SIMULATE THE NUCLEATION OF ELECTROLYTES WITH EXPLICIT SOLVENTS VIA TWO APPROACHES

by

Xinyi Li

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2020

Date of final oral examination: 10/27/20

The dissertation is approved by the following members of the Final Oral Committee:
JR Schmidt, Professor, Chemistry
Arun Yethiraj, Professor, Chemistry
Reid Van Lehn, Assistant Professor, Chemical and Biological Engineering
Gilbert Nathanson, Professor, Chemistry

This thesis is dedicated to my advisor, Prof. JR Schmidt, for his invaluable guidance and endless support through my PhD career.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Prof. JR Schmidt for his patient guidance, enthusiastic encouragement and useful critiques of this research work. JR has been a phenomenal mentor through my entire graduate career. His wide-ranging interests, as well as curiosity and passion towards science have constantly inspired me on my learning path. His meticulous attitude towards research, and desires to a complete understanding of research problems have taught me how to conduct high-quality research. And none of the work contained herein would have been possible without his valuable guidance, scientific insights, and the many helpful and stimulating discussions we have had in the past years. Additionally, JR's primary concern has always been the students' growth and well-being. His advice on both research as well as on my career have been priceless. I cannot image having a better advisor and mentor for my Ph.D study.

Besides my advisor, I also appreciate the support of other committee members: Prof. Arun Yethiraj, Prof. Reid Van Lehn and Prof. Gilbert Nathanson for their insightful comments and encouragement, and for their questions which have always provided a valuable new viewpoint on my research. In addition, I would like to thank my former committer memeber, Prof. Qiang Cui for the insightful conversations we have had in my junior years.

I would like to acknowledge many graduate students and postdocs in the chemistry department that have created a warm and friendly working atmosphere. Every member of my group, Benjamin, Eric, Mary, Tingting, Chenyang, Aurora, Nina, Tesia, Kai, Zhongyi and Ajay has been great to work with. It was always a pleasure coming to work everyday with such lovely and engaging people. I also wish to pay special regards to Jesse, who was a great mentor when I was an undergraduate exchange student and introduced me to the world of molecular simulation. Additionally, I would like to thank my office mates, Hyuntae, Kyeong-jun and Chang Yun, who have always been great friends to me and taught me many things both in research and life. Other than friends in theoretical chemistry institute (TCI), I also would like to thank Yuzhou, who has been a best friend to me since college and

always willing to help me when I am in need.

Special thanks to my parents for their endless love and support. I am so appreciative to have a family who always support my decisions.

At the end, I want to express my gratitude to my girlfriend Yusi for her companion and love. I cannot imagine this 5 years long journey without her on my side, supporting me, encouraging me and giving me the opportunity to be loved.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Nucleation is a key step in the synthesis of crystalline materials, whether from the melt or from solution. Despite the importance of these early-stage nucleation processes, a detailed atomic-level understanding is often lacking due to the difficulties in probing the nuclei at the associated length- and time-scales. In this thesis, we propose two approaches to model the nucleation in atomic-level. We first introduce a hybrid grand canonical Monte Carlo/molecular dynamics (GCMC/MD) method for simulating the nucleation of weak electrolytes in explicit solvent. The approach is capable of efficiently simulating the nucleation of dilute solutions while including the atomistic influence of the surrounding solvent, and provides access to the full nucleation free energy surface and associated nucleation free energy barrier. After validating the method against a simple model system, we applied the approach to the nucleation of a low-solubility rock-salt structure in liquid water. We find that the calculated nucleation barriers, in conjunction with analytic rate theories, yields predicted nucleation rates that are in excellent agreement with brute-force MD simulations of the supersaturated solution. To further improve the efficiency and parallelism, we have developed a second simulation approach based on the graph structure of the nucleus. This approach models the nucleation for each individual cluster structure and averages the results according to their Boltzmann distribution, thus avoids complete structural sampling in one single simulation. The graph-based method was validated against lattice model and later tested for the low-solubility rock-salt system, which yields similar results with our hybrid GCMC/MD approach. We anticipate possible applications of above approaches to a wide variety of related weak electrolytes, including $CaCO_3$, zeolites, and metal-organic frameworks.

1 INTRODUCTION

Nucleation is defined as the initial step during most first-order phase transitions such as crystallization, condensation, or melting. It represents the emergence of a new thermodynamic phase driven by local fluctuation at an atomistic level. Among all types of nucleation, the solution-phase crystallization has attracted particular attention as it often plays a decisive role in nanomaterial synthesis, biomineralization and pharmaceutical formulation[40, 44, 12]. The resulting nucleus often controls the structure of the growing crystal and the final crystalline product from a potentially diverse set of possible polymorphs, each with distinct properties. Despite the importance of these early-stage nucleation processes, a detailed atomic-level understanding is often lacking due to the difficulties in probing the nuclei at the associated length- and time-scales; this challenge is particularly pronounced for complex nanomaterials and nanoporous materials such as zeolites and metal-organic frameworks (MOFs)[40]. Nucleation may occur via a variety of mechanisms including classical nucleation theory (CNT) or multi-step pathways, depending on the choice of solvent(s), temperature, supersaturation, or myriad other parameters.[12]

1.1 Nucleation Theories

CNT is the simplest while most widely used theory in describing the nucleation process. Despite its simplicity, it provides a fundamental understanding of nucleation and gives reasonable predictions for nucleation rates. CNT originated from the work of Volmer and Weber[42], Becker and Döring[5], and Frenkel[15]. The theory considers the nucleation under supersaturation as a process driven by two competing factors: the spontaneous tendency of a supersaturated solution to undergo phase transformation and the formation of energetically unfavorable interface. These two factors lead to a size-dependent nucleation free energy, which is expressed as

$$\Delta G = 4\pi r^2\gamma - \frac{4}{3}\pi r^3\Delta\mu, \tag{1.1}$$

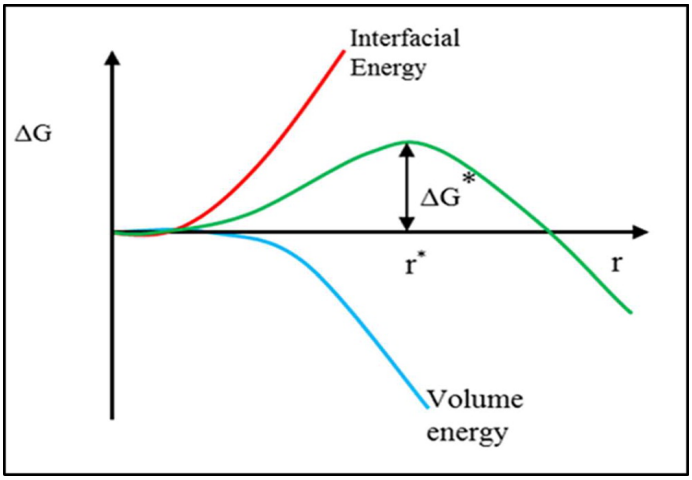


Figure 1.1: Schematic representation showing the dependence of nucleation free energy ΔG on the radius r according to classical nucleation theory. Free energy barrier and critical nucleus size are denoted as ΔG^* and r^*. Reprinted from ref [22]

where r is the size of the nucleus, γ is the positive free energy contribution arising from the formation of interface and $\Delta\mu$ is the negative chemical potential change per unit volume when transforming the solutes into crystalline phase. The positive interface term indicates the nucleation predicted by CNT is an activated process whose kinetics is determined by nucleation free energy barrier ΔG^*. And the cluster size corresponding to this barrier is so called "critical nucleus size" r^*. With the free energy barrier, the nucleation rate can be expressed in the form of the Arrhenius reaction rate equation as:

$$J = A\exp\left[-\frac{\Delta G^*}{k_b T}\right] \tag{1.2}$$

where k_b is the Boltzmann constant and T is the temperature. The pre-exponential factor A is determined from kinetic considerations.

One of the simplifying assumptions made in CNT is that the nucleus directly transforms to the thermodynamically most stable phase, without going through any intermediate state. However, when polymorphism is expected for a system, the early-statge nucleation may proceed through a metastable state. Multiple experiments have revealed the disordered transient precursor phases during the process

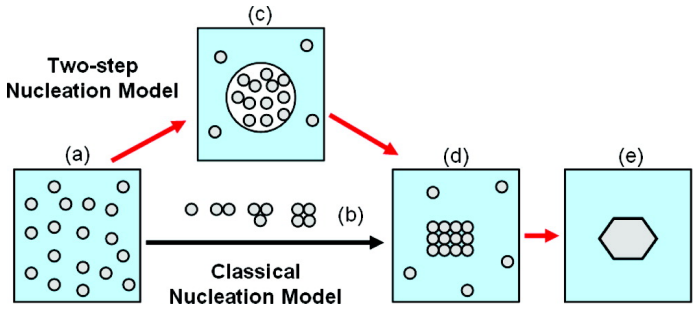


Figure 1.2: Alternative pathways leading from solution to solid crystal: (a) supersaturated solution; (b) ordered subcritical cluster of solute molecules, proposed by classical nucleation theory; (c) liquid-like cluster of solute molecules, dense precursor proposed by two-step nucleation theory; (d) ordered crystalline nuclei; (e) solid crystal. Reprinted from ref [12]

of biomineralization[44] and protein crystallization[45, 17]. Such nucleations are often described as multi-step pathways which form one or more amorphous or alternative crystalline intermediates before reaching a thermodynamically stable phase. The emergence of metastable intermediates usually generates a lower free energy barrier and possibly lead to a barrierless nucleation. Without considering the intermediate state, CNT fails to make quantitative predictions for such nucleations.

Instead, a nonclassical multiple-step mechanism (two-step mechanism in most cases)[41] was developed for those systems. It assumes the nucleation proceeds in two (multiple) steps: the formation of a droplet of a dense liquid induced by phase separation followed by the formation of a crystalline nucleus inside the droplet due to the structural ordering. When the sum of the free energy barriers from two (multiple) steps is lower than the one predicted by CNT, this two (multiple)-step mechanism is preferred.

Even for a simple crystal system, nucleation path may change significantly depending on a variety of conditions. To determine the underneath mechanism and better control the nucleation behavior, an atomistic-level understanding is required.

1.2 Nucleation of MOFs

MOFs are a class of nanoporous materials built of inorganic nodes (metal cations or oxide) bridged by organic linkers. With this general motif, hundreds of thousands[30] of MOF materials can be designed, which opens up the exciting possibility of generating tailored MOF materials for a wide range of applications via a rational "crystal engineering" approach[33]. The "crystal engineer" often requires significant insights into the fundamental processes governing MOF nucleation and growth, as well as the relationship between reaction parameters (choice of solvent(s), time, temperature, etc.) and synthetic outcome. Unfortunately, due to the lack of understanding for the underlying MOF nucleation mechanism(s), the above relationship is hard to predict. And the successful MOF synthesis is usually based on trial and error, chemical intuition, and/or large-scale screenings, rather than by rational design.[14, 35]

To enable targeted MOF synthesis without expensive experimental screenings, fundamental insights into MOF nucleation with atomistic details are needed, especially in the early-stage nucleation which is extremely challenging to explore in experiment due to the required spatial and temporal resolution.[3] For this reason, corresponding computational studies (i.e., molecular simulation) are necessary to offer a window into this process, and this has been a long-term goal of our research group.

Although molecular simulation provides a direct way to extract atomistic information during the nucleation, it also faces many challenges. The choice of solvent(s) usually plays a decisive role in determining the synthetic outcome, and potentially directs the crystallization via templating. Therefore simulation with explicit solvents is necessary for understanding the MOF nucleation. But explicitly including solvent molecules significantly increase the computing cost, especially for low-solubility materials under modest supersaturation. In addition, due to the complicated porous structure and the potential free energy barriers, the nucleation also suffers from ineffective sampling associated with polymorphism and rare events. Fortunately, all the above obstacles are also presented in the nucleation of weak

electrolytes, and we hope through the study of solution-phase weak electrolyte nucleation in this dissertation, we can provide efficient simulation strategies for understanding the early-stage MOF crystallization.

1.3 Molecular Modeling in Nucleation

Molecular simulation can probe the details of the nucleation process, including at its earliest stages, with atomistic resolution. Both molecular dynamics (MD) and Monte Carlo (MC) methods have been employed previously to understand nucleation[34]. For example, leveraging large-scale brute-force MD simulation, Patey and Chakraborty[10, 9] were able to predict the mechanism of NaCl nucleation at a high concentration. However, brute-force MD cannot easily be applied to sparingly soluble materials (e.g., weak electrolytes and other common nanoporous materials, such as zeolites or MOFs). In such cases, the saturated solution contains only an extremely low concentration of solutes. As such, the MD simulation would require both immense system sizes (to allow for sufficient solutes for even a modest sized nucleus) and simulation times (to allow for diffusive transport to the nucleus surface). In addition, outside of extreme supersaturation, nucleation is a highly activated process that cannot be effectively sampled by unbiased MD.

Biased/accelerated MD methods such as umbrella sampling[36, 37, 25] or metadynamics[27, 26, 4] can be used to circumvent the nucleation free energy barrier and thus are excellent candidates for the examination of nucleation of concentrated solutions. For example, umbrella sampling has been employed to study the nucleation of ice[29] and molten NaCl[39], while metadynamics was utilized to study the nucleation of an LJ liquid[38] and both aqueous urea[31] and NaCl solutions[16]. These methods rely upon the appropriate selection of order parameter(s) or collective variable(s), and the applied bias dictates that no direct dynamical information can be extracted.[34]

Path-sampling methods provide an alternative approach that does not require a priori selection of a collective variable. Transition path sampling[11, 6] was used by Zahn[46] to study nucleation of NaCl solution. Later, Panagiotopoulos and

co-workers[18] employed forward flux sampling[1, 2] to calculate the nucleation rate for the same system. Transition interface sampling[13, 28, 20] was also used by Jungblut[19] to obtain the rate of LJ nucleation.

Alternatively, "seeding" methods have also been employed to help circumvent the nucleation barriers and obtain a critical cluster size.[32, 8, 7, 21] However, the use of a pre-existing seed implies that it is impossible to calculate absolute nucleation barriers/rates or the details of early stage of nucleation. Further, neither biasing, path-sampling, nor seeding methods can easily be extended to weak electrolytes since they do not directly address the associated length- and time-scale challenges imposed by these dilute solutions.

To address those challenges, a "grafting" method was developed by Zahn and co-workers[23, 24] and utilized to simulate nucleation and growth of NaCl and CaF_2. In this approach, solvent is first removed, and a single ion is manually attached to an energetically-favored position on the cluster surface. The solvent is then returned, followed by energy relaxation. By iterating this procedure, a stable cluster is generated. Unfortunately, this Monte Carlo-like procedure is not reversible and thus violates detailed balance. Thus, although this scheme may generate a "representative" nucleus, it cannot (generally) be used to provide thermodynamic/kinetic data or associated nucleation barriers. (Note that it may be possible to assign approximate free energies to the generated clusters using a "two-phase" thermodynamic model, as has been done in simulations of calcium carbonate nucleation.[43])

In summary, the above simulation techniques either fail to address the length- and time-scale challenges imposed by dilute solutions, or violate detailed balance when estimating the free energy. In this dissertation, we present two methodologies developed from our group to model the nucleation of weak electrolytes, which successfully address the above challenges in crystal nucleation. Chapter 2 describes a hybrid grand canonical Monte Carlo/molecular dynamics (GCMC/MD) method and validates the method against simple Lennard-Jones (LJ) model. In Chapter 3, this method is applied to a a low-solubility rock-salt structure in liquid water, and the predicted nucleation rate is compared with the rate generated from brute-force MD simulations. To further improve the efficiency and parallelism, in Chapter

4, we develop a second simulation approach based on the graph structure of the nucleus. It is first validated against lattice model. Later we generalize this approach to nonequilibrium simulations and extend its application to atomistic systems. Overall conclusions and avenues for future research are the subject of Chapter 5.

REFERENCES

[1] Allen, Rosalind J, Daan Frenkel, and Pieter Rein ten Wolde. 2006. Simulating rare events in equilibrium or nonequilibrium stochastic systems. *The J. Chem. Phys.* 124(2):024102.

[2] Allen, Rosalind J, Chantal Valeriani, and Pieter Rein ten Wolde. 2009. Forward flux sampling for rare event simulations. *J. Phys.: Condens. Matter* 21(46): 463102.

[3] Attfield, Martin P, and Pablo Cubillas. 2012. Crystal growth of nanoporous metal organic frameworks. *Dalton Transactions* 41(14):3869–3878.

[4] Barducci, Alessandro, Giovanni Bussi, and Michele Parrinello. 2008. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* 100(2):020603.

[5] Becker, R, and W Döring. 1935. The kinetic treatment of nuclear formation in supersaturated vapors. *Ann. Phys* 24(719):752.

[6] Bolhuis, PeteráG. 1998. Sampling ensembles of deterministic transition pathways. *Faraday Discuss.* 110:421–436.

[7] Browning, Andrea Robben, Michael F Doherty, and Glenn H Fredrickson. 2008. Nucleation and polymorph selection in a model colloidal fluid. *Phys. Rev. E* 77(4):041604.

[8] Cacciuto, A, S Auer, and D Frenkel. 2004. Onset of heterogeneous crystal nucleation in colloidal suspensions. *Nature* 428(6981):404.

[9] Chakraborty, Debashree, and GN Patey. 2013. Evidence that crystal nucleation in aqueous nacl solution occurs by the two-step mechanism. *Chem. Phys. Lett.* 587:25–29.

[10] ———. 2013. How crystals nucleate and grow in aqueous nacl solution. *The J. Phys. Chem. Lett.* 4(4):573–578.

[11] Dellago, Christoph, Peter G Bolhuis, and David Chandler. 1998. Efficient transition path sampling: Application to lennard-jones cluster rearrangements. *The J. Chem. Phys.* 108(22):9236–9245.

[12] Erdemir, Deniz, Alfred Y Lee, and Allan S Myerson. 2009. Nucleation of crystals from solution: classical and two-step models. *Acc. Chem. Res.* 42(5): 621–629.

[13] van Erp, Titus S, Daniele Moroni, and Peter G Bolhuis. 2003. A novel path sampling method for the calculation of rate constants. *The J. Chem. Phys.* 118(17):7762–7774.

[14] Férey, Gérard. 2008. Hybrid porous solids: past, present, future. *Chemical Society Reviews* 37(1):191–214.

[15] Frenkel, Julius. 1939. A general theory of heterophase fluctuations and pretransition phenomena. *The Journal of Chemical Physics* 7(7):538–547.

[16] Giberti, Federico, Gareth A Tribello, and Michele Parrinello. 2013. Transient polymorphism in nacl. *J. Chem. Theory Comput.* 9(6):2526–2530.

[17] Gliko, Olga, Nikolaus Neumaier, Weichun Pan, Ilka Haase, Markus Fischer, Adelbert Bacher, Sevil Weinkauf, and Peter G Vekilov. 2005. A metastable prerequisite for the growth of lumazine synthase crystals. *Journal of the American Chemical Society* 127(10):3433–3438.

[18] Jiang, Hao, Amir Haji-Akbari, Pablo G Debenedetti, and Athanassios Z Panagiotopoulos. 2018. Forward flux sampling calculation of homogeneous nucleation rates from aqueous nacl solutions. *The J. Chem. Phys.* 148(4):044505.

[19] Jungblut, Swetlana, and Christoph Dellago. 2011. Heterogeneous crystallization on tiny clusters. *EPL (Europhys. Lett.)* 96(5):56006.

[20] Juraszek, J, G Saladino, TS Van Erp, and FL Gervasio. 2013. Efficient numerical reconstruction of protein folding kinetics with partial path sampling and pathlike variables. *Phys. Rev. Lett.* 110(10):108106.

[21] Kalikka, Janne, Jaakko Akola, Julen Larrucea, and RO Jones. 2012. Nucleus-driven crystallization of amorphous $ge_2sb_2te_5$: A density functional study. *Phys. Rev. B* 86(14):144113.

[22] Karthika, S, TK Radhakrishnan, and P Kalaichelvi. 2016. A review of classical and nonclassical nucleation theories. *Crystal Growth & Design* 16(11):6663–6681.

[23] Kawska, A, J Brickmann, O Hochrein, and Dirk Zahn. 2005. From amorphous aggregates to crystallites: modelling studies of crystal growth in vacuum. *Zeitschrift für anorganische und allgemeine Chemie* 631(6-7):1172–1176.

[24] Kawska, Agnieszka, Jürgen Brickmann, Rüdiger Kniep, Oliver Hochrein, and Dirk Zahn. 2006. An atomistic simulation scheme for modeling crystal formation from solution. *The J. Chem. Phys.* 124(2):024513.

[25] Kumar, Shankar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. 1992. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Comput. Chem.* 13(8):1011–1021.

[26] Laio, Alessandro, and Francesco L Gervasio. 2008. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.* 71(12):126601.

[27] Laio, Alessandro, and Michele Parrinello. 2002. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* 99(20):12562–12566.

[28] Moroni, Daniele, Titus S van Erp, and Peter G Bolhuis. 2004. Investigating rare events by transition interface sampling. *Physica A* 340(1-3):395–401.

[29] Radhakrishnan, Ravi, and Bernhardt L Trout. 2003. Nucleation of hexagonal ice (i_h) in liquid water. *J. Am. Chem. Soc.* 125(25):7743–7747.

[30] Rimer, Jeffrey D, and Michael Tsapatsis. 2016. Nucleation of open framework materials: Navigating the voids. *MRS Bulletin* 41(5):393.

[31] Salvalaglio, Matteo, Claudio Perego, Federico Giberti, Marco Mazzotti, and Michele Parrinello. 2015. Molecular-dynamics simulations of urea nucleation from aqueous solution. *Proceedings of the National Academy of Sciences* 112(1): E6–E14.

[32] Sanz, Eduardo, Carlos Vega, JR Espinosa, R Caballero-Bernal, JLF Abascal, and C Valeriani. 2013. Homogeneous ice nucleation at moderate supercooling from mol. simul. *J. Am. Chem. Soc.* 135(40):15008–15017.

[33] Seoane, Beatriz, Sonia Castellanos, Alla Dikhtiarenko, Freek Kapteijn, and Jorge Gascon. 2016. Multi-scale crystal engineering of metal organic frameworks. *Coordination Chemistry Reviews* 307:147–187.

[34] Sosso, Gabriele C, Ji Chen, Stephen J Cox, Martin Fitzner, Philipp Pedevilla, Andrea Zen, and Angelos Michaelides. 2016. Crystal nucleation in liquids: Open questions and future challenges in molecular dynamics simulations. *Chem. Rev.* 116(12):7078–7116.

[35] Stock, Norbert, and Shyam Biswas. 2012. Synthesis of metal-organic frameworks (mofs): routes to various mof topologies, morphologies, and composites. *Chemical reviews* 112(2):933–969.

[36] Torrie, Glenn M, and John P Valleau. 1974. Monte carlo free energy estimates using non-boltzmann sampling: Application to the sub-critical lennard-jones fluid. *Chem. Phys. Lett.* 28(4):578–581.

[37] ———. 1977. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* 23(2):187–199.

[38] Trudu, Federica, Davide Donadio, and Michele Parrinello. 2006. Freezing of a lennard-jones fluid: From nucleation to spinodal regime. *Phys. Rev. Lett.* 97(10):105701.

[39] Valeriani, C, E Sanz, and D Frenkel. 2005. Rate of homogeneous crystal nucleation in molten nacl. *The J. Chem. Phys.* 122(19):194501.

[40] Van Vleet, Mary J, Tingting Weng, Xinyi Li, and JR Schmidt. 2018. In situ, time-resolved, and mechanistic studies of metal–organic framework nucleation and growth. *Chem. Rev.* 118(7):3681–3721.

