universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

An Always-Accepting Transition Path Sampling Algorithm for Overdamped Dynamics: Application to CO_2 Clathrate Hydrate Nucleation

verfasst von | submitted by

Magdalena Häupl BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 876
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Physics
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Christoph Dellago
Mitbetreut von | Co-Supervisor:	Alessandro Coretti PhD

Acknowledgements

First, I would like to thank my supervisors Christoph Dellago and Alessandro Coretti. Christoph's deep understanding and inspiring approach to statistical physics has made me appreciate it since my first course on the topic. He has the outstanding ability of making the most complex of topic approachable, which also carried over into this thesis.

It is because of Alessandro that this project has become what it is. I have benefitted from his deep knowledge on the topic through many long discussions and he always believed in my work even when I was unsure about it. He played a big part in making this a great experience.

I would like to thank my boyfriend Konsti, who has supported me in this journey from the start. I am deeply inspired by him and he is always a helpful presence, both with my studies and especially beyond. I feel very grateful, to have him in my life.

There has been many other people that have supported me throughout my studies, and there could be much time taken to thank people for their individual qualities, but in this case I will keep it short. It is truly a gift to have friends that both make the time a little lighter and fun, but also genuinely are people that I look up to. I want to especially thank David, Szymon, Anni, Lilly and Isak for being with me from the start and Axel for lending helpful insight into impossible worksheets (such as classical mechanics).

Ich möchte mich außerdem bei meinen Eltern bedanken, ohne ihre konstante und vollständige Unterstützung wäre ich nicht da, wo ich bin. Ihr ständiger Glauben in mich und das Interesse an meiner Arbeit ist unendlich viel wert und macht schwere Tage leichter. Ganz wichtig ist es auch, mich bei meiner Schwester Johanna zu bedanken. Es gibt mir sehr viel Freude, sie für immer in meinem Leben zu wissen, als jemand, der mich beeindruckt, inspiriert und hinter mir steht.

Abstract

Rare event sampling attempts to resolve the critical challenge of quantifying transition pathways, despite them being exponentially improbable fluctuations that determine material evolution yet still evade equilibrium observation. As a foundational approach in rare-event sampling, Transition Path Sampling enables the study of transition regions through iterative trajectory generation. The method propagates new trajectories by selecting configurations from existing pathways and initiating simulations from these points. Standard Transition Path Sampling methods typically reject $\geq$60% of generated paths, which constitutes a substantial inefficiency when sampling complex systems. It is therefore the goal of this thesis to develop an algorithm that completely overcomes the need to reject paths, instead always accepting proposed trajectories. This thesis provides a derivation for this method from the detailed balance condition and evaluates its performance when applied to clathrate hydrate nucleation, a particularly complex rare event. We find that our algorithm greatly increases sampling efficiency, requiring 2.8 times fewer force evaluations for half-trajectory replacement and 1.4 times fewer for full-trajectory replacement than standard TPS. This efficiency gain enables previously unattainable sampling of the Transition Path Ensemble for clathrate hydrate nucleation, revealing a prominent crystalline nucleation channel at 260 K and 500 bar.

Kurzfassung

Das Sampling seltener Ereignisse zielt darauf ab, die entscheidende Herausforderung der Quantifizierung von Übergangspfaden zu lösen. Diese stellen exponentiell unwahrscheinliche Fluktuationen dar, welche die Materialentwicklung bestimmen, sich jedoch einer Beobachtung im Gleichgewicht entziehen. Als grundlegende Methode für das Sampling seltener Ereignisse ermöglicht das 'Transition Path Sampling' die Untersuchung von Übergangsregionen durch iterative Erzeugung von Trajektorien. Die Methode generiert neue Trajektorien, indem Konfigurationen aus bestehenden Pfaden ausgewählt und Simulationen von diesen Punkten aus gestartet werden. Standard Transition Path Sampling Methoden verwerfen typischerweise $\geq$60% der erzeugten Pfade, was eine erhebliche Ineffizienz bei der Untersuchung komplexer Systeme darstellt.

Daher ist es das Ziel dieser Arbeit, einen Algorithmus zu entwickeln, der die Notwendigkeit, Trajektorien zu verwerfen, vollständig überwindet. Diese Arbeit liefert eine Herleitung dieser Methode aus der detaillierten Gleichgewichtsbedingung und bewertet ihre Leistung bei der Anwendung auf die Nukleation von Klathrathydraten, einem besonders komplexen seltenen Ereignis.

Wir stellen fest, dass unser Algorithmus die Sampling-Effizienz erheblich steigert: Er benötigt 2,8-mal weniger Kraftauswertungen für den Austausch von Halb-Trajektorien und 1,4-mal weniger für den Austausch vollständiger Trajektorien im Vergleich zum Standard-TPS.

Dieser Effizienzgewinn ermöglicht eine bisher unerreichbares Sampling des Transition-Path-Ensembles für die Klathrathydrat-Nukleation. Dabei wird ein ausgeprägter kristalliner Nukleationskanal bei 260 K und 500 bar identifiziert.

Contents

1 Introduction

Rare events govern material evolution from the exponential tails of probability distributions, forming bridges between metastable states that escape equilibrium sampling, yet dictate macroscopic behavior. Protein folding, activation-limited chemical reactions, and phase transitions all exhibit rare barrier-crossing events, among many other examples. Nucleation, the dominant mechanism for first-order phase transitions like crystallization from liquid or droplet condensation from vapor, represents a fundamental example of this challenge: a fluctuation overcoming an energy barrier with exponentially low probability redirects an entire system. [23].

This process exhibits a fundamental timescale separation: the system undergoes rapid, local fluctuations continuously, but only those rare fluctuations exceeding the barrier trigger transitions. The barrier-crossing event itself unfolds on molecular rearrangement timescales ($\sim 10^{-12} - 10^{-9}$ s), while the preceding exploration of metastable states spans orders of magnitude longer.

Molecular simulation is an extremely powerful tool, especially for studying the length scales associated with nucleation, but it is fundamentally limited by the timescales accessible to simulation. Moreover, the time resolution required to resolve the molecular process under investigation imposes that most of the simulation time is spent in waiting for the event to happen (if ever). Addressing this challenge has driven significant innovation in computational rare-event methodologies, among which Transition Path Sampling (TPS) is a prominent example. This method addresses the sampling problem by focusing computational effort on reactive trajectories while minimizing exploration of thermodynamically stable basins. This is achieved through iterative exploration: new trajectories are generated from configurations along previously visited reactive pathways. These new trajectories are then either accepted or rejected, to assign the correct statistical weight, progressively enhancing sampling of the Transition Path Ensemble (TPE) [10, 17].

Gaining a sufficient understanding for the TPE requires a large number of paths. In complex systems, this imposes severe computational demands that risk overlooking critical transition pathways within the energy landscape. Despite this, traditional TPS methods involve rejecting and therefore discarding what are typically 60 - 80 % of all generated paths. This percentage of rejected paths constitutes valuable computational resources and time, ultimately producing what is regarded as waste.

In this thesis, we derive a method for TPS that eliminates this waste, exclusively

generating transition paths that will be accepted. Therefore, this new method is termed Always Accepting Transition Path Sampling (AA-TPS). It tackles the two relevant sources of rejection in standard TPS. The first is the restriction on which ones of the stable states must be reached, which is eliminated by the introduction of a widely applicable overdamped description of the system dynamics. The second is the partial rejection due to the acceptance criterion, which is replaced by an a posteriori reweighting of all generated paths to guarantee that the correct transition path ensemble is sampled. The reweighting factor is derived from the detailed balance condition.

The AA-TPS method is first probed on its ability to correctly sample a 2D potential landscape compared to state-of-the-art alternatives. To thoroughly investigate the performance of AA-TPS, and in particular to verify its scalability to real-world applications, the method is used to sample a system, prominent for its computational and theoretical complexity, namely the formation of CO_2 Clathrate Hydrates. Clathrate hydrates are crystalline inclusion compounds in which water molecules (host) form hydrogen-bonded cage-like structures that encapsulate small guest molecules under specific thermodynamic conditions [33]. Despite their resemblance to ice, CO_2 clathrate hydrates exhibit distinct nucleation and growth kinetics due to the dynamic coupling between host water cages and guest CO_2 molecules. This interplay creates a complex free-energy landscape with high-energy barriers and multiple competing nucleation pathways. This system represents a suitable opportunity to analyze the practical potential of AA-TPS beyond the theoretical advantages, while additionally offering insights into the free energy landscape of CO_2 clathrate hydrates.

This thesis is structured as follows: In chapter 2 the theoretical framework of TPS is laid out by first establishing the ensemble formalism and then presenting the sampling algorithms, One-way Shooting Transition Path Sampling (1wS-TPS) as the benchmarking method and then introducing AA-TPS. Chapter 3 introduces clathrate hydrates, covering fundamental structural properties, nucleation hypotheses, and the theoretical framework including overdamped dynamics and key order parameters. Chapter 4 presents and discusses TPS simulation outcomes, analyzing path correlations, nucleation pathways, and critical nucleus characteristics.

2 Theory

2.1 Transition Path Sampling

Transition Path Sampling (TPS) is a computational methodology developed to provide an unbiased understanding of systems characterized by rare transitions. [10] The situation that TPS aims to address is generally characterized by the features presented in the following. The system of interest has an unknown and generally rough energy barrier, separating two defined metastable states A and B. Transitions between these two states can be observed but are exceedingly rare on the timescale of simulation. Since we have knowledge about the system's evolution following deterministic or stochastic equations of motion, we can simulate the system. Different versions of TPS are therefore unified by their common approach to addressing these situations by evolving transition paths according to a Markov chain Monte Carlo approach, finally constructing an unbiased ensemble of paths, spanning from the initial state A to the final state B.

2.1.1 Constructing the Transition Path Ensemble

In TPS, our aim is to obtain a Transition Path Ensemble (TPE) that fully characterizes the particular transition of interest in the system. This transition is defined by the states it connects, and a trajectory is deemed reactive if it starts in state A and reaches state B at a later point. As the number of samples tends toward infinity, the collected ensemble converges to the equilibrium transition path distribution for the transition. Consequently, the TPE comprehensively describes all possible pathways for the transition. Analysis of these pathways can now yield the degrees of freedom governing the transition, revealing their individual importance to the process and their progression throughout the transition.

Let us now try to define this TPE, which we denote as $P_{\mathrm{AB}}(X)$, where X defines an ordered sequence of configurations with a defined length L:

$$X(L) \equiv \{x_0, x_1, \ldots, x_{L-2}, x_{L-1}\}. \tag{2.1}$$

Each configuration x_i represents a complete snapshot of the state at index i. The composition of this state x_i depends on the governing dynamics. For systems evolving under Hamiltonian or Langevin dynamics, x_i consists of both all positions

and all velocities at that point in time, so $x_t = (\vec{r}(t), \vec{p}(t))$. In the overdamped regime, where strong friction dominates and faster momentum degrees of freedom are integrated out, the system's state is fully characterized by positions alone, i.e. $x_t = (\vec{r}(t))$.

To be able to define the ensemble of all transition paths, we first need to define the probability $P[X(L)]$ of observing an arbitrary trajectory $X(L)$. A trajectory is constructed by first selecting an initial state x_0 followed by the sequential generation of new configurations, until a maximum path length L is reached. Assuming the underlying dynamics to be Markovian, each new state depends only on the previous one, leading to the following definition:

$$P[X(L)] = \rho(x_0) \prod_{i=0}^{L-2} p(x_i \to x_{i+1}). \tag{2.2}$$

When a new configuration is generated from a previous one, this process follows the short-time transition probability $p(x_i \to x_{i+1})$. This short time transition probability depends on the system's underlying dynamics. For Hamiltonian dynamics, the equations of motion are given as

$$\dot{r} = \frac{\partial H(r,p)}{\partial p}, \qquad \dot{p} = -\frac{\partial H(r,p)}{\partial r}, \tag{2.3}$$

which enforces deterministic evolution. Accordingly, for any given configuration x_i at index i there exists only one state that it maps onto at a later index $i+1$,

$$x_{i+1} = \phi(x_i). \tag{2.4}$$

Hence, the short-time transition probability for Hamiltonian dynamics is

$$p(x_i \to x_{i+1}) = \delta[x_{i+1} - \phi(x_i)]. \tag{2.5}$$

In the case of Brownian dynamics, an example of stochastic evolution, the governing equation is the overdamped Langevin equation:

$$m\gamma\dot{r} = -\frac{\partial V(r)}{\partial r} + \mathcal{F}, \tag{2.6}$$

where m is the mass of the particle and γ is the friction coefficient.

The stochastic behavior is introduced through the Gaussian random force $\mathcal{F}$, which therefore has the following properties

$$\langle \mathcal{F}(t) \rangle = 0, \tag{2.7}$$

$$\langle \mathcal{F}(t)\mathcal{F}(0) \rangle = 2m\gamma k_{\mathrm{B}} T \delta(t), \tag{2.8}$$

where k_B is the Boltzmann constant, T is the temperature. The variance is given by the fluctuation-dissipation theorem. From this, we find that the Brownian short-time transition probability

$$p(x_i \to x_{i+1}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{\left(x_{i+1} - x_i + \frac{1}{\gamma m}\frac{\partial V}{\partial x}\right)^2}{2\sigma^2} \right\}, \tag{2.9}$$

with the variance σ^2 as

$$\sigma^2 = \frac{2k_\mathrm{B}T}{m\gamma}\Delta t, \tag{2.10}$$

where Δt is the time step.

What turns a simple trajectory into a transition path, however, is the fact that it connects the two relevant stable states A and B. To impose this condition we rewrite Eq. (2.2) as

$$P_\mathrm{AB}\left[X(L)\right] = \frac{1}{Z_\mathrm{AB}(L)} h_\mathrm{A}(x_0) h_\mathrm{B}(x_{L-1}) \rho(x_0) \prod_{i=0}^{L-2} p(x_i \to x_{i+1}), \tag{2.11}$$

where the characteristic functions h_A and h_B are introduced. They are defined as

$$h_k(x) = \begin{cases} 1 & \text{if } x \in k, \\ 0 & \text{if } x \notin k, \end{cases} \tag{2.12}$$

with $k \in \{\mathrm{A}, \mathrm{B}\}$. Additionally, since we are now choosing to consider only a subset, namely only reactive paths, of our earlier defined probability distribution $P[X(L)]$, the distribution needs to be normalized by the *path partition function* Z_AB,

$$Z_\mathrm{AB}(L) \equiv \int Dx(L) h_A(x_0) P[x(L)] h_B(x_{L-1}), \tag{2.13}$$

in which we integrate over all frames using the definition

$$\int Dx(L) \equiv \int \ldots \int \mathrm{d}x_0 \mathrm{d}x_1 ... \mathrm{d}x_{L-1}. \tag{2.14}$$

Since Eq. (2.11) requires the trajectory to always be of length L with the endpoint x_{L-1} lying inside the state B, this version of TPS is commonly referred to as 'fixed-length' TPS.

Due to thermal fluctuations, systems often briefly exit the stable states after initial crossing events. Consequently, trajectories that successfully reach B may subsequently leave it by time $t = L\Delta t$, causing technically reactive paths to be

rejected. It is therefore reasonable to define a condition that defines a path as reactive as soon as it reaches the target state. With this definition, we can simply stop the simulation as soon as it hits one of the two states. Unlike before, this requires continuous monitoring during the simulation to detect state crossings.

For this reason a 'flexible-length' ensemble can be defined as

$$P_{\mathrm{AB}}[X(L)] \propto H_{\mathrm{AB}}(X)\rho(x_0)\prod_{i=0}^{L-2} p(x_i \to x_{i+1}). \tag{2.15}$$

Here, $H_{\mathrm{AB}}(X)$ is a characteristic function, similar to h_{A} and h_{B} in Eq. (2.11), ensuring that only reactive paths are considered. Accordingly, it is unity only if the trajectory X successfully transitions from state A to state B. In this case, that means that the configuration x_0 is in state A and x_{L-1} in state B, while all other configurations in between should lie outside of A and B.

