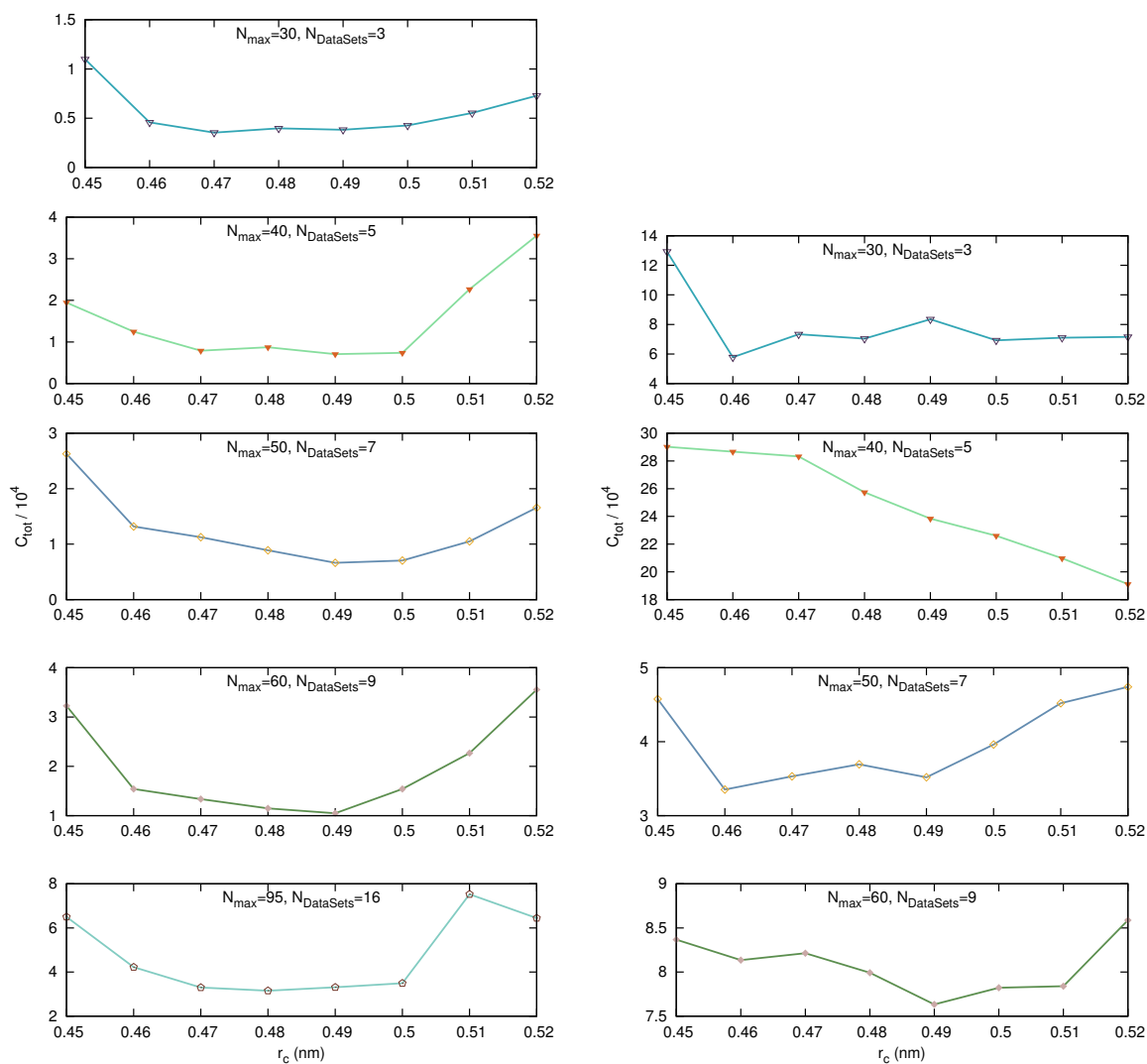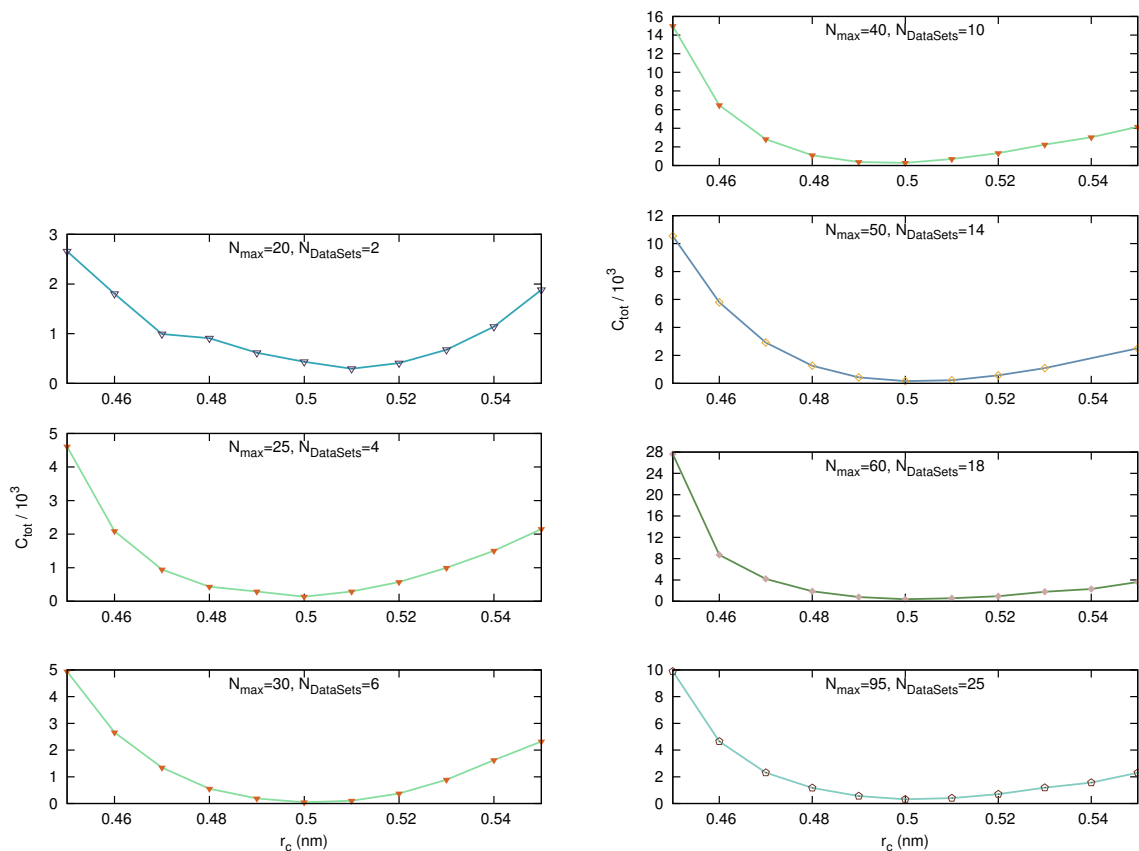
Figure C.2: The convergence criterion C_{tot} (over a range of inter-atomic cutoff distances r_c used to define member of a cluster) from the global fitting procedure using cluster size distributions from simulations at both concentration (284 mM and 510 mM).