[41] Vekilov, Peter G. 2004. Dense liquid precursor for the nucleation of ordered solid phases from solution. *Crystal Growth & Design* 4(4):671–685.

[42] Volmer, Martin, and A Weber. 1926. Keimbildung in übersättigten gebilden. *Zeitschrift für physikalische Chemie* 119(1):277–301.

[43] Wallace, Adam F, Lester O Hedges, Alejandro Fernandez-Martinez, Paolo Raiteri, Julian D Gale, Glenn A Waychunas, Stephen Whitelam, Jillian F Banfield, and James J De Yoreo. 2013. Microscopic evidence for liquid-liquid separation in supersaturated $caco_3$ solutions. *Science* 341(6148):885–889.

[44] Weiner, Steve, and Lia Addadi. 2011. Crystallization pathways in biomineralization. *Annual review of materials research* 41:21–40.

[45] ten Wolde, Pieter Rein, and Daan Frenkel. 1997. Enhancement of protein crystal nucleation by critical density fluctuations. *Science* 277(5334):1975–1978.

[46] Zahn, Dirk. 2004. Atomistic mechanism of nacl nucleation from an aqueous solution. *Phys. Rev. Lett.* 92(4):040801.

2 MODELING NUCLEATION VIA HYBRID GCMC/MD SIMULATION

2.1 Introduction

Obtaining thermodynamic data on nucleation processes requires sampling over the nucleate cluster size and configurations in accord with the associated Boltzmann factor. Frenkel and co-workers proposed a straightforward approach to accomplish this sampling using grand canonical Monte Carlo (GCMC) in conjunction with a cluster size distribution theory.[72] Chen et al. later applied this GCMC technique to the nucleation of Lennard-Jones system.[52, 49] Wu and Deem[73] also employed this method to obtain the nucleation free energy surface of zeolite. In the latter case, the authors utilized a continuum dielectric to account for the role of the surrounding solvent but did not include the influence of explicit solvent. Such solvent can play an important role not only in solvating/stabilizing the surface of the growing nucleus, but also in templating the growth of the structure, especially during the synthesis of porous structures.[53, 54] The same approach was later applied to a calcium carbonate solution,[60] using a dielectric continuum to represent the influence of the surrounding solvent and yielding thermodynamic properties (e.g., equilibrium constants) in agreement with experiment and consistent with classical nucleation theory.

Building on this prior work, in this chapter we present a methodology to model the nucleation of weak electrolytes (such as sparingly soluble salts) in the presence of explicit solvent. We use a combination of sampling and biasing approaches (hybrid GCMC/MD, aggregation-volume-bias,[50, 51, 52] expanded ensemble[64, 56, 57]) to circumvent the challenges of dilute systems, mass transport, and nucleation barriers. We first benchmark and verify the methodology against a simple Lennard-Jones model system, and then apply it to the nucleation of a low-solubility rock-salt structure.[48] In the latter case (Chapter 3), we validate the approach by comparing the calculated nucleation rate against that observed via

large-scale brute-force MD and find excellent agreement.

2.2 Theory

2.2.1 Cluster Size Distribution

Our methodology is based on the cluster size distribution theory which relates the nucleation free energy with the distribution of cluster sizes. It was first proposed by ten Wolde and Frenkel[72] and we present the derivation here with minor adaptions.

For a grand canonical system at temperature T, volume V and chemical potential μ, the partition function is written as

$$\Xi(\mu, V, T) \equiv \sum_{N=0}^{\infty} exp(\beta\mu N) Q(N, V, T) \tag{2.1}$$

Here, N is the number of particles in the system and Q is the canonical ensemble partition function for a system of size N:

$$Q(N, V, T) = \frac{1}{\Lambda^{3N} N!} \int dr^N exp[-\beta U(r^N)], \tag{2.2}$$

where U is the potential energy of the system, $\beta \equiv 1/k_b T$ is the reciprocal temperature, k_b is Boltzmann constant and Λ is the thermal de Broglie wavelength of the particle.

For simplicity, we only consider the vapor-liquid nucleation here. The same derivation and conclusion can be extended to the solution-phase nucleation. Assuming a cluster criterion to define liquid clusters, the partition function of this

system can be expressed as

$$\begin{aligned}\Xi(\mu, V, T) = \sum_{N_l}^{\infty} exp(\beta\mu N_l) \sum_{N_v}^{\infty} exp(\beta\mu N_v) \frac{1}{{\Lambda_l}^{3N_l} N_l!} \frac{1}{{\Lambda_v}^{3N_v} N_v!} \\ \times \int dr^{N_v} \int dr^{N_l} \mathscr{W}(r^{N_v}; r^{N_l}) \times exp[-\beta U(r^{N_v}; r^{N_l})]\end{aligned} \tag{2.3}$$

In the above expression, subscription l denotes the particle in liquid state and v denotes the particle in vapor state. $\mathscr{W}(r^{N_v}; r^{N_l})$ is a weight function. It takes value of one if N_l number of particles in liquid state and zero otherwise. It is important to be noticed here, the general weight function $\mathscr{W}$ consists of single-cluster weight functions. Here, we use n to represent the size of each single cluster and the total number of clusters of size n is denoted as N_n. We also index the cluster of size n by $j_n = 1, 2 \cdots N_n$. Then, the weight function $\mathscr{W}$ can be expressed as:

$$\mathscr{W}_{N_l} = \sum \prod_{n} \prod_{j_n=1}^{N_n} w_{j_n}(r^n), \tag{2.4}$$

where each w_{j_n} represents the weight function for cluster j_n. It equals one if the cluster satisfies the cluster criterion and zero otherwise. $\sum$ represents all cluster distributions of N_l liquid-like particles. With these definitions, we can rewrite the partition function in terms of each single cluster:

$$\begin{aligned}\Xi(\mu, V, T) = \sum_{N_1=0}^{\infty} \sum_{N_2=0}^{\infty} \cdots \sum_{N_{n_{max}}=0}^{\infty} \frac{1}{N_1! N_2! \cdots N_{n_{max}}!} \prod_{n=1}^{n_{max}} \left(exp(\beta\mu n) n^3 / [\Lambda^{3n} n!]\right)^{N_n} \\ \times \sum_{N_v=0}^{\infty} exp(\beta\mu N_v) \frac{1}{{\Lambda_v}^{3N_v} N_v!} \prod_{n} \left[\int dr'^{n-1}\right]^{N_n} \\ \times \int \prod_{n=1}^{n_{max}} \prod_{j_n=1}^{N_n} dR_{j_n} w_{j_n}(R_{j_n}, r'^{n-1}; r^{N_v}) exp[-\beta U(R; r^{N_v})].\end{aligned} \tag{2.5}$$

Here, the entire partition function is represented by partition functions of individual

clusters. For each cluster, the cartesian coordinates are transformed into center-of-mass coordinates. R_{j_n} is the center-of-mass for cluster j_n and r's are positions after transformation. The term n^3 is the Jacobian determinant from coordinate transformation. In the above expression, we also introduce an artificial parameter n_{max}, which defines the maximum size of the cluster.

For a specific configuration of clusters, a potential of mean force $W(r^{N_\iota};\mu)$ can be defined by considering all vapor configurations:

$$\begin{aligned} exp[-\beta W(r^{N_\iota};\mu)] \equiv \sum_{N_\nu=0}^{\infty} exp(\beta\mu N_\nu)\frac{1}{\Lambda^{3N_\nu}N_\nu!}\times\int dr^{N_\nu}\prod_{n=1}^{n_{max}}\prod_{j_n=1}^{N_n} w_{j_n}(R_{j_n},r'^{n-1};r^{N_\nu}) \\ \times exp[-\beta U(r^{N_\iota};r^{N_\nu})]. \end{aligned} \tag{2.6}$$

The potential of mean force takes the average of interactions from vapor-like particles. In the case of solution, this potential of mean force can also be defined by averaging over degrees of freedom of solvent molecules. With the definition of the potential of mean force, the partition function can be written as

$$\begin{aligned} \Xi(\mu,V,T) = \sum_{N_1=0}^{\infty}\sum_{N_2=0}^{\infty}\cdots\sum_{N_{n_{max}}=0}^{\infty}\frac{1}{N_1!N_2!\cdots N_{n_{max}}!}\prod_{n=1}^{n_{max}}\left(exp(\beta\mu n)n^3/[\Lambda^{3n}n!]\right)^{N_n} \\ \times\prod_{n=1}^{n_{max}}\left[\int dr'^{n-1}\right]^{N_n}\int\prod_{n=1}^{n_{max}}\prod_{j_n=1}^{N_n} dR_{j_n} exp[-\beta W(r^{N_\iota};\mu)]. \end{aligned} \tag{2.7}$$

The potential of mean force $W(r^{N_\iota};\mu)$ consists of intercluster interactions and intracluster interactions. Assuming the intercluster interaction is pair-wise additive and only depends on positions of the center-of-mass of the clusters (which is valid for dilute systems that are being considered in this dissertation), $W(r^{N_\iota};\mu)$ can be

written as:

$$W(r^{N_\iota};\mu) = W_0 + \sum_n \sum_{j_n=1}^{N_n} W_n(r^{n,j_n};\mu) + \frac{1}{2}\sum_{n,n'}\sum_{j_n,j_{n'}} W_{n,n'}(R_{n,j_n}, R_{n',j_{n'}};\mu). \quad (2.8)$$

Where W_0 is the potential of mean force without clusters, W_n represents the intra-cluster interaction for each cluster and $W_{n,n'}$ is the pair-wise intercluster interaction. With this assumption, the partition function can be further expanded as:

$$\begin{aligned}\Xi(\mu,V,T) =& exp(-\beta W_0)\sum_{N_1=0}^{\infty}\sum_{N_2=0}^{\infty}\cdots\sum_{N_{n_{max}}=0}^{\infty}\frac{1}{N_1!N_2!\cdots N_{n_{max}}!}\prod_{n=1}^{n_{max}}(exp(\beta\mu n)n^3/[\Lambda^{3n}n!])^{N_n}\\ &\times\prod_{n=1}^{n_{max}}\left[\int dr'^{n-1}exp[-\beta W_n(r'^{n-1};\mu)]\right]^{N_n}\\ &\times\int\prod_{n=1}^{n_{max}}\prod_{j_n=1}^{N_n}dR_{j_n}exp[-\beta W_{n,n'}(R_{n,j_n},R_{n',j_{n'}};\mu)].\end{aligned} \quad (2.9)$$

For a system like LJ vapor or solution of weak electrolytes, it is safe to make a further approximation to ignore the intercluster interactions because of the low density of clusters. And Eq. 2.9 can be simplified as:

$$\begin{aligned}\Xi(\mu,V,T) =& exp(-\beta W_0)\sum_{N_1,N_2,\cdots,=0}^{\infty}\prod_n[exp(\beta\mu n N_n]\\ &\times\prod_n\frac{1}{N_n!}\left[\frac{Vn^3}{\Lambda^{3n}n!}\int dr'^{n-1}exp[-\beta W_n(r'^{n-1};\mu)]\right]^{N_n}.\end{aligned} \quad (2.10)$$

In the above equation, the term inside the last bracket is in fact the partition function for a single cluster of size n:

$$Z_n = \frac{Vn^3}{\Lambda^{3n}n!}\int dr'^{n-1}exp[-\beta W_n(r'^{n-1};\mu)]. \quad (2.11)$$

And the entire partition function can be simplified as:

$$\begin{aligned}\Xi(\mu, V, T) &= exp(-\beta W_0) \sum_{N_1, N_2, \cdots, =0}^{\infty} \prod_n \frac{[exp(\beta \mu n) Z_n]^{N_n}}{N_n!} \\ &= exp(-\beta W_0) \prod_n \sum_{N_n=0}^{\infty} \frac{[exp(\beta \mu n) Z_n]^{N_n}}{N_n!} \\ &= exp(-\beta W_0) \prod_n exp(exp[\beta \mu n] Z_n) \\ &= exp(-\beta W_0) exp(\sum_n exp[\beta \mu n] Z_n).\end{aligned} \tag{2.12}$$

Now consider the density of clusters of size n',

$$\begin{aligned}\langle N_{n'} \rangle &= \frac{1}{\Xi(\mu, V, T)} exp(-\beta W_0) \sum_{N_{n'}=0}^{\infty} N_{n'} \frac{[exp(\beta \mu n') Z_{n'}]^{N_{n'}}}{N_{n'}!} \prod_{n \neq n'} \sum_{N_n=0}^{\infty} \frac{[exp(\beta \mu n) Z_n]^{N_n}}{N_n!} \\ &= \frac{\sum_{N_{n'}=0}^{\infty} N_{n'} \frac{[exp(\beta \mu n') Z_{n'}]^{N_{n'}}}{N_{n'}!}}{\sum_{N_{n'}=0}^{\infty} \frac{[exp(\beta \mu n') Z_{n'}]^{N_{n'}}}{N_{n'}!}}.\end{aligned} \tag{2.13}$$

Since $e^x = \sum_n \frac{x^n}{n!}$ and $xe^x = \sum_n n \frac{x^n}{n!}$, $\langle N_{n'} \rangle$ can be simplified as:

$$\langle N_n \rangle = Z_n exp[\beta \mu n] \tag{2.14}$$

Here we replace n′ by n. The left-hand side of Eq. 2.14 is the cluster size distribution and the right-hand side is the partition function for the cluster. If we define the free energy of nucleus at size n as

$$F_n \equiv k_b T \ln Z_n, \tag{2.15}$$

Eq. 2.14 can be rewritten as

$$\langle N_n \rangle = \exp\left[-\beta \left(F_n - n\mu\right)\right] = \exp(-\beta \Delta F). \tag{2.16}$$

The same derivation can be applied to isothermal–isobaric system and Helmholtz free energy change ΔF can be replaced by Gibbs free energy change ΔG.

2.2.2 Grand Canonical Monte Carlo Cluster Sampling

In conjunction with the assumption that cluster-cluster interactions are negligible (appropriate for dilute solutions, such as saturated weak electrolytes), the previous section established the relation between cluster size distribution and nucleation free energy as

$$P_n = \frac{\langle N_n \rangle}{N} = exp[-\beta \Delta G_n] \tag{2.17}$$

Here, P_n is probability of observing a cluster of size n, N_n is the number of clusters of size n, N is the total number of particles in the system and ΔG_n is the Gibbs free energy of cluster of size n (where the reference state is the homogeneous phase). (Note that the neglect of cluster-cluster interactions also implicitly involves the neglect of solute-cluster interactions, as an isolated solute is a merely a cluster of size 1. Nonetheless, these interactions are very modest for sparingly soluble solutes.) Sampling the nucleation free energy barrier is thus equivalent to measuring the cluster size distribution. We utilize a grand canonical Monte Carlo (GCMC)-based approach to sample the distribution of cluster sizes and configurations within the nucleating system. Such an approach requires a definition of a "cluster" to distinguish between agglomerated particles and those in free solution. We utilize Stillinger's[66] cluster criterion, whereby any two solutes within a certain cutoff distance are considered belong to a cluster.

For dilute solutions, the cluster size distribution cannot easily be sampled via conventional unbiased MD simulation due to both system size and transport limitations. However, given the above assumptions, it is possible to obtain the cluster size distribution by sampling over the size(s) and configurations of a single solvated cluster using GCMC. Via GCMC, solute molecules can be attached/detached to/from the surface of the cluster, and cluster size distribution can be efficiently

sampled. The associated acceptance rules are expressed as in conventional GCMC:

$$acc(N \to N+1) = min\left[1, \frac{V}{\Lambda^3(N+1)} exp\{\beta[\mu - U(N+1) + U(N)]\}\right] \quad (2.18)$$

$$acc(N \to N-1) = min\left[1, \frac{\Lambda^3 N}{V} exp\{-\beta[\mu + U(N-1) - U(N)]\}\right] \quad (2.19)$$

Here, V is the volume of the system, Λ is the thermal de Broglie wavelength, and $U(N)$ represents the energy of a solution system containing solute cluster of size N, and μ is the solute chemical potential in the (super)-saturated solution. The neglect of cluster-cluster interactions implies that the system is restricted to sampling a single cluster. This "single cluster" criterion is enforced throughout the simulation. This general approach has been used previously by Chen and Siepmann[52, 49] to study the Lennard-Jones system vapor-liquid and vapor-solid nucleation, and Wu and Deem[73] to examine the nucleation of zeolite. Here we extended this approach to nucleation in explicit solvent.

2.2.3 Aggregation-Volume-Bias Monte Carlo

Traditional GCMC employs random insertion and deletion moves to add/remove particles from the system. While this simple approach works reasonably well for homogeneous systems, it is highly inefficient for adding/removing particles to a growing cluster. In that case, the single cluster criterion would be almost inevitably violated by a solute inserted at a random location.

Here, we utilize aggregation-volume-bias Monte Carlo (AVBMC)[50, 51, 52] to bias the insertion/deletion moves toward the surface of the cluster. AVBMC chooses an existing solute atom (within the cluster) as a reference and inserts/deletes another solute in the vicinity of the reference solute. The AVBMC acceptance rule is expressed as:

$$acc(N \to N+1) = min\left[1, \frac{N \times V_{in} \times exp\{\beta[\mu - U(N+1) + U(N)]\}}{\Lambda^3 \times (N+1) \times (N_{in}+1)}\right] \quad (2.20)$$

$$acc(N \to N-1) = min\left[1, \frac{\Lambda^3 \times N \times N_{in} \times exp\{-\beta[\mu + U(N-1) - U(N)]\}}{(N-1) \times V_{in}}\right] \tag{2.21}$$

Here, V_{in} is volume of surrounding area of the reference particle based on Stillinger's cluster criterion and N_{in} is number of other particles in this region.

The AVBMC protocol ensures that single cluster criterion is satisfied during insertion step, but it may be violated in a deletion step. Thus, after each deletion step, the new configuration will be checked. If the single cluster criterion is violated, the deletion step will be rejected immediately. (This rejection does not violate detailed balance, since the resulting clusters would lie outside the phase space of the single cluster.) The combination of GCMC/AVBMC allows for efficient sampling of cluster sizes and configurations.

2.2.4 Expanded Ensemble and Wang-Landau Sampling

In contrast to prior work, we include explicit solvent molecules in our simulation and simulate charged electrolytes. These factors increase the challenge to efficiently inserting/deleting solutes, due to steric effects (overlapping solute-solvent) and large energy fluctuations (compared to k_bT, due to strong solute-solvent and solute-solute interactions).

Here, we adopt the expanded ensemble method[64, 56, 57] to scale the solute interaction during the course of the insertion/deletion process. In the expanded ensemble method, different states (i.e., solute numbers) are connected via fictitious intermediate states. These unphysical intermediates exist solely to smooth the transition between the (physically meaningful) endpoints and to enhance the sampling of cluster configurations.

Here, we label the physical states by the number of solutes (or ion pairs) in the cluster (N), while intermediate states are labeled by $N + \lambda$, and $\lambda(0 < \lambda < 1)$ is a scale parameter that represents a partially interacting (and partially charged) solute. The number of intermediate states is arbitrary but influences the efficiency of the simulation.

Within the expanded ensemble, we can rewrite the AVBMC-GCMC acceptance rule as:

$$
\begin{aligned}
&acc(\lambda_m \rightarrow \lambda_{m+1}) = \\
&min\left[1, (\frac{N \times V_{in}}{\Lambda^3})^{\lambda_{m+1}-\lambda_m} \frac{((N+\lambda_m) \times (N_{in}+\lambda_m))^{\lambda_m}}{((N+\lambda_{m+1}) \times (N_{in}+\lambda_{m+1}))^{\lambda_{m+1}}}\right. \\
&\left. exp\{\beta[(\lambda_{m+1}-\lambda_m)\mu - U(\lambda_{m+1}) + U(\lambda_m)]\}\right]
\end{aligned}
\tag{2.22}
$$

$$
\begin{aligned}
&acc(\lambda_m \rightarrow \lambda_{m-1}) = \\
&min\left[1, (\frac{N \times V_{in}}{\Lambda^3})^{\lambda_{m-1}-\lambda_m} \frac{((N+\lambda_m) \times (N_{in}+\lambda_m))^{\lambda_m}}{((N+\lambda_{m-1}) \times (N_{in}+\lambda_{m-1}))^{\lambda_{m-1}}}\right. \\
&\left. exp\{\beta[(\lambda_{m-1}-\lambda_m)\mu - U(\lambda_{m-1}) + U(\lambda_m)]\}\right]
\end{aligned}
\tag{2.23}
$$

Here the subscript m represents the mth intermediate state.

The expanded ensemble modifies the transition path between different states, but does not influence the free energy difference between the states. As such, it is unable to efficiently sample states with large variations in free energy, as are expected during nucleation. Thus, to achieve a relatively uniform sampling for all states/cluster sizes, we also employed Wang-Landau sampling method[70], which adds a progressive bias to frequently visited states to achieve a uniform probability distribution; the unbiased cluster free energy distribution can be trivially extracted from the applied (converged) Wang-Landau bias.

2.2.5 Hybrid GCMC/MD

While the GCMC approach allows for an efficient sampling of cluster sizes, it is also necessary to sample both the cluster and solvent configurations. MD provides an efficient approach to sampling over a wide variety of such configurations. (Note that the GCMC insertion/deletion already implicitly samples various cluster configurations, although with limited efficiency.) Briefly, we employ a hybrid GCMC/MD

scheme utilizing alternating steps of GCMC insertion/deletion and MD sampling. We employ constant temperature MD, consistent with the GCMC simulations. The combination of GCMC and MD steps collectively samples all of the relevant phase space.

Note that unlike in AVBMC, where the single cluster criterion is rigorously enforced, the MD can cause this criterion to be violated through particle diffusion or cluster fragmentation. To prevent this, we add constraints designed to enforce the Stillinger's cluster criterion. We utilize constraints based on a minimum spanning tree (MST) algorithm[67]. Upon full insertion/deletion of a ion pair into the cluster, a graph is regenerated using particles as graph nodes and particle-particle distance as graph edges. The MST for the graph is then calculated. For each edge in the MST, a soft-wall constraint potential is added

$$U(r) = k \times max(0, r - r_{max})^2 \tag{2.24}$$

As such, if the distance of two particles of the same edge is within the maximum distance given by the Stillinger's cluster criterion, no interaction will be created between them. However, if two particles diffuse away, a large attractive force will pull them back. As opposed to a hard wall potential, this restraint is non-singular but does allow for small violations of the cluster criterion. To prevent this small violation, we check the cluster criterion after each MD trajectory. In the (unlikely) event of a violation, the trajectory is rolled back and rerun using a new set of initial velocities sampled from a Boltzmann distribution. As shown below, we find that our results are largely insensitive to the details of the constraint, and any artifacts can be further reduced by updating the MST more frequently.

2.2.6 Chemical Potential Calculation

We calculate the chemical potential of the crystalline solid solute phases via the Einstein molecule method.[69, 58]. The solution-phase chemical potentials are calculated via particle insertion. [71] In the particle insertion method, an excess particle is repeatedly and randomly inserted into the N particle system and the

excess chemical potential is given by the ensemble average of the energy difference between (N+1)-particle system and N-particle. As in the case of cluster growth, we also utilize an expanded ensemble and Wang-landau sampling for efficiency. Laaksonen et al. used this approach to calculate the solvation free energy of alkali halide ion pairs.[62, 63, 47]

2.3 Results and Discussion

All MD simulations were carried out using the GPU-accelerated OpenMM software package,[55] in conjunction with a custom C++ interface to carry out the GCMC simulations.

2.3.1 Lennard-Jones Liquid/Vapor Nucleation

We first benchmark our methodology via calculation of the free energy surface for LJ vapor-liquid nucleation, comparing against the earlier work of Chen and Siepmann[52]. In their work, AVBMC/GCMC was employed to sample the cluster size distribution, but (given the absence of solvent) no MD or translational MC moves were carried out to enhance the configurational sampling. Here, we examine the role of hybrid GCMC/MD sampling (and of the required MST restraints) on the calculated free energy.