2.1.2 Sampling the Transition Path Ensemble via One-Way Shooting

Transition Path Sampling (TPS) is often described as 'Monte Carlo in Path Space'. In common Monte Carlo simulations, we generate the desired distribution by repeatedly proposing new configurations, which will be accepted or rejected according to a certain acceptance probability. Through this acceptance probability, detailed balance is enforced, which ensures that the simulation eventually samples the equilibrium distribution, assuming the system is ergodic. By counting how often a configuration is visited and how long the system remains in that configuration (e.g. newly proposed configurations are rejected) the configuration is given its correct statistical weight. Monte Carlo techniques can indeed be extended to the sampling of trajectories. Instead of performing a random walk through configuration space, we now transition between entire trajectories in path space. While various ways of transitioning between trajectories have been proposed, the most common version is the so-called 'shooting algorithm'. This involves picking a configuration, termed 'shooting-point', along an existing transition path and using this as an initial state, for subsequent simulations, eventually resulting in a new transition path. The crucial step is now to find an acceptance criterion for these trajectories that allows us to still sample the correct distribution. To obtain the equilibrium transition path distribution, it is possible to start from the detailed balance condition,

$$f(x)p(x \to y) = f(y)p(y \to x). \tag{2.16}$$

This condition expresses the idea that in equilibrium all transitions from x to y must be exactly compensated by transitions from y to x. Accordingly, $f(x)$ is the

probability distribution at point x and $p(x \to y)$ is the transition probability from point x to point y, with the right-hand side (RHS) simply representing the reverse transition.

To shift from configurations to trajectories, we reformulate the condition as follows:

$$\begin{aligned} P_{\mathrm{AB}}[X] P_{\mathrm{gen}}\left[X \to X'\right] \times P_{\mathrm{acc}}\left[X \to X'\right] = \\ P_{\mathrm{AB}}[X'] P_{\mathrm{gen}}\left[X' \to X\right] \times P_{\mathrm{acc}}\left[X' \to X\right]. \end{aligned} \tag{2.17}$$

We have replaced $f(x)$ with the probability distribution $P_{\mathrm{AB}}[X]$ of a path X that connects states A and B, which is eventually the distribution we want to sample. The transition probability is now split into two terms, resulting from the fact that to transition from an old path X to a new path X', the latter needs to be both generated with probability $P_{\mathrm{gen}}[X \to X']$ and accepted with probability $P_{\mathrm{acc}}[X \to X']$. The RHS again represents the inverse process.

The equation above can be rearranged to reveal the defining condition on the acceptance probabilities,

$$\frac{P_{\mathrm{acc}}\left[X \to X'\right]}{P_{\mathrm{acc}}\left[X' \to X\right]} = \frac{P_{\mathrm{AB}}[X'] P_{\mathrm{gen}}\left[X' \to X\right]}{P_{\mathrm{AB}}[X] P_{\mathrm{gen}}\left[X \to X'\right]} \tag{2.18}$$

According to this relation, appropriate acceptance probabilities can be chosen. To obtain a valid acceptance probability, it will have to satisfy this condition and also be normalized to lie within the interval $P_{\mathrm{acc}}\left[X \to X'\right] \in [0,\, 1]$.

As in standard Monte-Carlo, a popular and simple choice for this acceptance probability is the *Metropolis-Hastings* criterion,

$$\begin{aligned} P_{\mathrm{acc}}\left[X \to X'\right] &= \min\left\{1, \frac{P_{\mathrm{AB}}[X'] P_{\mathrm{gen}}\left[X' \to X\right]}{P_{\mathrm{AB}}[X] P_{\mathrm{gen}}\left[X \to X'\right]}\right\} \\ &= H_{\mathrm{AB}}[X'(L')] \min\left\{1, \frac{P[X'] P_{\mathrm{gen}}\left[X' \to X\right]}{P[X] P_{\mathrm{gen}}\left[X \to X'\right]}\right\} \end{aligned} \tag{2.19}$$

Going from the first to the second line, we insert the path probabilities P_{AB} from Eq. (2.15), with the path partition functions canceling, and since the old path X' has to be reactive, $H_{\mathrm{AB}}[X(L)] = 1$.

Following this or other criteria obtained from Eq. (2.18) in computation ensures that every reactive path obtained follows the detailed balance condition and the ensemble sampled is therefore an equilibrium transition path ensemble. We have arrived at a rough outline for the procedure of sampling the transition ensemble, which consists of first generating a new path X' from an old path X; if this path X'

is reactive, that is, it goes from A to B, it is accepted according to the probability given by the acceptance criterion (2.19). If it is not reactive or it does not satisfy Eq. (2.19), X' is rejected and, importantly, the old path X is added to the ensemble again, to increase its statistical weight. If X' is reactive and satisfies Eq. (2.19), it is accepted and it is added to the ensemble and will serve as the old path X in the next iteration.

The generation of a new trajectory X' from an existing path X constitutes the core computational operation in Transition Path Sampling. Across different TPS implementations, this process is generally implemented through a *shooting move*, which we now explain both conceptually and mathematically.

This consists of the choice of a configuration along the trajectory X, with a certain selection probability p_{sel}, potentially perturbing this configuration, and then generating a new trajectory from this *shooting point* $x_s = (r_s, v_s)$ through time integration. One way to obtain a new trajectory X' is to start one simulation by integrating forward from $x_s = (r_s, v_s)$ until one of the two states or a maximum path length L_{max} is reached. To obtain a full trajectory, the shooting point velocity is then reversed, and the new initial configuration is given by $\bar{x}_s = (r_s, \bar{v}_s) = (r_s, -v_s)$. A new simulation is started and if this simulation reaches the state opposite to the first shooting move, a reactive trajectory has been generated.

The algorithm detailed here is what is usually referred to as *Two-Way Shooting*.

This method ensures an almost entirely new path at every acceptance, which is generally beneficial for faster exploration of the TPE. However, it has the drawbacks of being, at the same time, very computationally expensive, requiring a high number of force evaluations. Moreover, it often generates paths that are not reactive, since it relies on both paths reaching their respective, opposite states. To address these issues it can be beneficial to replace only a part of the old trajectory.

This is what is generally called One-way Shooting Transition Path Sampling (1wS-TPS). In this method, generating the new path requires only one shooting move, rather than two, through the following procedure. After a shooting point x_s was chosen, we split the old trajectory $X(L) = \{x_0, x_1, ..., x_{L-1}\}$ into two parts separated by the shooting point. Assuming that X starts in A and ends in B we remain with a backward segment $X_{\text{bw}} = \{x_0, x_2, ..., x_{s-1}\}$, with $x_0 \in$ A, and a forward segment $X_{\text{fw}} = \{x_{s+1}, ..., x_{L-1}\}$, with $x_{L-1} \in$ B. We then choose to regenerate either the forward segment X_{fw} or the backward segment X_{bw}, each with a probability most often set to 50%. If we choose to shoot forward, the initial configuration for the shooting move remains the shooting point x_s, if we choose to shoot backward we invert the velocities and shoot from $\bar{x}_s$. If the new trajectories reach their predefined target states (B for forward, A for backward) the old and new segments are stitched together.

This work utilizes 1wS-TPS. Therefore, to define the generation probability P_{gen},

we need the selection probability p_{sel} of the shooting point x_s and a numerical integration of the equilibrium dynamics from the shooting point, either forward or backward, until a state is reached.

We separate the following scenarios:

(a) We draw a 'forward-shot' so the shooting point velocity is not inverted. From the old trajectory X we retain the backward segment $X_{\text{bw}} = \{x_0, ..., x_{s-1}\}$. An integration is started from x_s and will only be considered for acceptance if the resulting path X'_{fw} reaches state B. The new trajectory now is obtained by stitching together the two pieces $X' = X_{\text{bw}} + X'_{\text{fw}} = \{x_0, ..., x_{s-1}, x'_{s'}, ..., x'_{L'-1}\}$. We note that $x_s = x'_{s'}$, $s = s'$, and define the length of the partial trajectory as $\ell = L' - s'$. This scenario is depicted in Fig. 3.3a

(b) We draw a 'backward-shot' so the shooting point velocity is inverted. From the old trajectory X we retain the forward segment $X_{\text{fw}} = \{x_{s+1}, ..., x_{L-1}\}$, with the shooting point being included in the new partial segment. An equilibrium trajectory is started from $\bar{x}_s$ and will only be considered for acceptance if it reaches state A. Since trajectories are defined to span from A to B, the newly generated partial trajectory is relabeled as $X'_{\text{bw}} = \{x'_0, x'_1, \ldots, x'_{s'}\}$. The new trajectory is stitched together as $X'(L') = X'_{\text{bw}} + X_{\text{fw}} = \{x'_0, \ldots, x'_{s'}, x_{s'+1}, \ldots, x_{L'-1}\}$, where $L' = s' + 1 + (L' - s' - 1) = \ell + (L - s - 1)$ since $s' + 1 = \ell$ and $L' - s' = L - s$. We note that in this case, $s' \neq s$ but $x'_{s'} = x_s$. This scenario is depicted in Fig. 3.3b

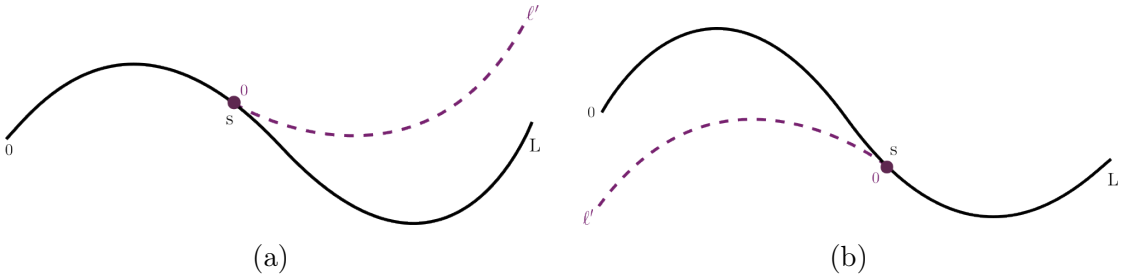


Figure 2.1: Visual description of a 'forward-shot' (a) and a 'backward-shot' (b). We depict an old path X (black) on which a shooting point is chosen at index s from which a new trajectory segment (purple) is generated. The new path X' is assembled from segments of the original path and the new segment. Specifically, in forward shooting (a), X' comprises the original segment from x_0 to x_{s-1} (black) and the new segment (purple). In backward shooting (b), X' consists of the new segment (purple) followed by the original segment from x_s to x_{L-1} (black).

First we address situation (a), where the new path replaces the forward segment of the trajectory, which gives the following generation probability

$$P_{\text{gen}}\left[X \to X'\right] = p_{\text{sel}}(x_s|X) \prod_{j=0}^{\ell-2} p(x'_j \to x'_{j+1}). \tag{2.20}$$

Here $p_{\text{sel}}(x_s|X)$ is the probability of selecting the shooting point x_s from the original trajectory X and the product denotes the iterative transitions from one configuration x'_j to the next x'_{j+1}. Importantly, the shooting point is not being perturbed in 1wS-TPS, since otherwise there would be discontinuities when stitching together the partial trajectories.

We plug this into the ratio of acceptance probabilities given by Eq. (2.18),

$$\begin{aligned}\frac{P[X']P_{\text{gen}}[X' \to X]}{P[X]P_{\text{gen}}[X \to X']} = &\frac{\rho_{\text{eq}}(x'_0) \prod_{i=0}^{s'-1} p(x'_i \to x'_{i+1}) \prod_{i=s'}^{L'-2} p(x'_i \to x'_{i+1})}{\rho_{\text{eq}}(x_0) \prod_{i=0}^{s-1} p(x_i \to x_{i+1}) \prod_{i=s}^{L-2} p(x_i \to x_{i+1})} \\ &\times \frac{p_{\text{sel}}(x'_{s'}|X')\, \rho_{\text{eq}}(x_s) \prod_{j=0}^{\ell-2} p(x_j \to x_{j+1})}{p_{\text{sel}}(x_s|X)\, \rho_{\text{eq}}(x'_{s'}) \prod_{j=0}^{\ell'-2} p(x'_j \to x'_{j+1})}.\end{aligned} \tag{2.21}$$

To clarify the structure, note that the path probability has been divided at the shooting point index s into two segments, the one before the shooting point $\prod_{i=0}^{s-1} p(x_i \to x_{i+1})$ (from initial state x_0 to shooting point x_s) and the one after the shooting point $\prod_{i=s}^{L-2} p(x_i \to x_{i+1})$ (from shooting point x_s to final state x_{L-1}) (and similarly for the new trajectory X').

Canceling terms and substituting back into Eq. (2.19) yields the final result for the Metropolis-Hastings acceptance criterion in case (a) as

$$P_{\text{acc}}[X \to X'] = H_{\text{AB}}[X'(L')] \min\left\{1, \frac{p_{\text{sel}}(x'_{s'}|X')}{p_{\text{sel}}(x_s|X)}\right\}. \tag{2.22}$$

In case (b), where the new path replaces the backward segment of the trajectory, the terms do not cancel trivially. As stated earlier, constructing X' requires reindexing to obtain a consistent enumeration of configurations for a path from A to B. Similarly, it is now necessary to reverse the new path segment mathematically.

This path reversal operation connects fundamentally to the concept of microscopic reversibility. Hamiltonian mechanics reveals its inherent reversibility through the

existence of paired solutions to its equations of motion. For any trajectory γ_{F} defined for $t \in [0, T]$ with initial conditions $\Gamma_{\mathrm{F}}(0) = (r(0), p(0))$, where $\Gamma_{\mathrm{F}}(t) = (\mathbf{r}(t), \mathbf{p}(t))$ represents the phase space position at time t, there exists a time-reversed solution γ_{R}. This reversed trajectory is defined on the interval $t \in [T, 0]$ with initial conditions $\Gamma_{\mathrm{R}}(T) = (\mathbf{r}(T), -\mathbf{p}(T))$, evolves as $\Gamma_{\mathrm{R}}(t) = (\mathbf{r}(t), -\mathbf{p}(t))$, and constitutes a valid solution to the equations of motion.

This implies that if we start from a time-reversed final phase space point $\Gamma_{\mathrm{R}}(\tau)$ and integrate backwards in time, we will trace out the same trajectory, only in reverse order and with inverted momenta. Newtonian dynamics are therefore fully invariant under time-reversal when forces are independent of velocities.

While this deterministic time-reversal symmetry provides intuition for microscopic reversibility, it is not a necessary condition. Crucially, microscopic reversibility extends to stochastic dynamics, where we define it probabilistically:

$$P(\gamma_{\mathrm{F}}) = P(\gamma_{\mathrm{R}}). \tag{2.23}$$

This states, that the probability of observing a forward trajectory γ_{F} must be equal to the probability of observing the time-reversed counterpart γ_{R}. Reformulating this in terms of state transitions yields the dynamic property of detailed balance:

$$\rho_{\mathrm{eq}}(x)p(x \to y) = \rho_{\mathrm{eq}}(\bar{y})p(\bar{y} \to \bar{x}). \tag{2.24}$$

Here x and y are points in phase-space, whereas $\bar{x}$ and $\bar{y}$ are their time-reversed counterpart. This reversal is achieved by applying a time-reversal operator, the standard form of this is

$$\mathcal{T} : x = (r, p) \mapsto \bar{x} = \mathcal{T}x = (r, -p). \tag{2.25}$$

Using relation Eq. (2.24), we reverse the newly generated segments in Eq. (2.21) to obtain

$$\begin{aligned} \rho_{\mathrm{eq}}(x_s) \prod_{j=0}^{\ell-2} p(x_j \to x_{j+1}) &= \rho_{\mathrm{eq}}(\bar{x}_{\ell-1}) \prod_{j=1}^{\ell-1} p(\bar{x}_j \to \bar{x}_{j-1}), \\ \rho_{\mathrm{eq}}(x'_{s'}) \prod_{j=0}^{\ell'-2} p(x'_j \to x'_{j+1}) &= \rho_{\mathrm{eq}}(\bar{x}'_{\ell'-1}) \prod_{j=1}^{\ell'-1} p(\bar{x}'_j \to \bar{x}'_{j-1}). \end{aligned} \tag{2.26}$$

Finally, after shifting the indices and reversing once again on the RHS, we arrive at

$$\begin{aligned}\rho_{\mathrm{eq}}(x_s)\prod_{j=0}^{\ell-2}p(x_j\to x_{j+1}) &= \rho_{\mathrm{eq}}(x_0)\prod_{j=0}^{s-1}p(x_j\to x_{j+1}),\\ \rho_{\mathrm{eq}}(x'_{s'})\prod_{j=0}^{\ell'-2}p(x'_j\to x'_{j+1}) &= \rho_{\mathrm{eq}}(x'_0)\prod_{j=0}^{s'-1}p(x'_j\to x'_{j+1}).\end{aligned} \tag{2.27}$$

Substituting these terms in Eq. (2.21) and after trivial cancellations, we obtain the final result for the acceptance criterion in case (b) as

$$P_{\mathrm{acc}}[X\to X'] = H_{\mathrm{AB}}[X'(L')]\min\left\{1,\frac{p_{\mathrm{sel}}(x'_{s'}|X')}{p_{\mathrm{sel}}(x_s|X)}\right\}. \tag{2.28}$$

We find that in both scenarios, the acceptance criterion collapses to the same result, which depends only on the selection probability of both paths.