C.2 Cluster size distributions and global fitting for the aqueous phase simulations
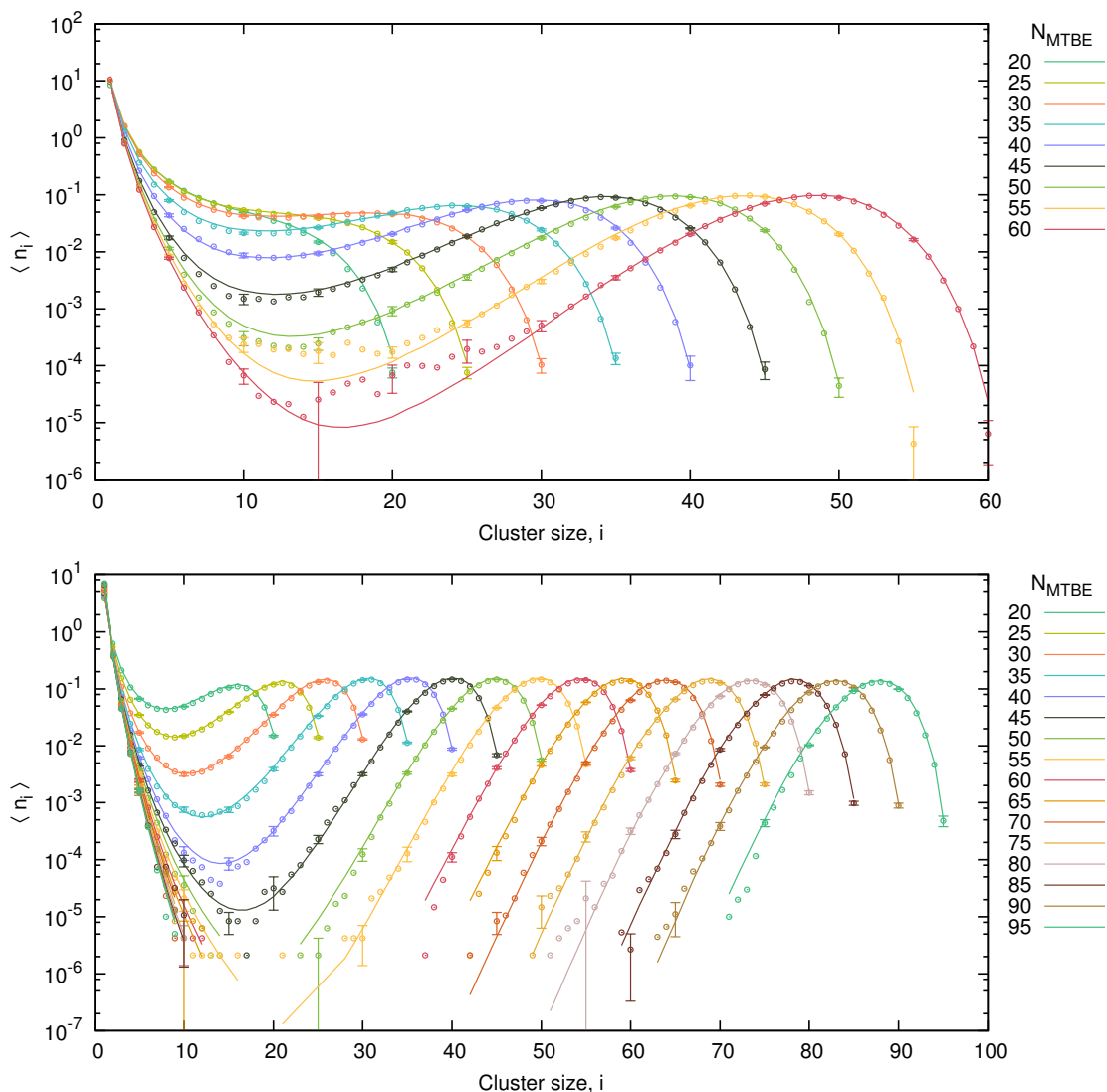
Figure C.3: Equilibrium cluster size distributions from MD simulation data (points) and the global fitting routine (lines) are displayed for the vapor-phase MTBE for simulations with a concentration of 284 mM over systems sizes of 20-60 molecules (top) and a concentration of 510 mM over systems sizes of 20-95 molecules (bottom). Error bars are given every fifth cluster size as the standard error calculated by block averaging over 19 50 ns segments of the trajectory.
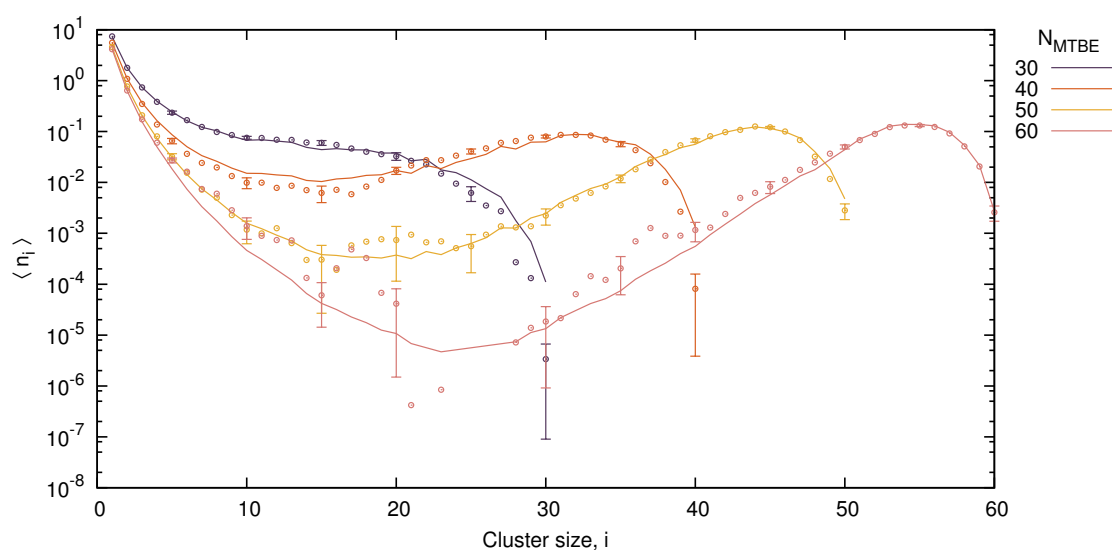
Figure C.4: Equilibrium cluster size distribution from MD simulation data (points) and the global fitting routine (lines) are displayed for MTBE solvated in Tip-4p water at concentrations of 761, 1019, 1275 and 1537 mM for system sizes of 30-60 MTBE. Error bars are given every fifth cluster size as the standard error calculated by block averaging over 19 25 ns segments of the trajectory.

Figure C.5: A plot of the convergence criterion in Equation 4.63 for fitting to the aqueous phase simulations (top) with * marking the iterations sampled in the subsequent plot of the ΔG_i of association at intervals over the 500 fitting iterations (bottom) for a monomer concentration of 184 mM.

Appendix D Simulations of NaCl aggregation from solution: Solvent determines topography of free energy landscape
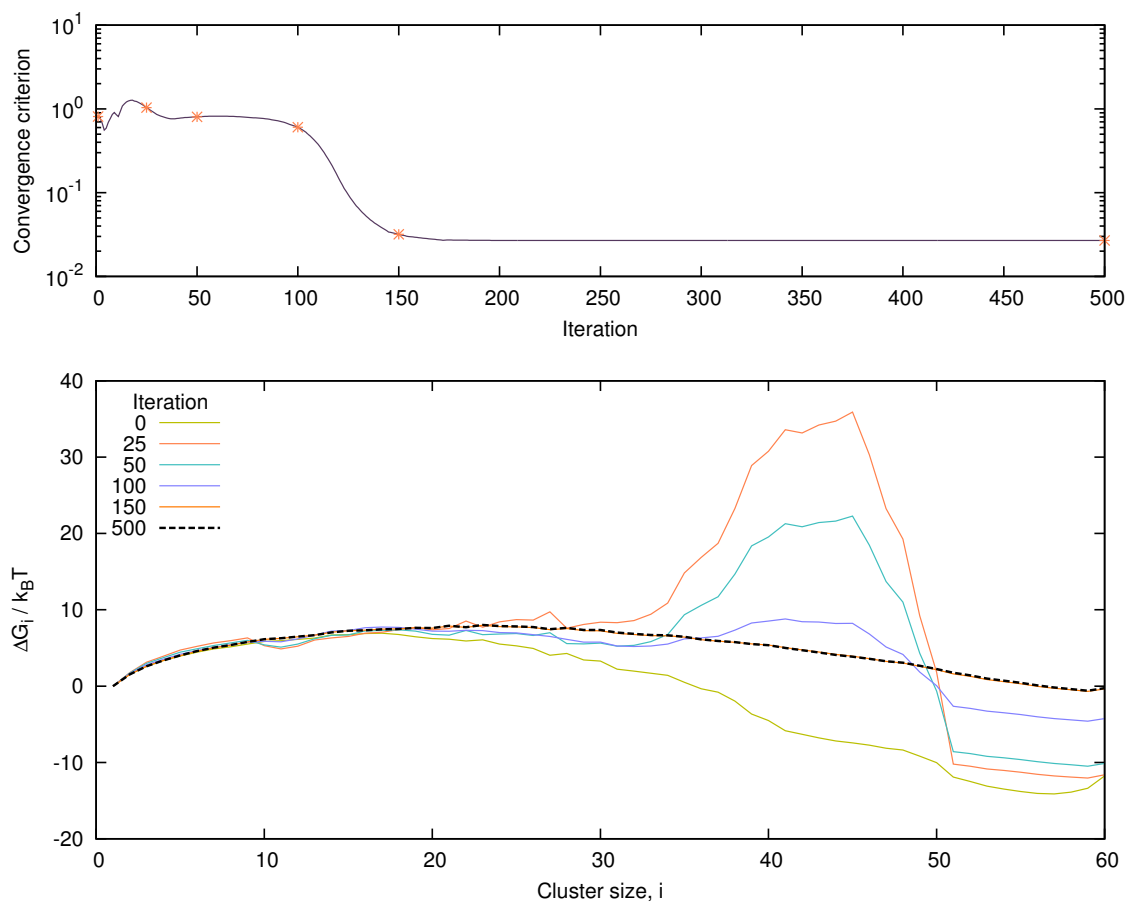
D.1 Cluster definition and the convergence criteria