For LJ vapor-liquid nucleation, no periodic boundary condition (PBC) or cutoffs were applied. Reduced temperatures (T^*) of 0.7 and 0.8 were chosen, along with a cluster criterion of 1.5 σ. The MD simluations utilized Langevin dynamics with a 2.0 fs time step. Between every 2 GCMC steps, 10 steps of MD were carried out. Only 1 intermediate state (with scaling parameter 0.25) was utilized in the expanded ensemble. MST restraints were updated whenever a real (integer) state was sampled by GCMC. The cluster size was sampled from 1 to 200 using 2 windows (1-120 and 80-200). The initial cluster structure for the second window was taken from the snapshot of first window simulation. Four replicate simulations were carried out for each window. Wang-Landau sampling was employed in each simulation

until the Wang-Landau factor converges to 10^{-3}. Subsequently, the cluster size distribution was obtained via 10^7 GCMC steps. WHAM analysis[61] was employed to combine the free energy of different windows using the WHAM package[59]. To better benchmark against reference data[52] and understand the potential artificial effects from MST restraints, GCMC simulations without MD were also carried for this simple system. The results are shown in the Figure 2.1a.

It is clear that both our GCMC and hybrid GCMC/MD results well reproduce the earlier work. However, we find that the hybrid GCMC/MD sampling is five to six times efficient compared to sampling only using GCMC (in terms of both wall time and numbers of GCMC steps required), likely due to the enhanced configurational sampling afforded by the MD. Figure 2.1a does not show any obvious discrepancy between results of hybrid GCMC/MD and GCMC sampling, which indicates that at this temperature ($T^* = 0.7$), the MST restraints do not introduce strong artificial effects on the sampling of cluster configurations.

To better understand MST restraints, we conducted the same type of simulation at higher temperatures ($T^* = 0.75$ and 0.8); the results are shown in the Figure 2.1b and 2.1c. In these cases, we find that our GCMC result essentially perfectly matches the reference data. However, the hybrid GCMC/MD result shows a small deviation of $\sim 2\ k_bT$ at large cluster size, suggesting a small artifact arising from the MST restraints. And this deviation is more pronounced at $T^* = 0.8$. We speculate that, at a low temperature, particles do not have a strong tendency to explore new configurations, and configurations generally (over each of the short MD trajectory) evolve within the framework defined by MST. In contrast, at relatively high temperature, the particles are artificially restricted to a subset of phase space. We find that the influence of the MST restraints can be essentially eliminated by increasing the frequency at which the MST (and the associated constraints) are updated. Updating the MST restraints after every GCMC cycle yields the result in Figure 2.1c, which is in essentially quantitative agreement with the reference data. Note that systems with stronger interactions (e.g., salts) are unlikely to suffer from constraint-induced artifacts even with infrequent MST updates.

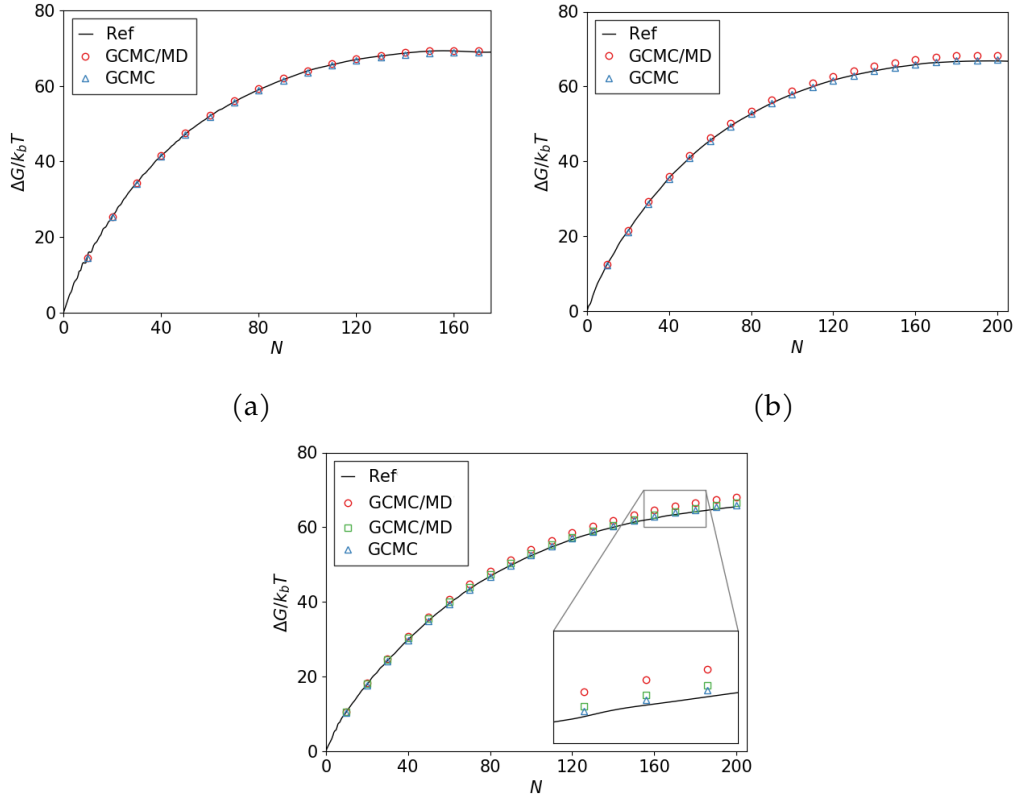


Figure 2.1: (a) Free energy surface for LJ vapor-liquid nucleation at $T^* = 0.7$, $n_v = 5.75 \times 10^{-3}$. The black curve is the reference free energy surface calculated by Chen and Siepmann[52]. The red circles and green triangles correspond to free energies predicted by hybrid GCMC/MD sampling and GCMC sampling, respectively. (b) Free energy surface for LJ vapor-liquid nucleation at $T^* = 0.75$, $n_v = 8.2 \times 10^{-3}$. The black curve is the reference free energy surface calculated by Chen and Siepmann[52]. The red circles and green triangles correspond to free energies predicted by hybrid GCMC/MD sampling and GCMC sampling, respectively. (c) Free energy surface for LJ vapor-liquid nucleation at $T^* = 0.8$, $n_v = 1.1 \times 10^{-2}$. The black curve is the reference free energy surface[52]. The red circles represent free energies predicted by hybrid GCMC/MD, with MST updated upon sampling a physical state. The green squares represent the results with more frequent MST updates after every GCMC step. The blue triangles are results predicted by pure GCMC sampling.

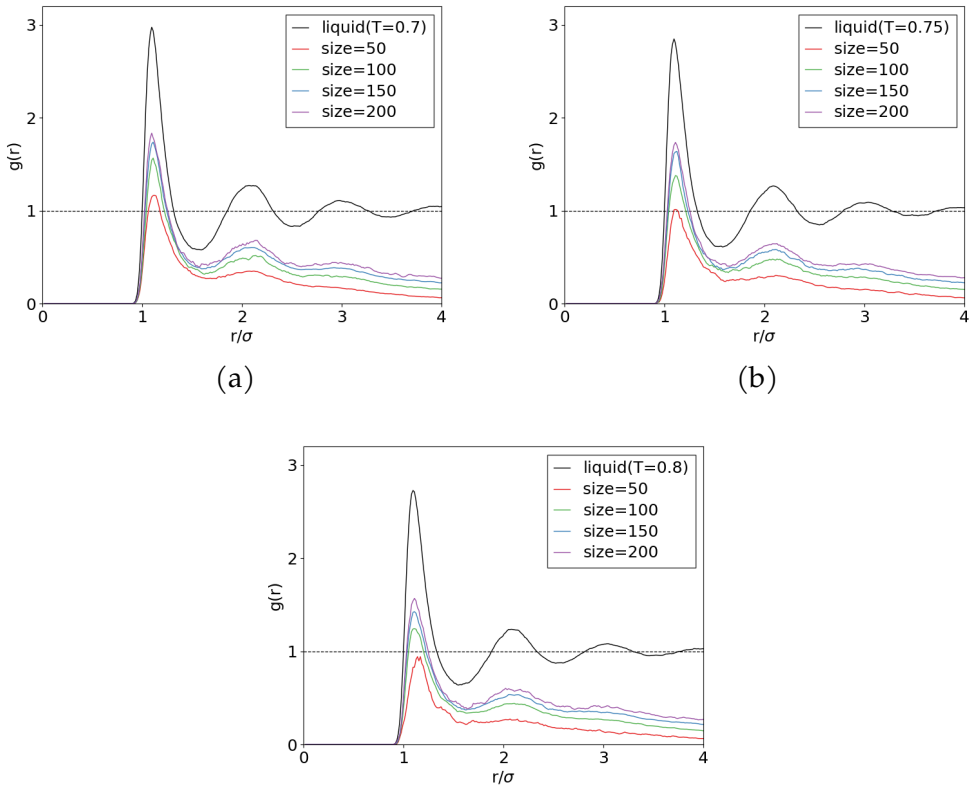
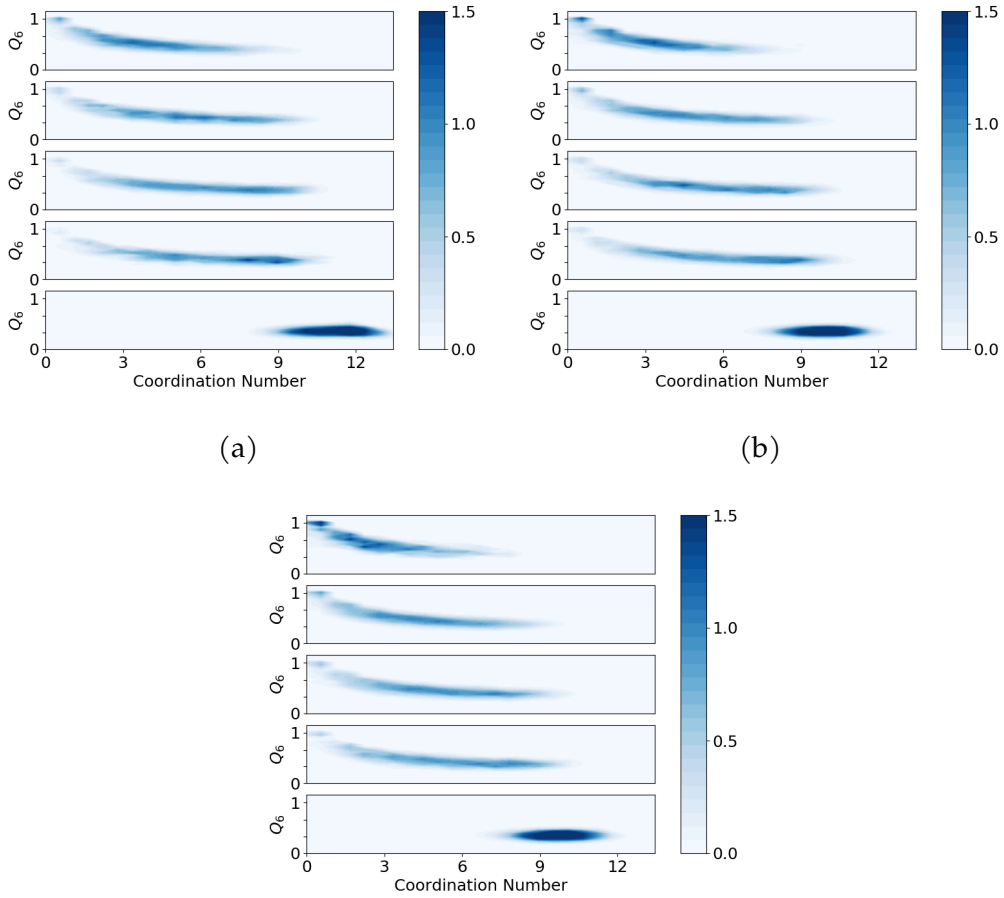


Figure 2.2: Radial distribution functions for liquid and cluster structures at (a) $T^* = 0.7$, (b) $T^* = 0.75$ and (c) $T^* = 0.8$. In all three plots, the black curves are the RDFs for homogeneous liquid structures. The red, green, blue and purple correspond to the cluster structures at size 50,100,150 and 200, respectively.

To characterize the structures of nuclei, we calculated the radial distribution functions (RDFs) for clusters taken from the hybrid GCMC/MD simulations at corresponding temperatures, and compared them against RDFs for LJ liquid. The results are shown in Figure 2.2. We can observe that the positions of the first two peaks are almost identical in all RDFs, which indicates even at small cluster size (50), the nuclei have already expressed the liquid-like characters. Note here, the exact matching between the peak heights of RDFs for nuclei and liquid is not expected here, since the surface particles in nuclei are not fully surrounded by neighboring

particles.



(a) (b)

(c)

Figure 2.3: Particle probability density with respect to coordination number and the Steinhardt parameter, Q_6 at (a) $T^* = 0.7$, (b) $T^* = 0.75$ and (c) $T^* = 0.8$. For each plot, from top to bottom, the subplots correspond to clusters of size 50, 100, 150 and 200 and the homogeneous liquid.

To further support the this conclusion, we also calculated the Steinhardt parameter Q_6[65] and coordination number for particles in clusters using PLUMED[68]; the resulting particle probability distributions with respect to both Q_6 and coordination number are shown in Figure 2.3, plotted alongside the corresponding results

for the bulk liquid. From those plots, we can observe that the particle probability distribution of cluster shifts towards the distribution of liquid as the cluster size increases. And at size 200, some of the particles in the core of the cluster even share the same environment with liquid particles. This excellently agrees with ten Wolde and Frenkel's observation.[72] From those analysis, we can conclude that LJ vapor-liquid nucleation within reduced temperature 0.7~0.8 can be described as a one-step procedure without going through other intermediate states, as predicted by CNT. This observation is consistent with Chen's and ten Wolde's works.[52, 72]

2.3.2 Lennard-Jones Liquid/Solid Nucleation

We also applied this approach to LJ vapor-solid nucleation at $T^* = 0.6$. For LJ vapor-solid nucleation, PBCs were applied in conjunction with a cutoff of 4.5 σ; no long-range interaction were considered. Other simulation details are identical to the vapor-liquid simulation. LJ solid at this condition forms FCC (face-centered cubic) lattice. We calculated the crystalline chemical potential using 4000 LJ particles. The spring strength of Einstein crystal was set as 8000 k_bT/A^2. In the free energy perturbation step of Einstein molecule method, 20ns NVT simulation was carried out for Einstein crystal. During the thermodynamic integration step, 20 λ values were chosen from Gaussian quadrature on 0 to 1. For each λ value, a 4ns NVT simulation was carried out.

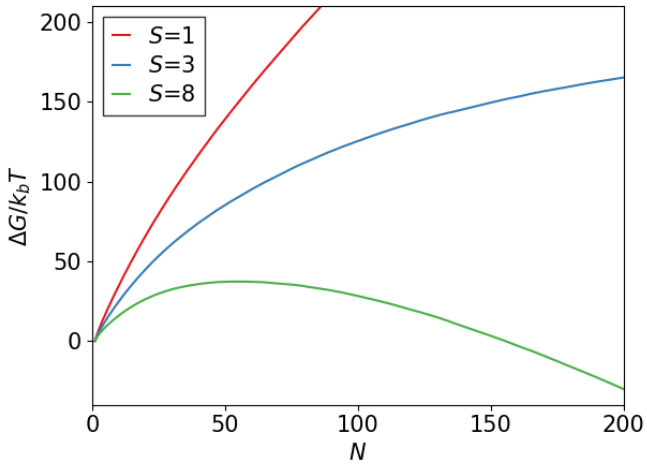
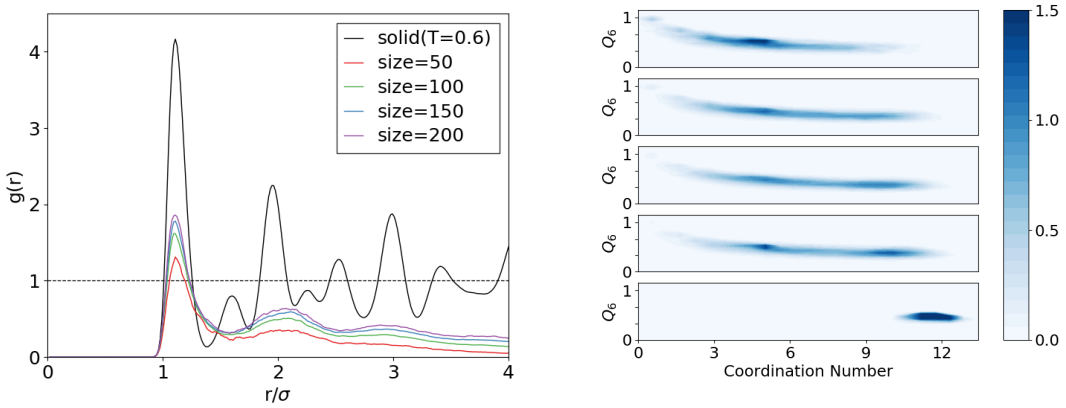


Figure 2.4: Free energy surface for LJ vapor-solid nucleation at $T^* = 0.6$. Red, blue and green lines are free energy surfaces corresponding to supersaturation of $S = 1, 3$ and 8. N represents the cluster size in all three plots.

We calculate the chemical potential of the solid (and saturated vapor, taken as Ar) under these conditions to be $\mu_{LJ} = -14.737 \pm 0.001$ k_bT. The calculated free energy surface at saturation, as well as for supersaturation ratios of $S = 3$ and 8 are shown in Figure 2.4. At saturation, the free energy monotonically increases with the growth of cluster size, consistent with a positive surface term (since there is no driving force for bulk nucleation). At a supersaturation ratio of $S = 3$, the free energy surface becomes more flat, exhibiting a critical cluster size beyond 200. At higher supersaturations ($S = 8$), the critical cluster size is located at $N = 55$ corresponding to a 37.3 ± 0.1 k_bT free energy barrier. It should be noticed here, since we ignore the intercluster interactions, at fixed temperature, the free energy barrier and critical cluster size is only determined by supersaturation, or pressure in this case.



(a) (b)

Figure 2.5: (a) RDFs for LJ crystal and clusters at $T^* = 0.6$. The black curve is the RDF for bulk crystal. The red, green, blue and purple correspond to the cluster structures at size 50,100,150 and 200, respectively. (b) Particle probability density with respect to coordination number and the Steinhardt parameter, Q_6: (top to bottom) clusters of size 50, 100, 150 and 200, and the bulk crystal.

We also plot the RDFs and particle probability distribution for clusters and crystal at $T^* = 0.6$, as shown in Figure 2.5. The RDFs of clusters clearly represent the structure of liquid rather than crystal. For this reason, we believe the vapor-solid nucleation at moderate undercooling (13% with respect to the bulk-phase triple point of 0.689[49]) proceeds through an intermediate state instead of directly transforming to the crystalline phase. This is also indicated in the particle probability distribution. In the plot of particle probability distribution, we can observe that at size 200, some particles in the core of the cluster present similar coordination numbers as in FCC structure, but the average Q6 value for those particles is only 0.39 while the Q6 for FCC crystal is 0.52. Therefore, those clusters cannot be considered as crystalline precursors. And the vapor-solid nucleation should be described by a two-step model.

2.4 conclusion

We have developed a hybrid grand canonical Monte Carlo/molecular dynamics (GCMC/MD) approach in combination with aggregation-volume-bias, expanded ensemble and Wang-Landau sampling. It circumvents the challenges of dilute systems, mass transport, and nucleation barriers in crystal nucleation. The method was benchmarked against Lennard-Jones model system and excellent agreement was achieved when comparing with reference data. In the next chapter, we will extend its application to an ionic system with explicit solvent.

REFERENCES

[47] Åberg, K Magnus, Alexander P Lyubartsev, Sven P Jacobsson, and Aatto Laaksonen. 2004. Determination of solvation free energies by adaptive expanded ensemble molecular dynamics. *The J. Chem. Phys.* 120(8):3770–3776.

[48] Alejandre, José, and Jean Pierre Hansen. 2007. Ions in water: From ion clustering to crystal nucleation. *Phys. Rev. E* 76(6):061505.

[49] Chen, Bin, Hyunmi Kim, Samuel J Keasler, and Ricky B Nellas. 2008. An aggregation-volume-bias monte carlo investigation on the condensation of a lennard-jones vapor below the triple point and crystal nucleation in cluster systems: an in-depth evaluation of the classical nucleation theory. *The J. Phys. Chem. B* 112(13):4067–4078.

[50] Chen, Bin, and J Ilja Siepmann. 2000. A novel monte carlo algorithm for simulating strongly associating fluids: Applications to water, hydrogen fluoride, and acetic acid. *The J. Phys. Chem. B* 104(36):8725–8734.

[51] ———. 2001. Improving the efficiency of the aggregation-volume-bias monte carlo algorithm. *The J. Phys. Chem. B* 105(45):11275–11282.

[52] Chen, Bin, J Ilja Siepmann, Kwang J Oh, and Michael L Klein. 2001. Aggregation-volume-bias monte carlo simulations of vapor-liquid nucleation barriers for lennard-jonesium. *The J. Chem. Phys.* 115(23):10903–10913.

[53] Cooper, Emily R, Christopher D Andrews, Paul S Wheatley, Paul B Webb, Philip Wormald, and Russell E Morris. 2004. Ionic liquids and eutectic mixtures as solvent and template in synthesis of zeolite analogues. *Nature* 430(7003):1012.

[54] Ding, Ran, Chao Huang, Jingjing Lu, Junning Wang, Chuanjun Song, Jie Wu, Hongwei Hou, and Yaoting Fan. 2015. Solvent templates induced porous

metal–organic materials: conformational isomerism and catalytic activity. *Inorg. Chem.* 54(4):1405–1413.

[55] Eastman, Peter, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. 2017. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* 13(7): e1005659.

[56] Escobedo, Fernando A, and Juan J de Pablo. 1995. Monte carlo simulation of the chemical potential of polymers in an expanded ensemble. *The J. Chem. Phys.* 103(7):2703–2710.

[57] ———. 1996. Expanded grand canonical and gibbs ensemble monte carlo simulation of polymers. *The J. Chem. Phys.* 105(10):4391–4394.

[58] Frenkel, Daan, and Anthony JC Ladd. 1984. New monte carlo method to compute the free energy of arbitrary solids. application to the fcc and hcp phases of hard spheres. *The J. Chem. Phys.* 81(7):3188–3193.

[59] Grossfield, Alan. 2012. Wham: the weighted histogram analysis method. *version* 2(9):06.

[60] Henzler, Katja, Evgenii O Fetisov, Mirza Galib, Marcel D Baer, Benjamin A Legg, Camelia Borca, Jacinta M Xto, Sonia Pin, John L Fulton, Gregory K Schenter, et al. 2018. Supersaturated calcium carbonate solutions are classical. *Sci. Adv.* 4(1):eaao6283.

[61] Kumar, Shankar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. 1992. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Comput. Chem.* 13(8):1011–1021.

[62] Lyubartsev, Alexander P, Aatto Laaksonen, and Pavel N Vorontsov-Velyaminov. 1996. Determination of free energy from chemical potentials: Application of the expanded ensemble method. *Mol. Simul.* 18(1-2):43–58.

[63] Lyubartsev, AP, OK Fo/rrisdahl, and A Laaksonen. 1998. Solvation free energies of methane and alkali halide ion pairs: An expanded ensemble molecular dynamics simulation study. *The J. Chem. Phys.* 108(1):227–233.

[64] Lyubartsev, AP, AA Martsinovski, SV Shevkunov, and PN Vorontsov-Velyaminov. 1992. New approach to monte carlo calculation of the free energy: Method of expanded ensembles. *The J. Chem. Phys.* 96(3):1776–1783.