2.1.3 Always Accepting Transition Path Sampling

It is generally regarded as a key step in *One-Way Shooting* TPS to keep track of velocities when going from one transition path to the next, in order to generate physically continuous trajectories. What this enforced velocity continuity has always implied is that the state that needs to be reached is being decided before the simulation of the new trajectory in the shooting move is even started. Every time this simulation reaches a state other than this predefined target state, the path is not reactive and will be thrown away. This means that a great amount of computational resources and time will go to waste.

In many real scenarios, the time needed for the momenta to decorrelate is far shorter than the timescales of the transition under investigation. This separation validates integrating out momenta, forming the basis of overdamped dynamics. For this reason, when dealing with an overdamped system, there is no need to maintain consistency of velocities between trajectory frames. This subtle detail indeed carries great implications. It is then possible to omit the restriction on the final state of the generated trajectory: After any of the states is reached, the new trajectory is obtained by stitching the new segment with the corresponding part on the old one that yields a path that connects A to B. This would mean that every single shooting move will result in a reactive trajectory.

This new algorithm, capable of generating reactive trajectories at every trial, is laid out in the following:

1. On a path that stretches from A to B, we pick a shooting point x_s according to some selection probability $p_{\mathrm{sel}}(x_s|X)$.

2. This shooting point acts as a separator of the old path X, splitting it into $X_{\mathrm{bw}} = \{x_0, \ldots, x_{s-1}\}$ and $X_{\mathrm{fw}} = \{x_{s+1}, \ldots, x_{L-1}\}$. The shooting point is excluded in both cases.

3. Starting from the shooting point x_s an equilibrium simulation is started, which is stopped once it hits one of the two stable states (or the maximum path length $L_{\max}$). This new trajectory is denoted as $X'_{\mathrm{fw}} = \{x'_0, \ldots, x'_{\ell-1}\}$.

   a) If the new trajectory hits state B, the new path is obtained as $X'(L') = X_{\mathrm{bw}} + X'_{\mathrm{fw}} = \{x_0, \ldots, x_{s-1}, x'_{s'}, \ldots, x'_{L'-1}\}$ with $x'_{s'} = x_s$, $s' = s$ and $L' = s + \ell$. (This scenario maps onto a 'forward-shot' in standard *One-Way Shooting* TPS and can therefore be compared to Fig. 3.3a)

   b) If the new trajectory hits state A, the order of points in X'_{fw} is reversed and relabeled to obtain $X'_{\mathrm{bw}} = \{x'_0, x'_1, \ldots, x'_{s'}\}$. Then we also relabel the points on X_{fw} as $X_{\mathrm{fw}} = \{x_{s'+1}, \ldots, x_{L'-1}\}$. The new path is now given as $X'(L') = X'_{\mathrm{bw}} + X_{\mathrm{fw}} = \{x'_0, \ldots, x'_{s'}, x_{s'+1}, \ldots, x_{L'-1}\}$ where $L' = s' + 1 + (L' - s' - 1) = \ell + (L - s - 1)$ since $s' + 1 = \ell$ and $L' - s' = L - s$. Note that in this case, $s' \neq s$ but $x'_{s'} = x_s$. (This scenario maps onto a 'backward-shot' in standard *One-Way Shooting* TPS and can therefore be compared to Fig. 3.3b)

Finding the acceptance criterion for this method generally follows the derivation in section 2.1.2. Importantly, for this method all trajectories are reactive, so $H_{\mathrm{AB}}[X'(L')] = 1$ for all $X'(L')$ that the algorithm generates. The acceptance criterion therefore takes the familiar form

$$P_{\mathrm{acc}}[X \to X'] = \min\left\{1, \frac{p_{\mathrm{sel}}(x'_{s'}|X')}{p_{\mathrm{sel}}(x_s|X)}\right\}. \tag{2.29}$$

The form of this acceptance criterion implies that whenever $p_{\mathrm{sel}}(x'_{s'}|X')$ is lower than $p_{\mathrm{sel}}(x_s|X)$, the path has a chance of being rejected. This acceptance/rejection scheme is an essential part of any Monte Carlo technique, with both the acceptance and rejection serving a fundamental purpose in assigning each sample its correct relative statistical weight. It is however not the only way of doing so. With the aim of accepting every proposed reactive trajectory while finding an appropriate statistical weight $\Omega[X(L)]$, we write down an auxiliary path distribution as

$$\widetilde{P}_{\mathrm{AB}}[X(L)] \propto H_{\mathrm{AB}}[X(L)]\widetilde{P}[X(L)] = H_{\mathrm{AB}}[X(L)]\Omega[X(L)]P[X(L)]. \tag{2.30}$$

We can derive an explicit expression for this statistical weight once again from detailed balance. To this end, we derive the acceptance criterion from detailed balance for the correct sampling of the probability $\widetilde{P}_{\mathrm{AB}}[X(L)]$ now in the form of

$$\widetilde{P}_{\mathrm{AB}}[X(L)]P_{\mathrm{gen}}\left[X(L)\to X'(L')\right]\times P_{\mathrm{acc}}\left[X(L)\to X'(L')\right]= \widetilde{P}_{\mathrm{AB}}[X'(L')]P_{\mathrm{gen}}\left[X'(L')\to X(L)\right]\times P_{\mathrm{acc}}\left[X'(L')\to X(L)\right]. \tag{2.31}$$

Since the goal is to accept every possible proposed path, we require that $P_{\mathrm{acc}}[X\to X']=1$ for all pairs of paths X, X'. Plugging this into Eq. (2.31), we arrive at

$$\frac{P_{\mathrm{acc}}\left[X(L)\to X'(L')\right]}{P_{\mathrm{acc}}\left[X'(L')\to X(L)\right]}=\frac{\Omega[X']P[X']P_{\mathrm{gen}}[X'\to X]}{\Omega[X]P[X]P_{\mathrm{gen}}[X\to X']}=1, \tag{2.32}$$

This can be rearranged to find the appropriate weights

$$\frac{\Omega[X']}{\Omega[X]}=\frac{P[X]P_{\mathrm{gen}}[X\to X']}{P[X']P_{\mathrm{gen}}[X'\to X]}. \tag{2.33}$$

The ratio on the RHS has been derived for this algorithm in Eq. (2.29). Therefore, substituting in Eq. (2.33) yields

$$\frac{\Omega[X']}{\Omega[X]}=\frac{p_{\mathrm{sel}}(x_s|X)}{p_{\mathrm{sel}}(x'_{s'}|X')}. \tag{2.34}$$

Inserting a very general form for the selection probability as

$$p_{\mathrm{sel}}(x_s|X)=\frac{\omega(x_s)}{\sum_{x_i\in X}\omega(x_i)}, \tag{2.35}$$

where $\omega(x)>0$ for all x, gives the following condition for the weights:

$$\frac{\Omega[X']}{\Omega[X]}=\frac{\sum_{x'_i\in X'}\omega(x'_i)}{\sum_{x_i\in X}\omega(x_i)}. \tag{2.36}$$

A natural choice for the statistical weight $\Omega[X]$ of a path X is therefore

$$\Omega[X]=\sum_{x_i\in X}\omega(x_i). \tag{2.37}$$

where ω is a weight function for the configurations on a path that characterizes the selection probability and is a parameter of the particular algorithm.

We have in this way found a statistical weight that only depends on the current path, through which the correct transition path ensemble $P_{\mathrm{AB}}[X(L)]$ can be recovered via

$$P_{\mathrm{AB}}[X(L)]=\widetilde{P}_{\mathrm{AB}}\cdot\Omega[X]^{-1} \tag{2.38}$$

Finding this statistical weight and thereby recovering the correct transition path ensemble makes it possible to go through the sampling mechanism without ever rejecting paths due to the acceptance probability. This constitutes a substantial boost in the number of paths adding to the TPE, utilizing all generated path without letting anything go to waste.

2.1.4 Choosing a shooting point selection probability

In both described algorithms, the acceptance criterion has been derived as a function of the shooting point selection probability. To perform computation using this acceptance criterion, it is therefore necessary to concretely define this.

A standard and simple choice for this selection probability is choosing it uniformly along the whole transition path, which means

$$p_{\text{sel}}(x_i|X(L)) = \frac{1}{L}. \tag{2.39}$$

Plugging this into the Metropolis-Hastings criterion, as given in Eq. (2.29) for the Always Accepting Algorithm (or for a very similar result into Eq. (2.28) for One-Way Shooting), we arrive at the following

$$P_{\text{acc}}[X \to X'] = \min\left\{1, \frac{L}{L'}\right\}. \tag{2.40}$$

Due to the uniform choice of the shooting point, this method often starts new shooting moves from configurations far away from the barrier top. This results in the undesirable scenario where new trajectories must re-ascend the energy barrier, either partially or fully, yielding a high number of non-reactive paths. One method proposed by Jung et al. [22] leans on the fact that in many systems of interest there already exists at least some partial information about the system and its transition landscape. The core idea is to steer the shooting point selection in a way that maximizes the production of reactive paths. This is achieved by restricting the selection of shooting points to a defined region, either using a 'hard cutoff', where points are chosen uniformly within a range, or by sampling from a probability distribution centered around the desired area. When chosen well, this region will focus shooting points on top of the kinetically relevant barrier.

When imposing a 'hard cutoff', we define the shooting range S as $S = \{x|q_0 < q(x) < q_1\}$, where $q(x)$ denotes a one-dimensional order parameter. This shooting range will lie between the stable states, with $q_0 \geq q_A$ and $q_0 \leq q_B$ and all transition paths need to pass through S at least once. In this case, the acceptance probability takes a similar form as Eq. (2.40), but instead of the full length of the trajectory, now only the points inside the shooting range $X_S = \{x_k|x_k \in X; q_0 < q(x_k) < q_1\}$ are being considered. Therefore, the acceptance probability now depends on the

number of frames inside the shooting range on the old path $n = |X_S|$ and on the new path $n' = |X'_S|$.

$$P_{\text{acc}}\left[X \to X'\right] = \min\left[1, \frac{n}{n'}\right]. \tag{2.41}$$

Shooting only from this predefined region can however lead to issues in more complex transition landscapes consisting of two or more possible transition channels. If shooting points are confined to the saddle region, the sampling may become trapped in individual transition channels, as these pathways are often separated by high barriers at this stage. It can therefore be of great value to shoot from outside of the predefined shooting range S with a certain probability. This idea is generalized by the concept of a weighted shooting range [22]. In this approach, we introduce a certain weight function $w(x)$ with $w(x) \geq 0$, which assigns every point x_i on the trajectory X a certain probability of being chosen as the shooting point

$$p_{\text{sel}}(x_i|X) = \frac{w(x_i)}{\sum_{x_k \in X} w(x_k)}. \tag{2.42}$$

With this selection probability we can write down the Metropolis acceptance criterion in the case of a weighted shooting range by plugging into Eq. (2.29) as

$$P_{\text{acc}}\left[X \to X'\right] = \min\left[1, \frac{\sum_{x_k \in X} w(x_k)}{\sum_{x'_k \in X'} w(x'_k)}\right]. \tag{2.43}$$

One suitable choice for such a weight function, that offers great flexibility, is the generalized normal distribution

$$\mathcal{GN}(x, \mu, \alpha, \beta) \propto e^{-(|x-\mu|/\alpha)^\beta}, \tag{2.44}$$

where $\mu \in \mathbb{R}$ denotes the mean, $\alpha \in \mathbb{R}_{>0}$ the scale, and $\beta \in \mathbb{R}_{>0}$ the shape of the function. Plugging this weight function into Eq. (2.43) we arrive at the following term for the acceptance criterion:

$$P_{\text{acc}}\left[X \to X'\right] = \min\left[1, \frac{\sum_{x_k \in X} e^{-(|x_k-\mu|/\alpha)^\beta}}{\sum_{x'_k \in X'} e^{-(|x'_k-\mu|/\alpha)^\beta}}\right]. \tag{2.45}$$

Sampling according to this procedure enables the efficient generation of a greater number of reactive paths while maintaining a manageable level of path correlation, providing a flexible, straightforward, and effective TPS framework.

Selecting an appropriate shooting point distribution requires careful consideration. While a shooting from the top approach can greatly improve efficiency, it also poses the risk of locking the system into one reaction channel, by confining the shooting points into a region where transitions are hardly possible. Conversely, relaxed

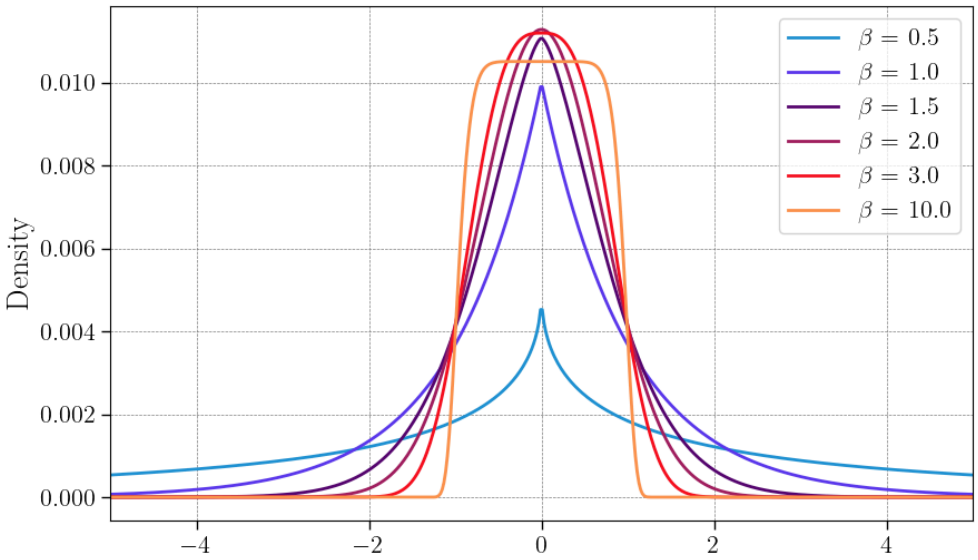


Figure 2.2: The generalized normal distribution $\mathcal{GN}(x, \mu, \alpha, \beta)$ for $\mu = 0$, $\alpha = 1$ and a range of β values.

shooting ranges, e.g. uniform shooting, may lead to less total channel switches by rejecting a higher percentage of paths. The choice of how confined or relaxed this selection probability should be is therefore non-trivial and dependent on the energy landscape being explored.

2.2 Implementation

The transition path distribution ρ_{TP} (sometimes also noted as $\rho(x|\mathrm{TP})$), defined as the probability density of reactive trajectories connecting states A and B under the system's native dynamics (or in other words, the probability density of configurations that are on reactive paths), serves as the fundamental target for TPS methods. To verify that the AA-TPS can correctly sample this distribution, we compare its resulting distribution $\rho_{\mathrm{TP}}(AA)$ to that of the benchmark method 1wS-TPS $\rho_{\mathrm{TP}}(1wS)$. Both methods are implemented using shooting from the top. These methods are both implemented in python and are tested on a single particle system in a 2D potential before their application to the clathrate hydrate system. The potential used in this 2D model is given by

$$U(\mathbf{x}) = \left\{ \frac{1}{4} \left[\left((x^{(0)})^2 + (x^{(1)})^2 - 4 \right)^2 + (x^{(1)})^2 \right] \right\} \tag{2.46}$$

Both underlying Molecular Dynamics (MD) simulations are run in OpenMM [15] using the Langevin middle integrator with a friction of $1\,\mathrm{ps}^{-1}$, a timestep of 2 fs, a mass of 1 u and a temperature of 300 K. The order parameter is chosen to be the x coordinate, with the stable states being defined as A: $x < -2$ and $B : x > 2$.

The transition path distribution resulting from 1wS-TPS, $\rho_{\mathrm{TP}}(1wS)$, is first compared to the equilibrium simulation results $\rho_{\mathrm{TP}}(EQ)$, as seen in Fig. 2.3. For 1wS-TPS a total number of 100,000 trials were done, with 38,932 successful trajectories, while for equilibrium sampling one trajectory of 100,000,000 frames was run, resulting in 12,140 transition paths.