The convergence criterion was calculated as,

$$C_{\text{tot}} = \frac{1}{N_{\text{max}}} \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} C_{i,j} \tag{D.1}$$

where N_{max} is the number of unique cluster compositions observed in the simulation, N_A and N_B are the maximum number of monomers from the two different components in the system (in this case Na$^+$ and Cl$^-$), and $C_{i,j}$ is the convergence criterion for a cluster composed of i and j monomers of the two components. The later is defined as a weighted geometric average of the standard deviations in the cluster size frequency from fit $\langle n_{i,j,k} \rangle_{\text{fit}}$ and the simulation $\langle n_{i,j,k} \rangle_{\text{sim}}$ for simulation k over the $J_{i,j}$ simulations where the cluster presents:

$$C_{i,j} = \exp \left(\sum_{k=1}^{J_{i,j}} w_{i,j,k} \log \left(\frac{\langle n_{i,j,k} \rangle_{\text{fit}} - \langle n_{i,j,k} \rangle_{\text{sim}}}{\langle n_{i,j,k} \rangle_{\text{sim}}} \right)^2 \right) \tag{D.2}$$

The weighting appropriated to each data point in a simulation is given by the normalized relative statistical certainty of a data point:

$$w_{i,j,k} = w_{i,j,k}^{\text{rel}} \left(\sum_{k=1}^{J_{i,j}} w_{i,j,k}^{\text{rel}} \right)^{-1}, \qquad w_{i,j,k}^{\text{rel}} = \frac{\langle n_{i,j,k} \rangle_{\text{sim}}}{\sigma \left(\langle n_{i,j,k} \rangle_{\text{sim}} \right)} \tag{D.3}$$

In Fig. D.1 we show the total convergence criterion C_{tot} for the global fits to data for NaCl cluster formation in the solvent environments where the fraction of methanol, x_{MeOH} is 0.0, 0.4, and 0.5.

It can clearly be seen that there is no apparent trend in the quality of fit that would indicate an optimal cutoff that minimizes the deviation in the fit. We discussed in the
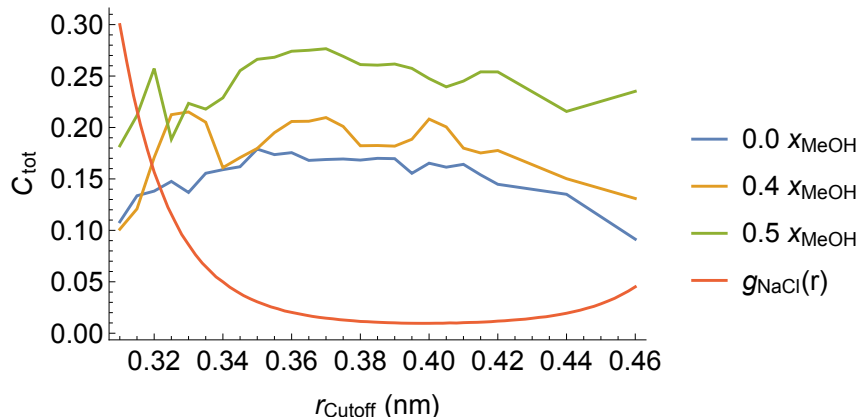
Figure D.1: The total convergence criterion C_{tot} for NaCl cluster formation in pure SPC/E water and the MeOH mixtures for a series of radial cutoff distances r_{cutoff} plotted along side the radial distribution function for the Na$^+$ and Cl$^-$.

paper that the quality of the fit where $x_{MeOH} = 0.0$ is impacted by non-ideal effects that the fitting method does not explicitly account for. In the case of $x_{MeOH} = 0.5$ the sampling proves to be somewhat challenging since the ordered NaCl crystal is observed but the frequencies of transitions between the ordered and amorphous state is low. However the trend does not even present in the case of $x_{MeOH} = 0.4$ where non-ideal effects and sampling challenge do not present a challenge to the global fit.

We chose a cluster definition in terms of a radial cutoff distance $r_{cutoff} = 0.34$ based on the tail from the first peak of the radial distribution function (red line, Fig. D.1) for Na$^+$ and Cl$^-$ ions.

D.2 Implementation of the PEACH method

In the partition-enabled analysis of cluster histograms (PEACH) method, an initial guess at the equilibrium associations constants is obtained using a geometric average over the simulation association constants generated using the law of mass action $\left(K'_{i,j} = c_{i,j} / \left(c^i_{1,0} c^j_{0,1} \right) \right)$ weighted by the standard error in the cluster size frequencies across all the simulations where that cluster size manifests. The equilibrium

constant is related to the standard state partition functions $q_{i,j}^\circ (V^\circ, T)$ as,

$$K_{i,j} = \frac{q_{i,j}^\circ}{\left(q_{1,0}^\circ\right)^i \left(q_{0,1}^\circ\right)^j} \tag{D.4}$$

where for our purposes the standard state is defined as 1 ion/nm³. For each system size we define the partition function $q_{i,j} = q_{i,j}^\circ V/V^\circ$.

Following our previous work,[59] the relationship between the grand canonical partition function $\Xi(z_A, z_B, V, T)$ and the canonical partition function $Q(N_A, N_B, V, T)$,

$$\Xi(z_A, z_B, V, T) = \exp\left(\sum_{i,j}^\infty q_{i,j} z_A^i z_B^j\right) = \sum_{N_A, N_B = 0}^\infty z_A^{N_A} z_B^{N_B} Q(N_A, N_B, V, T) \tag{D.5}$$

where z_A and z_B are the chemical activity of the distinct chemical components, enables us to use an iterative loop of partial derivatives to generate the canonical partition functions:

$$Q(N_A, N_B, V, T) = \frac{1}{N_A! N_B!} \left(\frac{\partial^{N_A}}{\partial z_A} \frac{\partial^{N_B}}{\partial z_B} \Xi(z_A, z_B, V, T)\right)_{z_A, z_B = 0} \tag{D.6}$$

Using the explicitly exact relationship between the partition functions and the canonical partition functions $Q(N_A, N_B, V, T)$ for a system with a total of N_A and N_B parts of each molecule type, the average cluster frequency is,

$$\langle n_{i,j} \rangle_{N_A, N_B} = \frac{q_{i,j}\, Q(N_A - i, N_B - j, V, T)}{Q(N_A, N_B, V, T)} \tag{D.7}$$

The PEACH method thus generates a set of average cluster frequencies that are then used in tandem with the raw simulation cluster size frequencies to adjust the standard state partition functions to better reproduce the simulated average cluster size frequencies. Assuming that the user has a good cluster definition and that the system does not have egregious non-ideal effects, the PEACH method converges to an optimal fit in the partition functions which can then be used to generate a final set of equilibrium associations constants, $K_{i,j}$.

D.3 Quality of the fits

The following figures show the quality of the global fits from the PEACH method for all of the cluster sizes as 2-D plots. For each data set, there are two surfaces shown. The left plot is the log of the cluster size frequency from simulation. The right plot shows the quality of the fit in two ways:

1. The relative difference in magnitude between the cluster size frequency from the fit and that of the raw data. This is shown using a color spectrum, cool colors denoting an underestimation in the fit and warm colors indicating the opposite. Green indicates data points where the relative magnitudes of the frequencies are close to equal.

$$\log_{10}\left(\frac{\langle n_{i,j}\rangle_{\text{fit}}}{\langle n_{i,j}\rangle_{\text{sim}}}\right) \tag{D.8}$$

2. The error in the fit relative to the quality of the data point as denoted by the standard error for each data point $\sigma_{\text{Err}} = \sigma\left(\langle n_{i,j}\rangle_{\text{sim}}\right)$. This quality evaluation is only indicated for data points where the error has exceeded $\pm 1\sigma_{\text{Err}}$ where an open circle indicates a relative overestimation in the fit and filled black point a relative under estimation of the fit.

$$\frac{\langle n_{i,j}\rangle_{\text{fit}} - \langle n_{i,j}\rangle_{\text{sim}}}{\sigma_{\text{Err}}} \tag{D.9}$$

D.3.1 Pure SPC/E solvated NaCl aggregation

Quality of fit plots are provided for the simulations of NaCl cluster formation in pure SPCE water ($x_{\text{MeOH}} = 0.0$) for systems with 20 (Fig. D.5), 30 (Fig. D.4), and 45 (Fig. D.3) NaCl pairs. It should again be noted that in the simulations used in the PEACH fit, the formation of an ordered crystal did not occur. There is however a systematic trend in the errors which can be seen in the quality of fit.