[65] Steinhardt, Paul J, David R Nelson, and Marco Ronchetti. 1983. Bond-orientational order in liquids and glasses. *Phys. Rev. B* 28(2):784.

[66] Stillinger Jr, Frank H. 1963. Rigorous basis of the frenkel-band theory of association equilibrium. *The J. Chem. Phys.* 38(7):1486–1494.

[67] Strachan, A, and CO Dorso. 1997. Fragment recognition in molecular dynamics. *Phys. Rev. C* 56(2):995.

[68] Tribello, Gareth A, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. 2014. Plumed 2: New feathers for an old bird. *Comput. Phys. Commun.* 185(2):604–613.

[69] Vega, Carlos, and Eva G Noya. 2007. Revisiting the frenkel-ladd method to compute the free energy of solids: The einstein molecule approach. *The J. Chem. Phys.* 127(15):154113.

[70] Wang, Fugao, and DP Landau. 2001. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* 86(10):2050.

[71] Widom, Ben. 1963. Some topics in the theory of fluids. *The J. Chem. Phys.* 39(11):2808–2812.

[72] ten Wolde, Pieter Rein, and Daan Frenkel. 1998. Computer simulation study of gas–liquid nucleation in a lennard-jones system. *The J. Chem. Phys.* 109(22): 9901–9918.

[73] Wu, Minghong G, and Michael W Deem. 2002. Monte carlo study of the nucleation process during zeolite synthesis. *The J. Chem. Phys.* 116(5):2125–2137.

3 APPLICATION OF HYBRID GCMC/MD APPROACH FOR IONIC SYSTEM IN SOLUTION PHASE

3.1 Introduction

Despite its accurate prediction for nucleation in LJ system, our hybrid GCMC/MD approach has not been validated for crystal nucleation in solution phase, with explicit consideration of solvents and Coulomb interactions. To this end, we employed a low-solubility rock-salt model[74] which serves as a simple but complete representation of actual crystals. The model was adapted from an existed NaCl force field[85, 78] with modification to reduce solubility. Accurate description of the nucleation for this model requires the method to fully address the challenges from explicit solvents, nucleation barriers and ineffective phase space samplings associated with the ionic nucleus.

In the following sections, we first extend our derivation of hybrid GCMC/MD methodology to ionic systems, in which cations and anions are considered as different species and inserted separately. The solubility of this rock-salt model is estimated by equating the solution- and solid-phase chemical potentials obtained by particle insertion and Einstein molecular method, respectively. The nucleation free energy is evaluated from our hybrid GCMC/MD approach and structural analysis is employed on the nuclei to help identify the nucleation pathway. The nucleation rate is calculated from the free energy barrier and compared with the results from brute-force MD simulations. Eventually, we extrapolate the rate calculation to moderate supersaturation and compare it with experimental NaCl nucleation.

3.2 Theory

3.2.1 GCMC for Electrolytes

In the above sections, we developed the AVBMC-EE-GCMC algorithm for a system of single species. For systems consisting of two-component electrolytes such as NaCl, this methodology needs to be adapted accordingly. During the insertion of each ion pair, the interaction of cation is gradually turned on first. The pairing anion will be inserted only after the the cation has been fully inserted. Correspondingly, in the deletion step, the cation will be removed following the deletion of anion. Note that the free energies of the (physically meaningful) states are unaffected by the use of non-charge-neutral intermediates states.

We adapted the original AVBMC procedure as follows (using a cation as an example): (i) Select a random anion as a reference particle. (ii) Calculate the volume of the region surrounding the reference anion based on Stillinger's cluster criterion (V_{in}) and count number of other cations ($N_{in,c}$) in this region. (iii) Randomly insert a new cation in this region/remove a random particle from this region. (iv) Calculate the energy difference (ΔE) between new state and old state. The corresponding acceptance rules are expressed as follows (Detailed derivation is shown in Section 3.2.2):

$$acc(N_c \to N_c + 1) = min\left[1, \frac{V_{in} \times exp\{\beta[\frac{\mu}{2} - U(N_c + 1) + U(N_c)]\}}{{\Lambda_c}^3 \times (N_{in,c} + 1)}\right] \tag{3.1}$$

$$acc(N_c \to N_c - 1) = min\left[1, \frac{{\Lambda_c}^3 \times N_{in,c} \times exp\{-\beta[\frac{\mu}{2} + U(N_c - 1) - U(N_c)]\}}{V_{in}}\right] \tag{3.2}$$

Here, N_c is the number of cations in the cluster, Λ_c is the thermal de Broglie wavelength for the cation and μ is the chemical potential for each ion pair. Note that it is impossible to rigorously decompose μ into specific cation/anion contributions. Here, we arbitrarily utilize $\frac{\mu}{2}$ for the cation/anion, which influences the energy of the (charged) intermediate states, but not the physically meaningful, neutral states.

We find that our results are invariant to the details of this insertion process or the order of ion insertion (see Section 3.3.3). For expanded ensemble system, the above acceptance rules can be written as follows:

$$\begin{aligned} acc(\lambda_m \to \lambda_{m+1}) = min\Bigg[1, \quad & (\frac{V_{in}}{{\Lambda_c}^3})^{\lambda_{m+1}-\lambda_m} \frac{(N_{in,c}+\lambda_m)^{\lambda_m}}{(N_{in,c}+\lambda_{m+1})^{\lambda_{m+1}}} \\ & exp\{\beta[(\lambda_{m+1}-\lambda_m)\frac{\mu}{2} - U(\lambda_{m+1}) + U(\lambda_m)]\}\Bigg] \end{aligned} \tag{3.3}$$

$$\begin{aligned} acc(\lambda_m \to \lambda_{m-1}) = min\Bigg[1, \quad & (\frac{V_{in}}{{\Lambda_c}^3})^{\lambda_{m-1}-\lambda_m} \frac{(N_{in,c}+\lambda_m)^{\lambda_m}}{(N_{in,c}+\lambda_{m-1})^{\lambda_{m-1}}} \\ & exp\{\beta[(\lambda_{m-1}-\lambda_m)\frac{\mu}{2} - U(\lambda_{m-1}) + U(\lambda_m)]\Bigg] \end{aligned} \tag{3.4}$$

3.2.2 Derivation of AVBMC-GCMC for Ion Pairs

In this section, we give a detailed derivation for Eq. 3.1 and 3.2. In Monte Carlo algorithm, the condition of detailed balance is imposed as the following equation:

$$\mathcal{K}(o \to n) \quad = \quad \mathcal{K}(n \to o), \tag{3.5}$$

where $\mathcal{K}(o \to n)$ is the flow of being in the configuration from an old state o to a new state n. And it is defined as the product of the probability of being in the current state o, $\mathcal{N}(o)$, the transition probability of generating a move from state o to state n, $\alpha(o \to n)$ and the probability of accepting this move, $acc(o \to n)$:

$$\mathcal{K}(o \to n) \quad = \quad \mathcal{N}(o) \times \alpha(o \to n) \times acc(o \to n), \tag{3.6}$$

According to cluster size distribution (Eq. 2.14), the probability of finding the cluster of size N is given by

$$\mathcal{N}(N) = Z_N exp[\beta\mu N], \tag{3.7}$$

where Z_N is defined as

$$Z_N = \frac{1}{(\Lambda_c\Lambda_a)^{3N}N!^2\Lambda_{solv}{}^{3N_{solv}}N_{solv}!}\int dr^{6N+3N_{solv}}w(r^{6N})exp[-\beta U(r^{6N},r^{3N_{solv}})] \tag{3.8}$$

for ion pairs in solution. In this expression, N is the size of the cluster/number of ion pairs and N_{solv} is the number of the solvents. Λ is the thermal de Broglie wavelength and subscription c, a and $solv$ stands for cation, anion and solvent. w in the integrand is the weight function corresponding to the cluster criterion. The number of solvents is constant during the simulation. Thus,

$$\mathcal{N}(N) \propto \frac{exp[\beta\mu N]}{(\Lambda_c\Lambda_a)^{3N}N!^2}\int dr^{6N+3N_{solv}}w(r^{6N})exp[-\beta U(r^{6N},r^{3N_{solv}})] \tag{3.9}$$

In our procedure of inserting an ion pair, the cation is first inserted according to the following scheme:

1. Select a random anion as a reference particle.
2. Calculate the volume of the region surrounding the reference anion based on Stillinger's cluster criterion (V_{in}).
3. Randomly insert a new cation into this region.

The total volume for inserting the cation is $N_a \times V_{in}$ where N_a is number of anions and it equals to N. So the probability of generating an insertion step for cation is given by

$$\alpha(N_c \to N_c + 1) \quad = \quad \frac{dr^3}{N_a \times V_{in}}\frac{1}{N_c+1}, \tag{3.10}$$

where $dr^3/N_a \times V_{in}$ is the probability from selecting a specific position in volume $N_a \times V_{in}$ and $1/N_c + 1$ is included here to remove the distinguishability brought by the new inserted cation.

For the reverse step, the deletion of one cation from $N_c + 1$ cations is proceeded as follows:

1. Select a random anion as a reference particle.
2. Calculate the volume of the region surrounding the reference anion based on Stillinger's cluster criterion (V_{in}) and count number of cations ($N_{in,c} + 1$) in this

region.
3. Remove a random cation from this region.
The probability of generating such deletion step is given by

$$\alpha(N_c + 1 \to N_c) \quad = \quad \frac{1}{N_a}\frac{1}{N_{in,c}+1}, \tag{3.11}$$

where $1/N_a$ is from selecting a reference anion from N_a anions and $1/N_{in,c}+1$ is from selecting a specific cation from all cations around this reference anion.

Following the detailed balance and Metropolis algorithm[84, 81], the acceptance probability is given by

$$\begin{aligned} acc(N_c + 1 \to N_c) &= min\left[1, \frac{N(N_c+1)}{N(N_c)}\frac{\alpha(N_c \to N_c+1)}{\alpha(N_c+1 \to N_c)}\right] \\ &= min\left[1, \frac{V_{in} \times exp\{\beta[\frac{\mu}{2} - U(N_c+1) + U(N_c)]\}}{{\Lambda_c}^3 \times (N_{in,c}+1)}\right] \end{aligned} \tag{3.12}$$

And the acceptance probability for deleting a cation is

$$acc(N_c \to N_c - 1) = min\left[1, \frac{{\Lambda_c}^3 \times N_{in,c} \times exp\{-\beta[\frac{\mu}{2} + U(N_c-1) - U(N_c)]\}}{V_{in}}\right] \tag{3.13}$$

Note we arbitrarily utilize $\frac{\mu}{2}$ for the cation/anion and it will not influence the physically meaningful state.

For the insertion and deletion of anion, following the same procedure, we will get

$$acc(N_a + 1 \to N_a) \quad = \quad min\left[1, \frac{V_{in} \times exp\{\beta[\frac{\mu}{2} - U(N_a+1) + U(N_a)]\}}{{\Lambda_a}^3 \times (N_{in,a}+1)}\right] \tag{3.14}$$

$$acc(N_a \to N_a - 1) = min\left[1, \frac{{\Lambda_a}^3 \times N_{in,a} \times exp\{-\beta[\frac{\mu}{2} + U(N_a-1) - U(N_a)]\}}{V_{in}}\right] \tag{3.15}$$

The only difference is in the derivation, the number of the reference cations N_c equals N+1 instead of N.

3.3 Results and Discussion

For our low-solubility rock-salt system, we adopted Alejandre and Hansen's set of force fields,[74] consisting of a modified NaCl model[85, 78] in conjunction with the SPC/E model[76]. In order to enhance aggregation (i.e., reduce solubility), Alejandre and Hansen added an LJ interaction site to the H atoms to reduce the Cl-H attraction; we retain this modification.[74] The final force field parameters are listed in Table 3.1. All LJ interactions are calculated using Lorentz-Berthelot rule. In contrast to ref 74, we did not omit H-H and H-O LJ interaction, for computational convenience. As shown in next section, unlike the original NaCl force field, the modified model is a low-solubility weak electrolyte and thus an ideal candidate for the present study.

PBCs were employed with LJ cutoff of 1.5 nm. No long-range corrections were considered. Particle mesh Ewald (PME) summations were carried out for Coulomb interactions. Temperature was chosen at room temperature (298.15 K), enforced using a Langevin thermostat. During the hybrid GCMC/MD simulation, the pressure was fixed at 1 atm using a Monte Carlo barostat. The Stillinger's cluster criterion was chosen at 0.35 nm and only connections between cations and anions are considered. A total 2500 water molecules were positioned around NaCl cluster. Between every 2 GCMC steps, 10 steps of MD were carried out with a 2 fs time step. Cations and anions are inserted and deleted individually. For each ion, four intermediates with quadratically distributed scaling parameters were utilized for the expanded ensemble. Thus, for each ion pair, there were 9 intermediate states in total, including a state containing fully inserted cation. MST restraints were updated whenever a physical (integer) state was sampled by GCMC. The cluster size was sampled from 1 to 40 using 4 windows (1-16, 8-24, 16-32 and 24-40). Except for the first window, the initial cluster structure for each window was taken from a snapshot of the simulation of the previous window. four replicate simulations were carried

Table 3.1: LJ and Coulomb parameters of NaCl and water. σ and ϵ are in Å and kJ/mol, respectively.

	charge	σ	ϵ
Na	+1	0.2584	0.4184
Cl	-1	0.4036	0.4184
O	-0.8476	0.3166	0.6498
H	+0.4238	0.065	0.1663

out for each window. Wang-Landau sampling was employed in each simulation until the Wang-Landau factor converged to 10^{-4}. The biases were collected and analyzed using WHAM.

The chemical potential for solid- and solution-phase salt were also calculated. For particle insertion, most simulation details are unchanged from the hybrid GCMC/MD, except that the volume is fixed during the entire simulation. After Wang-Landau sampling, 2×10^6 MC steps were carried out to refine the chemical potential. For the Einstein molecule approach, 4000 NaCl ion pairs were used. The spring strength of Einstein crystal was set as 8000 $k_b T/A^2$. In the free energy perturbation step, a 20 ns NVT simulation was carried out for Einstein crystal. During thermodynamic integration, 20 λ values were chosen from Gaussian quadrature on 0 to 1. For each λ value, a 4ns NVT simulation was carried out.

3.3.1 MST restraints for NaCl

In Chapter 2, by comparing the nucleation free energy surface predicted by hybrid GCMC/MD method with reference data for the LJ systems, we indirectly proved that the artificial effects of MST restraints can be reduced to a minimum level as long as the restraints are frequently updated. To verify whether this conclusion is also valid for the NaCl system, here we carried out 400 ps simulations for 100 NaCl clusters of size 5 with MST restraints both fixed and updated every 200 fs, and corresponding trajectories of the average value of the radius of gyration (R_g) are plotted in Figure 3.1.

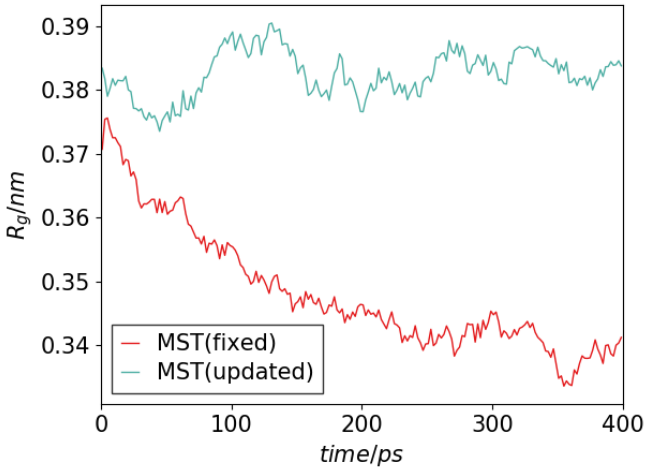


Figure 3.1: Trajectories of the average value of R_g for NaCl clusters at size 5 with MST restraints fixed (red) and updated every 200 fs (green).

From the plot, we observe that the fixed MST restraints make the clusters more compact due to the artificial effects as indicated in the gradually reduced R_g. And we think the reason for such effects is that there are different numbers of "graphs" (MSTs) allowed for compact vs extended clusters (in particular, more graphs allowed for the former), and enforcing a single graph thus biases the simulation toward compact structures. An alternative way of explaining this is that there is more overlap in the phase space of the graphs for the compact structures, systematically biasing the results that way. Fortunately such effects can be eliminated by updating MST restraints every 200 fs as indicated in the green curve. In addition to the indirect evidence from Chapter 2, the observations here directly proves that the frequently updated MST restraints will not affect the structure of the nucleus.

3.3.2 Solubility Estimation for NaCl

Prior to examining the nucleation of our rock-salt system, we first calculated the salt model in SPC/E-H water. We find the solubility by equating the solution- and solid-phase chemical potentials. By thermodynamic integration to an Einstein crystal,

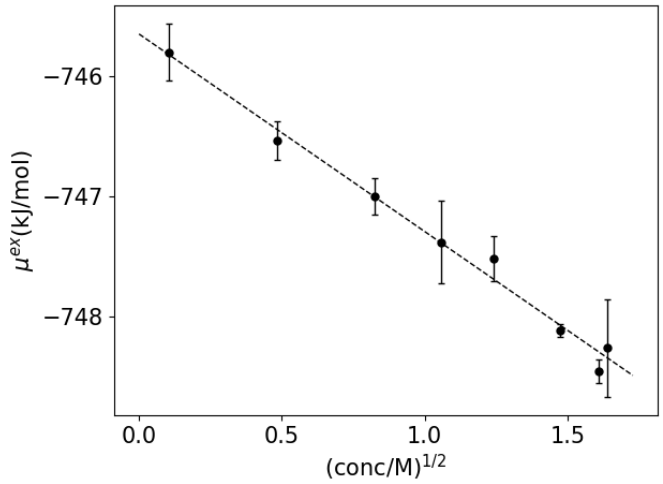


Figure 3.2: Excess chemical potential μ^{ex} of solute plotted with the square root of concentration. Black dots are the excess chemical potentials calculated via particle insertion. The dashed line is the linear fit of the excess chemical potential with the square root of concentration.

we obtained the chemical potential $\mu = -825.29$ kJ/mol for the solid. To obtain the solute chemical potential of this model as a function of concentration, we calculated the excess chemical potential, μ^{ex}, for the solute in a range of concentrations and extrapolated it using Debye-Hückel theory. The total chemical potential is given by adding on the ideal gas chemical potential, μ^{id}.

According to Debye-Hückel theory, in dilute limit, the logarithm of activity coefficient is linear with the square root of ionic strength (and concentration). Thus, we linearly fit excess chemical potential and square root of concentration as shown in Figure 3.2. By extrapolation, we estimated the solubility of this model as 0.027 M, with a similar solubility to LiF, another rock-salt structure weak electrolyte. This salt force field is thus ideal for the present study since its low solubility ensures that cluster-cluster (and cluster-solute) interactions should be negligible (a key assumption), but the solubility remains high enough that our estimated nucleation rates can be validated against large-scale brute-force MD simulation (vide infra).

The chemical potentials of the solute at other concentrations were also calculated

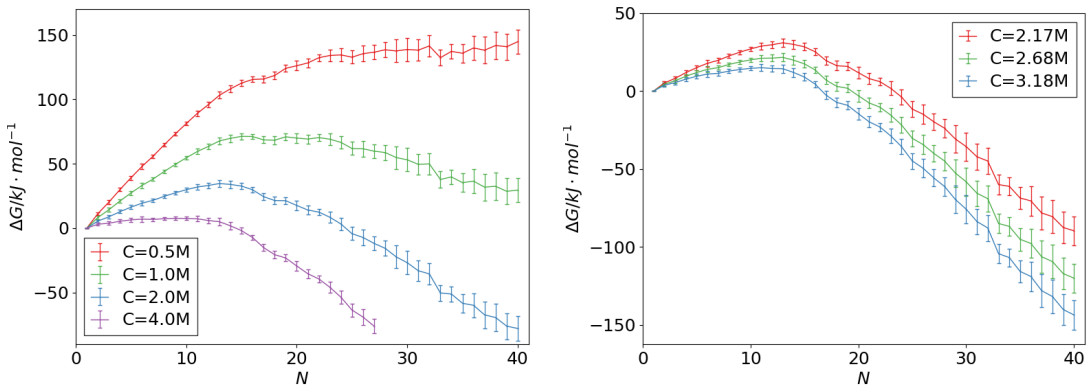
Table 3.2: Chemical potentials of solute at different concentrations. All units in kJ/mol.

C(M)	0.5	1.0	2.0	2.16	2.68	3.18	4.0
μ^{id}	-65.0	-61.5	-58.1	-57.7	-56.6	-55.8	-64.7
μ^{ex}	-746.8	-747.3	-748.0	-748.1	-748.3	-748.6	-758.9
μ^{tot}	-811.8	-808.8	-806.1	-805.8	-805.0	-804.4	-803.6

and used as input to generate nucleation free energy surfaces at various supersaturations. Selected concentrations and corresponding chemical potentials are listed in Table 3.2

3.3.3 Nucleation Free Energy Surface

Having determined the solute chemical potentials, we then used our hybrid GCMC/MD method to calculate the nucleation free energy surface for the salt in water; results are shown in Figure 3.3. In the left panel, the free energy surfaces at solute concentrations of 0.5, 1.0, 2.0 and 4.0 M are plotted. The free energy barrier for nucleation, $\Delta G^{\neq}$, at 2.0 M is 34.7 ± 2.7 kJ/mol, and is reduced to 7.6 ± 1.3 kJ/mol under large supersaturation (4.0 M). Consistent with these barriers, we find that large-scale brute-force MD under these same conditions leads to nucleation which is essentially unobservable (within microsecond time scales) and extremely rapid (within a nanosecond), respectively. Therefore, we calculated also several additional free energy surfaces at intermediate concentrations of 2.17, 2.68 and 3.18 M, corresponding to 100,125 and 150 NaCl ion pairs in 2500 waters. The resulting nucleation free energy surfaces are plotted in the right panel of Figure 3.3. The corresponding critical cluster size and free energy barriers are listed in Table 3.3.



(a) (b)

Figure 3.3: NaCl nucleation free energy surfaces at various concentrations. (a) Nucleation free energy surfaces at 0.5, 1.0, 2.0 and 4.0 M. (b) Nucleation free energy surfaces at 2.17, 2.68 and 3.18 M. N is the cluster size (number of ion pairs) in both plots.

To understand the effect of inserting sequence, we also compared the free energy surfaces obtained through different inserting sequences(Figure 3.4) at concentration of 2.68 M. From this comparison, we can conclude that the details of the insertion procedure can be ignored and the ordering (and character) of the intermediate states does not affect the sampling, as expected, even for non charge-neutral states.