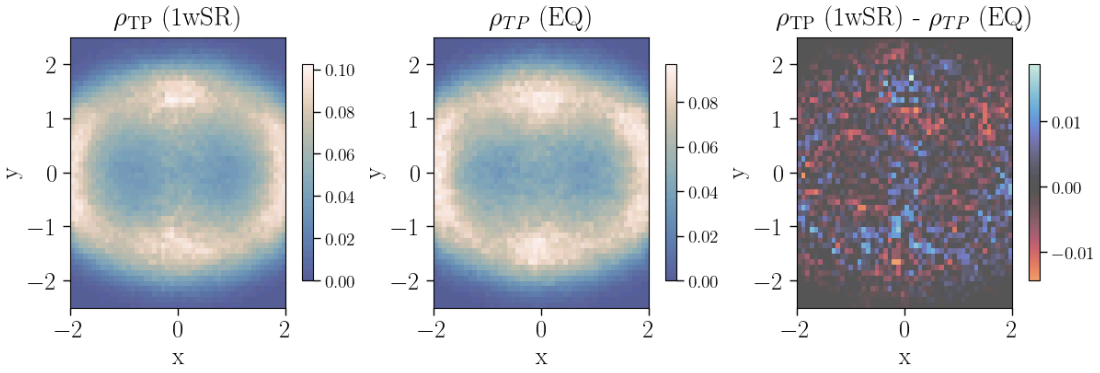


Figure 2.3: Comparison of the TPE $\rho_{\mathrm{TP}}(1wS)$ from 1wS-TPS (left) versus the equilibrium reference TPE $\rho_{\mathrm{TP}}(EQ)$ (middle), with their difference $\rho_{\mathrm{TP}}(1wS) - \rho_{\mathrm{TP}}(EQ)$ shown in the right panel.

The close agreement between the 1wS-TPS TPE and equilibrium simulation results is demonstrated by the small, uniformly distributed errors. We use this confirmation, to assess the performance of AA-TPS by comparing the resulting TPE $\rho_{\mathrm{TP}}(AA)$ against a TPE generated using One-Way Shooting $\rho_{\mathrm{TP}}(1wS)$, after switching to overdamped dynamics, using the Brownian integrator. This simulation is done using a friction of $1\,\mathrm{ps}^{-1}$, a timestep of 2 fs, a mass of 1 u and $\mathrm{k_B T} = 1\,\mathrm{kJ\,mol}^{-1}$.

To obtain the correct TPE, AA-TPS requires a reweighting of the generated trajectories, as outlined at the end of Sec. 2.1.3. In this case, this is done by by resampling the trajectories randomly with a probability equal to their normalized weight, given by $1/\sum_{x_i \in X} w(x_i)$ (see Eq. 2.38).

The results are shown in Fig. 2.4.

Once again, we see strong agreement between the two TPEs, with only small, evenly distributed errors. This shows that AA-TPS is able to sample the TPE in agreement with the standard 1wS-TPS TPE.

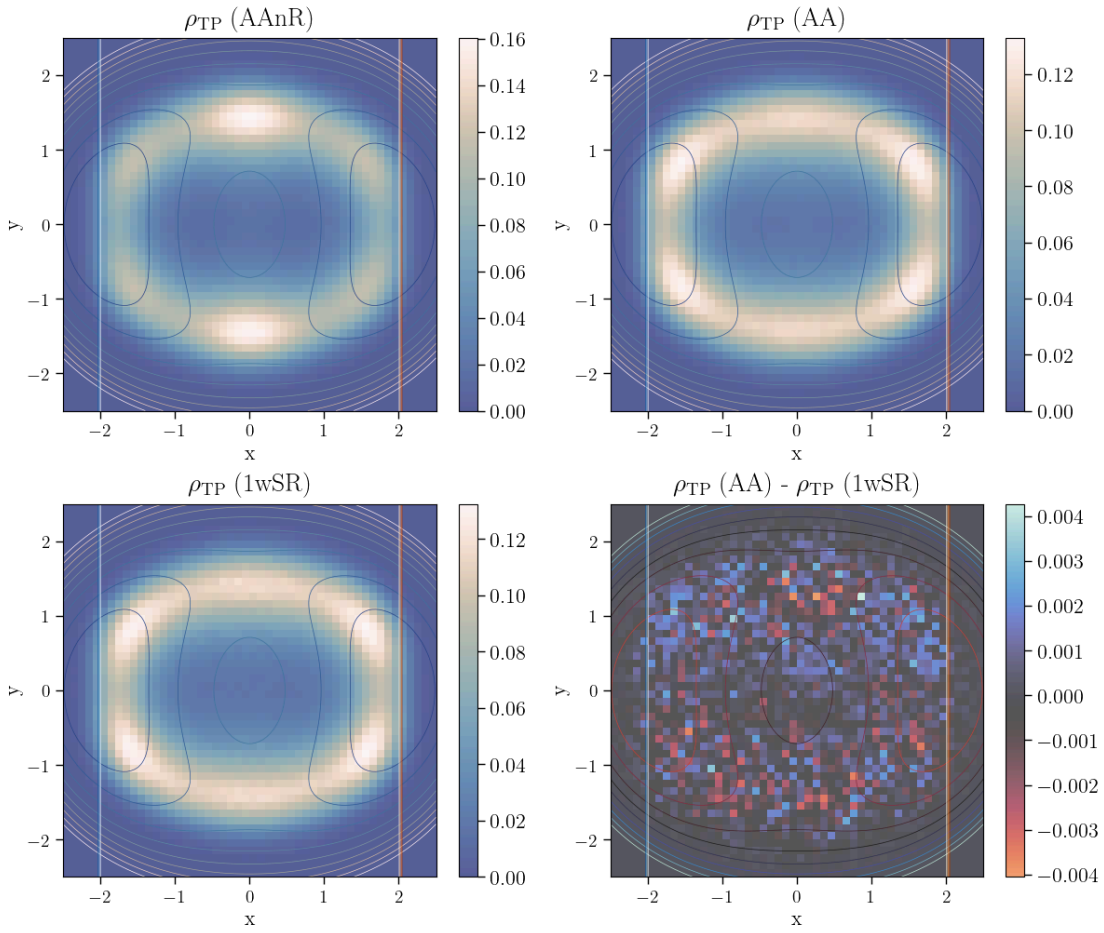


Figure 2.4: Comparison of the TPE $\rho_{\mathrm{TP}}(AA)$ obtained using AA-TPS without reweighting (upper left) and with resampling (upper right), versus the 1wS-TPS TPE $\rho_{\mathrm{TP}}(1wS)$ (lower left). The lower right panel shows their difference $\rho_{\mathrm{TP}}(AA) - \rho_{\mathrm{TP}}(1wS)$.

3 Clathrate Hydrates

3.1 Introduction to Clathrate Hydrates

In 1810, Sir Humphry Davy made a pivotal discovery while studying chlorine hydrate, directing chlorine gas into liquid water under high pressures, a crystalline solid formed when chlorine gas interacted with water at low temperatures. He demonstrated that this material consisted of cage-like water structures encapsulating chlorine molecules, overturning the previous assumption that it was solid chlorine [14]. This is generally regarded as the first official description of a phenomenon now referred to as 'Clathrate Hydrates'. This name refers to a class of solid inclusion compounds comprised of water cages (hosts) trapping small molecules ($<$ 9 Angstrom) (guests), such as carbon dioxide, methane or propane. These cages take the shape of various polyhedral cavities, connected through their faces or vertices to form a larger structure. These compounds can be found both naturally occurring, typically on coastal seafloors and in permafrost regions, and artificially made, through the recreation of these natural conditions by subjecting a water-gas solution to specific factors such as relatively low temperatures and high pressures [33].

These cages are stabilized by the guest molecules through van der Waals interactions. Accordingly, while empty cages do exist, the overall stability of the hydrate depends on the percentage of cavities which are filled by guests [25, 35]. The cages are generally described using the notation C^n, where C represents the number of sides on a face, and n indicates how many times such a face appears in the cage. For example, the commonly observed cage denoted as $5^{12}6^2$ consists of 12 pentagonal faces and 2 hexagonal faces.

In the crystalline phase, there are 3 different structures that have been most commonly identified, denoted sI, sII and sH hydrates. These structures are identified through their specific symmetries, where both sI and sII display a cubic crystal structure each made up by two different types of cages, while sH has a hexagonal crystal structure with 3 different cage types identified. The kind of structure that is most abundantly found depends on the guest molecule, with CO_2 clathrates typically crystallizing in the sI structure [33]. More than a century after their discovery, in 1934, hydrate research first gained more widespread industrial interest, since it was reported by Hammerschmidt that these hydrates are the reason for

blockages of gas pipelines [19].

This topic of flow assurance, which first brought attention to clathrates, is still today a main focus of research. As subsea gas and oil production moves into deeper waters, the associated pressure and temperature conditions are becoming increasingly extreme. This is an environment which is highly conducive to the nucleation and growth of clathrate blockages, that pose a significant threat to the natural environment [29].

Gas hydrates occur naturally in widespread deposits around the world, typically found in marine sediments on continental margins and in and beneath Arctic permafrost regions. These deposits receive attention through multiple avenues, raising climate concerns due to the risk of destabilization from rising ocean and air temperatures, which could release trapped gases, while also presenting a potential new source of energy. Being one of the most abundant resources of methane on our planet, the exploitation of clathrates for a relatively clean burning fuel source, which emits less carbon dioxide than regular fuels, seems like an attractive choice [35].

However, the disturbance of these deposits causes a multitude of risks which raise ethical and climate concerns. For one, the decomposition of these clathrates could lead to the release of methane into the atmosphere, which is powerful greenhouse gas. Additionally, clathrates act as a stabilizer in soft sediments and it has been suggested, that disturbing them by warming or drilling would weaken the sea floors [8, 29].

More recently there have been advancements into the study of Hydrate-based Carbon Capture and Storage (HCCS), which presents a range of advantages over conventional carbon capture techniques. HCCS is generally regarded as highly efficient because of the higher storage density compared to standard storage methods. It has multiple areas of implementation, being able perform direct air capture, be applied on power plant flue gases and also industrial emissions. Moreover, it provides a safer option for storage because of the increased stability of hydrates compared to gaseous or supercritical states [2].

Over the years, considering the new areas of interest, the relevant research has moved away from thermodynamic study and kinetic insight has become increasingly relevant. This shift in focus has highlighted the importance of understanding the fundamental mechanisms behind hydrate formation, particularly the nucleation process, which remains one of the most challenging and least understood aspects of clathrate science.

3.2 Clathrate Hydrate Nucleation

The foundational framework for understanding nucleation is Classical Nucleation Theory (CNT), with its core concepts established in the 19th century by Gibbs [18] and formalized in the early 20th century. It relies on the postulate that a range of first-order phase transitions occur through the formation of a spherical nucleus of the new phase, which forms through thermal fluctuations.

The core concept describes nucleation as a competition between two opposing free energy contributions, one positive and one negative, whose interplay creates an energy barrier that the system must overcome to reach the new, stable phase.

The forming nucleus creates an energetic cost of interface between the two phases, resulting in a thermodynamic barrier, proportional to R^2, with R being the radius of the growing cluster, that opposes the emergence of the new phase. However, if thermal fluctuations allow the nucleus to reach a critical size, the energetic benefit associated with the volume of the more stable phase $\propto R^3$ provides a compensating driving force. Once this balance tips, the new phase becomes thermodynamically favored and grows to macroscopic dimensions. An expression for the free energy $\Delta F(R)$ as a function of the radius of the nucleus can be estimated as

$$\Delta F(R) = -\frac{4\pi R^3}{3}\rho|\Delta\mu| + 4\pi R^2\gamma. \tag{3.1}$$

where ρ is the number density of the nucleating phase, $\Delta\mu$ is the difference in chemical potential and γ is the surface tension. From this free energy landscape, we can also find the critical nucleus size as

$$R^* = \frac{2\gamma}{\rho|\Delta\mu|}, \tag{3.2}$$

which leads to the following correlation between the critical barrier height and the critical nucleus size:

$$\Delta F(R^*) = \frac{4}{3}\pi R^{*2}\gamma. \tag{3.3}$$

Despite its frequent failure to accurately predict nucleation rates [23], CNT remains a widely used framework for understanding nucleation processes. It explains the process of nucleation as a Kramers problem, where the system is trapped in a metastable state and must overcome a free energy barrier to transition to a more stable state, providing fundamental intuition for why nucleation acts as a rare event.

We know, however, that the nucleation of clathrates does not occur through a one-dimensional process, but rather through a complex, multi-dimensional process that involves multiple transition channels. Over the years a number of theories,

largely phenomenological, have been proposed to explain the nucleation of clathrate hydrates [13]. In the following, will briefly discuss the most relevant ones here.

The fundamental theory of clathrate hydrate nucleation is the Labile Cluster Theory, which was first proposed by Sloan and Koh [34]. This theory describes the onset of nucleation as labile clusters diffusing through water as individual entities. The labile clusters are believed to be short-lived and can easily dissociate, but if they agglomerate and reach a critical size, they become stable and grow into larger clusters, eventually leading to the formation of a hydrate. This theory already includes some understanding of the kinetics of nucleation, as it suggests that the formation of labile clusters is a stochastic process that is influenced by the specific guest molecule.

More recently there has been a shift towards a more detailed understanding of the nucleation process, driven by molecular simulation, with the introduction of the Local Structuring Theory [28]. This theory suggests that the nucleation of clathrate hydrates is driven by the local structuring of guest molecules into a configuration similar to that of clathrate hydrates. This suggests, that guest molecule concentration and structure are a first factor, with water structure following second, which is in direct contention with the Labile Cluster Theory.

To reconcile this discrepancy, a new theory was proposed, aiming to combine the two previous theories, which is referred to as the Blob nucleation hypothesis [21]. This explains the nucleation as a multi-step process, which starts through the concentration of guest molecules in the water phase, which then leads to the formation of amorphous clusters, or 'blobs'. These blobs present an intermediate, step between decomposition and formation, until they transform into a metastable clathrate nucleus. Finally, this amorphous nucleus turns into a stable crystalline clathrate.

These theories present an evolution in clathrate nucleation understanding, facilitated by the development of molecular simulations methods, making it possible to uncover multistep pathways beyond experimental capabilities. Atomistic MD simulations provide essential resolution of nucleation processes at molecular length- and time-scales. However, this resolution necessitates inherent compromises in accessible system sizes and simulation durations. While it is possible to perform straightforward MD simulation of clathrate hydrates [6, 16, 36], it is advantageous to reconcile the issue of exponential waiting times associated with rare events, such as nucleation, making use of enhanced sampling methods [24]. Various rare event sampling techniques have previously been used to investigate the nucleation of ice and even clathrate hydrates, including TPS [4, 5, 9, 24]. Enhanced sampling methods generally depend on the choice of a reaction coordinate or an order parameter, whereas TPS's relative insensitivity to this choice presents a significant advantage over other methods [24, 37]. TPS provides natural dynamics which is focused

on the transition barrier, eliminating the issue of long waiting times inflating conventional simulations. Despite this, these transition paths can be extremely long and computationally costly to produce. In standard TPS methods, it is a fundamental component of the sampling procedure to reject roughly 60 - 80% of all produced paths, constituting a significant waste of computational resources. With less than half of all produced paths contributing to the TPE, gathering statistically significant samples becomes prohibitively difficult, limiting the ability to make conclusive mechanistic inferences.

3.3 Sampling via Always-Accepting Transition Path Sampling

The aim of this work is to sample the TPE of CO_2 clathrate hydrates using the AA-TPS algorithm. This method represents a solution to the critical inefficiency of low path acceptance rates, thereby maximizing productive trajectory yield while minimizing wasted computational effort. This algorithm requires the system to be described by overdamped dynamics. It is therefore necessary to justify that for the system at hand there exists a relevant timescale at which overdamped dynamics will fully recover the dynamics of the full (non-overdamped) Langevin system and that this timescale is the one at which the formation of clathrates occurs.

To do this, we assume that there exists a timescale τ_v at which the velocities have fully decorrelated. Given an initial time τ_0, by $\tau_0 + \tau_v$ all memory of the initial velocity state is erased, and any observable will behave as if velocities are drawn freshly from the Boltzmann distribution.

In dense, relatively hot systems, frequent particle collisions cause rapid velocity decorrelation, leading to a short τ_v. On the other hand, with particles being trapped in their dense environment, positional diffusion is severely restricted, leading to much slower positional decorrelation with long τ_r. For observables whose variations are significant only on timescales larger than τ_v, it is therefore justified to collapse the full phase space $\Gamma = \{x_i, v_i\}$ into a restricted, position only phase space $\Gamma* = \{x_i\}$.

To test this hypothesis, we calculate the normalized autocorrelation function $\langle x_j(0)x_j(\tau)\rangle$ of the velocity v and the position r as:

$$\langle x_j(0)x_j(\tau)\rangle \approx \frac{1}{T_{\max}} \int_0^{T_{\max}} \frac{(x_j(t) - \bar{x}_j)(x_j(t+\tau) - \bar{x}_j)}{\sigma_{xj}^2} dt, \tag{3.4}$$

where $T_{\max}$ is the observation time, $\bar{x}_j$ the equilibrium average of the variable x for a particle j and σ_{xj}^2 is the variance. This quantity is then averaged over all particles j to obtain the ensemble-averaged result.