Fig. D.4 for the simulations with 30 NaCl pairs showcases this trend very well. At high concentrations the large cluster are underestimated and the small clusters over

estimated by the fit. As the concentration of the system is lowered, the trend changes like a seesaw so that the smaller cluster are underestimated and the large clusters overestimated. We posited in the main paper that this is the result of a non-ideal crowding effect that occurs at these high concentrations that the current PEACH method does not account for. The global fit attempts to find the best fit resulting in this seesawing trend in the quality of the fit.

Table D.1:   Coordination numbers for Na$^+$ and Cl$^-$ ions in pure SPC/E water as calculated through integration over the first solvent shell of the radial distribution functions.

[NaCl] (m)	$\langle n_{Na-O} \rangle$	$\langle n_{Cl-O} \rangle$
0.56	5.3	7.2
1.11	5.0	7.1
1.67	4.7	7.0
2.22	4.5	6.9
2.78	4.3	6.7
3.33	4.1	6.6
3.88	3.9	6.5
r_{shell} (nm)	0.32	0.39
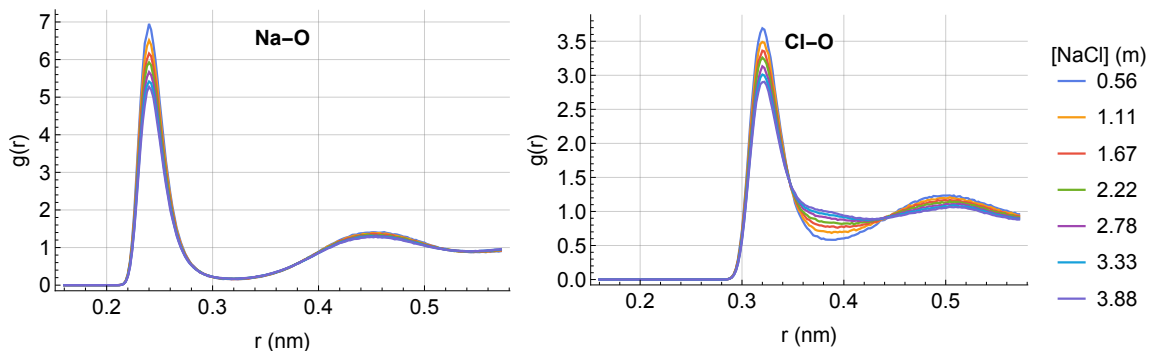
Figure D.2:   The radial distribution functions between the Na$^+$ (left) and Cl$^-$ (right) ions and the oxygen atom on the SPC/E water for simulations of 30 NaCl pairs solvated in pure SPC/E water.

On a cautionary note, although the data presented in Table D.1 indicates a concentration dependent trend, it cannot be assumed that this is a trend held by the monomer-solvent coordination numbers. The radial distribution functions that were calculated (Fig. D.2) average over all ions in the system, including those in large

clusters. The depletion effect that we see in the coordination numbers may simply be a result of including ions that are coordinated to other ions and thus would have a lower solvent coordination number simply due to geometric factors.

Figure D.3: Quality of fit 2-D plots for NaCl cluster formation in systems with 45 NaCl pairs in pure SPC/E water ($x_{MeOH} = 0.0$). The legends on the first plot apply to all other 2-D surface plots in this figure. The linear plot on the left shows the results for only the neutral clusters where the line represents the results of the global fit, the points are the simulated data, and the grey shaded region signifies the standard error.