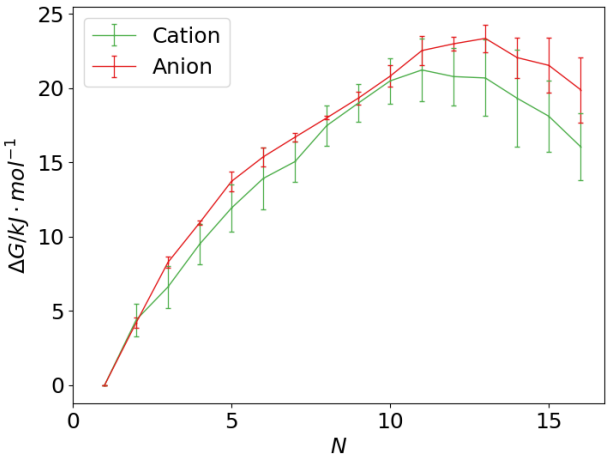
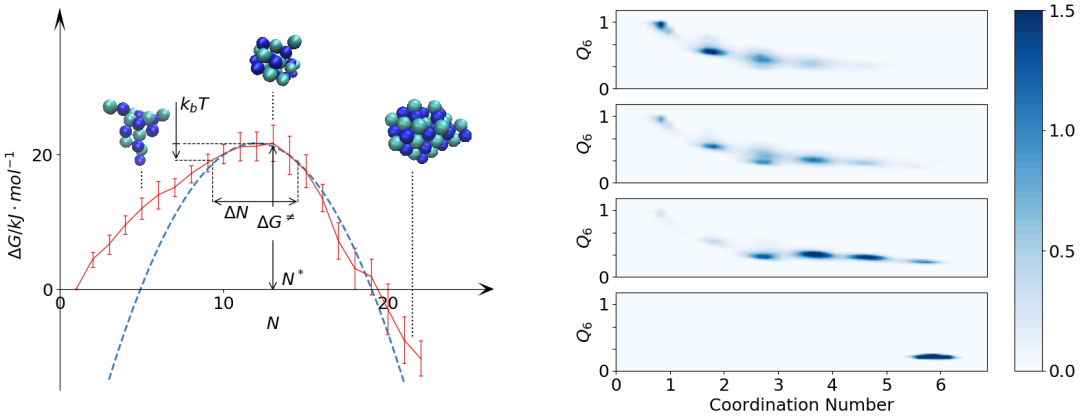


Figure 3.4: Free energy surfaces obtained by inserting cation first(red) and inserting anion first(green).

3.3.4 Structural Analysis for Clusters

Snapshots of the cluster structures at different nucleation stages were saved from the GCMC/MD simulation. In Figure 3.5a, snapshots corresponding to prenucleation, the critical cluster size, and postnucleation are superimposed on the free energy surface from the 2.68 M simulation. The first snapshot were taken at cluster size $N = 8$. At this prenucleation stage, Na^+ and Cl^- form a loose and amorphous structure. Although no clear ordered bulk structure exists inside the cluster, the single layer of NaCl still preserves a rock-salt (100) surface structure. For both the critical cluster and postnucleation stage, clear rock-salt structures were observed. To further quantify the evolving coordination and local order of the growing cluster, we calculated the Steinhardt parameter[86] Q_6 and coordination number for ions in clusters using PLUMED[87]; the resulting ion probability distribution with respect to both Q_6 and coordination number is shown in Figure 3.5b, plotted alongside the corresponding results for the bulk crystal (with a bulk coordination number of 6 and small Q_6). From this plot, we observe that, with increasing cluster size, the fraction of bulk-like ions increases, exhibiting both full octahedral coordination and



(a) (b)

Figure 3.5: (a) Nucleation free energy surface analysis and cluster structures. The red solid line is the nucleation free energy surface at 2.68 M. The blue dashed line is a parabolic fit in the region close to the free energy barrier at critical cluster size. Snapshots of clusters with N $= 8, 13$ and 24 ion pairs are shown on top of the free energy curve. Na^+ and Cl^- are colored in blue and green, respectively. (b) Ion probability density with respect to coordination number and the Steinhardt parameter, Q_6: (top to bottom) clusters of size $8, 13$ and 24, and the bulk crystal.

appropriate local order. Therefore, we believe that, at this concentration (and for this salt model), nucleation is governed by CNT rather than a two-step transition mechanism (which would involve a transformation from an initial amorphous to final crystalline structure). This conclusion is consistent with results from Juang et al.,[82], who found that nucleation in aqueous NaCl solutions shifts from a one-step mechanism to a two-step mechanism on crossing the spinodal.

3.3.5 Nucleation Rate Estimation from Free Energy Surface

Given the calculated free energy barriers, we can estimate the nucleation rate via the analytic theory[75, 77]:

$$r = \beta Z C exp\left(-\frac{\Delta G^{\neq}}{k_b T}\right) \tag{3.16}$$

Here, C represents concentration, and Z is the Zeldovich factor, which characterizes the flatness of free energy surface at the critical cluster size, which is related to a recrossing factor of the nucleation process:

$$Z = \sqrt{-\frac{1}{2\pi k_b T}\frac{\partial^2 \Delta G_n}{\partial n^2}\bigg|_{n=n^*}} \tag{3.17}$$

We applied parabolic fit of the free energy surface near critical cluster size as shown in Figure 3.5, to estimate the curvature of ΔG with respect to the cluster size. β is the growth rate of the critical cluster which can be estimated via[83]:

$$\beta = 4\pi r^* \frac{D_i}{\Omega}\frac{x_i^0}{y_i^e} \tag{3.18}$$

where r^* is radius of critical cluster and D_i is the diffusion for rate-limiting agent i. Here we estimate r^* by assuming a spherical critical cluster, and take the diffusion constants of the ions from the sum of their MD-calculated values. Ω is the volume corresponding to one atomic site, which can be directly calculated from Stillinger's cluster criterion. x_i^0 and y_i^e are the respective atomic fraction in the solution phase and solid phase. Calculated Z and β values and the resulting calculated nucleation rates are listed in Table 3.3. Note that the rates are given in units of #nucleation events/(simulation box·ns), which facilitates the direct comparison of the nucleation rates with those observed from large-scale brute-force MD in the next section.

Table 3.3: Nucleation rate calculation for low-solubility rock-salt nucleation

C(M)	n^*	$\Delta G^{\neq}$(kJ/mol)	Z	$\beta(ns^{-1})$	$r(ns^{-1})$
2.17	13	31.0	0.17	28.0	2×10^{-3}
2.68	13	21.7	0.17	36.7	1.2×10^{-1}
3.18	13	14.4	0.17	37.7	2.8

3.3.6 Nucleation Rate Estimation from Molecular Dynamics

Under relatively high supersaturation, it is possible to compare (and thus benchmark) the above predicted nucleation rates directly against large-scale MD simulations. We thus conducted long MD simulations, recording the size of the largest cluster along the trajectory as an indicator of nucleation at three concentrations (2.17, 2.68 and 3.18 M). Four replicate MD simulations were carried out for each concentration. To evaluate the nucleation rate, we recorded the time at which nucleation occurs during the trajectory (if any). The nucleation time must be sufficiently long enough to allow for quasi-equilibration prior to nucleation such that the observed nucleation time is independent of the initial ion configuration. Only the 2.68 M simulation yields a useful measure of rate, since the 2.17 M and 3.18 M simulations led to no nucleation and nearly immediate nucleation, respectively. Therefore, we conducted a total of 32 200 ns MD trajectories at 2.68 M. Nucleation events were observed in nine trajectories, with corresponding nucleation times of 30, 50, 65, 80, 80, 125, 130, 145, and 160 ns; 8 (of the 32) representative trajectories are plotted in Figure 3.6.

Our GCMC/MD simulation at this concentration predicted a critical cluster size as 13. From the MD trajectory, we can estimate the critical size by examining the local maximum cluster size prior to nucleation. This value fluctuates between 10~20 with few exceptions, consistent with the critical cluster size calculated via GCMC.

Assuming that the nucleation events obey a Poisson distribution, the probability of k nucleation events happening in n simulations within a time interval t can be

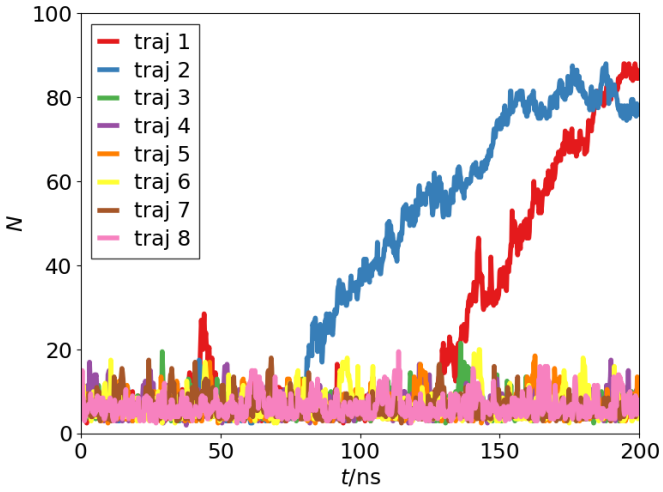


Figure 3.6: Selected MD trajectories of largest cluster sizes.

expressed as:

$$P(k,t,n) = e^{-nrt}\frac{(nrt)^k}{k!} \tag{3.19}$$

We picked 10 time intervals from 0~20 ns to 0~200 ns and counted the number of nucleation events within each time interval for 32 trajectories. The nucleation rate was then estimated from these data points using maximum likelihood estimation (see Figure 3.7). The nucleation rate we extracted from this process is 1.4x10^{-3} /(simulation box· ns) which differences by an order of 2 from that calculated via our GCMC/MD method (1.2x10^{-1}).

Note that quantitative agreement should not necessarily be expected in this case due to the slightly different approximations made in the two methods. In particular, the GCMC simulations neglect the cluster-solute interactions. Although these interactions are quite modest for dilute solutions, they are not entirely negligible (especially at large supersaturations), and their neglect may artificially reduce the nucleation rate. In addition, the concentrations (and thus chemical potential) of the solute is not fixed during the MD simulation, and decreases during the course of aggregation. If there is a cluster of size 13 existing in the simulation, then the effective concentration is reduced by ~10.4%, which would reduce the chemical

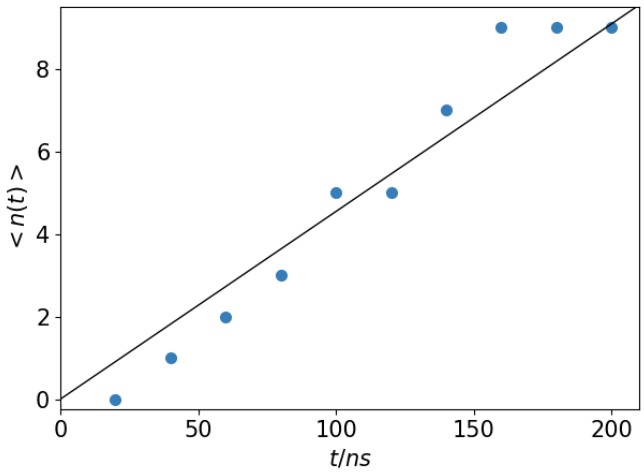


Figure 3.7: Nucleation rate analysis for MD trajectories. The black solid line is the expected number of nucleations according to Poisson distribution of the estimated nucleation rate. The blue dots are actual observed number of nucleations in 32 trajectories within different time intervals.

potential by ~0.22 k_bT for each ion pair. And correspondingly, the free energy barrier would be increased by in total 0.22×13=2.86 k_bT and lead the nucleation rate to be reduced by ~ 17 times. Considering there may be more than one agglomerate in the system before one of them reaching the critical cluster size, the nucleation rate will possibly be reduced by more than ~ 17 times, which is close to the difference between estimated rates from hybrid GCMC/MD and brute-force MD. In contrast to brute-force MD, GCMC provides a more faithful description of nucleation from bulk solution, where the solute concentration is essentially fixed. Nonetheless, the agreement between the two disparate methods is quite satisfying.

3.3.7 Nucleation under Modest Supersaturation

It is interesting to note that Desarnaud et al. find that spontaneous primary nucleation and growth of NaCl are observed at a supersaturation ratio of ~ 1.6 (implying a limit to the supersaturation), much lower than the supersaturation ratios utilized in the above MD simulations (~ 100).[79] This large discrepancy can be largely

Table 3.4: Nucleation rate estimates for rock-salt solution at dilute concentrations.

C(M)	n^*	$\Delta G^{\neq}$(kJ/mol)	Z	$\beta(ns^{-1})$	$r(L^{-1}s^{-1})$
0.1	332	943.44	0.020	4.9	3×10^{-135}
0.2	94	363.77	0.046	10.6	1×10^{-32}
0.3	55	231.96	0.066	11.5	3×10^{-9}
0.4	40	172.83	0.082	12.6	1×10^{2}
0.5	32	138.79	0.095	15.8	2×10^{8}

resolved by a simple comparison of the time- and length-scales of the experiment as compared to simulation, where (in the former case) only a single nucleation event in the macroscopic sample is sufficient to induce crystallization.

In contrast to MD, our GCMC approach provides a convenient approach to estimate nucleation barriers and rates at far smaller concentration/supersaturation, making more direct connection with experiment. We extrapolated the nucleation free energy surface to larger cluster size and estimated the rate for concentration 0.1, 0.2, 0.3, 0.4 and 0.5 M, with the results given in Table 3.4 (with rates now given as nucleation events per $L \cdot s$).

From Table.3.4, we conclude that at concentration of about 0.3-0.4 M, we expect to see nearly immediate nucleation in a macroscopic sample (of this model salt). The concentration still corresponds to a relatively high supersaturation ratio of 15, still roughly an order of magnitude higher than what is achievable for NaCl in experiment. We can understand this discrepancy by examining the solid-solution surface tension of the model salt ($0.18\ N{\cdot}m^{-1}$) as compared to that of NaCl for experiment ($0.08\ N{\cdot}m^{-1}$). Since the current salt force field yields a gas-phase (100) surface energy of close to the experimental value for NaCl[80] [165 vs 170 $kJ/(mol{\cdot}nm^2)$], we attribute the likely cause of the anomalously high solid-solution surface tension to the added Lennard-Jones interaction for H atom, originally utilized by Alejandre and Hansen. This additional term reduces the net attraction between the water H and Cl^-, thus increasing the surface tension. The markedly different solid-solution surface tensions yield dramatically different surface energies as the size of the cluster grows, contributing an additional $\sim 30 k_b T$ to the nucleation barrier at the

critical cluster size and thus dramatically stabilizing the supersaturated solution of the model salt as compared to NaCl. We anticipate that a more sophisticated salt (polarizable) force field, exhibiting accurate solubility and surface energy, would likely exhibit a far lower maximum supersaturation.

3.4 conclusion

We have utilized a hybrid grand canonical Monte Carlo/molecular dynamics (GCMC/MD) approach that allows us to model the nucleation of low-solubility materials/weak electrolytes in the presence of explicit solvent. Using this approach, we examined the free energy surface and associated barriers for the nucleation of an aqueous solution of a sparingly soluble salt, finding calculated nucleation rates that are in excellent accord with large-scale brute-force MD simulations. Upon extrapolating the free energy barriers to smaller supersaturation, and accounting for the errors in the salt solution surface tension, we find absolute nucleation rates that are not inconsistent with those observed for related salts in aqueous solution. We believe that our hybrid GCMC/MD method is one of the first methodologies that is able to address the system size, mass transport, and rare event sampling challenges that are inherent in the study of the nucleation of weak electrolytes in an explicit solvent. We anticipate that this approach can be easily extended to study other weak electrolytes/low-solubility materials, such as calcium carbonate and lithium fluoride, and potentially even nanoporous materials, where nucleation may involve solvent incorporation and/or templating.

[74] Alejandre, José, and Jean Pierre Hansen. 2007. Ions in water: From ion clustering to crystal nucleation. *Phys. Rev. E* 76(6):061505.

[75] Becker, R., and W. Döring. 1935. Kinetische Behandlung der Keimbildung in übersättigten Dämpfen. *Ann. Phys.* 416:719–752.

[76] Berendsen, HJC, JR Grigera, and TP Straatsma. 1987. The missing term in effective pair potentials. *J. Phys. Chem.* 91(24):6269–6271.

[77] Clouet, Emmanuel. 2010. Modeling of nucleation processes. *arXiv preprint arXiv:1001.4131*.

[78] Dang, Liem X. 1995. Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-crown-6 ether: a molecular dynamics study. *J. Am. Chem. Soc.* 117(26):6954–6960.

[79] Desarnaud, Julie, Hannelore Derluyn, Jan Carmeliet, Daniel Bonn, and Noushine Shahidzadeh. 2014. Metastability limit for the nucleation of nacl crystals in confinement. *The J. Phys. Chem. Lett.* 5(5):890–895.

[80] Gutshall, Paul L, and Gordon E Gross. 1965. Cleavage surface energy of nacl and mgo in vacuum. *J. Appl. Phys.* 36(8):2459–2460.

[81] Hastings, W Keith. 1970. Monte carlo sampling methods using markov chains and their applications.

[82] Jiang, Hao, Pablo G Debenedetti, and Athanassios Z Panagiotopoulos. 2019. Nucleation in aqueous nacl solutions shifts from 1-step to 2-step mechanism on crossing the spinodal. *The J. Chem. Phys.* 150(12):124502.

[83] Martin, Georges. 1979. The theories of unmixing kinetics of solid solutions. *CEA Centre d'Etudes Nucleaires de Saclay*.

[84] Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. 1953. Equation of state calculations by fast computing machines. *The journal of chemical physics* 21(6):1087–1092.

[85] Smith, David E, and Liem X Dang. 1994. Computer simulations of nacl association in polarizable water. *The J. Chem. Phys.* 100(5):3757–3766.

[86] Steinhardt, Paul J, David R Nelson, and Marco Ronchetti. 1983. Bond-orientational order in liquids and glasses. *Phys. Rev. B* 28(2):784.

[87] Tribello, Gareth A, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. 2014. Plumed 2: New feathers for an old bird. *Comput. Phys. Commun.* 185(2):604–613.

4 MODELING NUCLEATION VIA GRAPH-BASED APPROACH

4.1 Introduction

The hybrid GCMC/MD methodology successfully addresses the challenges of dilute systems, mass transport and explicit solvents in the solution-phase nucleation. But the underlying Markov chain Monte Carlo (MCMC) algorithm determines its substandard performance in terms of efficiency and parallelism, and thus limits its extension to more complicated systems. Due to the intrinsically serial property of the MCMC algorithm, it is not straight to parallel our hybrid GCMC/MD sampling program among multiple machines and benefit from high throughput computing. And the MCMC sampling over the nucleate cluster size is indeed a one-dimensional random walk which scales quadratically with the number of states. Therefore it is not optimal to extend this approach to larger cluster sizes, especially when the modeling system employs a relatively large number of intermediate states for expanded ensemble.

In addition to the above issues, the hybrid GCMC/MD also suffers from ineffective samplings. The hybrid GCMC/MD relies on the unbiased MD moves coupled with frequent insertions/deletions to sample the configurations of the nucleus. When the nucleus is involved with polymorphism and charges, MD without bias potential is usually not capable of sampling over the entire configurational phase space.[90] Our hybrid GCMC/MD approach manages to circumvent this issue by employing MC moves. The frequent insertions/deletions help to boost the reconstruction of the nucleus and allow multiple structures to be sampled during a single GCMC/MD simulation. But when the nucleus reaches to a certain size with a clear structure, rebuilding the nucleus to another structure may require the deletion of most existed particles and reduce the cluster to an early nucleation stage. Such moves are redundant in the purpose of structural sampling and significantly waste the computing resources.

To address the above challenges, in this chapter, we present a simulation approach based on the graph structure of the nucleus. Naive MD simulation cannot

work because it will only sample at most a sub-ensemble of cluster structures, therefore our idea is to decompose this ensemble of cluster structures into various sub-ensembles and sample each of them individually. Since the phase space of the nucleus in crystallization usually consists of several representative graph structures and each graph structure can be considered as a well defined sub-ensemble, the above idea can be easily applied here. In this graph-based method, multiple simulations are carried out and each of them only focuses on the nucleation associated with one specific structure. And the overall nucleation behavior can be recovered by combining the contribution from each single simulation with proper weight.

In the following sections, we first develop a non-graph-based method to estimate the nucleation free energy via a step-by-step procedure. In each step, a single thermodynamic integration (TI) is carried out for calculating the associated free energy. This approach is benchmarked against LJ nucleation and compared with the results predicted by our hybrid GCMC/MD method. Starting from this non-graph-based approach, the graph-based one is derived by considering multiple structurally specified simulations for each insertion step. And the free energy change obtained in each single simulation will contribute to the overall free energy calculation. To validate this approach, we test it on the nucleation in lattice model and the results are compared against values from GCMC simulations. According to the derivation, the graph-based method requires a clear "graph" definition, which is not easily achievable for actual atomistic systems. To avoid this restriction, we provide another justification for this approach from Jarzynski Equality[92, 91] and consider it as a nonequilibrium method. From this perspective, the graph-based approach no longer requires a rigorous "graph" definition and can be easily extended to atomistic systems. In the end, we employ the same low-solubility rock-salt model from Chapter 3 as a test system and compare the results from the graph-based approach and the hybrid GCMC/MD method.

4.2 Non-graph-based Thermodynamic Integration for Nucleation

4.2.1 Theory

Consider a cluster of size N at temperature T in a dilute system where the intercluster interactions can be safely ignored. The canonical ensemble partition function of this cluster is given by:

$$Z_N = \frac{1}{N!\Lambda^{3N}} \int dr^{3N} e^{-\beta U(r^{3N})} w(r^{3N}), \tag{4.1}$$

where Λ is the thermal de Broglie wavelength of the particle, $\beta \equiv 1/k_b T$ is the reciprocal temperature, k is Boltzmann constant, U is the energy of the system, and w is the weight function representing the cluster criterion. As in previous chapters, the cluster criterion is taken from Stillinger's[96] definition. In the case of solution-phase nucleation, the integral of solvent degrees of freedom should also be included. The free energy difference between clusters of size N and $N+1$ can be expressed as:

$$\begin{aligned} \Delta G &= -k_b T \ln \frac{Z_{N+1}}{Z_N} \\ &= -k_b T \frac{1}{N\Lambda^3} \frac{\int dr^{3(N+1)} e^{-\beta U(r^{3(N+1)})} w(r^{3(N+1)})}{\int dr^{3N} e^{-\beta U(r^{3N})} w(r^{3N})} \end{aligned} \tag{4.2}$$

And nucleation free energy surface can be obtained by calculating this ΔG for each size N.

Here we introduce a fictitious state denoted as $N+1'$, in which the cluster consists of $N+1$ particles but one of them is a non-interacting particle ("ghost particle"). It should be noticed here the ghost particle is also part of the cluster and must satisfy the cluster criterion. The partition function of state $N+1'$ can be

written as:

$$Z_{N+1'} = \frac{1}{(N+1)!\Lambda^{3(N+1)}} \int dr^{3(N+1)} e^{-\beta U(r^{3N})} w(r^{3(N+1)}), \tag{4.3}$$

where the energy U is calculated over N real particles. With this intermediate state, the free energy change ΔG can be decomposed into two contributions:

$$\begin{aligned} \Delta G &= -k_b T \ln \frac{Z_{N+1'}}{Z_N} \frac{Z_{N+1}}{Z_{N+1'}} \\ &= -k_b T \ln \frac{Z_{N+1'}}{Z_N} - k_b T \ln \frac{Z_{N+1}}{Z_{N+1'}} \end{aligned} \tag{4.4}$$

We further define ΔG_1 and ΔG_2 as

$$\begin{aligned} \Delta G_1 &= -k_b T \ln \frac{Z_{N+1'}}{Z_N} \\ &= -k_b T \ln \frac{1}{(N+1)\Lambda^3} \frac{\int dr^{3(N+1)} e^{-\beta U(r^{3N})} w(r^{3(N+1)})}{\int dr^{3N} e^{-\beta U(r^{3N})} w(r^{3N})} \end{aligned} \tag{4.5}$$

$$\begin{aligned} \Delta G_2 &= -k_b T \ln \frac{Z_{N+1}}{Z_{N+1'}} \\ &= -k_b T \frac{\int dr^{N+1} e^{-\beta U(r^{3(N+1)})} w(r^{3(N+1)})}{\int dr^{N+1} e^{-\beta U(r^{3N})} w(r^{3(N+1)})} = \left\langle e^{-\beta \Delta u} \right\rangle_{N+1'} \end{aligned} \tag{4.6}$$

And the free energy change ΔG of inserting a particle into the cluster is the summation of ΔG_1 and ΔG_2. Instead of evaluating ΔG directly using GCMC as we proposed in the hybrid GCMC/MC approach, here we calculate this free energy through two separate terms - ΔG_1 and ΔG_2.