For a clathrate hydrate system the velocity and position autocorrelation functions are given also for different stages along the crystallization transition. On the timescale of picoseconds (small enough to resolve in fine detail the clathrate formation process), the velocity autocorrelation function almost immediately decays to zero, indicating that the velocities are fully decorrelated. The position autocorrelation function, however, shows a much slower decay. This means that, at these timescales, the system can be described as a Brownian process. When producing trajectories, is it then ideal to save configurations with an interval (e.g. 200 ps) at which velocities have long decorrelated, while the positions are still highly correlated.

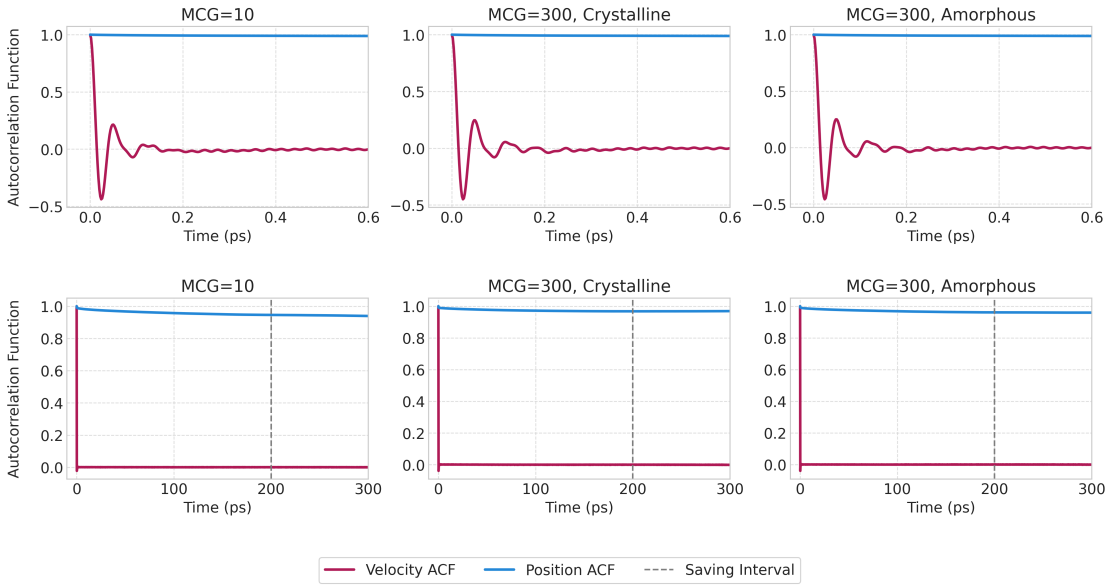


Figure 3.1: Autocorrelation functions (ACF) of the velocity and position of a clathrate hydrate system at different stages along the crystallization transition. The top panel shows the relaxation on a short timescale of up to 0.6 ps and the bottom panel shows the decorrelation on a longer timescale of up to 300 ps. The gray, dashed vertical line suggests a saving interval of 200 ps at which the velocities are fully decorrelated while the positions are still highly correlated.

3.4 Order Parameter

3.4.1 Mutually Coordinated Guest Parameter

A great strength of TPS is that there is no need for a reaction coordinate, giving specific information about how far along the reaction pathway the system is at a

given point. Instead, choosing which transition we want to observe relies on the definition of an initial state A and a final state B, two (meta)stable states of the system. A convenient choice is finding a suitable lower dimensional order parameter and requiring it to lie within certain limits. This choice of a state definition is a crucial step in making sure the TPS works as expected and that the TPE can deliver the desired information. Achieving this relies mostly on two criteria for the stable states. Firstly, the two states have to be entirely non-overlapping to guarantee that every generated path is reactive. Secondly, not to miss out on crucial reactive pathways, it is important to make the states large enough to account for typical equilibrium fluctuations [17].

The two relevant stable states which need to be distinguished when characterizing crystallization are the liquid and the solid state. Unlike first-order phase transitions, as the gas-liquid transition, where one-dimensional parameter is enough to describe the transition, the symmetry of the system is broken when going from the liquid to the solid state [30]. An added difficulty comes from the fact that when dealing with the amorphous state of Clathrate Hydrates, there is no global symmetry that can be quantified. Therefore we need to rely on single particle resolution, i.e. local order parameters. One parameter that has found previous success when studying clathrates is the Mutually Coordinated Guest (MCG) order parameter.

In a 2009 study, Walsh et al. conducted detailed microsecond-long simulations on the formation of methane hydrates [36]. In this work, they observed that a first precursor of nucleation is a pair of methane molecules being absorbed on both sides of a pentagonal ring of water molecules. This information has inspired Barnes et al. to define an order parameter based on the largest cluster built up of guest molecules that resemble this constellation [7].

A cluster is defined here as a collection of MCG monomers, which are guest molecules that meet the following criteria:

1. All guest molecules which are within cutoff distance ($R_g^{cut} = 9$ Angstrom) from one another are considered to be connected. Every such connection is referred to as an edge. See Fig. 3.2a for an illustration of this step.

2. Not all edges are retained. An edge between two guest molecules is only kept if at least N_w water molecules (with $N_w = 5$ in this study) are found that satisfy both of the following conditions:

   - Each water molecule lies within cutoff distance ($R_w^{cut} = 6$ Angstrom) from both guest molecules.
   - Each water molecule lies within the overlapping volume of two cones projected toward one another along the axis connecting the guest molecules, each spanning 45° from that axis.

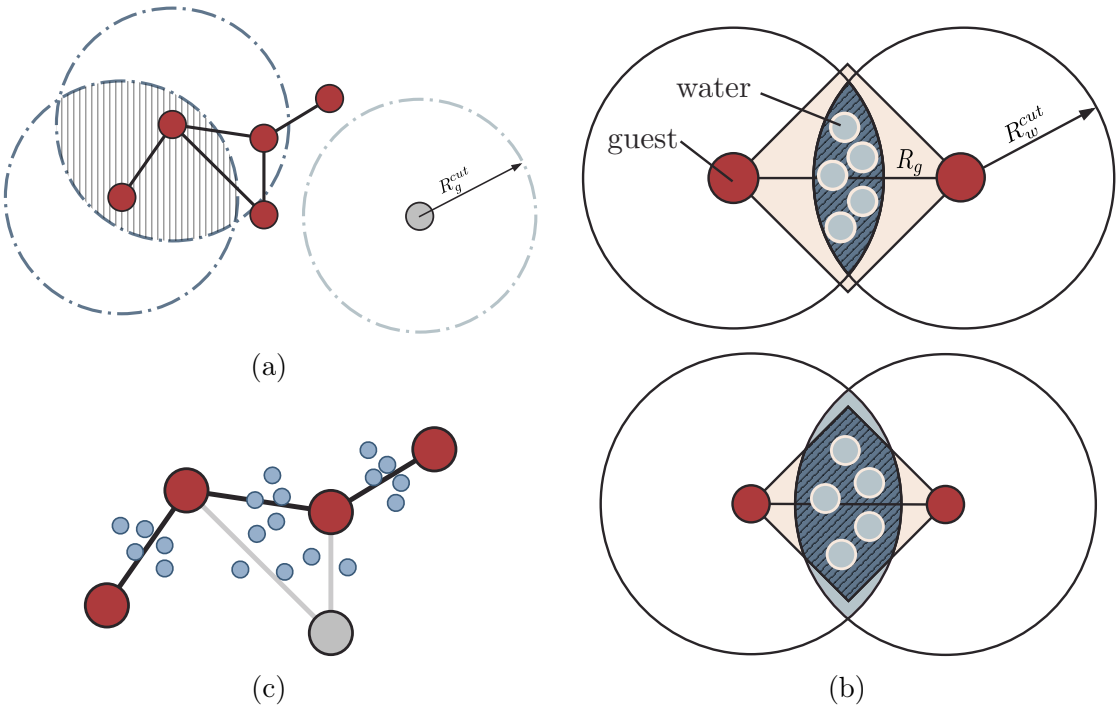


Figure 3.2: This visualization outlines the three-step procedure for determining the MCG order parameter, which quantifies the size of the largest cluster. (a) Guest Condition: Guest molecules (spheres) within pairwise distance $R_g^{cut} = 9\,\text{Å}$ (dashed blue circles) form preliminary edges (black lines). (b) Water Condition: Each edge requires at least $N_w = 5$ water molecules simultaneously located within $R_w^{cut} = 6\,\text{Å}$ of both guests (transparent spheres) and inside overlapping conical volumes spanning 45° radially outwards from the inter-guest axis (dark blue shading). (c) Edges are retained only when sufficient bridging waters exist. Guests with at least N connections ($N = 1$ for MCG-1) become MCG monomers. The largest connected cluster (red circles) defines the order parameter.

The combination of these two conditions form a volume, displayed in Fig. 3.2b for two different guest distances as the shaded blue region, in which all N_w water molecules separating the guest molecules must be located. If less than N_w waters are found within this region, the edge is removed, as seen in Fig. 3.2c.

3. Finally, all guest molecules that have at least N connecting nodes, where N is defined by the MCG-N order parameter (N=1 for MCG-1 as used in this study), will be denoted as MCG monomers.

The MCG order parameter is now taken to be the size of the largest cluster of MCG monomers found through this procedure in the system.

3.4.2 Cage Number Difference

The nucleation landscape of clathrate hydrates features a complex free energy surface governed by competing cage formation pathways. For CO_2 clathrates, we know that there exist at least two different channels of nucleation. The system can either solidify into an amorphous solid, or into a the crystalline phase, generally sI [33].

Previous research has established that these nucleation channels are defined by their populations of characteristic cages specific to each structure [31, 38].

For CO_2 clathrates, the amorphous pathway is characterized by a cage type rarely prominent in other hydrate systems: the $4^15^{10}6^2$ cage. The reason for this cage to be unusually common for the CO_2 guest molecule, is known to be linked to the size and especially the linear shape of CO_2 being compatible with the elongated geometry of the cage [4, 20]. The crystalline pathway, on the other hand, shows a high abundance of $5^{12}6^2$ cages, which are the most common cages in sI hydrates [11].

Having identified two cages whose abundance is each linked with one of the two nucleation pathways, we define a Cage Difference (CD) metric as the numerical difference between $5^{12}6^2$ and $4^15^{10}6^2$ cage populations:

$$\mathrm{CD} = N_{5^{12}6^2} - N_{4^15^{10}6^2} \tag{3.5}$$

This CD being above 0 indicates, that the system is currently in the crystalline nucleation pathway, while a CD below 0 indicates that the system is in the amorphous nucleation pathway. This parameter is comparable to past studies, which have instead used a ratio of cage numbers, as $\mathrm{CR} = N_{5^{12}6^2}/N_{4^15^{10}6^2}$ [4, 31, 38]. We choose to use the CD instead, since this preserves linear symmetry across the classification threshold and it eliminates sensitivity to division-by-zero artifacts and statistical outliers.

To be able to find this CD, it is necessary to be able to identify the relevant cages in the system. Our analysis employed a modified version of the GRADE code [26], extended to include identification of the $4^1 5^{10} 6^2$ cage type.

This code is able to identify cages in the system by employing a set of rules, which are based on the geometric properties of the cages. The code first identifies all water molecules within a cutoff distance of 3.5 Angstrom from each other and then identifies whether these molecules form a ring (square, pentagon, hexagon). Checking these rings for relevant overlaps with other rings leads to the discovery of certain 'cups', which are parts of the cages we want to identify. Specifically, we identify 5^6, $5^6 6^1$ and $4^1 5^4 6^2$ cups. $5^{12} 6^2$ cages consist of two $5^6 6^1$ cups and $4^1 5^{10} 6^2$ are made up of a combination of a 5^6 and a $4^1 5^4 6^2$ cup. By finding relevant overlaps between these cups, the code is able to identify the cages in the system.

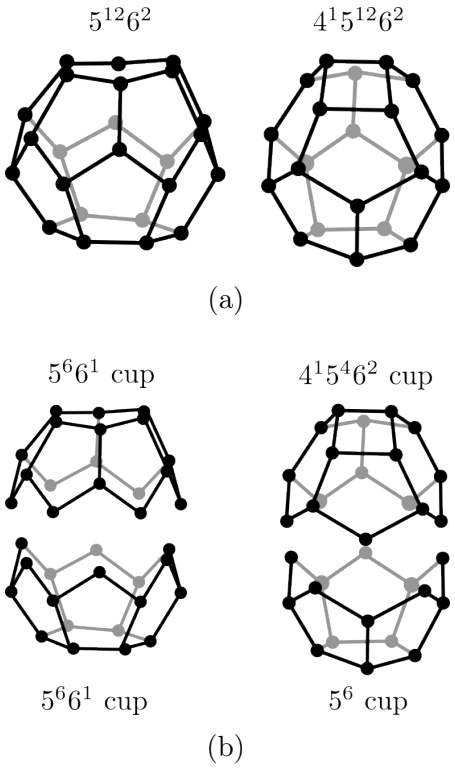


Figure 3.3: (a) Geometric representation of the two most relevant cages in the system, $5^{12} 6^2$ and $4^1 5^{10} 6^2$, with nodes representing the water molecules and (b) the cups making up these cages.

4 Results and Duscussion

4.1 Simulation Setup

4.1.1 System Parameters

In order to evaluate the performance of Always Accepting Transition Path Sampling (AA-TPS), we have used this method to sample the liquid to solid transition in a CO_2 Clathrate Hydrate system. To encourage their formation, the system contains 2944 water molecules and 512 CO_2 molecules, which corresponds to the stoichiometry of a 4x4x4 sI hydrate. All Molecular Dynamics (MD) simulations are performed using the simulation package OpenMM [15].

The non-bonded interactions in the system are modeled through Lennard-Jones interactions, which take the following shape

$$V_{\mathrm{LJ}}(r_{ij}) = 4\epsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right], \tag{4.1}$$

where r_{ij} is the distance between the particles i and j, σ_{ij} is the distance at which the potential is zero, and ϵ is the depth of the potential well.

For the water-water interactions, we used the TIP4P/Ice force field, which is a well-established model for water in the context of clathrate hydrates [1]. This model was specifically developed to better reproduce the melting behavior of water, over previous water models. As other TIP4P-type models, it is a rigid planar four-site interaction potential, which means that it models an additional fourth dummy site without mass but with additional charge.

For the guest-guest interactions we used the TraPPE force field which is a common choice for the simulation of CO_2 Clathrates [12, 27].

It is common to model intermolecular interactions through the Lorentz-Berthelot combination rules, blending the two distinct force fields through the following equations:

$$\sigma_{ij} = \frac{\sigma_i + \sigma_j}{2}, \tag{4.2}$$

$$\epsilon_{ij} = \sqrt{\epsilon_i \cdot \epsilon_j}. \tag{4.3}$$

However, as shown by Costandy et al. [12], the classic Lorenz-Berthelot combining rules applied to TIP4P water and TraPPE CO_2 force fields fail to predict the three-phase coexistence temperature (T_3) in CO_2 hydrates. Therefore, we implement modified guest-water interactions proposed in the same study, which apply the following equations to combine the ϵ parameters:

$$\epsilon_{O(CO_2)-O(H_2O)} = \chi\sqrt{\epsilon_{O(CO_2)} \cdot \epsilon_{O(H_2O)}}, \tag{4.4}$$

where the modification parameter χ was found to optimally reproduce the three-phase coexistence temperature (T_3) of CO_2 hydrates at $\chi = 1.08$.

Additionally, water oxygen-hydrogen and hydrogen-hydrogen distance were constrained using OpenMM at 0.09572 nm and 0.15139 nm, respectively as prescribed by the TIP4P/Ice force field. The van der Waals cutoff was set to 1 nm and the electrostatic interactions were calculated using the Particle Mesh Ewald method.

Similarly, the CO_2 bond length was constrained at 0.116 nm and the angle between the two bonds was harmonically restrained at 180° with a spring force of 1500 $kJmol^{-1}rad^{-2}$, as prescribed by the TraPPE force field.