Figure D.4: Quality of fit 2-D plots for NaCl cluster formation in systems with 30 NaCl pairs in pure SPC/E water ($x_{\mathrm{MeOH}} = 0.0$). The legends on the first plot apply to all other plots in this figure.
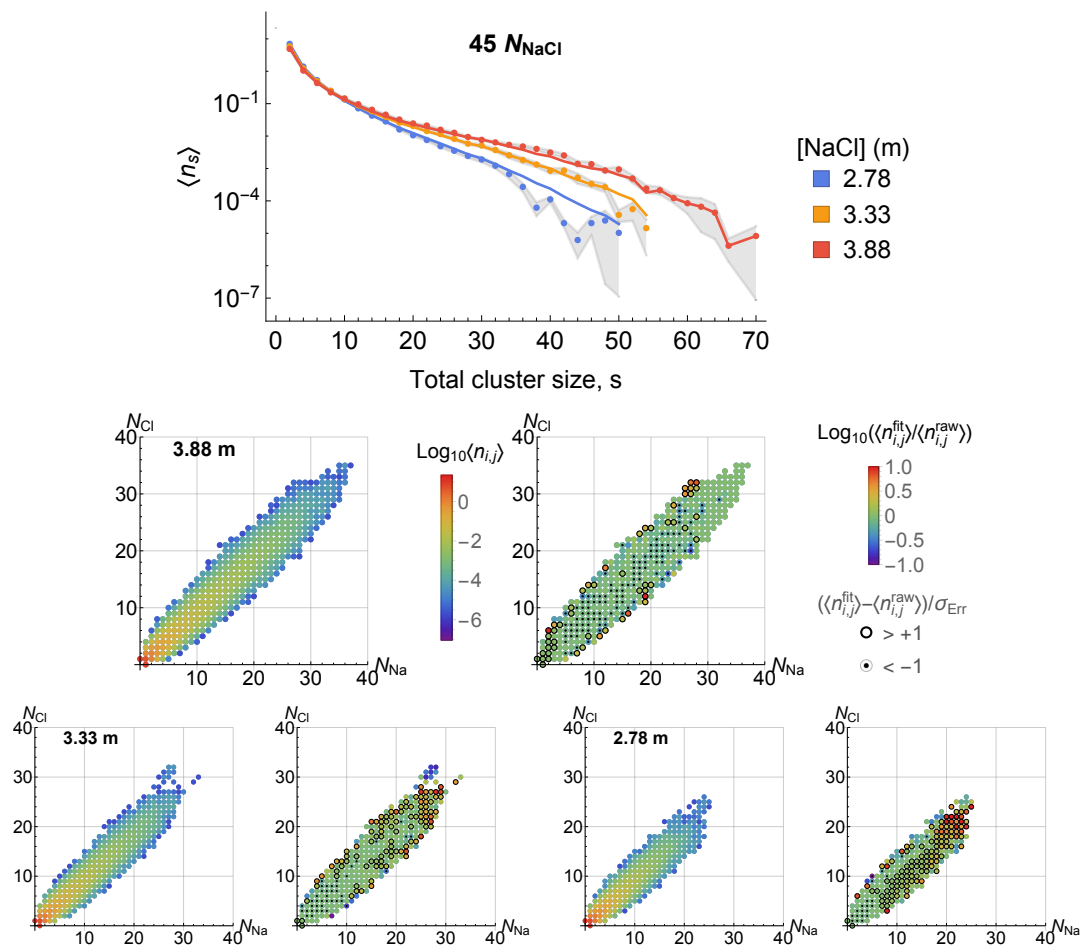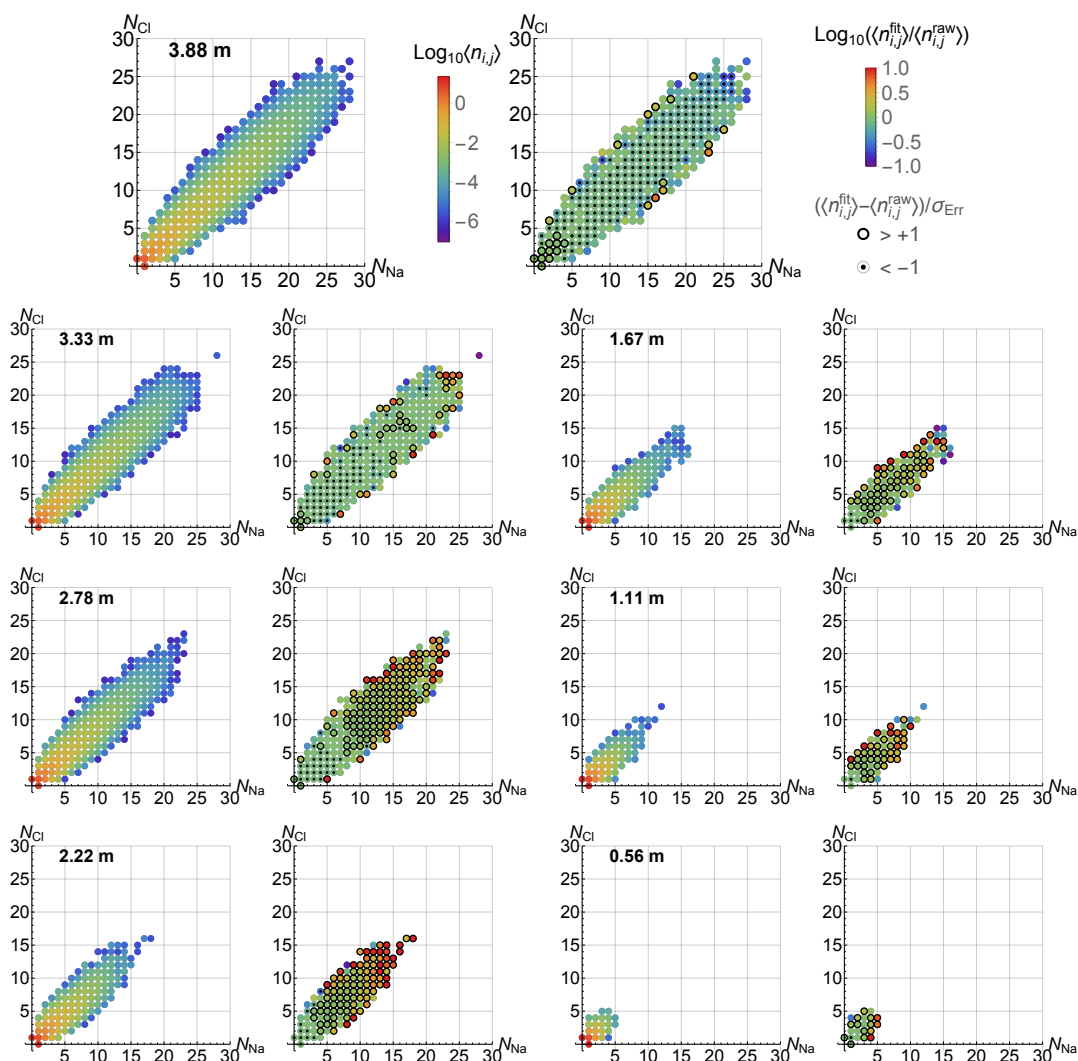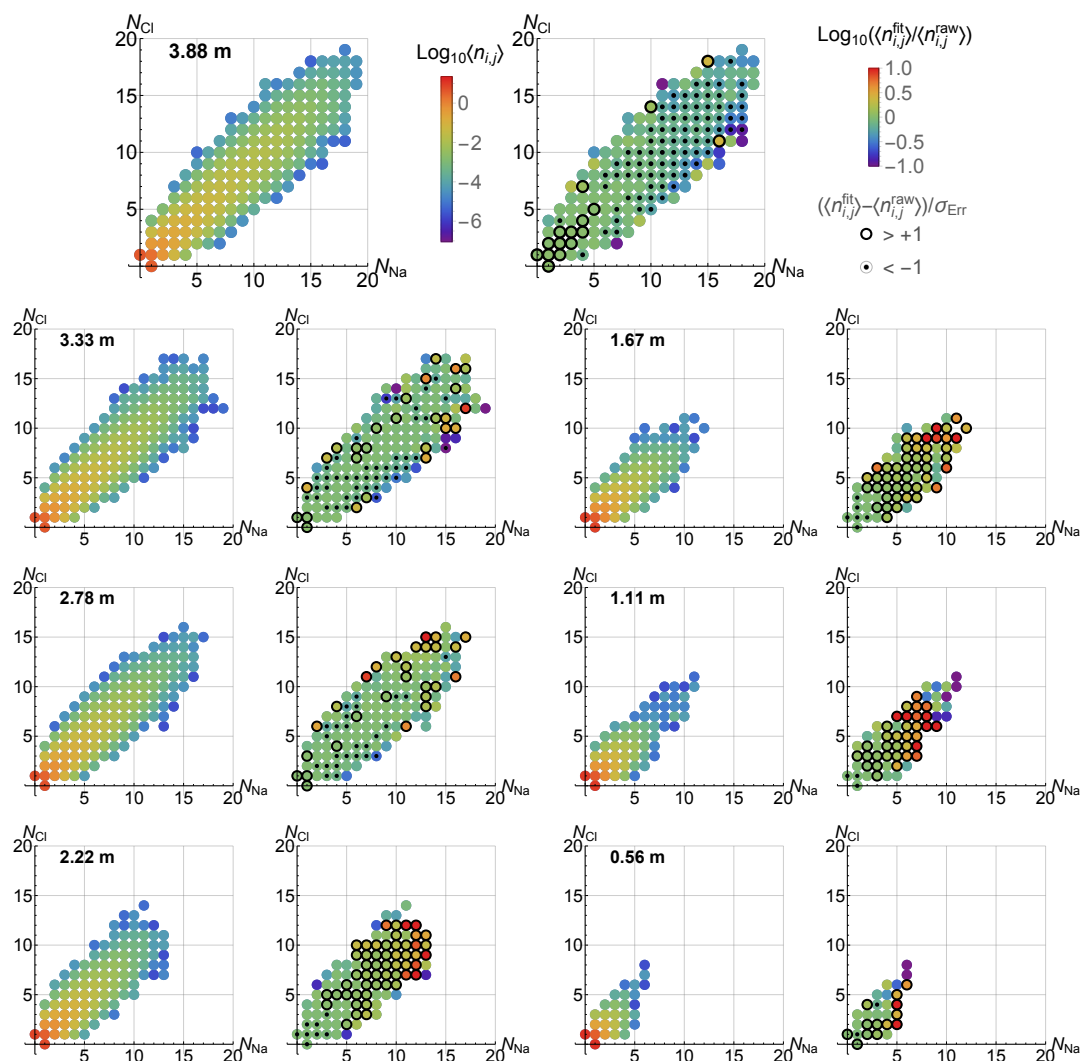
Figure D.5: Quality of fit 2-D plots for NaCl cluster formation in systems with 20 NaCl pairs in pure SPC/E water ($x_{\mathrm{MeOH}} = 0.0$). The legends on the first plot apply to all other plots in this figure.

D.3.2 NaCl aggregation in a solvent mixture of methanol and SPC/E water

The 2-D surface quality of fit plots are displayed for the systems with x_{MeOH} of 0.4 and 0.5 in Figures D.6 and D.7 respectively. It can be seen that in both sets of simulations, the quality of the fit is rather good and neither case presents systematic trends in error.