4.2.1.1 GCMC-Swap

Here we define the effective volume V_{eff} as:

$$V_{eff} = \frac{\int dr^{3(N+1)} e^{-\beta U(r^{3N})} w(r^{3(N+1)})}{\int dr^{3N} e^{-\beta U(r^{3N})} w(r^{3N})} \tag{4.7}$$

And ΔG_1 can be written as:

$$\Delta G_1 \quad = \quad -k_b T \ln \frac{V_{eff}}{(N+1)\Lambda^3} \tag{4.8}$$

Thus the essential part of estimating ΔG_1 is the correct calculation of V_{eff}. We may consider V_{eff} as the volume of the region around the cluster of size N and this volume can be easily estimated by numerical Monte Carlo. But this assumption ignores cases where the ghost particle lies in the 'interior' of the cluster, i.e., where the original N atoms alone do not satisfy the cluster criterion in the absence of the ghost particle.

To correctly include such configurations, we developed a new GCMC-Swap method, which is an adaption of naïve GCMC. This method consists of three types of moves: insertion, deletion and swap. In the insertion step, we insert a non-interacting particle into the system with a cluster of size N and decide whether to accept this insertion by checking the cluster criterion for $N+1$ particles. During the deletion step, the non-interacting atom will be removed. If the deletion leads to the break of cluster criterion (the remaining N particles do not satisfy the cluster criterion), this move will be rejected. In addition to insertion and deletion, there is a third type of Monte Carlo move called "swap". In the swapping move, we switch the non-interacting particle with a random interacting particle and decide whether to accept this move by the energy change. Such steps allow us to sample configurations which are ignored if we evaluate V_{eff} by numerical Monte Carlo and thus give a correct volume prediction.

4.2.1.2 Thermodynamic Integration

ΔG_2 is defined as the free energy difference between state $N+1'$ and $N+1$ and this can be obtained by thermodynamic integration (TI)[93] via gradually turning on the interaction of the ghost particle through a coupling parameter λ. The free

energy change during TI can be expressed as:

$$\Delta G_2 = \int_{\lambda=0}^{\lambda=1} d\lambda \left\langle \frac{\partial U(\lambda)}{\lambda} \right\rangle_\lambda \tag{4.9}$$

This integration will be carried out numerically by Gaussian quadrature. To avoid singularity when λ approaches 0, LJ interaction will be replaced by a soft-core potential[95]:

$$U_{LJ}(\lambda) = 4\lambda\varepsilon \left[\left(\frac{\sigma}{r_{eff}} \right)^{12} - \left(\frac{\sigma}{r_{eff}} \right)^{6} \right], \quad r_{eff} = \sigma \left[0.5(1-\lambda) + \left(\frac{r}{\sigma} \right)^{6} \right]^{1/6} \tag{4.10}$$

For the initial configuration, a ghost particle is attached to the cluster of size N. During simulation, the Stillinger's cluster criterion is enforced on total N+1 particles and the MST restraints are established based on the graph structure of the cluster to prevent it from falling apart. After every few MD steps, the graph will be regenerated and MST restraints will be updated accordingly to minimize the artificial effects from the restraints. To successfully apply this method, TI is expected to be able to effectively sample all cluster sub-ensembles, as well as configurations with the non-interacting particle in the interior of the cluster. This assumption is only valid for simple systems. For strong-interaction systems, more advanced sampling techniques are required.

4.2.2 Results: Lennard-Jones Nucleation

We benchmarked our non-graph-based methodology via calculation of the free energy surface for LJ vapor-liquid nucleation, comparing against our previous hybrid GCMC/MD approach[94]. For LJ vapor-liquid nucleation, no PBC or cutoffs were applied. Reduced temperature (T^*) of 0.7 was chosen, along with a Stillinger's cluster criterion of 1.5 σ. The MD simulations utilized Langevin dynamics with a 2.0 fs time step. MST restraints were updated every 200 fs. The cluster size was sampled from 1 to 60 through 59 consecutive insertions. Four replicate thermodynamic integrations and GCMC-Swap simulations were carried out for each insertion.

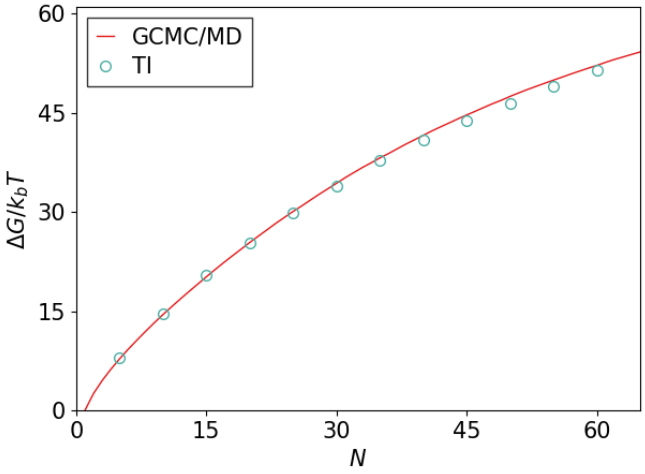


Figure 4.1: Free energy surface for LJ vapor-liquid nucleation at $T^* = 0.7$, $n_v = 5.75 \times 10^{-3}$. The red curve is the reference free energy surface calculated by our hybrid GCMC/MD approach[94]. The green circles correspond to free energies predicted by non-graph-based method. N represents the cluster size in the plots.

Each GCMC-Swap simulation consists of 10^6 MC steps with a 10% probability of selecting insertion moves, a 10% probability of selecting deletion moves and an 80% probability of selecting swapping moves. During each thermodynamic integration, 10 λ values were chosen from Gaussian quadrature on 0 to 1. For each λ value, a 200 ps NVT simulation was carried out and energy derivative was sampled every 200 fs. The results are shown in Figure 4.1.

The non-graph based method well reproduce the result from earlier work. In addition, we also compare the efficiency of two approaches. For simple systems like LJ, the thermodynamic integration approach does not necessarily provide more efficiency than hybrid GCMC/MD approach, but it provides a better scalability. Via step-wise procedure, the new approach scales only linearly with the cluster size. While the hybrid GCMC/MD is intrinsically a random work sampling and scales quadratically with the number of states. This difference is more pronounced when expanded ensemble is employed in hybrid GCMC/MD method. Also, since the free energy can be directly calculated by thermodynamic integration instead of

being estimated from cluster distribution, this new approach does not suffer from rare event sampling issue caused by nucleation free energy barriers.

4.3 Graph-based Method

4.3.1 Theory

The non-graph-based method works for non-solution-phase, weak-interaction systems like Lennard-Jones model. But for nucleation involved with explicit solvents and strong interactions, this approach cannot guarantee that all sub-ensembles of clusters are completely sampled. This ineffective sampling will lead to an inaccurate estimation of ΔG_2 as the TI cannot explore all phase space with unbiased MD. additionally, the evaluation of V_{eff} will also become impractical in such systems. GCMC-Swap method requires a reasonable acceptance ratio for swapping moves. But with strong interactions and explicit solvents, most swapping moves will be rejected due to steric effects (overlapping solute-solvent) and large energy fluctuations (compared to k_bT, due to strong solute-solvent and solute-solute interactions).

To address the sampling issue, here we introduce a graph-based approach. For the solution-phase crystallization, the configurations of the nucleus are usually dominated by several representative graph structures. Although the comprehensive sampling over all of them is not achievable in one single unbiased MD simulation, the phase space defined by each graph can be easily explored. Our idea is to describe the entire nucleation by combining the nucleation corresponding to each representative graph structure. And such structure-dependent nucleations can be easily modeled with relatively short simulations. The graph-based idea is illustrated in Figure 4.2.

As before, we start from the partition function of the cluster at size N. Since the entire ensemble of configurations can be categorized into several representative

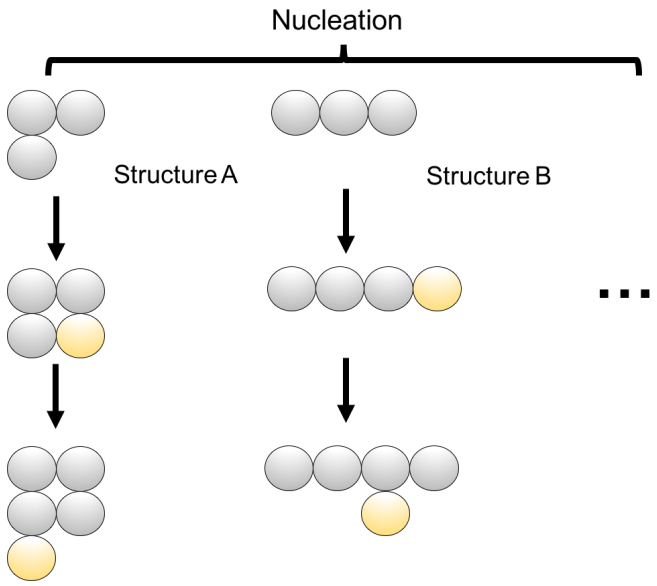


Figure 4.2: Schematic representation of the graph-based approach. The entire nucleation can be considered as the combination of nucleation for each graph structure.

graph structures, the partition can be rewritten as:

$$Z_N = \frac{1}{N!\Lambda^3} \sum_i \int_i dr^{3N} e^{-\beta U(r^{3N})} w(r^{3N}), \tag{4.11}$$

where subscription i is the index for graph structures and the integral of i includes all configurations defined by the corresponding graph. Here, we denote the partition function for each graph structure as:

$$Z_{N,i} = \int_i dr^{3N} e^{-\beta U(r^{3N})} w(r^{3N}) \tag{4.12}$$

Then the total partition function can be written as the sum of those individual partition functions:

$$Z_N = \sum_i Z_{N,i} \tag{4.13}$$

With those definitions, the free energy change of the cluster growing from size N to $N+1$ is given by:

$$\begin{aligned} \Delta G &= -k_b T \ln \frac{Z_{N+1}}{Z_N} \\ &= -k_b T \ln \frac{\sum_j Z_{N+1,j}}{\sum_i Z_{N,i}}, \end{aligned} \tag{4.14}$$

where i is the index of structures for cluster at size N and j is the one for cluster at size $N+1$. For convenience, we name cluster of size N as "parent" and cluster of size $N+1$ as "child". And each child structure j grows from the corresponding parent structure j^{parent}. Then the Eq. 4.14 can be rewritten as:

$$\Delta G = -k_b T \ln \sum_j \frac{Z_{N,j^{parent}}}{\sum_i Z_{N,i}} \frac{Z_{N+1,j}}{Z_{N,j^{parent}}}, \tag{4.15}$$

The first term $Z_{N,j^{parent}}/\sum_i Z_{N,i}$ in the summation is the probability of structure j^{parent} among all parent structures for the cluster of size N:

$$P_{N,j^{parent}} = \frac{Z_{N,j^{parent}}}{\sum_i Z_{N,i}} \tag{4.16}$$

And the second term $Z_{N+1,j}/Z_{N,j^{parent}}$ gives the free energy change for the specific parent cluster j^{parent} growing into the specific child cluster j:

$$\Delta G(j^{parent} \to j) = -k_b T \ln \frac{Z_{N+1,j}}{Z_{N,j^{parent}}} \tag{4.17}$$

Now, the overall nucleation free energy can be expressed in terms of the nucleation free energy of each graph structure:

$$\Delta G = -k_b T \ln \sum_j P_{N,j^{parent}} \exp[-\beta \Delta G(j^{parent} \to j)] \tag{4.18}$$

To estimate $\Delta G(j^{parent} \to j)$, for each j^{parent} cluster, we introduce a fictitious

state (denoted as $N+1', j^{parent}$) where an additional non-interacting atom is attached to the cluster and the corresponding partition function can be written as:

$$Z_{N+1',j^{parent}} = \frac{1}{(N+1)!\Lambda^{3(N+1)}} \int_{j^{parent}} dr^{3(N+1)} e^{-\beta U(r^{3N})} \widetilde{w}(r^{3(N+1)}) \tag{4.19}$$

The energy U is calculated over N interacting particles. Notice here the cluster criterion for this state is different than the criterion we use to define clusters of size N+1 and we use $\widetilde{w}$ instead of w to denote the weight function. This criterion requires the N real particles from structure j^{parent} to satisfy the commonly defined criterion for cluster of size N and the additional $N+1$th ghost particle to be attached to them. With this intermediate state, we can expand Eq. 4.17 as:

$$\Delta G(j^{parent} \rightarrow j) \quad = \quad -k_b T \ln \frac{Z_{N+1',j^{parent}}}{Z_{N,j^{parent}}} \frac{Z_{N+1,j}}{Z_{N+1',j^{parent}}} \tag{4.20}$$

For the first term inside the natural logarithm, we define the the effective volume for structure j^{parent} as:

$$V^{eff}_{j^{parent}} \quad = \quad \frac{\int_{j^{parent}} dr^{3(N+1)} e^{-\beta U(r^{3N})} \widetilde{w}(r^{3(N+1)})}{\int_{j^{parent}} dr^{3N} e^{-\beta U(r^{3N})} w(r^{3N})} \tag{4.21}$$

Then the term $Z_{N+1',j^{parent}}/Z_{N,j^{parent}}$ can be expressed as

$$\frac{Z_{N+1',j^{parent}}}{Z_{N,j^{parent}}} \quad = \quad \frac{V^{eff}_{j^{parent}}}{(N+1)\Lambda^3} \tag{4.22}$$

The second term $Z_{N+1,j}/Z_{N+1',j^{parent}}$ can also be expanded as follows:

$$\begin{aligned}\frac{Z_{N+1,j}}{Z_{N+1',j^{parent}}} &= \frac{\int_j dr^{3(N+1)} e^{-\beta U(r^{3(N+1)})} w(r^{3(N+1)})}{\int_{j^{parent}} dr^{3(N+1)} e^{-\beta U(r^{3N})} \widetilde{w}(r^{3(N+1)})} \\ &= \frac{\int_j dr^{3(N+1)} e^{-\beta U(r^{3N})} \widetilde{w}(r^{3(N+1)})}{\int_{j^{parent}} dr^{3(N+1)} e^{-\beta U(r^{3N})} \widetilde{w}(r^{3(N+1)})} \\ &\quad \times \frac{\int_j dr^{3(N+1)} e^{-\beta U(r^{3(N+1)})} w(r^{3(N+1)})}{\int_j dr^{3(N+1)} e^{-\beta U(r^{3N})} \widetilde{w}(r^{3(N+1)})},\end{aligned} \tag{4.23}$$

where the introduced integral represents a sub-ensemble of the intermediate state $N+1', j^{parent}$. In this sub-ensemble, the ghost particle is restricted in a sub-region to make the cluster of size N+1 form child structure j. Therefore the first fraction in the above expression is in fact the probability of the selecting structure j among all the child structures of parent j^{parent}. Since the new inserted particle is a non-interacting one, the probability of selecting j is the volume fraction for the ghost atom to form structure j and here we denote it as P_j. The second faction in the above expression can also be further expanded with a new phase integral for a cluster which consists of N+1 real particles and satisfies the cluster criterion corresponding to weight function $\widetilde{w}$:

$$\begin{aligned}\frac{\int_j dr^{3(N+1)} e^{-\beta U(r^{3(N+1)})} w(r^{3(N+1)})}{\int_j dr^{3(N+1)} e^{-\beta U(r^{3N})} \widetilde{w}(r^{3(N+1)})} &= \frac{\int_j dr^{3(N+1)} e^{-\beta U(r^{3(N+1)})} \widetilde{w}(r^{3(N+1)})}{\int_j dr^{3(N+1)} e^{-\beta U(r^{3N})} \widetilde{w}(r^{3(N+1)})} \\ &\quad \times \frac{\int_j dr^{3(N+1)} e^{-\beta U(r^{3(N+1)})} w(r^{3(N+1)})}{\int_j dr^{3(N+1)} e^{-\beta U(r^{3(N+1)})} \widetilde{w}(r^{3(N+1)})}\end{aligned} \tag{4.24}$$

The first fraction in the right-hand side is the free energy change of turning on the interaction of the ghost particle:

$$< e^{-\beta \Delta U} >_{j,N+1'} = \frac{\int_j dr^{3(N+1)} e^{-\beta U(r^{3(N+1)})} \widetilde{w}(r^{3(N+1)})}{\int_j dr^{3(N+1)} e^{-\beta U(r^{3(N+1)})} \widetilde{w}(r^{3(N+1)})} \tag{4.25}$$

This value can be obtained through thermodynamic integration or other free energy calculation methods. Since the sampling is restricted in the specific structure j, with appropriate restraints, the free energy can be easily calculated with unbiased simulation. And the second fraction in Eq. 4.24 can be written as :

$$\begin{aligned}\frac{\int_j dr^{3(N+1)} e^{-\beta U(r^{3(N+1)})} w(r^{3(N+1)})}{\int_j dr^{3(N+1)} e^{-\beta U(r^{3(N+1)})} \widetilde{w}(r^{3(N+1)})} &= \frac{\int_j dr^{3(N+1)} e^{-\beta U(r^{3(N+1)})} \widetilde{w}(r^{3(N+1)}) \frac{w(r^{3(N+1)})}{\widetilde{w}(r^{3(N+1)})}}{\int_j dr^{3(N+1)} e^{-\beta U(r^{3(N+1)})} \widetilde{w}(r^{3(N+1)})} \\ &= \left\langle \frac{w(r^{3(N+1)})}{\widetilde{w}(r^{3(N+1)})} \right\rangle_{j,\widetilde{w}} \equiv M_j \end{aligned} \tag{4.26}$$

This is the ratio between weight function $\widetilde{w}$ and w, and we denote it as M_j. The ratio M can be obtained through the analysis of the cluster structure j. With those definition, the free energy change $\Delta G(j^{parent} \to j)$ can be eventually expressed as:

$$\Delta G(j^{parent} \to j) = -k_b T \ln \frac{V^{eff}_{j^{parent}}}{(N+1)\Lambda^3} P_j < e^{-\beta \Delta U} >_{j,N+1'} M_j \tag{4.27}$$

and Eq. 4.14 can be rewritten as:

$$\Delta G = -k_b T \ln \sum_j P_{N,j^{parent}} \frac{V^{eff}_{j^{parent}}}{(N+1)\Lambda^3} P_j < e^{-\beta \Delta U} >_{j,N+1'} M_j \tag{4.28}$$

The summation in the above equation is over all child structure j s and we can replace it by $\sum_{j^{parent}} \sum_{j \in j^{parent}}$. Doing that implicitly requires that the child structure j can only grow from one specific parent structure j^{parent}. If we consider all particles are indistinguishable in structure j, this assumption seems invalid because by inserting the new particle into different locations, the same child structure j can be generated from multiple different parent structures. However, in actual simulation, particles are distinguishable and structure j can be considered as one specific permutation of some indistinguishable structure, and it can only grow from one specific parent permutation j^{parent}. Therefore the above assumption is indeed satisfied in the

actual simulation. By replacing the summation, the free energy can be expressed as:

$$\Delta G = -k_b T \ln \sum_{j^{parent}} P_{N,j^{parent}} \frac{V_{j^{parent}}^{eff}}{(N+1)\Lambda^3} \sum_{j \in j^{parent}} P_j < e^{-\beta \Delta U} >_{j,N+1'} M_j \tag{4.29}$$

With Eq. 4.29, we can design a full procedure to calculate ΔG for size N to $N+1$: (i) Select a number of parent structures according to their equilibrium probability $P_{N,j^{parent}}$. This probability can be calculated by the free energy of structure j^{parent} from previous step of inserting the Nth particle. (ii) For each parent structure j^{parent}, calculate the corresponding effective volume $V_{j^{parent}}^{eff}$. The details for calculating effective volume will be explained in Section 4.3.1.1. (iii) Select the child structure j from all child structures of parent j^{parent}. The corresponding probability is the volume fraction for the new particle to form structure j. Thus we can uniformly sample a random position for the $N+1$th particle in the region defined by $V_{j^{parent}}^{eff}$ and determine the child structure j accordingly. (iv) Restrain the cluster within structure j and turn on the interaction of the ghost atom to obtain the free energy change using TI or other free energy calculation methods. (v) Calculate ratio M for structure j. The details will be explained in Section 4.3.1.2. (vi) Calculate the free energy change for the growth of each child structure by

$$\Delta G(j^{parent} \rightarrow j) = -k_b T \ln \frac{V_{j^{parent}}^{eff}}{(N+1)\Lambda^3} < e^{-\beta \Delta U} >_{j,N+1'} M_j. \tag{4.30}$$

And the free energy will be used to estimated $P_{N+1,j^{parent}}$ for the next insertion step. It should be noticed here for the current step, the parent clusters have already been selected according to equilibrium distribution. Therefore, the equilibrium distribution of children only depends on $\Delta G(j^{parent} \rightarrow j)$ from the current step and not other steps before the current step. (vii) Calculate the ΔG by taking exponential

average of all $\Delta G(j^{parent} \to j)$s according to Eq. 4.29:

$$\Delta G = -k_b T \ln \frac{1}{n} \sum_{j=1}^{n} exp(-\beta \Delta G(j^{parent} \to j)), \tag{4.31}$$

where n is the number of parallel simulations we launch for each step of nucleation. It should be noticed here, since we select parent clusters according to its Boltzmann distribution and select children according to volume fraction, the probability $P_{N,j^{parent}}$ and P_j are not explicitly expressed in Eq.4.30 and Eq. 4.31 but rather presented implicitly in the bias used in the selection of the parent and child clusters.

4.3.1.1 Volume Contribution

Accurately calculating $V^{eff}_{j^{parent}}$ is essential for free energy calculation. Unlike the effective volume defined in previous section, $V^{eff}_{j^{parent}}$ here can be calculated by numerical Monte Carlo efficiently.

In Eq. 4.21, the effective volume is defined as the ratio of two phase integrals. In the numerator, $\widetilde{w}$ inside the integrand is the weight function corresponding to the requirement that N real particles satisfy the cluster criterion of size N and the ghost particle is attached to them to keep total $N+1$ particles satisfy the cluster criterion of size $N+1$. Thus, Eq. 4.21 can be rewritten as:

$$\begin{aligned} V^{eff}_{j^{parent}} &= \frac{\int_{j^{parent}} dr^{3N} e^{-\beta U(r^{3N})} w(r^{3N}) dr^3 w(r^{3(N+1)})}{\int_{j^{parent}} dr^{3N} e^{-\beta U(r^{3N})} w(r^{3N})} \\ &= \int_{j^{parent}} dr^3 w(r^{3(N+1)}) \big| w(r^{3N}), j^{parent} \end{aligned} \tag{4.32}$$

And the effective volume is actually the volume the ghost particle can take to still keep the cluster criterion satisfied. We can obtain this volume by numerical Monte Carlo through the following procedure: We launch a number of trial insertions into the simulation system. For each random insertion, we check the cluster criterion and consider the insertion as a successful one if it is satisfied. The effective volume is given by the product of the total volume and the ratio of successful insertions.

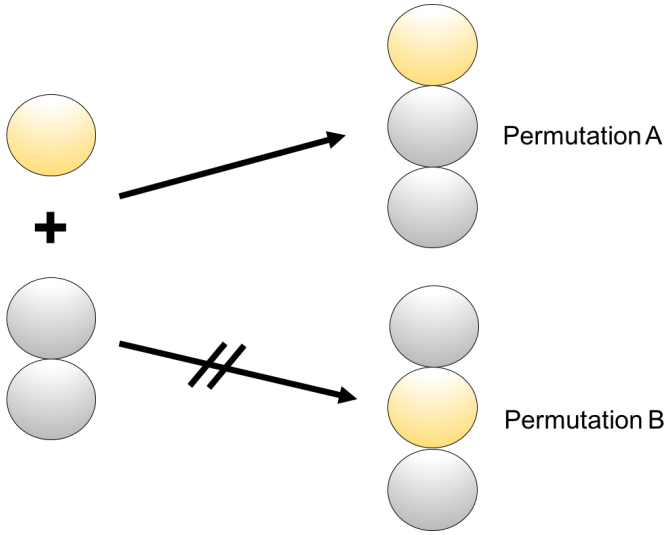


Figure 4.3: The illustration of missing permutations during the step-by-step nucleation.