The simulations are performed using a cubic box with periodic boundary conditions. The equations of motion are solved using the Velocity Verlet with Velocity Randomization (VVVR) integrator [32] with a timestep of 2 fs. As the name suggests, this integrator is a modification of the Velocity Verlet (VV) integrator, which is a second-order symplectic integrator. The modification consists of a randomization step that is applied to the velocities at a given frequency of 1 ps^{-1} (or once per 500 timesteps), which keeps the system at a given temperature. Additionally to temperature control, we also keep the pressure constant, allowing the system to sample from the Isothermal–isobaric (NPT) ensemble. This is achieved using a Monte Carlo barostat, which stochastically proposes isotropic scaling of the box size at a given frequency of 0.25 ps^{-1} (or once per 2000 timesteps). Proposed volume changes are accepted or rejected based on a Monte Carlo criterion, thereby maintaining constant pressure.

4.1.2 Obtaining the initial trajectory

As a first step, the system is equilibrated at a temperature of 300 K and a pressure of 400 bar, since the spontaneous formation of clathrates hydrates is likely to occur even with brute force MD simulations, without the need for enhanced sampling methods. At this temperature, the system is in the liquid phase where the CO_2 supersaturation significantly exceeds natural aqueous solubility limits, resulting in spontaneous bubble formation. This bubble acts as a driving force for the clathrate formation, as it provides a source of CO_2 molecules to the assembling clathrates cages during the nucleation process. The resulting configuration is used as an

initial state for another 'equilibration' run, lasting for 4 ns now at a temperature of 250 K and a pressure of 400 bar. From this equilibration, multiple initial MD simulations were started at the same temperature and pressure, lasting for 880 ns. In these trajectories the formation of clathrate hydrates is observed, making it possible for these trajectories to serve as a starting point for TPS.

Since the goal is to perform TPS at a temperature of 260 K and a pressure of 500 bar, conditions aligned with prior simulation studies [4, 5], and of interest for industrial contexts, direct use of trajectories equilibrated at 250 K as initial paths for TPS would likely incur a significant computational cost due to the slow relaxation to the true transition path ensemble. For this reason, we pick multiple configurations from the 250 K trajectory, at different sizes of the growing cluster, measured by the MCG parameter, as explained in Section 3.4.1. From each of these configurations, 10 different simulations are started. When the transition into both the liquid and the solid phase is observed from a given configuration, the trajectory segments are stitched together to obtain a final initial trajectory, which spans from the liquid to the solid state. Not only does this protocol generate suitable initial paths for TPS but also constitutes a small-scale committor analysis, providing preliminary insights into the free energy landscape along the MCG reaction coordinate. This provides valuable insight, since when performing 'shooting from the top' TPS (see Section 2.1.4) we want to place the center of the shooting range on the top of the barrier. From this committor analysis, this center of the shooting range μ is found to be at a MCG value of 175.

4.1.3 Transition Path Sampling Procedure

Starting from the Initial Trajectory generated as detailed in the previous section, we start the equilibration sampling procedure using One-way Shooting TPS. We start 12 parallel simulations, each starting from the same initial trajectory. This simulation runs for 4 days on 12 NVIDIA A100 GPUs. Averaged across these runs, each simulation produced 6.7 accepted transition paths. From each 12 TPS runs, the final accepted transition path serves as an initial path for a new batch of TPS simulations. All other transition paths from this initial sampling procedure are discarded. Using this 12 new initial equilibrated paths, we can run both the 1wS-TPS and the AA-TPS simulations, to compare their performances. We perform three consecutive batches of 12 TPS simulations. Each batch runs for 4 days on the same hardware as the equilibration run, and the simulations in each batch use the final paths from the previous batch as their starting points. The TPS algorithms share the same parameters defining the shooting range, which are chosen to be $\mu = 175$, as mentioned in the previous section, $\alpha = 15$ and $\beta = 2.5$ (see Eq. (2.44)). Both methods save a configuration every 200 ps in the produced trajectories.

In total, 1wS-TPS was able to generate 280 accepted trajectories, with 401 rejected trajectories, which comes out to a percentage of accepted paths of 41.1%, while AA-TPS generated 732 accepted trajectories, with 0 rejected trajectories.

4.2 Path Correlations

To estimate the performance of the two TPS methods, we want to analyze the correlation between the transition paths sampled by the two methods. Generally, if a method decorrelates paths quickly, a larger area of path-space will most likely be explored, making the method more effective. This means that, just like in standard configuration Monte Carlo, accepting a high number of new paths alone is not a complete indicator for how good a method is. If all the accepted paths are 'close' to each other, the method is not providing a good exploration of the TPE. There are many ways to define how 'close' two paths are. The two TPS methods we are comparing both replace a part of the old trajectory at every accepted iteration. Therefore, a reasonable way of estimating this similarity between paths is by tracking how much of the previous path remains the same when moving to a new transition path. The exact procedure to define this oldest path segment will be laid out in the following.

Using each generated transition path as a potential initial path, we systematically process all subsequent paths while tracking the evolving Oldest Path Segment (OPS), the segment which is still retained from the initial path. Through this procedure, we can analyze features of the OPS, as for example its length and the ages (the number of iterations) at which the OPS completely disappears with a new accepted transition path. With every trial we also track how many force evaluations were needed to produce it. This allows us to estimate the real computational cost of the two methods, with force evaluations being the crucial factor when evaluating methods performance.

The evolution of the average length of the OPS divided by the average initial path length $\langle L_{\mathrm{OPS}}\rangle/\langle L_0\rangle$ as a function of the average force evaluations $\langle n_{\mathrm{F}}\rangle$ is shown in Fig. 4.1. The inset is plotted on the same axes, but it now shows the average normalized length of only surviving paths, meaning only non-zero lengths are counted. This highlights those cases in which a portion of an initial path is still residual after many accepted path, an information that is otherwise lost in the main plot.

To obtain the average force evaluations $\langle n_{\mathrm{F}}\rangle$, the trials at which the OPS length was recorded are multiplied with the average number of force evaluations per trial, which is roughly 43.7×10^6 in the case of AA-TPS and 47.8×10^6 for 1wS-TPS. We need to consider with this analysis that the maximum age of the OPS is capped by the number of trials performed in the TPS simulations. Across the 12 TPS

simulations, the AA-TPS method generated an average of 61.0 sequential trials, while the 1wS-TPS method averaged 56.75.

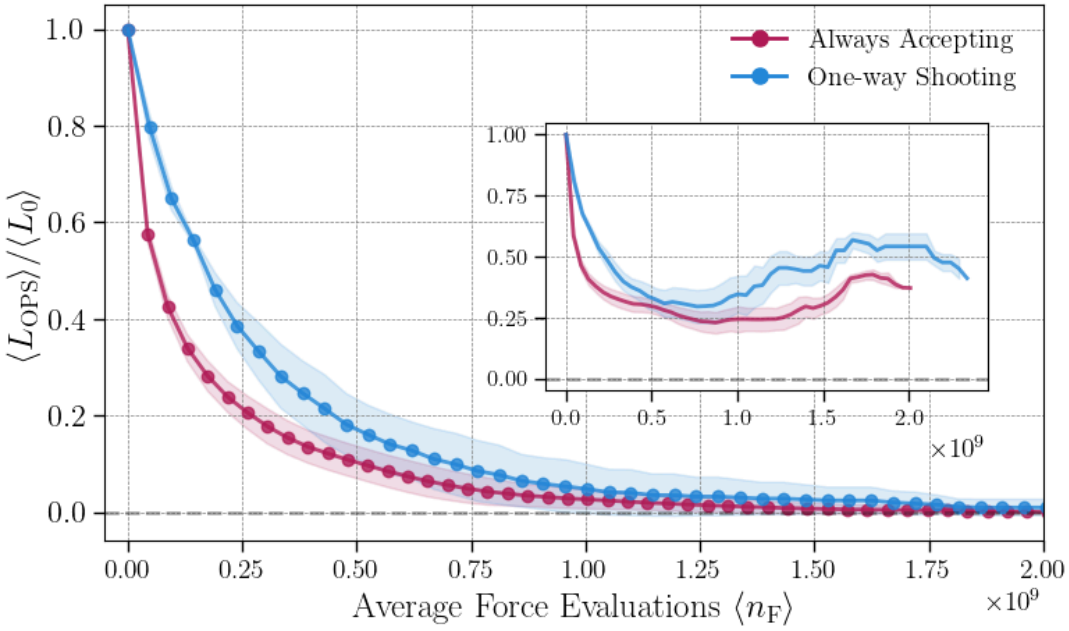


Figure 4.1: Average length of the OPS divided by the average initial path length $\langle L_{\mathrm{OPS}} \rangle / \langle L_0 \rangle$ as a function of the average force evaluations $\langle n_{\mathrm{F}} \rangle$. The inset similarly shows the normalized average length for average force evaluations, however only for surviving paths. The error was calculated through blocked averages.

This analysis reveals AA-TPS achieves faster decay of the OPS length with significantly lower computational cost than 1wS. Importantly, obtaining a path that retains 50% of the original trajectory takes 1.7×10^8 average force evaluations for 1wS-TPS, whereas AA-TPS only needs 0.6×10^8 average force evaluations. This means that 1wS-TPS requires 2.8 times as many force evaluations as AA-TPS to replace at least half of the path. Reducing the OPS to 10% of the initial path length requires approximately 5.1×10^8 force evaluations for AA-TPS compared to 7.1×10^8 for 1wS, a 1.4 times reduction in computational demand with the new algorithm. This substantial difference directly impacts resource allocation and sampling time efficiency.

Interestingly, the inset shows the OPS length of only the surviving paths, which shows a minimal surviving length of about 25% of the original path length for AA-TPS and 30% for 1wS-TPS. At correspondingly high force evaluations, when

few paths are alive, both methods still show quite high path lengths, that rise once again as shorter paths are being completely eliminated. This implies that even at high force evaluations, when most OPS have already been eliminated, the remaining paths can still be at lengths around 56% of the original path for 1wS-TPS and of 42% for AA-TPS. This suggests that, while a significant number of OPSs shrink and then die out on shorter timescale, there is a group of highly persistent OPS for both methods, that remain at substantial lengths, until being replaced.

In Fig. 4.2 we see the number of remaining OPS as a function of the average force evaluations $\langle n_{\mathrm{F}} \rangle$, along with the average OPS age, defined as the force evaluations necessary for the complete replacement of the path segment.

Consistent with the earlier trend, we once again observe that AA-TPS outperforms 1wS-TPS in terms of the number of surviving paths. It takes approximately 3.4×10^8 force evaluations for AA-TPS to reach a point where only 50% of the initial paths remain, while 1wS-TPS requires about 5.3×10^8 force evaluations to achieve the same result. As the fraction of surviving paths decreases, the observed difference between methods decreases. Additionally, the average age, representing the number of force evaluations required to completely replace an OPS, is comparable for both AA-TPS and 1wS-TPS, with AA-TPS showing a slight advantage. Still, we notice once again that 1wS-TPS takes many more force evaluations to replace the longest remaining OPS.

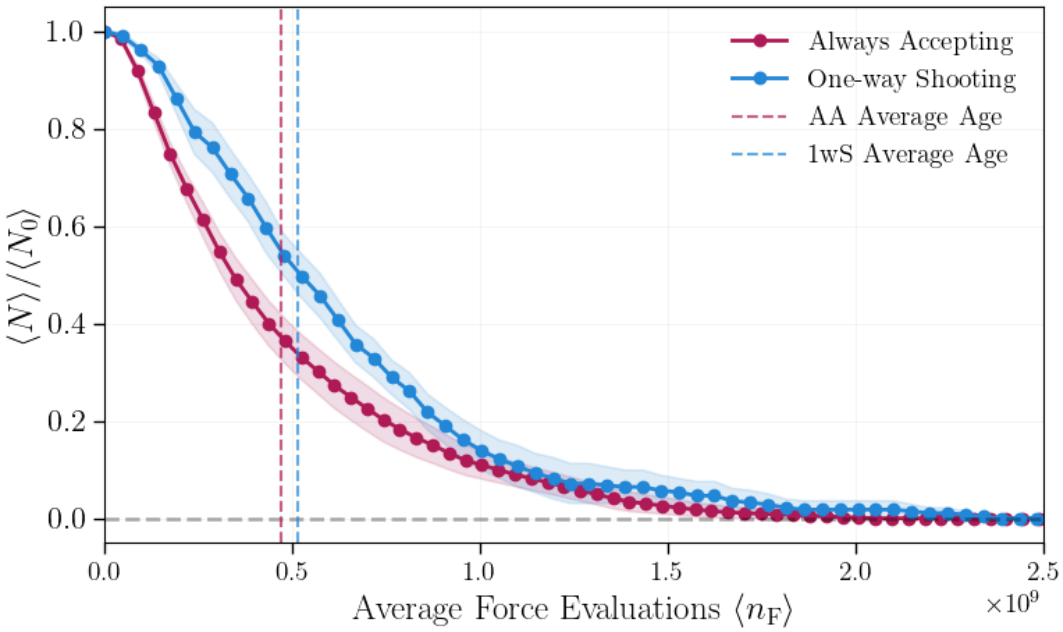


Figure 4.2: Solid lines show the number of surviving paths divided by the number of initial paths $\langle N \rangle / \langle N_0 \rangle$ as a function of the average force evaluations $\langle n_F \rangle$. The dashed lines show the average age of the surviving paths, defined as the force evaluations needed for complete path segment replacement. The error was calculated through blocked averages.

Since the AA-TPS algorithm leads to a much higher percentage of accepted paths, it is not a surprise that the number of force evaluations is significantly lower for AA-TPS than for 1wS-TPS. Note that, a more efficient path production does not necessarily mean that a method is better at sampling path space. Yet, all the results presented point to the conclusion that AA-TPS replaces paths considerably faster than 1wS-TPS. This strongly implies that the AA-TPS algorithm is much more efficient at sampling path space than standard methods. Path correlation is however not the only measure of how well a method performs, with this being a highly non trivial question. Therefore, we will look more closely into the capability of AA-TPS to sample the complex TPE of clathrate hydrates in the following section.

4.3 Nucleation Channel Analysis

To identify the nucleation channels present in the system, the TPEs resulting from the two TPS methods are analyzed for their structural properties. As outlined in Section 3.4.2 we especially identify two cages present in the system, namely the $5^{12}6^2$ and $4^15^{10}6^2$ cages. The prevalence of the former is an indicator of the crystalline phase, while prevalence of the latter indicates an amorphous phase.

Firstly, we look at this number of cages for all trajectories every 2 ns, which is 10 times the saving frequency of 200 ps. The density of these cage counts for the two cage types identified are shown in Fig. 4.3. The left panel shows the histogram of the cage count density for the 1wS-TPS method, while the right panel shows the cage counts for the AA-TPS method.

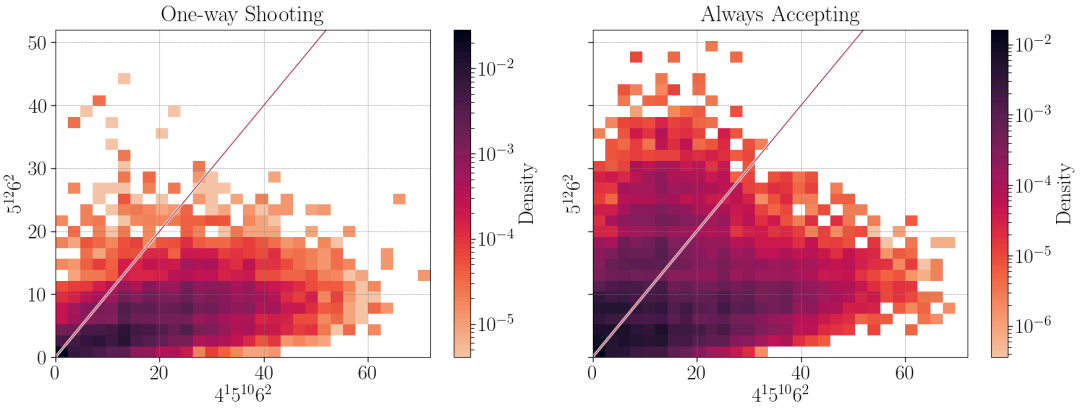


Figure 4.3: Histogram of the cage count densities of the $5^{12}6^2$ and $4^15^{10}6^2$ cages for the whole transition paths. The left panel shows the cage count density for 1wS-TPS, while the right panel shows the cage count density for AA-TPS. The red line indicates an equal number of both cage types.

The results from 1wS-TPS show a clear preference for the amorphous channel, with only slight population of the crystalline channel. This result is not surprising and consistent with previous findings, which also indicate that, at the given temperature of 260 K, the system is largely found in the amorphous phase [4]. It is, however, notable that there are structures which are far into the crystalline channel, with up to 4 times as many $5^{12}6^2$ cages as $4^15^{10}6^2$, which hints at the existence of a crystalline channel, not well sampled by 1wS-TPS. Comparing to the results from AA-TPS, we see a much more balanced distribution of the two cage types. This indicates that AA-TPS is samples more efficiently both nucleation channels, while 1wS-TPS is largely stuck in the amorphous channel. The amorphous channel still seems to be the preferred channel. However, the more efficient AA-TPS algorithm

is now able to sample the crystalline channel.