Figure D.6: Quality of fit 2-D plots for NaCl cluster formation in systems with 20 NaCl pairs in a mixture of SPC/E water and methanol ($x_{\mathrm{MeOH}} = 0.4$). The legends on the first plot apply to all other plots in this figure.
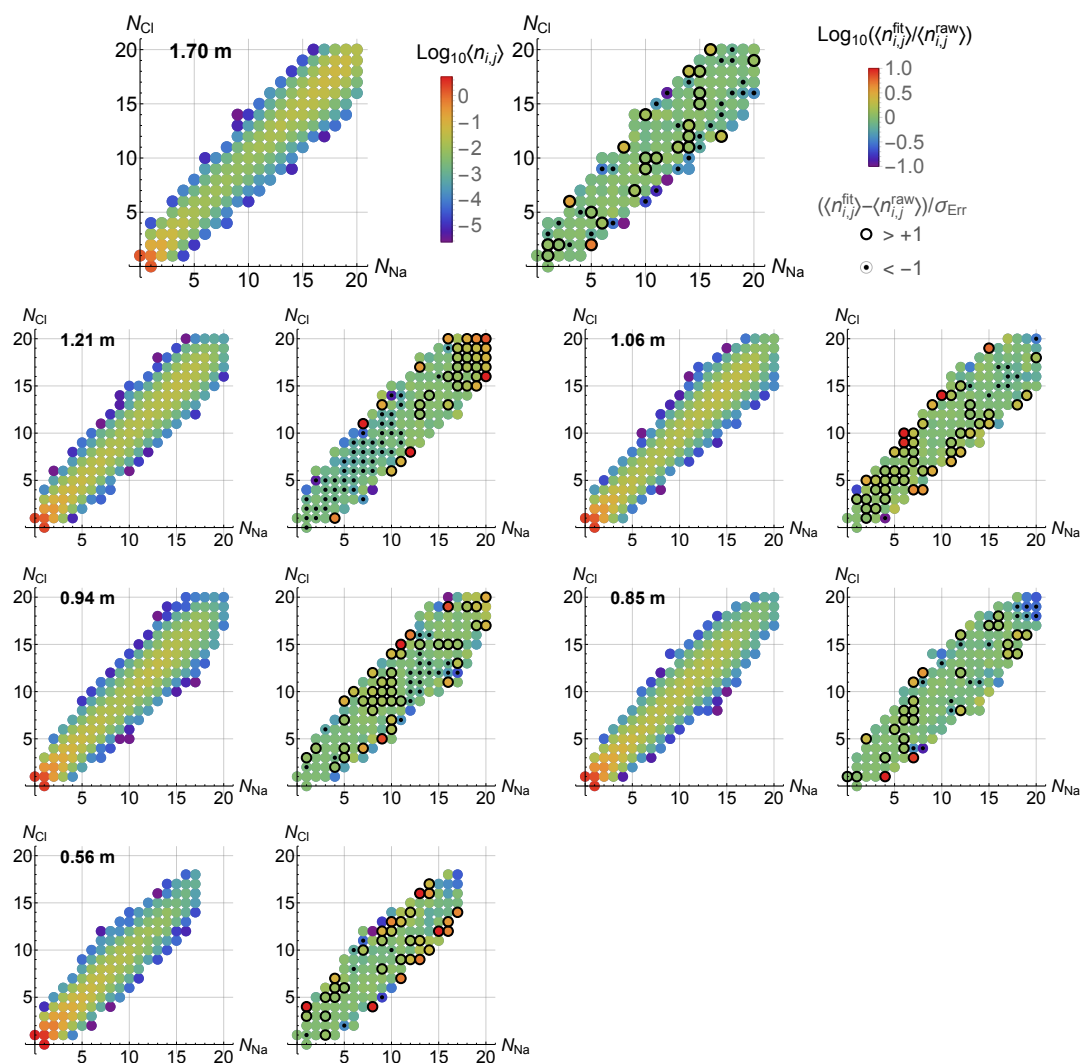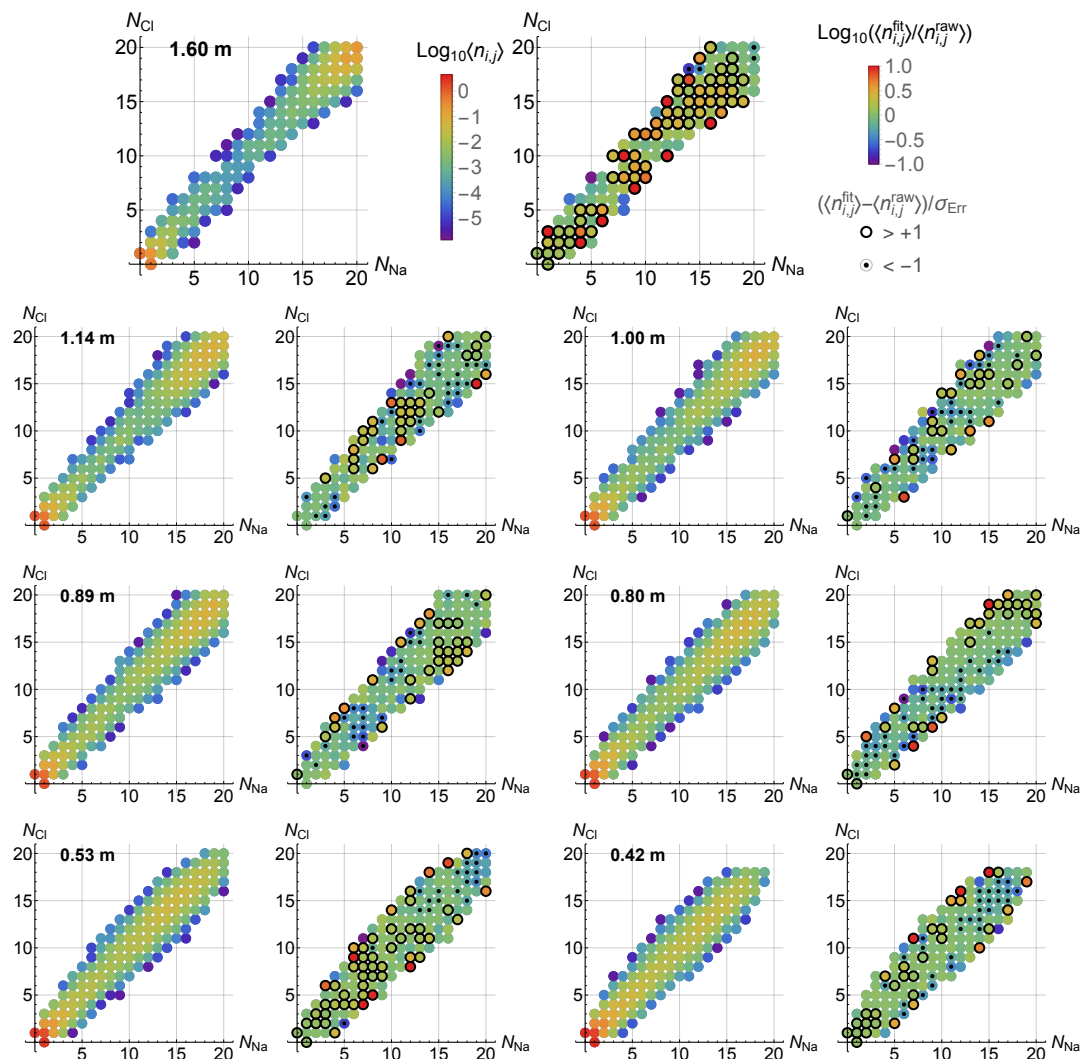
Figure D.7: Quality of fit 2-D plots for NaCl cluster formation in systems with 20 NaCl pairs in a mixture of SPC/E water and methanol ($x_{\mathrm{MeOH}} = 0.5$). The legends on the first plot apply to all other plots in this figure.

Bibliography

(1) Chandler, D., *Introduction to modern statistical mechanics*; Oxford University Press, Inc.: 198 Madison Avenue, New York, New York 10016-4314, 1987.

(2) Mathai, J. C.; Tristram-Nagle, S.; Nagle, J. F.; Zeidel, M. L. *J. Gen. Physiol.* **2008**, *131*, 69–76.

(3) Hockney, R. W.; Goel, S. P.; Eastwood, J. *J. Comp. Phys.* **1974**, *14*, 148–158.

(4) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(5) Abraham, M.; van der Spoel, D.; Lindahl, E.; Hess, B.; the GROMACS development team GROMACS User Manual version 2018.1., 2018.

(6) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.

(7) Frenkel, D.; Smit, B., *Understanding molecular simulation from algorithms to applications*, Second; Computational Science Series, Vol. 1; Academic Press: San Diego, California, USA, 2002.

(8) Patel, L. A.; Kindt, J. T. *Soft Matter* **2016**, *12*, 1765–1777.

(9) Lin, C. M.; Li, C. S.; Sheng, Y. J.; Wu, D. T.; Tsao, H. K. *Langmuir* **2012**, *28*, 689–700.

(10) Hossann, M.; Syunyaeva, Z.; Schmidt, R.; Zengerle, A.; Eibl, H.; Issels, R. D.; Lindner, L. H. *J. Control. Release* **2012**, *162*, 400–406.

(11) May, J. P.; Li, S. D. *Expert Opin. Drug Deliv.* **2013**, *10*, 511–527.

(12) Deamer, D. W.; Bramhall, J. *Chem. Phys. Lipids* **1986**, *40*, 167–188.

(13) Nagarajan, S.; Schuler, E. E.; Ma, K.; Kindt, J. T.; Dyer, R. B. *J. Phys. Chem. B* **2012**, *116*, 13749–13756.