In the case that the cluster still has relative flexibility within graph structure j, we need couple numeric Monte Carlo with MD simulations.

4.3.1.2 Estimate Ratio M

In previous derivation, we consider each graph structure j corresponding to a specific permutation to establish the relationship between parent and child clusters. But some permutations cannot be generated through the step-by-step particle attachment as illustrated in Figure 4.3. For a linear structure of 3 particles, the permutation A can be created by attaching the new atom (the yellow particle in Figure 4.3) to a cluster of size 2. But permutation B cannot be generated since its "parent" is not a well-defined cluster. To correctly describe the equilibrium phase space, we need to account for the contribution of such missing permutations. In that sense, ratio M_j defined in Eq. 4.26 is not related to the specific permutation that j corresponds to. Instead, it is a property of the general graph structure behind permutation j. In another word, the value of M is universal for all permutations of a defined graph.

In Eq. 4.26, we expressed M_j as the ratio between weight function $w(N+1)$ and $\widetilde{w}(N+1)$ for the general graph j corresponds to. The numerator indicates that as long as $N+1$ particles satisfy the cluster criterion, the cluster can be generated from any parent cluster or non-cluster structure. While the denominator requires the cluster to only grow from a "qualified" parent cluster. To account for the permutations missed in weight function $\widetilde{w}(N+1)$, we introduced ratio M and estimated this ratio through a detailed structural analysis: For $N+1$ particles in structure j, every time we do a trial deletion to remove one of them and check whether the remaining N particles satisfy the cluster criterion. Consider it as a successful deletion if the cluster is not broken by deletion and the ratio of successful trials is the multiplicative inverse of M_j. For systems where the phase space defined by each graph is relatively extensive, we need to couple this approach with MD simulations.

4.3.1.3 Biasing Strategy

In the original approach, we select parent clusters according to the Boltzmann distribution and select the child clusters depending on volume fraction. This sampling strategy works for simple models, but may fail in a strong-interaction system.

Selecting parent clusters according to Boltzmann distribution helps us bias more computing resources on preferred configurations. But for strong-interaction system, within limited number of samplings from Boltzmann distribution, high-energy parent configurations may not be selected for even once. However, those unstable parent configurations may lead to stable child structures and also make significant contributions in the free energy calculation. For this reason, bias should be considered on unstable parent structures to guarantee they are also sampled during the parent selection. One possible way to add such bias is to increase the temperature in the Boltzmann factor.

In child selection, we uniformly choose the position for the new inserted particle and most of the insertion will end up in unpreferred sites. Within limited simulations, we may not be able to generate the stable child structure. So when inserting

the $N+1$th particle, we need to bias on preferred positions with prior knowledge such as coordination number or electronic site potential[89].

Suppose we select the parent structure based on biased probability $\widetilde{P}_{N,j^{parent}}$ and select the child structure with probability $\widetilde{P}_j$, the Eq. 4.30 needs to be rewritten as:

$$\Delta G(j^{parent} \to j) \quad = \quad -k_b T \ln \frac{P_{N,j^{parent}}}{\widetilde{P}_{N,j^{parent}}} \frac{V^{eff}_{j^{parent}}}{(N+1)\Lambda^3} \frac{P_j}{\widetilde{P}_j} < e^{-\beta \Delta U} >_{j,N+1'} M_j. \tag{4.33}$$

to include the reweighting factors and Eq. 4.31 also needs to be adapted accordingly.

4.3.2 Results: Nucleation in Lattice Model

4.3.2.1 Lattice Model in Vapor

Lattice model serves as a perfect model system for validating the graph-based methodology. Despite its simplicity, it is a powerful model to describe the ordered crystal structure. Also, the clusters in lattice model can be easily categorized into distinct graph structures with each graph corresponding to a specific particle arrangement. This automatically satisfies the assumption we made in Eq. 4.11 that the phase space of the cluster consists of several representative structures.

Here we took a 10×10 two-dimensional lattice model system with PBCs applied. Particles in the lattice only interact with their nearest neighbors. And two particles occupying the same lattice site is not allowed. For clusters in the lattice model, the cluster criterion is adapted from Stillinger's definition, whereby only nearest neighbors are considered belong to a cluster. The cluster size was sampled from 1 to 9 which is limited by the system size. Two systems with different interaction parameters ($1\ k_bT/20\ k_bT$) for nearest neighbors were considered and thermal de Broglie wavelength was chosen as 3.16 and 2.24×10^3, respectively.

The benchmark data is generated by GCMC. Only insertion and deletion steps were applied. Both moves were only considered to be acceptable when the cluster criterion was satisfied. Four replicate simulations were carried out in conjunction

with 10^7 MC steps in each. For the system with the interaction parameter of 20 k_bT, Wang-Landau sampling was employed until the Wang-Landau factor converges to 10^{-3}.

In the graph-based simulation, no translational or rotational MC moves was employed for lattice model to avoid the sampling across different graph structures. Cluster size was sampled from 1 to 9 through 8 steps. For each step, 400 parent structures were selected from the ending configurations of the previous step according to Boltzmann distribution (except for size 1, the cluster is just one particle). Since no structural sampling was employed for the growth of each individual structure, both effective volume $V_{j^{parent}}$ and free energy change $< e^{-\beta \Delta U} >_{j,N+1'}$ can be directly calculated. $V_{j^{parent}}$ was easily estimated as the number of empty neighboring sites around the parent structure. A random position among those neighbouring sites was selected for the insertion of the $N+1$th particle and the corresponding energy change is the value for $< e^{-\beta \Delta U} >_{j,N+1'}$. After insertion, ratio M was calculated and the free energy for each child was generated for next insertion. Four replicate simulations were carried out for each system.

To further improve the accuracy, for the system with the interaction parameter of 1 k_bT, simulations were also carried out with 4000 individual growths for each cluster size. For the system with the interaction parameter of 20 k_bT, simulations with different biasing strategies were also carried out to enhance the sampling. In this simulation, parent clusters were selected by $exp(\ln P_{N,j^{parent}}/10)$ instead of unbiased $P_{N,j^{parent}}$ from Boltzmann distribution to favor high energy parent clusters. In child selection, the probability is weighted by 10^{cn} where cn stands for coordination number of this position to guarantee the stable child structures were selected. Both biases were reweighted afterwards.

For lattice model with an interaction parameter of 1 k_bT, the results predicted by graph-based approach are shown in the Figure 4.4. With 400 individual graph-dependent growths calculated for each cluster size, the new graph-based approach well reproduce the results from GCMC with a ~0.1 k_bT uncertainty. We believe this uncertainty comes from the inaccurate description for the equilibrium cluster structure distribution. Since there is no configurational sampling, each distinct

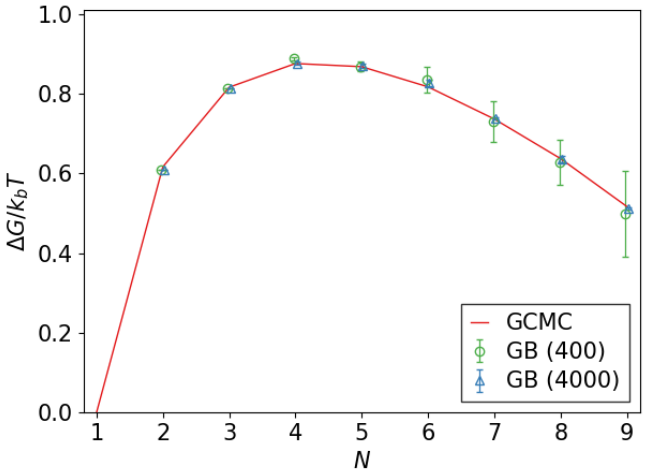


Figure 4.4: Free energy surface of the nucleation in lattice model with an interaction parameter of 1 k_bT. The red curve is the reference free energy surface calculated by GCMC. The green circles correspond to free energies predicted by our graph-based approach with 400 individual growths for each cluster size. The blues triangles represent results with 4000 individual growths for each cluster size.

particle arrangement of the cluster should be considered as a representative graph structure. And for weak-interaction system, all those representative structures share similar energies and the corresponding contributions deserve to be accurately described. However, during each nucleation step, 400 sampled parent structures may not able to accurately reflect the actual Boltzmann distribution for those many representative structures. And this type of error is accumulated during the entire simulation.

To solve this problem, we simply increased the number of sampled parent structures to 4000 and the corresponding standard deviation of the free energy surface dropped to 0.008 k_bT. This value is smaller than the standard deviation (0.02 k_bT) from the results predicted by GCMC. Considering there were 10^7 MC moves in each GCMC simulation, which requires 10^7 energy calculation, our approach provides a significant improvement in efficiency by only conducting 32000 energy calculations.

Since most electrolytes are strongly interacted, to better mimic those systems, we also investigated nucleation in lattice model with a high interaction parameter ($20\ k_bT$). The results are shown in Figure 4.5. Green circles represent free energy surface obtained from unbiased sampling, in which 400 parent structures were sampled from Boltzmann distribution and insertion positions of new particles were uniformly selected. Unlike weak-interaction system, the results given by unbiased graph-based method are systematically higher than the benchmark data predicted by GCMC. We can understand this discrepancy by tracking the evolution of the cluster. Because the interaction between nearest neighbours is $20\ k_bT$, at each cluster size, the energetically unfavored configurations have at least $20\ k_bT$ higher energy compared to stable ones, which leads to 5×10^8 times less probability in Boltzmann distribution. In practical, with limited number of samplings, such structures will never be selected as parent structures for the next nucleation step. And the evolution of the cluster is dominated by a few paths in which the child cluster always grows from the most stable parent cluster. This also explains the small uncertainty in the results since only a few paths were sampled. Although such unstable configurations do not make significant contributions in free energy calculation of the current cluster size and can be safely ignored. But by ignoring those configurations, we also eliminate possible evolution paths which lead to stable configurations in later nucleation steps. Thus in later steps, even the probability of the energetically stable configurations can not be accurately calculated.

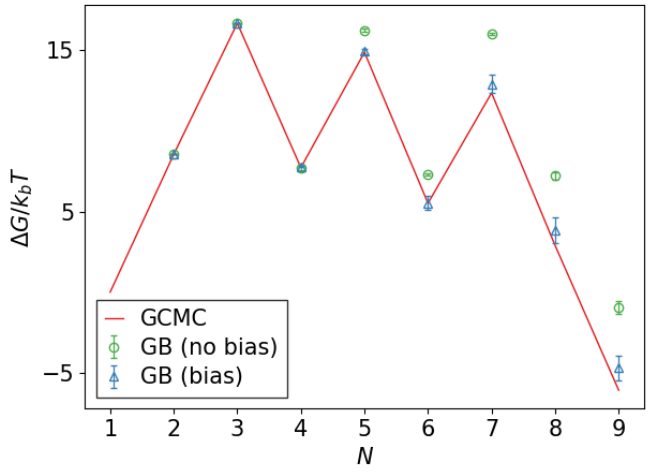


Figure 4.5: Free energy surface of the nucleation in lattice model with an interaction parameter of 20 k_bT. The red curve is the reference free energy surface calculated by GCMC. The green circles correspond to free energies predicted by our graph-based approach without bias. The blues triangles represent results predicted by our graph-based approach with bias.

To address this problem, we adopted biasing methods in selecting parent clusters and child clusters. For sampling parent structures, we selected the structure according to $exp(\ln P_{N,j^{parent}}/10)$ instead of Boltzmann probability $P_{N,j^{parent}}$. This bias equivalently increases the temperature of the system by 10 times and makes high-energy configurations more probable. In addition, we also added bias in selecting insertion positions to prefer stable child structures. This bias may not be necessary here but would benefit the simulation for complicated system where generating a stable configuration is a rare event. The corresponding results of the biased method are presented as blue triangles in Figure 4.5 and well reproduce the results from GCMC simulation.

4.3.2.2 Lattice Model in Solvents

In th previous section, the correctness of the graph-based approach was validated by simple lattice model with no explicit solvent considered. And the free energies

were simply energies. To be more analogous to the realistic case, in this section, we include explicit solvents in lattice model to address the solvent effects. And accordingly, the free energies are estimated through samplings over solvent configurations. Same lattice model from the previous section was adopted here. In addition, there were 30 solvent particles randomly positioned in the lattice. The interaction parameters were chosen as 1, 0.5 and 0.5 $k_b T$ for solute-solute, solute-solvent and solvent-solvent interactions. Thermal de Broglie wavelength was chosen as 3.16.

The benchmark data is generated by GCMC. In addition to insertion and deletion moves, a translational MC move for solvent particles was also employed. In this move, a trial displacement is generated for a random solvent particle and the energy change is used to determine whether to accept this move. In each GCMC simulation, 10^7 insertion and deletion moves were carried out for the cluster and each followed by 30 translational moves for the solvents.

In the graph-based simulation, most simulation parameters and procedures were inherited from previous section (400 parent structures were selected for each step from Boltzmann distribution and no other bias was employed). The only difference is in the estimation of term $< e^{-\beta \Delta U} >_{j,N+1'}$. Although no translational or rotational MC moves was employed for solute particles, sampling over solvent configurations is still required. Thus, term $< e^{-\beta \Delta U} >_{j,N+1'}$ cannot be simply calculated by energy change anymore. Here, we employed free energy perturbation (FEP)[99] for evaluating this term. And in each FEP, 1000 samplings of energy change ΔU were carried out and each followed by 30 translational MC moves for solvent particles. The results are shown in Figure 4.6

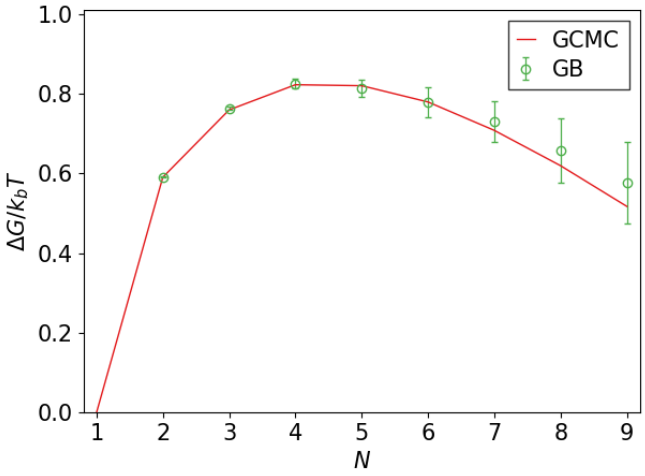


Figure 4.6: Free energy surface of the nucleation in lattice model with explicit solvents. The red curve is the reference free energy surface calculated by GCMC. The green circles correspond to free energies predicted by our graph-based approach.

Even with explicit solvents, our approach is still able to reproduce the free energy results from GCMC. The uncertainty is same with the uncertainty from the simulation without explicit solvent. Thus we believe this uncertainty is still from the error accumulated in the distribution of cluster structures and can be reduced by simply increasing the number of parent clusters sampled at each cluster size.

4.4 Understanding Graph-based Approach from Jarzynski Equality

4.4.1 Theory

In the previous section, we derived the graph-based approach for modeling the nucleation as in Eq. 4.29. However, this derivation indicates two strong requirements which cannot always be satisfied in practical simulations. As in Eq .4.11, in order to properly decompose the entire partition function into several components according to the graph structures, we need to guarantee that the phase space defined by

each graph structure does not overlap with others. This requirement can be easily satisfied in a simple system like lattice model, in which each graph is defined as the arrangement of the particles. And the overlap between phase spaces can be avoided by freezing the degrees of freedom for the cluster. However, in a practical atomistic system, totally freezing the movement of the cluster with sampling only over solvent degrees of freedom is unrealistic. Therefore, a proper definition of the graph is essential. If the graph structure is defined rigorously, the corresponding phase space for each graph will be limited and overlap can be avoided. But rigorous definition generates immense representative graph structures which require a significant amount of computing resources to calculate the free energies for all of them. However, if we employ a relatively relaxed definition, the non-overlap requirement will be violated considering the extensive phase space for each graph.

In addition to the above requirement, we also implicitly assume that the sub-ensemble defined by each graph structure can be effectively sampled. Otherwise, the free energy obtained for each structure as in Eq .4.30 is not an accurate equilibrium free energy and cannot be used to determine the Boltzmann distribution. This assumption may be valid for a limited sub-ensemble corresponding to a rigorously defined graph, but it cannot be easily satisfied when a relaxed graph definition is employed, in which each graph corresponds to an extensive phase space.

In this section, we provide another perspective from the Jarzynski Equality[92, 91] to understand this graph-based approach. With Jarzynski Equality, we can prove that even the above two requirements are violated, our approach is still valid by recovering the free energy from nonequilibrium process.

4.4.1.1 "Pruning" Method from Jarzynski Equality

It has been shown that for two thermodynamic states A and B, the irreversible work performed on the system during the nonequilibrium process switching state A to state B does not equal to the free energy difference, due to the lag developed between the nonequilibrium phase space and the instantaneous equilibrium distribution. And the average of irreversible works from an ensemble of nonequilibrium

trajectories provides an upper bound for the free energy change as a result of dissipation:

$$\langle W \rangle \quad \geqslant \quad \Delta G, \tag{4.34}$$

where W is the irreversible work and ΔG is the free energy change. Fortunately, Jarzynaki has shown that in fact one can recover the free energy change by taking the exponential average of nonequilibrium works:

$$\left\langle e^{-\beta W} \right\rangle \quad = \quad e^{-\beta \Delta G}, \tag{4.35}$$

where the exponential average is taken from an ensemble of trajectories starting from a canonical distributed initial conditions. This remarkable equation provides an ingenious way to estimate equilibrium quantity in terms of nonequilibrium measurements.

Now considering a three-states system with an initial state A, an ending state B and an intermediate state C. A group of nonequilibrium trajectories are carried out and the corresponding irreversible work for each trajectory is denoted as $W_{i,A\to B}$ where i is the index of the trajectory. Each trajectory from state A to state B can be treated as a two-step process consists of trajectories of $A \to C$ and $C \to B$, thus $W_{i,A\to B}$ can be written as:

$$W_{i,A\to B} \quad = \quad W_{i,A\to C} \quad + \quad W_{i,C\to B} \tag{4.36}$$

According to Jarzynaki Equality, the free energy change from state A to B and A to C can be written as:

$$\begin{aligned} \Delta G_{A\to B} \quad &= \quad -\frac{1}{\beta} \ln \sum_i exp(-\beta W_{i,A\to B}) \\ \Delta G_{A\to C} \quad &= \quad -\frac{1}{\beta} \ln \sum_i exp(-\beta W_{i,A\to C}) \end{aligned} \tag{4.37}$$

And $\Delta G_{C\to B}$ is given by:

$$\begin{aligned}\Delta G_{C\to B} &= \Delta G_{A\to B} - \Delta G_{A\to C} \\ &= -\frac{1}{\beta}\ln\sum_i \frac{exp(-\beta W_{i,A\to C})}{\sum_j exp(-\beta W_{j,A\to C})} exp(-\beta W_{i,C\to B}) \\ &= -\frac{1}{\beta}\ln\sum_i \tilde{P}_{i,A\to C} exp(-\beta W_{i,C\to B}),\end{aligned} \tag{4.38}$$

where

$$\tilde{P}_{i,A\to C} = \frac{exp(-\beta W_{i,A\to C})}{\sum_j exp(-\beta W_{j,A\to C})}. \tag{4.39}$$

Therefore the free energy change from state C to B is the exponential average of $W_{i,C\to B}$ weighted by probability $\tilde{P}_{i,A\to C}$. It is important to be noticed here, although $\tilde{P}$ resembles the Boltzmann factor, it is determined by nonequilibrium works rather than equilibrium free energies.

Eq. 4.38 gives no additional information if a sufficient number of irreversible trajectories can be carried out from the initial state to the final state. However, for energetically well-separated states, the required number of repetitions to obtain converged free energy is prohibitively high.[98] Echeverria and her co-workers[88] addressed this problem by segmenting nonequilibrium process into multiple components with several intermediate states. But it requires for each intermediate state, the canonical distribution can be recovered via equilibrium sampling, which is not practically achievable for strong-interaction systems like crystals. Here, we adopted the idea from Echeverria's work and developed a "pruning" method to efficiently sample well-separated states via Eq. 4.38 and without requiring equilibrium samplings for intermediate states.

Considering two well-separated states A and B with m intermediate states $C_1, C_2 \cdots C_m$, instead of carrying out nonequilibrium simulations directly from state A to B, we first launch m_1 simulations from A to C_1 and the free energy $\Delta G_{A\to C_1}$ is given by Eq. 4.37. Next, from the m_1 trajectories, we select m_2 of them based on the probability $\tilde{P}_{i,A\to C_1} = exp(-\beta W_{i,A\to C_1})/\sum_j exp(-\beta W_{j,A\to C_1})$ and

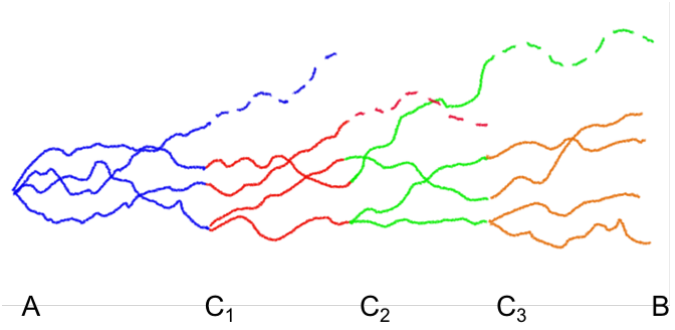


Figure 4.7: Schematic representation of the "pruning" method for a system with three intermediate states. Solid lines represent the nonquilibrium trajectories. Dashed lines are trajectories which are eliminated.

continue the simulation to reach state C_2. According to Eq. 4.38, the free energy $\Delta G_{C_1 \to C_2}$ can be calculated by taking the exponential average of $W_{i,C_1 \to C_2}$. We repeat this procedure until final state B to get free energies for all states. Notice that for selecting initial conditions at state C_n, the probability $\tilde{P}_{i,A \to C_n}$ is given by the normalized product of $\tilde{P}_{i,A \to C_{n-1}}$ and $\tilde{P}_{i,C_{n-1} \to C_n}$:

$$\begin{aligned} \tilde{P}_{i,A \to C_n} &= \frac{exp(-\beta W_{i,A \to C_n})}{\sum_j exp(-\beta W_{j,A \to C_n})} \\ &= \frac{exp(-\beta W_{i,A \to C_{n-1}}) exp(-\beta W_{i,C_{n-1} \to C_n})}{\sum_j exp(-\beta W_{j,A \to C_{n-1}}) exp(-\beta W_{j,C_{n-1} \to C_n})} \\ &= \frac{\tilde{P}_{i,A \to C_{n-1}} \tilde{P}_{i,C_{n-1} \to C_n}}{\sum_j \tilde{P}_{j,A \to C_{n-1}} \tilde{P}_{j,C_{n-1} \to C_n}} \end{aligned} \tag{4.40}$$

Since the trajectories has been already weighted by $\tilde{P}_{i,A \to C_{n-1}}$, at state C_n, we only need to select trajectories based on $\tilde{P}_{i,C_{n-1} \to C_n}$ which only depends on the work done in the previous step.

By selecting trajectories according to the nonequilibrium work at intermediate states, the above method tends to eliminate the trajectories which make trivia contributions to the free energies of later states, and bias more on trajectories which generate lower dissipation. Also, since two adjacent states are energetically similar,

the number of nonequilibrium trajectories required to achieve a converged work distribution is significantly reduced compared with the amount of repetitions needed for direct switching between initial state and final state.