Instead of looking at the whole paths, it is also interesting to look at the cage counts at the endpoints of the transition paths, which are shown in Fig. 4.4. Since the definition of the stable states considers the system solid for a MCG of 300, these points on the trajectory contain a largest cluster of roughly this size. Again, the left panel shows the cage count density for AA-TPS and the right panel shows the cage count density for 1wS-TPS. The endpoints are especially interesting, since they give the strongest indication for the structure in which the clathrate will eventually grow into. Once again, results similar to the one presented in Fig. 4.3 are obtained, with AA-TPS sampling the crystalline channel significantly more than 1wS-TPS.

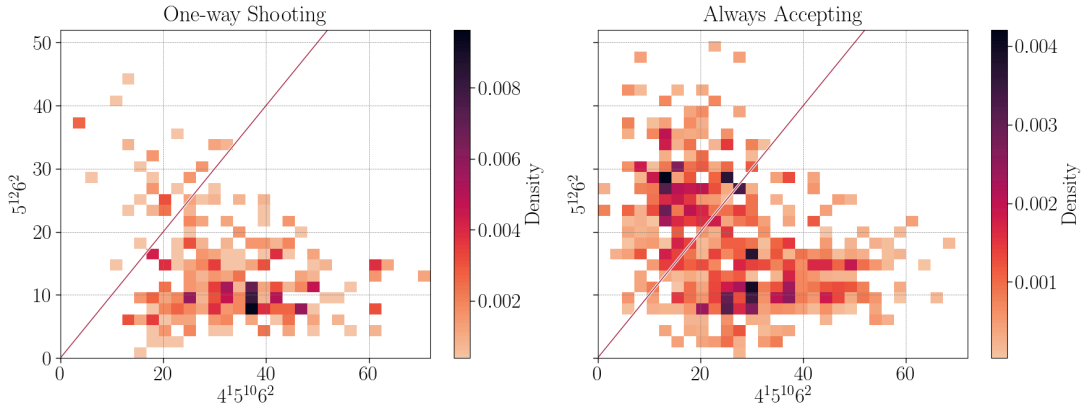


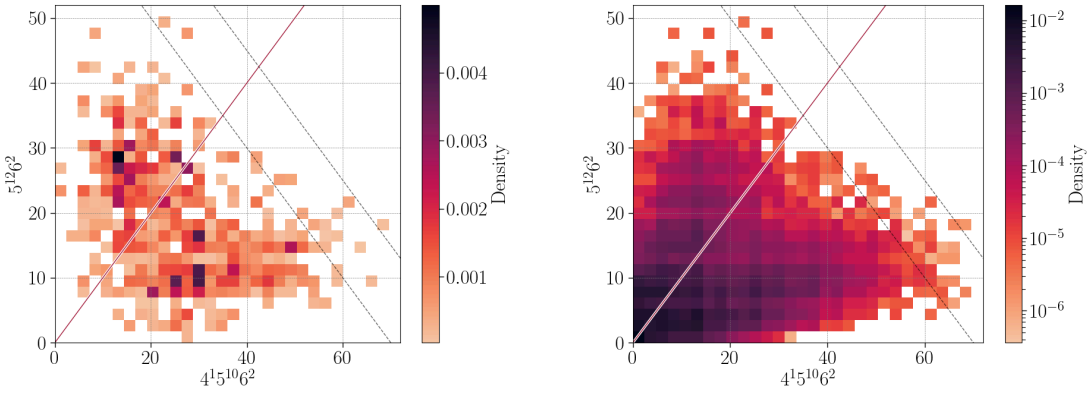
Figure 4.4: Histogram of the cage count densities of the $5^{12}6^2$ and $4^15^{10}6^2$ cages at the endpoints of the transition paths, roughly 300 MCG. The left panel shows the cage counts for 1wS-TPS, while the right panel shows the cage counts for AA-TPS. The red line indicates an equal number of both cagetypes.

Notably, we find many structures situated right at the boundary between the crystalline and amorphous channels (indicated by the red line), suggesting that the separation between the two is not yet sufficient to fully prevent transitions between them. If at 300 MCG this separation between the channel was fully established, you would expect a much lower density along the threshold between amorphous and crystalline.

It is, however, reasonable that MCG might be poorly suited as an order parameter for characterizing this separation since it indiscriminately counts all cage structures, regardless of type. This is evident in Fig. 4.4, while systems can contain clusters of around 300 MCG, the number of relevant cages (namely $5^{12}6^2$ and $4^15^{10}6^2$) can be as low as 20.

Therefore, in Fig. 4.5, we mark constant numbers of total cages N_{tot}, where $N_{\text{tot}} = N_{5^{12}6^2} + N_{4^1 5^{10} 6^2}$, of 70 (left) and 85 (right) with gray dashed lines. This analysis relies exclusively on the results obtained using AA-TPS, since this alone makes this analysis possible.

We observe that for a total number of cages of roughly 70 to 85, a noticeable separation emerges. In this range, there is a considerable population of the two channel deep into the states, whereas no structures are present around the threshold region. This behavior indicates that N_{tot} serves as a more effective order parameter for analyzing this channel separation than MCG. Accordingly, we identify the potential onset of channel separation at around $N_{\text{tot}} = 70$ cages.



(a) (b)

Figure 4.5: Histogram of the cage count densities of the $5^{12}6^2$ and $4^1 5^{10} 6^2$ cages for around 300 MCG (left) and for the whole trajectories (right), both are results of AA-TPS. A constant numbers of total cages $N_{\text{tot}} = N_{5^{12}6^2} + N_{4^1 5^{10} 6^2}$ of 70 (left) and 85 (right) is marked with gray dashed lines.

Finally, we are interested in the evolution of the structural makeup of the forming cluster. We can analyze this by looking at a histogram of the CD density for increasing MCG values, which is shown in Fig. 4.6. The left panel shows the CD for the 1wS-TPS method, while the right panel shows the CD for the AA-TPS method. To assure some confidence in the determination of amorphous vs. crystalline, we impose a cutoff of at least 5 cages for both cage types.

The AA-TPS method once again beautifully captures the structures for growing clusters, displaying a conical shape for increasing MCG, which is not well defined in the case of 1wS-TPS. We observe that some trajectories form crystalline phases immediately upon nucleation, which only turn more crystalline as the MCG increases. This statement seems to form an apparent contrast to the Blob Nucleation

Hypothesis, which requires initial amorphous precursors prior to crystallization. This disagreement, however, stems from differing structural definitions.

Our amorphous vs. crystalline classification depends exclusively on local cage numbers, while true crystallinity requires detection of long-range translational symmetry. [30] Using our definition we can state that the cages needed to build a crystalline structure are already present in the system, however potentially not yet arranged in a way that would resemble an sI structure. The results presented are therefore not a contradiction to the Blob Hypothesis and rather add onto the existing knowledge, by showing that the cages needed to form a crystalline structure are already present in the system, even at low MCG values.

In this figure, we once again nicely see the preference for the amorphous channel, with the densely populated areas tending more and more towards $4^1 5^{10} 6^2$ channels for growing MCG values. Additionally, we notice that there is no visible separation between the two channels, suggesting once again that this divergence of channels is not well defined by the MCG value alone.

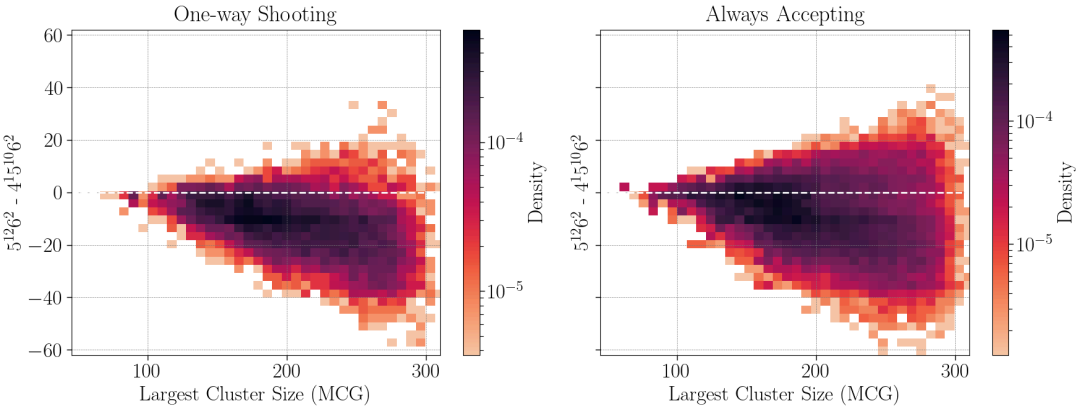


Figure 4.6: Histogram of the cage number difference density as $5^{12}6^2 - 4^1 5^{10} 6^2$ for growing largest cluster sizes indicated by the MCG parameter. The left panel shows results for 1wS-TPS, while the right panel shows the results for AA-TPS.

Conclusively, we can state that the AA-TPS method has made it possible, for the first time, to achieve an understanding of the clathrate hydrates at a temperature of 260 K, that includes their existence in the crystalline form. We have shown that there exist two channels, which indicate a separation appearing roughly at 70 total cages $N_{\text{tot}} = N_{5^{12}6^2} + N_{4^1 5^{10} 6^2}$. To further investigate this separation, it is necessary to study longer trajectories, letting the simulations go to higher MCG values than the ones studied in this work. To meaningfully study larger cluster, is is however necessary to increase the system size, since for the current system, a

cluster of MCG 300 already takes up a significant portion of the simulation box.

Additionally, we find information about the onset of cluster growth, where from the start there is a significant population of the crystalline channel. It is therefore possible to state that the clusters necessary for either crystalline or amorphous growth are already present in the system. Whether or not early clusters are a significant indication of their evolution is not directly measured. The lack of a clear separation of channels at an early stage might be an indication that early cages do not yet lock the system into a certain later structure, but this could be an interesting topic for future research.

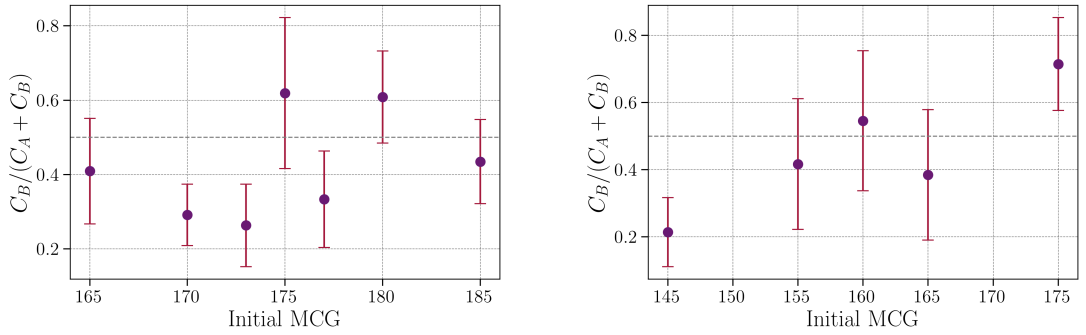
4.4 Critical Nucleus

To further analyze the nucleation process, we perform a Critical Nucleus Analysis, which is a method to identify the size of the critical nucleus in terms of the MCG order parameter. The critical nucleus is defined as the size of the cluster at which the top of the free energy barrier is reached. Before reaching the top of the barrier, any cluster formed is more likely to dissociate through thermal fluctuations, than to grow further. After reaching the top of the barrier, the cluster is more likely to grow larger, than to dissociate. Right at the top of the barrier, the cluster is in a point of unstable equilibrium, where the rate of dissociation and the rate of growth are equal.

We use this information to identify the critical nucleus size, by starting simulations from a range of initial configurations along the MCG order parameter. Initial configurations are chosen randomly from the transition paths samples during the TPS runs: for each MCG value, 20 initial configurations are picked and an equal number of simulations is started. These simulations are run for 200 ns and are then analyzed for their evolution, again using the MCG parameter. In particular, the goal of our analysis is to identify differences in the value of the critical nucleus size in units of the MCG for paths that end up in the crystalline or amorphous phase. It is possible, in fact, that an ordered and very symmetric structure like the crystalline one necessitates of a smaller nucleus to trigger the transition, compared to the amorphous case in which the level of symmetry is lower. For this reason, 20 configurations are selected separately among paths that reached the amorphous phase and among those that reached a crystalline phase.

The number of paths reaching the liquid state (state A), denoted as C_{A} and the number of paths reaching the solid state (state B), denoted as C_{B}, are counted. The ratio of paths landing in the solid state C_{B} over all paths that reached a state is then plotted against the initial MCG value, which is shown in Fig. 4.7. We expect to find the critical nucleus at the point where this ratio is equal to 0.5, meaning that half of the paths reached the solid state and half of the paths reached

the liquid state.



(a) (b)

Figure 4.7: Results of the Critical Nucleus Analysis for both amorphous and crystalline clathrates at 260 K. Panel (a) shows the results for the amorphous phase, while panel (b) shows the results for the crystalline phase. The error was calculated through blocked averages.

The results do not show a clear critical nucleus size. It is possible to identify a general trend, which points to a critical nucleus size for the amorphous phase lying roughly around 175 MCG. For the crystalline phase, the critical nucleus size seems to be around 160 MCG. While the uncertainty in these values remains high, the results point at the critical nucleus size being larger for the amorphous phase than for the crystalline phase as expected by physical intuition. Longer simulations and and analysis of bigger system sizes will shed more light on these preliminary conclusions.

CNT predicts a direct relation between the radius of the critical nucleus R^* and the height of the free energy barrier G^* as stated in Eq. 3.3. This suggests that if the amorphous phase has a larger critical nucleus size, it will also exhibit a higher free energy barrier that the system needs to grow into an amorphous solid. This information competes with our prior findings, stating that the TPE sampled demonstrates higher population of the amorphous phase than the crystalline phase after the critical nucleus size is reached. Critical Nucleus sizes are however not the only indicator for the population of specific channels. Instead it is possible, that early growth favors the formation of amorphous precursors due to low formation barriers, while reorganization into crystalline structures would be energetically costly. This kinetic trapping could explain the higher population of amorphous solids, despite the crystalline phase's smaller critical nucleus size.

The quantitative uncertainty in critical nucleus measurements limits exact interpretations. While a trend is observed in the size difference, definitive conclusions

about nucleation barriers require further investigation.

5 Conclusion

In this thesis, we have introduced a novel approach to TPS, termed AA-TPS, which addresses the limitations of traditional TPS methods. By eliminating the restriction on target states and employing a reweighting scheme based on detailed balance, AA-TPS generates exclusively accepted transition paths, significantly enhancing sampling efficiency. Beyond the theoretical derivation, we have proven the method's ability to correctly reproduce the equilibrium distribution in a 2D potential landscape. This promising result provides initial validation of the method's efficacy in sampling representative transition path ensembles. Comprehensive benchmarking against experimental observables, however, remains reserved for full-scale applications in realistic systems.

The TPS of the clathrate hydrates proves to be computationally costly. It is therefore an immediate success of AA-TPS to produce 2.6 times as many accepted paths as 1wS-TPS in the same amount of sampling time of roughly 12 days. We show that this increase in accepted paths also translates to a superior sampling of the TPE by quantifying the path decorrelation. Analysis of the OPS leads to the conclusion that, in terms of the average force evaluations necessary, the AA-TPS method produces decorrelated paths much faster than 1wS-TPS, reducing to a 50% overlap with the previous path 2.8 times faster. Quantitative comparisons confirm that our approach yields decorrelated trajectories and explores previously inaccessible path-space regions at smaller computational cost than the standard protocol 1wS-TPS.

This finding is further reflected in the results for the TPE of the CO_2 clathrate hydrates. Our AA-TPS study identifies a crystalline nucleation pathway for CO_2 hydrates at 260 K and 500 bar that, to the best of our knowledge, had remained mostly undetected in prior enhanced sampling studies, while experimental evidence suggested thermodynamic plausibility of crystalline formation [3]. A prior study by Arjun et al. [4] of CO_2 clathrates using TPS has performed 564 trials at the same conditions and indeed found no sifnificant population of the crystalline channel, concluding that this pathway is negligible under these conditions. This mirrors the results achieved using our benchmark method of 1wS-TPS, at a comparable amount of trials of 681, demonstrating that AA-TPS achieves improved sampling not only relative to our benchmark but also against established methods under identical conditions.