(14) Dyer, R. B.; Gai, F.; Woodruff, W. H.; Gilmanshin, R.; Callender, R. H. *Acc. Chem. Res.* **1998**, *31*, 709–716.

(15) Andersson, M.; Hammarstrom, L.; Edwards, K. *J. Phys. Chem.* **1995**, *99*, 14531–14538.

(16) Risselada, H. J.; Marrink, S. J. *Soft Matter* **2009**, *5*, 4531–4541.

(17) Spangler, E. J.; Kumar, P. B. S.; Laradji, M. *Soft Matter* **2012**, *8*, 10896–10904.

(18) Hirst, L. S.; Ossowski, A.; Fraser, M.; Geng, J.; Selinger, J. V.; Selinger, R. L. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 3242–3247.

(19) Wu, H. L.; Sheng, Y. J.; Tsao, H. K. *J. Chem. Phys.* **2014**, *141*, 124906.

(20) Sknepnek, R.; Vernizzi, G.; de la Cruz, M. O. *Soft Matter* **2012**, *8*, 636–644.

(21) Bowick, M. J.; Sknepnek, R. *Soft Matter* **2013**, *9*, 8088–8095.

(22) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750–760.

(23) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812–24.

(24) Nagle, J. F.; Wilkinson, D. A. *Biophys. J.* **1978**, *23*, 159–175.

(25) Lewis, R.; Pohle, W.; McElhaney, R. N. *Biophys. J.* **1996**, *70*, 2736–2746.

(26) Fraile, M. V.; Patron-Gallardo, B.; Lopez-Rodriguez, G.; Carmona, P. *Chem. Phys. Lipids* **1999**, *97*, 119–128.

(27) Singh, H.; Emberley, J.; Morrow, M. R. *Eur. Biophys. J. Biophy.* **2008**, *37*, 783–792.

(28) Yang, L.; Kindt, J. T. *J. Phys. Chem. B* **2016**, *120*, 11740–11750.

(29) Nagle, J. F.; Zhang, R. T.; Tristram-Nagle, S.; Sun, W. J.; Petrache, H. I.; Suter, R. M. *Biophys. J.* **1996**, *70*, 1419–1431.

(30) Petrache, H. I.; Dodd, S. W.; Brown, M. F. *Biophys. J.* **2000**, *79*, 3172–92.

(31) Nagle, J. F. *Biophys. J.* **1993**, *64*, 1476–1481.

(32) Sun, W.; Suter, R. M.; Knewtson, M. A.; Worthington, C. R.; Tristram-Nagle, S.; Zhang, R.; Nagle, J. F. *Phys. Rev. E* **1994**, *49*, 4665–4676.

(33) Sun, W. J.; Tristram-Nagle, S.; Suter, R. M.; Nagle, J. F. *Biophys. J.* **1996**, *71*, 885–891.

(34) Rand, R. P.; Parsegian, V. A. *Biochim. Biophys. Acta* **1989**, *988*, 351–376.

(35) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

(36) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(37) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(38) Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013**, *29*, 845–854.

(39) Åqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021–8024.

(40) Theodorou, D. N.; Suter, U. W. *J. Phys. Chem.* **1985**, *82*, 955.

(41) Vymetal, J.; Vondrasek, J. *J. Phys. Chem. A* **2011**, *115*, 11455–11465.

(42) Lamberg, A.; Taniguchi, T. *Soft Matter* **2014**, *10*, 257–261.

(43) Seifert, U.; Berndl, K.; Lipowsky, R. *Phys. Rev. A* **1991**, *44*, 1182–1202.

(44) Gordon, V. D.; Beales, P. A.; Zhao, Z.; Blake, C.; MacKintosh, F. C.; Olmsted, P. D.; Cates, M. E.; Egelhaaf, S. U.; Poon, W. C. K. *J. Phys. Condens. Matter* **2006**, *18*, L415.

(45) Schneider, S.; Gompper, G. *Europhys. Lett.* **2005**, *70*, 136.

(46) Marrink, S. J.; Risselada, J.; Mark, A. E. *Chem. Phys. Lipids* **2005**, *135*, 223–244.

(47) Katira, S.; Mandadapu, K. K.; Vaikuntanathan, S.; Smit, B.; Chandler, D. *eLife* **2015**.

(48) Wang, X.; Deserno, M. *J. Chem. Phys.* **2015**, *143*, 164109.

(49) McQuarrie, D. A., *Statistical mechanics*; University Science Books: Mill Valley, CA, 2000.

(50) Maibaum, L.; Dinner, A. R.; Chandler, D. *J. Phys. Chem. B* **2004**, *108*, 6778–6781.

(51) Johnston, M. A.; Swope, W. C.; Jordan, K. E.; Warren, P. B.; Noro, M. G.; Bray, D. J.; Anderson, R. L. *J. Phys. Chem. B* **2016**, *120*, 6337–51.

(52) Chen, A. A.; Pappu, R. V. *J. Phys. Chem. B* **2007**, *111*, 6469–6478.

(53) Giberti, F.; Tribello, G. A.; Parrinello, M. *J. Chem. Theory. Comput.* **2013**, *9*, 2526–2530.

(54) Wedekind, J.; Reguera, D.; Strey, R. *J. Chem. Phys.* **2006**, *125*.

(55) Salvalaglio, M.; Mazzotti, M.; Parrinello, M. *Faraday Discuss.* **2015**, *179*, 291–307.

(56) Ahlgren, S.; Ono, K. *Notices Am. Math. Soc.* **2001**, *48*, 978–984.

(57) Auluck, F. C. *Math. Proc. Camb. Philos. Soc.* **1953**, *49*, 72–83.

(58) Bureaux, J. *Math. Proc. Camb. Philos. Soc.* **2014**, *157*, 469–487.

(59) Zhang, X.; Patel, L. A.; Beckwith, O.; Schneider, R.; Weeden, C. J.; Kindt, J. T. *J. Chem. Theory Comput.* **2017**, *13*, 5195–5206.

(60) Patel, L. A.; Kindt, J. T. *J. Chem. Theory Comput.* **2017**, *13*, 1023–1033.

(61) Kindt, J. T. *J. Chem. Theory. Comput.* **2013**, *9*, 147–52.

(62) Kalikmanov, V. I. In *Lecture Notes in Physics*, Beiglbock, W., Ehlers, J., Hepp, K., Weidenmuller, H., Eds.; Springer: Netherlands, 2013.

(63) Girshick, S. L.; Chiu, C. P. *J. Chem. Phys.* **1990**, *93*, 1273–1277.

(64) Ten Wolde, P.; Frenkel, D. *J. Chem. Phys.* **1998**, *109*, 9901–9918.

(65) Oh, K. J.; Zeng, X. C. *J. Chem. Phys.* **1999**, *110*, 4471–4476.

(66) Chen, B.; Siepmann, J. I. *J. Phys. Chem. B* **2001**, *105*, 11275–11282.

(67) Kathmann, S. M.; Schenter, G. K.; Garrett, B. C.; Chen, B.; Siepmann, J. I. *J. Phys. Chem. C* **2009**, *113*, 10354–10370.

(68) Chen, B.; Siepmann, J. I.; Oh, K. J.; Klein, M. L. *J. Chem. Phys.* **2001**, *115*, 10903–10913.

(69) Loeffler, T. D.; Sepehri, A.; Chen, B. *J. Chem. Theory. Comput.* **2015**, *11*, 4023–32.

(70) Factorovich, M. H.; Molinero, V.; Scherlis, D. A. *J. Chem. Phys.* **2014**, *140*, 064111.

(71) Factorovich, M. H.; Molinero, V.; Scherlis, D. A. *J. Am. Chem. Soc.* **2014**, *136*, 4508–14.

(72) Barducci, A.; Bussi, G.; Parrinello, M. *Phys. Rev. Lett.* **2008**, *100*, 020603.

(73) Salvalaglio, M.; Perego, C.; Giberti, F.; Mazzotti, M.; Parrinello, M. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E6–E14.

(74) Agarwal, V.; Peters, B. *J. Chem. Phys.* **2014**, *140*, 084111.

(75) Kindt, J. T. *Phys. Procedia* **2014**, *53*, 63–70.

(76) Zoghbi, A.; Stojmenovic, I. *Int. J. Comput. Math* **1998**, *70*, 319–332.