Note that we can also select trajectories according to other probability as long as we reweight them at the end of simulation. Such probabilities vary for different systems and should help us better prune the trajectories.

4.4.1.2 Graph-based Approach as A Nonequilibrium Method

Now we can consider our graph-based approach as an example of the above method. In graph-based approach, we simulate the nucleation of a specific structure, which is actually a nonequilibrium process considering that the system always lags behind the equilibrium distribution which include all other structures in phase space. This nonequilibrium process starts with cluster of size 1 and ends at maximum cluster size, and cluster sizes in between are considered as intermediate states. The step-by-step nucleation procedure represents an implementation of the "pruning" method from Jarzynski Equality. In graph-based method, at each size N, we select parent structures according the Boltzmann distribution to carry out the next insertion step and the distribution is obtained by the free energy of each structure as in Eq .4.30. In fact the structure-based free energy is a nonequilibrium work considering only limited configurational sampling is carried out. And selecting parent structures according to corresponding free energies is actually pruning trajectories based on nonequilibrium works.

Notice here, the above idea requires Jarzynski Equality to be applied to grand canonical ensemble. More specifically, the procedure of calculating $\Delta G(j^{parent} \rightarrow j)$ in Eq. 4.30 needs to obey Jarzynski Equality. The calculation of term $< e^{-\beta \Delta U} >_{j,N+1'}$ in Eq. 4.30 is in fact carried out in canonical ensemble or isothermal-isobaric ensemble where Jarzynski Equality has been rigorously proved[91, 92, 97]. And for effective volume and ratio M, the values are estimated by Monte Carlo. Since Jarzynski has been proved in Monte Carlo simulation[91], we believe it can also be applied for those two terms and eventually be applied in the calculation of

$\Delta G(j^{parent} \rightarrow j)$.

With the above point of view, some requirements of the graph-based method are relaxed. Since we can view the phase space defined by each graph structure as a phase space sampled in a nonequilibrium simulation, the non-overlap rule is no longer required. And by treating the structure-based free energy as an irreversible work, it is not necessary to estimate this value accurately through equilibrium sampling. That is to say, we can employ a relaxed graph definition without enforcing sufficient sampling over the relatively large phase space defined by each graph structure. Without those requirements, we can simply extend this graph-based method to realistic systems with arbitrary graph definitions. Or more practically, we can define the phase space of a graph as the configurations can be sampled in a limited-time simulation.

4.4.2 Results: Lattice Model and NaCl

4.4.2.1 Lattice model in Solvent with Nonequilibrium Sampling

We have developed a "pruning" method from Jarzynski Equality and consider our graph-based simulation as an example of this nonequilibrium approach. To validate this idea, we employed an explicit nonequilibrium step in the graph-based simulation for lattice model and compared the results with the predictions from GCMC. In addition to proving the above argument, for comparison, we also carried out graph-dependent simulations directly from the initial state to the ending state without pruning trajectories (selecting parent clusters) and applied Jarzynski Equality to obtain the free energy surface. By comparing these results with results from graph-based (pruning) method, we can verify the correctness of the "pruning" method.

The same lattice model and simulation parameters from Section 4.3.2.2 was adopted here. To include an explicit nonequilibrium step in the approach, a short FEP simulation was adopted in the calculation of term $< e^{-\beta \Delta U} >_{j,N+1'}$ to prevent fully equilibrating the solvent configuration. In each FEP, only 10 samplings of energy change ΔU were carried out and each followed by 30 translational MC

moves for solvent particles. And with this nonequilibrium sampling, even for the same solute configuration, the uncertainty of the term $< e^{-\beta \Delta U} >_{j,N+1'}$ is ~0.5 $k_b T$ (compared with 0.04 $k_b T$ if 1000 energy samplings were carried out). Since we allow a greater uncertainty in the calculation of nonequilibrium work, in order to eventually obtain a converged free energy, at each cluster size, 1000 parent structures were selected according to the probability defined in Eq. 4.39 and corresponding simulations were carried out. The error bars were obtained by repeating the same procedure for four times. For comparison, we also carried out 1000 nonequilibrium trajectories directly from cluster size 1 to cluster size 9 without pruning trajectories at intermediate states. The free energy surface was calculated using Jarzynski Equality and error bars were generated from four duplicated simulations. The results are shown in Figure 4.8.

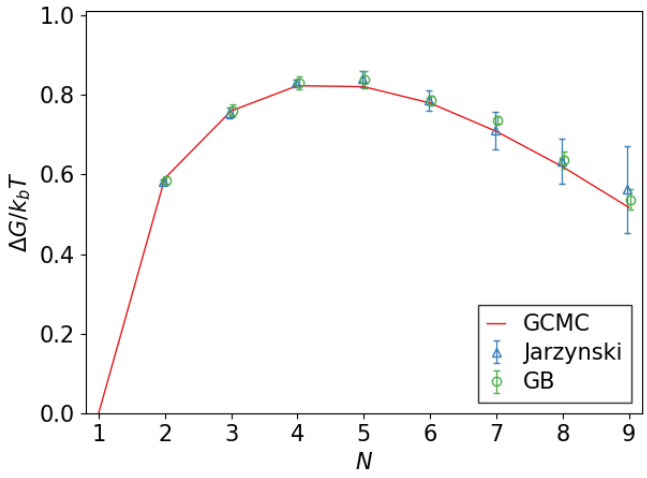


Figure 4.8: Free energy surface of the nucleation in lattice model with explicit solvents. The red curve is the reference free energy surface calculated by GCMC. The blue triangles represent the free energies predicted by original Jarzynski Equality. The green circles correspond to free energies predicted by our graph-based approach with nonequilibrium sampling employed.

With explicit nonequilibrium sampling employed, the graph-based method is still able to predict the correct nucleation free energy surface. It indicates that we

can safely treat this approach as a nonequilibrium approach from the point of view of Jarzynski Equality. Also by comparing the results from graph-based method with the predictions from original Jarzynski Equality, we can conclude that the "pruning" method not only gives a correct free energy prediction, but also reduce the uncertainty by ~5 times ($0.1\ k_bT$ vs $0.02\ k_bT$) using the same computing cost.

4.4.2.2 NaCl Nucleation

To extend the graph-based approach to weak electrolyte systems with explicit solvents, we benchmarked the approach via the low-solubility rock-salt model, comparing against results from our hybrid GCMC/MD approach. Since the rigorous graph definition is not applicable for a realistic atomistic system, here we take the perspective from Jarzynski Equality and treat each graph-based nucleation as a nonequilibrium process. With this point of view, rigorous graph definition is not necessarily required and the restraints to form a such specific graph is also no longer needed. But we can still give a relaxed graph definition and consider the approach in a graph-based manner. In the following simulations, no graph-related restraints is established on the cluster except MST restraints to help enforce the cluster criterion. Therefore each limited-time nonequilibrium simulation will automatically explore a sub-ensemble which contains one or more cluster structures. We can define all structures corresponding to each sub-ensemble as a graph and describe the nonequlibrium simulation in the language of the graph-based approach. Notice here, we make the above argument only to make the simulation correspond to the graph-based method. Alternatively, we can simply consider it as nonequilibrium simulation that samples (in a finite amount of time) a limited sub-ensemble.

The same rock-salt model and parameters were adopted here from Chapter 3 and MD simulations of time step 2 fs were carried out. In all simulations, PBCs were employed with LJ cutoff of 1.5 nm. No long-range corrections were considered. Particle mesh Ewald(PME) were carried out for Coulomb interactions. Temperature was chosen at room temperature (298.15K), enforced using a Langevin thermostat. The pressure was fixed at 1 atm using a Monte Carlo barostat. The Stillinger's

cluster criterion was chosen at 0.35 nm and enforced by MST restraints. Only connections between cations and anions were considered. The MST restraints were updated every 200 fs according to the adjacency matrix to avoid artificial effects on the cluster. A total of 2500 water molecules were positioned around the NaCl cluster. The insertion of cations and anions are implemented as separate steps with cations inserted first. The cluster size for ion pairs was sampled from 1 to 17. At each intermediate state (including the states with an additional cation), 160 parent clusters were selected from the ending conditions of previous trajectories and the child configurations are generated accordingly by attaching a new cation/anion.

In order to prevent eliminating parent clusters which may lead to non-trivial contributions in later step, instead of based on probability $\tilde{P}_i$ as defined in Eq. 4.39, we sampled parent clusters according to a weighted probability $\tilde{P'}_i$. In the early stage of nucleation (cluster size < 7), it is necessary to include all parent clusters including the ones corresponding to trajectories with large nonequilibrium works. Therefore we employed a bias which can be expressed as:

$$\tilde{P'}_i = \frac{exp[3 \ln \tilde{P}_i / \ln (\tilde{P}_{max}/\tilde{P}_{min})]}{\sum_j exp[3 \ln \tilde{P}_j / \ln (\tilde{P}_{max}/\tilde{P}_{min})]}, \tag{4.41}$$

Where $\tilde{P}_{max}$ is the largest probability among $\tilde{P}_i$ for all indices i and $\tilde{P}_{min}$ is the smallest one. By employing this bias, with 160 samplings, the least probable parent structure can still be sampled. When the cluster size is relatively large ($\geqslant 7$), some of the trajectories accumulate much more work than others and are probably not going to make any significant contributions to the later states. In such cases, we chose bias as:

$$\tilde{P'}_i = \frac{exp(\ln \tilde{P}_i/2)}{\sum_i exp(\ln \tilde{P}_i/2)}, \tag{4.42}$$

which simply doubles the temperature in Boltzmann factor. This increases the probability of selecting higher energy parent clusters relative to a standard canonical distribution. The corresponding reweighting factor $\tilde{P}_i/\tilde{P'}_i$ was included in the calculation of the free energy and also carried out to the probability estimation for the next step.

For each selected parent cluster, we estimated the effective volume by carrying out numerical Monte Carlo coupled with MD simulations. 200 ps simulation was carried out for each parent cluster and the configuration was sampled every 200 fs. For each sampled configuration, 10 trial insertions were generated into a region around the cluster. The region was defined as a sphere which centered at the center-of-mass of the cluster. The radius of the sphere was calculated as 0.35 nm (from Stillinger's cluser criterion) plus 1.3 times maximum radius of the parent cluster where the maximum radius was estimated over all 160 parent clusters generated from the previous step.

Among all the qualified child configurations generated in volume estimation, we could randomly choose one as the candidate child. But when the size of the cluster gets larger, randomly picking one insertion site would probably lead to an unfavorable child cluster structure. For this reason, we weighted the qualified child configurations based on their minimized energy. For each configuration, we turned on the interaction of the inserted cation/anion, removed all water molecules and minimized the energy of the cluster by optimizing the positions of ions. We ranked all configurations by the minimized energy and abandoned 5% most energetically unfavorable configurations. For the rest configurations, we calculated the energy difference ΔE between highest energy and lowest energy, and each configuration was weighted by

$$P_i = \frac{exp(-3\beta E_i/\Delta E)}{\sum_j exp(-3\beta E_j/\Delta E)} \tag{4.43}$$

After selecting the child configuration, the corresponding reweighting factor was employed in free energy calculation and inherited to the next nucleation step.

Corresponding to term $< e^{-\beta \Delta U} >_{j,N+1'}$ in Eq. 4.30, the nonequilibrium work of turning on the interaction of the inserted cation/anion was calculated by TI. During each TI, the LJ interaction was switched on first using a soft-core potential followed by the initiation of the Coulomb interaction. For each interaction, 10 λ values were chosen from Gaussian quadrature on 0 to 1. After the system switched to a new λ value, a 20 ps simulation was carried out to equilibrate the system followed by a 50 ps sampling (for LJ interaction) or 100 ps (for Coulomb interaction). The energy

derivative was sampled every 200 fs. To be noticed here, during the simulation, MST restraints were updated without including the new ion. To prevent the new ion moving apart from the cluster, a soft-wall potential defined in Eq. 2.24 was created between the inserted cation/anion and closest anion/cation before the simulation.

For the estimation of ratio M, a 200 ps simulation was carried out for each child cluster after fully turning on the interaction. The configuration was sampled every 200 fs. For each sampled cluster configuration, 10 trial deletions were generated and the cluster criterion was rechecked. The results are shown in Figure 4.9. The error bar for graph-based method was estimated by taking block average at each cluster size with 4 blocks and 40 data points in each block.

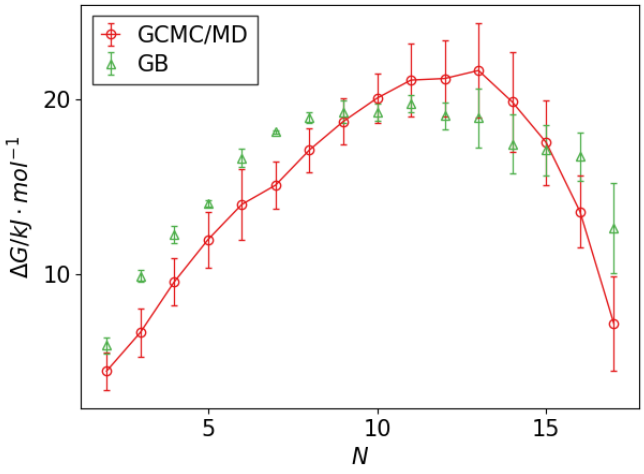


Figure 4.9: Free energy surface of the nucleation in rock-salt model. The red curve is the reference free energy surface calculated by the hybrid GCMC/MD approach. The green triangles correspond to free energies predicted by our graph-based approach.

The graph-based approach predicts a similar critical cluster size (11) and free energy barrier (19.8 kJ/mol^{-1}) with our hybrid GCMC/MD method (13, and 21.7 kJ/mol^{-1} respectively). And the corresponding nucleation rate difference is within one order of magnitude. At all other cluster sizes less 16, the difference in free energy is less than 1.5 k_bT. Considering both methods have significant error bars,

the agreement between two approaches is great. We noticed that beyond size 15, the graph-based method cannot predict a rapidly decreasing free energy anymore. That is because when cluster size is large, it is difficult to explore all child structures and find the most preferred configuration, especially when only 160 children are considered. Without including the stable structures, the estimated free energy will be much higher. Another possible explanation for this discrepancy is that the error accumulated during the trajectory pruning becomes non-trivial when the graph-based method proceeds to larger cluster sizes. And we hope to address those issues by migrating our program to a high throughput implementation and carry out more than 160 graph-dependent simulations for each cluster size with high throughput computing in the future.

4.5 conclusion

In this chapter, we presented a graph-based simulation approach which allows us to efficiently simulate the nucleation associated with polymorphism. This method addresses the sampling challenges from polymorphism by distributing the phase space into multiple parallel nonequilibrium samplings and benefits from high throughput computing. Using this approach, we examined the free energy surface and associated barriers for the nucleation in lattice model and found excellent agreement with the prediction from GCMC. We also extended this graph-based approach to low-solubility materials/weak electrolyte with Jarzynski Equality and well reproduced the results from the previous hybrid GCMC/MD approach. We anticipate that this approach can be easily extended to study other weak electrolytes/low-solubility materials, such as calcium carbonate and lithium fluoride, and can be applied to biomineralization and synthesis of crystalline materials.

[88] Echeverria, Ignacia, and L Mario Amzel. 2010. Helix propensities calculations for amino acids in alanine based peptides using jarzynski's equality. *Proteins: Structure, Function, and Bioinformatics* 78(5):1302–1310.

[89] Gale, Julian D, and Andrew L Rohl. 2003. The general utility lattice program (gulp). *Molecular Simulation* 29(5):291–341.

[90] Harding, John H, Dorothy M Duffy, Maria L Sushko, P Mark Rodger, David Quigley, and James A Elliott. 2008. Computational techniques at the organic-inorganic interface in biomineralization. *Chemical reviews* 108(11):4823–4854.

[91] Jarzynski, Christopher. 1997. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E* 56(5):5018.

[92] ———. 1997. Nonequilibrium equality for free energy differences. *Physical Review Letters* 78(14):2690.

[93] Kirkwood, John G. 1935. Statistical mechanics of fluid mixtures. *The Journal of chemical physics* 3(5):300–313.

[94] Li, Xinyi, and JR Schmidt. 2019. Modeling the nucleation of weak electrolytes via hybrid gcmc/md simulation. *Journal of chemical theory and computation* 15(11):5883–5893.

[95] Steinbrecher, Thomas, David L Mobley, and David A Case. 2007. Nonlinear scaling schemes for lennard-jones interactions in free energy calculations. *The Journal of chemical physics* 127(21):214108.

[96] Stillinger Jr, Frank H. 1963. Rigorous basis of the frenkel-band theory of association equilibrium. *The J. Chem. Phys.* 38(7):1486–1494.

[97] Williams, Stephen R, Debra J Searles, and Denis J Evans. 2007. Deterministic derivation of non-equilibrium free energy theorems for natural isothermal isobaric systems. *Molecular Physics* 105(8):1059–1066.

[98] Zuckerman, Daniel M, and Thomas B Woolf. 2002. Theory of a systematic computational error in free energy differences. *Physical Review Letters* 89(18): 180602.

[99] Zwanzig, Robert W. 1954. High-temperature equation of state by a perturbation method. i. nonpolar gases. *The Journal of Chemical Physics* 22(8):1420–1426.

5 CONCLUSION AND FUTURE DIRECTIONS

In this dissertation, we have presented two methodologies for simulating the nucleation of weak electrolytes with explicit solvents. With appropriately addressing the challenges of dilute systems, mass transport and ineffective sampling, our approaches successfully modeled nucleations in LJ system, lattice model and low-solubility rock-salt electrolyte. Since those challenges are common in solution-phase nucleation, the approaches developed herein can be extended to a wide range of physical systems, such as $CaCO_3$ and MOFs. However, there remain some issues that we hope to address in future work.

In Chapter 4, we utilized the graph-based method to predict the nucleation free energy for low-solubility rock-salt model, but at certain cluster size, the free energy cannot be accurately estimated due to the error accumulated during the trajectory pruning and the limited probability to find stable configurations. And we believe this problem can be solved by designing better biasing strategies and also increasing number of trajectories carried out at each cluster size. Therefore currently we are migrating the program to a high throughput implementation. With more computing power, we anticipate the predicted nucleation free energy surface from graph-based method can be extended to larger cluster sizes. Also, due to the strong interactions associated with the current low-solubility weak-electrolyte model, the uncertainty of the predicted results for this model is relatively large. For this reason, we plan to further validate this method against a simpler model with smaller charges and obtain relative accurate data to verify this method.

Also in Chapter 4, we developed a trajectory-pruning method based on Jarzynski Equality and benchmarked it on the nucleation of lattice model. The results indicate this method improves the accuracy of the original Jarzynski Equality with the same computing cost. But more importantly, for the free energy difference between two well-separated states, whose estimation usually requires a prohibitively high number of nonequilibirium trajectories using original Jarzynski Equality, this "pruning" method provides a much more affordable way to predict the value by

abandoning trajectories with trivial contributions at intermediate states. To validate this argument, a systematic study of this method is necessary, and more practical systems other than lattice model are required. Since this method is a general approach derived from Jarzynski Equality without any further assumptions, we anticipate a wide range of applications for this method in addition to nucleation.

Another possible improvement of the trajectory-pruning method is to recover the equilibrium distribution at the intermediate states. In the original graph-based approach derived in Section 4.3, the equilibrium distribution can be recovered from individual graph-dependent simulations via Boltzmann factor. But once we consider the graph-dependent simulations as trajectory-pruning simulations, the corresponding weight factor will be calculated by irreversible work and no longer considered as Boltzmann factor. Whether we can recover the equilibrium distribution from nonequilibrium simulations requires further theoretical study. Note here, successful demonstration of the above argument would give us access to other information in addition to the free energy. As in the case of our graph-based method, obtaining the accurate distribution of the cluster configurations is essential for structural analysis.

Assuming the above challenges can be addressed. Our next goal is to understand the nucleation of $CaCO_3$. Despite its widely appearance in geological deposition, biomineralization and marine sedimentation, details of the early-stage nucleation is often lacking, as both classical and nonclassical mechanisms have been utilized to describe the process.[104, 100, 102, 103, 101] Other than its unpredictable mechanism, the nucleation of $CaCO_3$ also presents more challenges for molecular simulation comparing to simple rock-salt model due to the high charge density, complex ion structure and potential polymorphism. A proper forcefield which can accurately describe multiple polymorphs and cluster-water interfaces is required. In addition to the forcefield, to employ the hybrid GCMC/MD and graph-based approach, a strategy of gradually introducing complex ions into the system is also necessary. In fact, hybrid GCMC/MD approach has been tested on the $CaCO_3$ nucleation, but due to the inability to address the challenge from the ineffective configurational sampling, the simulation cost is prohibitively high. Therefore, with appropriate

handling of this concern, we anticipate the graph-based method can effectively generate meaningful results.

Finally, we are interested in extending these simulation strategies to the nucleation of MOFs. Successful synthesis of promising MOF structures often relies on detailed control of solvent(s), temperature, supersaturation, or myriad other parameters. But the lack of atomistic understanding for MOF nucleation presents significant obstacles to adjust those experimental conditions. With the ability to simulate solution-phase nucleation, we expect that our methodologies are able to give a quantitative prediction for MOF nucleation. Especially, the ability to explicitly describe the solvents could allow our approaches to understand the solvent-template effect on the formation of such porous structures. Notice here, unlike the relative flexible structures for NaCl or $CaCO_3$ nuclei which may contain multiple morphologies, the structure of MOF nucleus is probably ordered as in lattice model and does not have a strong flexibility. For this reason, we believe the graph-based method can be easily applied to the nucleation of MOFs.

In summary, I list the possible future publications from the above work:

- With high throughput implementation of graph-based method, we hope to further verify this method and this will lead to a publication about the graph-based method development.
- We also plan to systematically study the trajectory-pruning method and apply it to other systems. This will lead to a publication about the trajectory-pruning method development.
- Later we will apply the graph-based method to the nucleation of MOFs/$CaCO_3$ and work on corresponding publications.

REFERENCES

[100] Gebauer, Denis, Antje Völkel, and Helmut Cölfen. 2008. Stable prenucleation calcium carbonate clusters. *Science* 322(5909):1819–1822.

[101] Nielsen, Michael H, Shaul Aloni, and James J De Yoreo. 2014. In situ tem imaging of $caco_3$ nucleation reveals coexistence of direct and indirect pathways. *Science* 345(6201):1158–1162.

[102] Pouget, Emilie M, Paul HH Bomans, Jeroen ACM Goos, Peter M Frederik, Nico AJM Sommerdijk, et al. 2009. The initial stages of template-controlled $caco_3$ formation revealed by cryo-tem. *Science* 323(5920):1455–1458.

[103] Rieger, J, T Frechen, G Cox, W Heckmann, C Schmidt, and J Thieme. 2007. Precursor structures in the crystallization/precipitation processes of caco 3 and control of particle formation by polyelectrolytes. *Faraday discussions* 136: 265–277.

[104] Teng, H Henry, Patricia M Dove, Christine A Orme, and James J De Yoreo. 1998. Thermodynamics of calcite growth: baseline for understanding biomineral formation. *Science* 282(5389):724–727.

6 LIST OF PUBLICATIONS

1. Van Vleet, Mary J and Weng, Tingting and Li, Xinyi and Schmidt, JR. “In Situ, Time-Resolved, and Mechanistic Studies of Metal–Organic Framework Nucleation and Growth". *Chem. Rev.* 118. 7 (2018), p. 3681-3721.

2. Li, Xinyi and Schmidt, JR. “Modeling the Nucleation of Weak Electrolytes via Hybrid GCMC/MD Simulation". *J. Chem. Theory Comput.* 15. 11 (2019), p. 5881-5893.

3. Li, Xinyi and Schmidt, JR. “Modeling the Nucleation of Weak Electrolytes via Graph-based Simulation”. *in preparation*