The population of both nucleation channels, as revealed by analyzing the full

trajectories of the $5^{12}6^2$ and $4^15^46^2$ cages, suggests only a tentative separation of pathways. Even at a largest cluster size of 300 MCG, there is no clear divide between the amorphous and crystalline channel. It is however notable, that there is a hint at a separation between channels when instead looking at total cage numbers $N_{\text{tot}} = N_{5^{12}6^2} + N_{4^15^{10}6^2}$ of 70 and above. This observation implies that the system's choice between crystalline and amorphous nucleation is governed more by the relative abundance of specific cage types than by the size of the largest cluster. To arrive at a clearer understanding of the two nucleation pathways and their possible separation, it is likely necessary to go to higher largest cluster sizes, which additionally requires an increase in the system size.

Further, we find that even at small MCG there is a preference for the formation of the $4^15^46^2$ cage, with amorphous structures being immediately favorable. Nonetheless, right from the earliest stages analyzed, there is a significant population of $5^{12}6^2$ cages, indicating that the crystalline pathway is viable right from the emergence of a cluster.

Preliminary analysis suggests a possible structure dependence in critical nucleus size, with crystalline trajectories exhibiting smaller critical nuclei, at around 160 MCG, while in the amorphous phase it is roughly estimated to be around 175 MCG. This observation warrants further investigation given the statistical uncertainty in pathway classification.

In conclusion, our Always Accepting Transition Path Sampling (AA-TPS) methodology samples previously inaccessible regions of the free-energy landscape governing CO_2 hydrate nucleation, while maintaining computational complexity comparable to standard TPS. AA-TPS resolves nucleation pathways that remained prohibitively expensive to capture with existing enhanced sampling approaches.

Bibliography

[1] J. L. F. Abascal et al. "A Potential Model for the Study of Ices and Amorphous Water: TIP4P/Ice". In: *The Journal of Chemical Physics* 122.23 (June 2005), p. 234511. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.1931662. (Visited on 24/07/2024).

[2] Morteza Aminnaji et al. "CO2 Gas Hydrate for Carbon Capture and Storage Applications – Part 1". In: *Energy* 300 (1st Aug. 2024), p. 131579. ISSN: 0360-5442. DOI: 10.1016/j.energy.2024.131579. URL: https://www.sciencedirect.com/science/article/pii/S0360544224013525 (visited on 03/06/2025).

[3] Sai Bhargav Annavajjala et al. "A Review of CO2 Clathrate Hydrate Technology: From Lab-Scale Preparation to Cold Thermal Energy Storage Solutions". In: *Energies* 18.10 (10 Jan. 2025), p. 2659. ISSN: 1996-1073. DOI: 10.3390/en18102659. URL: https://www.mdpi.com/1996-1073/18/10/2659 (visited on 15/07/2025).

[4] A. Arjun and P. G. Bolhuis. "Molecular Understanding of Homogeneous Nucleation of CO_2 Hydrates Using Transition Path Sampling". In: *The Journal of Physical Chemistry B* 125.1 (14th Jan. 2021), pp. 338–349. ISSN: 1520-6106, 1520-5207. DOI: 10.1021/acs.jpcb.0c09915. URL: https://pubs.acs.org/doi/10.1021/acs.jpcb.0c09915 (visited on 21/04/2024).

[5] A. Arjun and Peter G. Bolhuis. "Homogenous Nucleation Rate of CO2 Hydrates Using Transition Interface Sampling". In: *The Journal of Chemical Physics* 154.16 (26th Apr. 2021), p. 164507. ISSN: 0021-9606. DOI: 10.1063/5.0044883. URL: https://doi.org/10.1063/5.0044883 (visited on 04/03/2025).

[6] Brian C Barnes and Amadeu K Sum. "Advances in Molecular Simulations of Clathrate Hydrates". In: *Current Opinion in Chemical Engineering*. Nanotechnology / Separation Engineering 2.2 (1st May 2013), pp. 184–190. ISSN: 2211-3398. DOI: 10.1016/j.coche.2012.12.002. URL: https://www.sciencedirect.com/science/article/pii/S2211339812000949 (visited on 22/07/2025).

[7] Brian C. Barnes et al. "Two-Component Order Parameter for Quantifying Clathrate Hydrate Nucleation and Growth". In: *The Journal of Chemical Physics* 140.16 (28th Apr. 2014), p. 164506. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.4871898. URL: https://pubs.aip.org/jcp/article/140/16/164506/841694/Two-component-order-parameter-for-quantifying (visited on 13/10/2024).

[8] Benoıt Beauchamp. "Natural gas hydrates: myths, facts and issues". In: *Comptes Rendus. Géoscience* 336.9 (2004), pp. 751–765. ISSN: 1778-7025. DOI: 10.1016/j.crte.2004.04.003. URL: https://comptes-rendus.academie-sciences.fr/geoscience/articles/en/10.1016/j.crte.2004.04.003/ (visited on 03/06/2025).

[9] Yuanfei Bi and Tianshu Li. "Probing Methane Hydrate Nucleation through the Forward Flux Sampling Method". In: *The Journal of Physical Chemistry B* 118.47 (26th Nov. 2014), pp. 13324–13332. ISSN: 1520-6106. DOI: 10.1021/jp503000u. URL: https://doi.org/10.1021/jp503000u (visited on 22/07/2025).

[10] Peter G. Bolhuis et al. "TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark". In: *Annual Review of Physical Chemistry* 53 (Volume 53, 2002 1st Oct. 2002), pp. 291–318. ISSN: 0066-426X, 1545-1593. DOI: 10.1146/annurev.physchem.53.082301.113146. URL: https://www-annualreviews-org.uaccess.univie.ac.at/content/journals/10.1146/annurev.physchem.53.082301.113146 (visited on 26/06/2025).

[11] Wenjiu Cai, Xin Huang and Hailong Lu. "Instrumental Methods for Cage Occupancy Estimation of Gas Hydrate". In: *Energies* 15.2 (2 Jan. 2022), p. 485. ISSN: 1996-1073. DOI: 10.3390/en15020485. URL: https://www.mdpi.com/1996-1073/15/2/485 (visited on 09/07/2025).

[12] Joseph Costandy et al. "The Role of Intermolecular Interactions in the Prediction of the Phase Equilibria of Carbon Dioxide Hydrates". In: *The Journal of Chemical Physics* 143.9 (4th Sept. 2015), p. 094506. ISSN: 0021-9606. DOI: 10.1063/1.4929805. URL: https://doi.org/10.1063/1.4929805 (visited on 23/05/2025).

[13] Jinlong Cui et al. "Fundamental Mechanisms and Phenomena of Clathrate Hydrate Nucleation". In: *Chinese Journal of Chemical Engineering*. SI: Natural Gas Hydrates 27.9 (Sept. 2019), pp. 2014–2025. ISSN: 1004-9541. DOI: 10.1016/j.cjche.2018.12.016. (Visited on 01/06/2025).

[14] Humphry Davy. "I. The Bakerian Lecture. On Some of the Combinations of Oxymuriatic Gas and Oxygene, and on the Chemical Relations of These Principles, to Inflammable Bodies". In: *Philosophical Transactions of the Royal Society of London* 101 (Jan. 1997), pp. 1–35. DOI: 10.1098/rstl.1811.0001. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rstl.1811.0001 (visited on 03/06/2025).

[15] Peter Eastman et al. "OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics". In: *PLOS Computational Biology* 13.7 (26th July 2017), e1005659. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005659. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005659 (visited on 07/07/2025).

[16] Niall J. English and J. M. D. MacElroy. "Perspectives on Molecular Simulation of Clathrate Hydrates: Progress, Prospects and Challenges". In: *Chemical Engineering Science*. 2013 Danckwerts Special Issue on Molecular Modelling in Chemical Engineering 121 (6th Jan. 2015), pp. 133–156. ISSN: 0009-2509. DOI: 10.1016/j.ces.2014.07.047. URL: https://www.sciencedirect.com/science/article/pii/S0009250914003923 (visited on 22/07/2025).

[17] Mauro Ferrario, Giovanni Ciccotti and K. Binder, eds. *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology*. Lecture Notes in Physics 703–704. Berlin ; New York: Springer, 2006. 2 pp. ISBN: 978-3-540-35270-9 978-3-540-35283-9.

[18] J. Willard Gibbs. *On the Equilibrium of Heterogeneous Substances: First [-Second] Part Pt. 2*. On the Equilibrium of Heterogeneous Substances. Offprint from the Transactions of the Connecticut Academy of Sciences. Connecticut Academy of Arts and Sciences, 1874. DOI: 10.5479/sil.421748.39088007099781.

[19] E. G. Hammerschmidt. "Formation of Gas Hydrates in Natural Gas Transmission Lines". In: *Industrial & Engineering Chemistry* 26.8 (1st Aug. 1934), pp. 851–855. ISSN: 0019-7866. DOI: 10.1021/ie50296a010. URL: https://doi.org/10.1021/ie50296a010 (visited on 03/06/2025).

[20] Zhongjin He, Praveen Linga and Jianwen Jiang. "What Are the Key Factors Governing the Nucleation of CO2 Hydrate?" In: *Physical Chemistry Chemical Physics* 19.24 (21st June 2017), pp. 15657–15661. ISSN: 1463-9084. DOI: 10.1039/C7CP01350G. URL: https://pubs.rsc.org/en/content/articlelanding/2017/cp/c7cp01350g (visited on 09/07/2025).

[21] Liam C. Jacobson, Waldemar Hujo and Valeria Molinero. "Amorphous Precursors in the Nucleation of Clathrate Hydrates". In: *Journal of the American Chemical Society* 132.33 (25th Aug. 2010), pp. 11806–11811. ISSN: 0002-7863.

DOI: 10.1021/ja1051445. URL: https://doi.org/10.1021/ja1051445 (visited on 08/07/2025).

[22] Hendrik Jung, Kei-Ichi Okazaki and Gerhard Hummer. "Transition Path Sampling of Rare Events by Shooting from the Top". In: *The Journal of Chemical Physics* 147.15 (21st Oct. 2017), p. 152716. ISSN: 1089-7690. DOI: 10.1063/1.4997378. PMID: 29055335.

[23] S. Jungblut and C. Dellago. "Pathways to Self-Organization: Crystallization via Nucleation and Growth". In: *The European Physical Journal E* 39.8 (10th Aug. 2016), p. 77. ISSN: 1292-895X. DOI: 10.1140/epje/i2016-16077-6. URL: https://doi.org/10.1140/epje/i2016-16077-6 (visited on 12/11/2024).

[24] *Kinetics and Thermodynamics of Multistep Nucleation and Self-Assembly in Nanoscale Materials, Volume 151 | Wiley*. https://www.wiley.com/en-us/Kinetics+and+Thermodynamics+of+Multistep+Nucleation+and+Self-Assembly+in+N p-9781118309513. (Visited on 22/07/2025).

[25] Carolyn A. Koh. "Towards a Fundamental Understanding of Natural Gas Hydrates". In: *Chemical Society Reviews* 31.3 (7th May 2002), pp. 157–167. ISSN: 1460-4744. DOI: 10.1039/B008672J. URL: https://pubs.rsc.org/en/content/articlelanding/2002/cs/b008672j (visited on 01/06/2025).

[26] Farbod Mahmoudinobar and Cristiano L. Dias. "GRADE: A Code to Determine Clathrate Hydrate Structures". In: *Computer Physics Communications* 244 (1st Nov. 2019), pp. 385–391. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2019.06.004. URL: https://www.sciencedirect.com/science/article/pii/S0010465519301833 (visited on 11/12/2024).

[27] Jeffrey J. Potoff and J. Ilja Siepmann. "Vapor–Liquid Equilibria of Mixtures Containing Alkanes, Carbon Dioxide, and Nitrogen". In: *AIChE Journal* 47.7 (2001), pp. 1676–1682. ISSN: 1547-5905. DOI: 10.1002/aic.690470719. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/aic.690470719 (visited on 07/07/2025).

[28] Ravi Radhakrishnan and Bernhardt L. Trout. "A New Approach for Studying Nucleation Phenomena Using Molecular Simulations: Application to CO2 Hydrate Clathrates". In: *The Journal of Chemical Physics* 117.4 (22nd July 2002), pp. 1786–1796. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.1485962. URL: https://pubs.aip.org/jcp/article/117/4/1786/185802/A-new-approach-for-studying-nucleation-phenomena (visited on 08/07/2025).

[29] Livio Ruffine et al. "Environmental Challenges Related to Methane Hydrate Decomposition from Climate Change Scenario and Anthropic Activities: State of the Art, Potential Consequences and Monitoring Solutions". In: *Earth-Science Reviews* 246 (Nov. 2023), p. 104578. ISSN: 00128252. DOI: 10.1016/j.earscirev.2023.104578. URL: https://linkinghub.elsevier.com/retrieve/pii/S0012825223002672 (visited on 03/06/2025).

[30] John Russo and Hajime Tanaka. "Crystal Nucleation as the Ordering of Multiple Order Parameters". In: *The Journal of Chemical Physics* 145.21 (13th Sept. 2016), p. 211801. ISSN: 0021-9606. DOI: 10.1063/1.4962166. URL: https://doi.org/10.1063/1.4962166 (visited on 12/05/2025).

[31] Sapna Sarupria and Pablo G. Debenedetti. "Homogeneous Nucleation of Methane Hydrate in Microsecond Molecular Dynamics Simulations". In: *The Journal of Physical Chemistry Letters* 3.20 (18th Oct. 2012), pp. 2942–2947. DOI: 10.1021/jz3012113. URL: https://doi.org/10.1021/jz3012113 (visited on 16/07/2025).

[32] David A. Sivak, John D. Chodera and Gavin E. Crooks. "Time Step Rescaling Recovers Continuous-Time Dynamical Properties for Discrete-Time Langevin Integration of Nonequilibrium Systems". In: *The Journal of Physical Chemistry B* 118.24 (19th June 2014), pp. 6466–6474. ISSN: 1520-6106. DOI: 10.1021/jp411770f. URL: https://doi.org/10.1021/jp411770f (visited on 07/07/2025).

[33] E.D. Sloan Jr. and C.A. Koh. *Clathrate Hydrates of Natural Gases, Thrid Edition*. Clathrate Hydrates of Natural Gases, Third Edition. 2007, p. 730. 1 p.

[34] E. D. Sloan Jr and F. Fleyfel. "A Molecular Mechanism for Gas Hydrate Nucleation from Ice". In: *AIChE Journal* 37.9 (1991), pp. 1281–1292. ISSN: 1547-5905. DOI: 10.1002/aic.690370902. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/aic.690370902 (visited on 08/07/2025).

[35] Niraj Thakre and Amiya K. Jana. "Physical and Molecular Insights to Clathrate Hydrate Thermodynamics". In: *Renewable and Sustainable Energy Reviews* 135 (1st Jan. 2021), p. 110150. ISSN: 1364-0321. DOI: 10.1016/j.rser.2020.110150. URL: https://www.sciencedirect.com/science/article/pii/S136403212030441X (visited on 01/06/2025).

[36] Matthew R. Walsh et al. "Microsecond Simulations of Spontaneous Methane Hydrate Nucleation and Growth". In: *Science* 326.5956 (20th Nov. 2009), pp. 1095–1098. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1174010. URL: https://www.science.org/doi/10.1126/science.1174010 (visited on 07/10/2024).

[37] Pramod Warrier et al. "Overview: Nucleation of Clathrate Hydrates". In: *The Journal of Chemical Physics* 145.21 (7th Dec. 2016), p. 211705. ISSN: 0021-9606. DOI: 10.1063/1.4968590. URL: https://doi.org/10.1063/1.4968590 (visited on 04/11/2024).

[38] Zhengcai Zhang, Matthew R. Walsh and Guang-Jun Guo. "Microcanonical Molecular Simulations of Methane Hydrate Nucleation and Growth: Evidence That Direct Nucleation to sI Hydrate Is among the Multiple Nucleation Pathways". In: *Physical Chemistry Chemical Physics* 17.14 (27th Mar. 2015), pp. 8870–8876. ISSN: 1463-9084. DOI: 10.1039/C5CP00098J. URL: https://pubs.rsc.org/en/content/articlelanding/2015/cp/c5cp00098j (visited on 16/07/2025).

List of Acronyms

TPE Transition Path Ensemble

TPS Transition Path Sampling

AA-TPS Always Accepting Transition Path Sampling

1wS-TPS One-way Shooting Transition Path Sampling

MCG Mutually Coordinated Guest

HCCS Hydrate-based Carbon Capture and Storage

CNT Classical Nucleation Theory

MD Molecular Dynamics

VVVR Velocity Verlet with Velocity Randomization

NPT Isothermal–isobaric

VV Velocity Verlet

OPS Oldest Path Segment

CD Cage Difference

RHS right-hand side