(77) Kusaka, I.; Wang, Z. G.; Seinfeld, J. H. *J. Chem. Phys.* **1998**, *108*, 3416–3423.

(78) Chen, B.; Siepmann, J. I.; Oh, K. J.; Klein, M. L. *J. Chem. Phys.* **2002**, *116*, 4317–4329.

(79) Squillace, P. J.; Zogorski, J. S.; Wilber, W. G.; Price, C. V. *Environ. Sci. Technol.* **1996**, *30*, 1721–1730.

(80) Powers, S. E.; Rice, D.; Dooher, B.; Alvarez, P. J. J. *Environ. Sci. Technol.* **2001**, *35*, 24A–30A.

(81) Stubbs, J. M.; Potoff, J. J.; Siepmann, J. I. *J. Phys. Chem. B* **2004**, *108*, 17596–17605.

(82) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(83) Lundsgaard, R.; Kontogeorgis, G. M.; Economou, I. G. *Fluid Phase Equilib.* **2011**, *306*, 162–170.

(84) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(85) Essman, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8592.

(86) Lee, J. K.; Barker, J. A.; Abraham, F. F. *J. Chem. Phys.* **1973**, *58*, 3166–3180.

(87) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637–649.

(88) Schenter, G. K.; Kathmann, S. M.; Garrett, B. C. *J. Chem. Phys.* **1999**, *110*, 7951.

(89) Stillinger, F. H. *J. Chem. Phys.* **1963**, *38*, 1486.

(90) Kraehenbuehl, M. A.; Gmehling, J. *J. Chem. Eng. Data* **1994**, *39*, 759–762.

(91) Nielsen, T. B.; Hvidt, S.; Keiding, S. R.; Petersen, C.; Westh, P.; Keiding, K. *Phys. Chem. Chem. Phys.* **2011**, *13*, 1182–1188.

(92)   Wang, X.; Pan, J.; Wu, J.; Liu, Z. *J. Chem. Eng. Data* **2006**, *51*, 1394–1397.

(93)   Cárdenas, H.; Cartes, M.; Mejía, A. *Fluid Phase Equilib.* **2015**, *396*, 88–97.

(94)   Joswiak, M. N.; Duff, N.; Doherty, M. F.; Peters, B. *J. Phys. Chem. Lett.* **2013**, *4*, 4267–72.

(95)   Wilhelmsen, O.; Bedeaux, D.; Reguera, D. *J. Chem. Phys.* **2015**, *142*, 171103.

(96)   Allen, M. P.; Tildesley, D. J., *Computer simulation of liquids*; Oxford University Press: Oxford, 1987.

(97)   Korochkova, E. A.; Boltachev, G. S.; Baidakov, V. G. *Russ. J. Phys. Chem.* **2006**, *80*, 445–448.

(98)   Ashbaugh, H. S.; Liu, L.; Surampudi, L. N. *J. Chem. Phys.* **2011**, *135*, 054510.

(99)   In, Gelbart, W., Ben-Shaul, A., Roux, D., Eds.; New York : Springer-Verlag: New York, 1994, pp 1–104.

(100) Widom, B., *Statistical mechanics: a concise introduction for chemists*; Cambridge University Press: Cambridge, New York, 2002.

(101) Irle, S.; Zheng, G.; Wang, Z.; Morokuma, K. *J. Phys. Chem. B* **2006**, *110*, 14531–14545.

(102) Sosso, G. C.; Chen, J.; Cox, S. J.; Fitzner, M.; Pedevilla, P.; Zen, A.; Michaelides, A. *Chem. Rev.* **2016**, *116*, 7078–7116.

(103) Ohtaki, H.; Fukushima, N. *Pure Appl. Chem.* **1991**, *63*, 1743–1748.

(104) Chakraborty, D.; Patey, G. N. *J. Phys. Chem. Lett.* **2013**, *4*, 573–578.

(105) Chakraborty, D.; Patey, G. N. *Chem. Phys. Lett.* **2013**, *587*, 25–29.

(106) Lanaro, G.; Patey, G. N. *J. Phys. Chem. B* **2016**, *120*, 9076–9087.

(107) Choi, J. H.; Cho, M. *J. Chem. Phys.* **2015**, *143*, 104110.

(108) Choi, J. H.; Cho, M. *J. Chem. Phys.* **2014**, *141*, 154502.

(109) Kim, S.; Kim, H.; Choi, J. H.; Cho, M. *J. Chem. Phys.* **2014**, *141*, 124510.

(110) Kawska, A.; Brickmann, J.; Kniep, R.; Hochrein, O.; Zahn, D. *J. Chem. Phys.* **2006**, *124*, 024513.

(111) Zahn, D. *ChemPhysChem.* **2015**, *16*, 2069–2075.

(112) Zahn, D. *J. Phys. Chem. B* **2007**, *111*, 5249–5253.

(113) Zahn, D. *Phys. Rev. Lett.* **2004**, *92*, 040801.

(114) Filion, L.; Hermes, M.; Ni, R.; Dijkstra, M. *J. Chem. Phys.* **2010**, *133*, 244115.

(115) Van Duijneveldt, J. S.; Frenkel, D. *J. Chem. Phys.* **1992**, *96*, 4655–4668.

(116) Torrie, G. M.; Valleau, J. P. *Chem. Phys. Lett.* **1974**, *28*, 578–581.

(117) Giberti, F.; Salvalaglio, M.; Parrinello, M. *IUCrJ* **2015**, *2*, 256–266.

(118) Giberti, F.; Salvalaglio, M.; Mazzotti, M.; Parrinello, M. *Chem. Eng. Sci.* **2015**, *121*, 51–59.

(119) Anwar, J.; Zahn, D. *Angew. Chem. Int. Ed. Engl.* **2011**, *50*, 1996–2013.

(120) Kashchiev, D. *J. Chem. Phys.* **1982**, *76*, 5098–5102.

(121) Erdemir, D.; Lee, A. Y.; Myerson, A. S. *Acc. Chem. Res.* **2009**, *42*, 621–629.

(122) Ostwald, W. *Z. Phys. Chem.* **1897**, *22*, 289–330.

(123) Zimmermann, N. E.; Vorselaars, B.; Quigley, D.; Peters, B. *J. Am. Chem. Soc.* **2015**, *137*, 13352–13361.

(124) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33–38.

(125) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(126) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(127) Chandrasekhar, J.; Spellmeyer, D. C.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1984**, *106*.

(128) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(129) Steinhardt, P. J.; Nelson, D. R.; Ronchetti, M. *Phys. Rev. B* **1983**, *28*, 784–805.

(130) Russo, J.; Tanaka, H. *J. Chem. Phys.* **2016**, *145*, 211801.

(131) Alexander, S. *Model. Simul. Mater. Sci. Eng.* **2012**, *20*, 045021.

(132) Shevchuk, R. *J. Phys. Chem. A* **2016**, *120*, 2783–2788.

(133) Mickel, W.; Kapfer, S. C.; Schroder-Turk, G. E.; Mecke, K. *J. Chem. Phys.* **2013**, *138*, 044501.

(134) Kondepudi, D.; Prigogine, I., *Modern thermodynamics: From heat engines to dissipative structures*; Wiley: New York, 1988.

(135) Parsegian, V. A.; Rand, R. P.; Rau, D. C. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 3987–3992.

(136) Mancinelli, R.; Botti, A.; Bruni, F.; Ricci, M. A.; Soper, A. K. *J. Phys. Chem. B* **2007**, *111*, 13570–13577.

(137) Beladjine, S.; Amrani, M.; Zanoun, A.; Belaidi, A.; Vergoten, G. *Comput. Theor. Chem.* **2011**, *977*, 97–102.

(138) Soniat, M.; Rick, S. W. *J. Chem. Phys.* **2012**, *137*, 044511.

(139) Hassan, S. A. *J. Phys. Chem. B* **2008**, *112*, 10573–10584.

(140) Soniat, M.; Pool, G.; Franklin, L.; Rick, S. W. *Fluid Phase Equilib.* **2016**, *407*, 31–38.

(141) Oh, K. J.; Gao, G. T.; Zeng, X. C. *Phys. Rev. Lett.* **2001**, *86*, 5080–5083.

(142) Matsumoto, M.; Saito, S.; Ohmine, I. *Nature* **2002**, *416*, 409–413.

(143) Ford, I. J. *Phys. Rev. E* **1997**, *56*, 5615–5629.

(144) Pinho, S. P.; Macedo, E. A. *J. Chem. Eng. Data* **2005**, *50*, 29–32.